CVPR 2018 速報

片岡裕雄，Qiu Yue，相澤宏旭，鈴木智之，吉田光太，原健翔
鈴木亮太，福井宏，福原吉博，山本晋太郎，板摺貴大，荒木諒介，重中秀介，
美濃口宗尊，井上和樹，夏目亮太，中嶋航大，浅野一真，秋本直郁

http://hirokatsukataoka.net/project/cc/index_cvpaperchallenge.html

概要

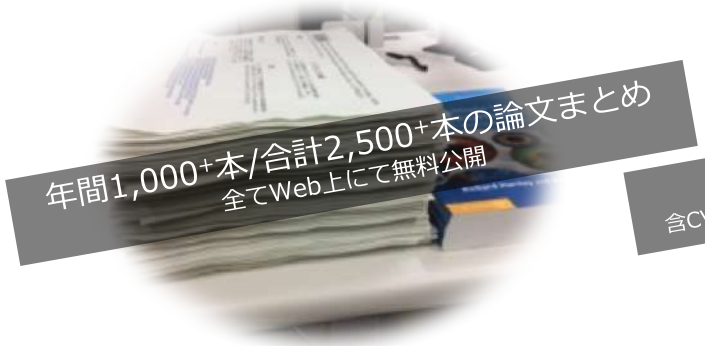
CV分野のトップ会議CVPR2018の参加速報

- cvpaper.challenge (次ページ) のメンバーで編集
  - CVPR 2018 完全読破チャレンジ (次々ページ) 実行中！
  - 今回，本会議2本，WS2本，コンペティション2件 (次々々ページ)
- 現在までの会議速報
  - CVPR 2016 速報: https://www.slideshare.net/HirokatsuKataoka/cvpr-2016
  - ECCV 2016 速報: https://www.slideshare.net/HirokatsuKataoka/eccv-2016
  - CVPR 2017 速報: https://www.slideshare.net/cvpaperchallenge/cvpr-2017-78294211
  - ICCV 2017 速報: https://www.slideshare.net/cvpaperchallenge/iccv-2017
- 全ての論文に目を通しているわけでは無いが，著者らができる限り聴講して議論を行った
- やはりDeep Neural Networks (DNN)の話が中心

cvpaper.challengeとは？

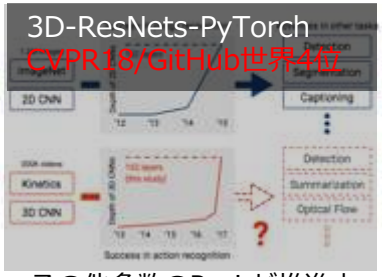
日本のCV分野を強くするチャレンジ！

◆ 論文読破・まとめ・発想・議論・実装・論文執筆に至るまで取り組む



Member List: 現在30+名
（産総研/電大/筑波大/早大/慶大/東大）

cvpaper.challengeの研究プロジェクト



その他多数のProj.が推進中

研究をより高い水準で組織化

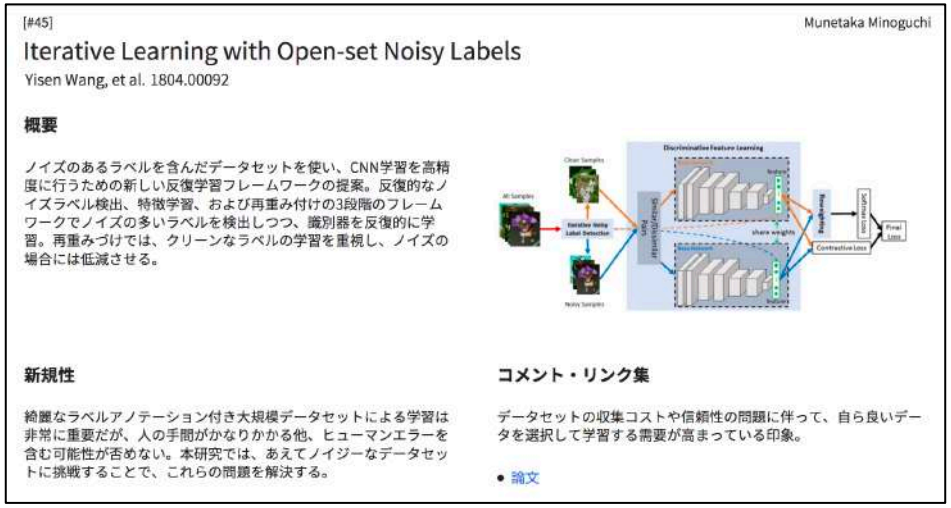
～集合知を発揮しやすい環境に～

CVPR 2018 完全読破チャレンジ

概要. CVPR2018に採択された1,000本弱の論文をcvpaper.challengeのメンバーおよびコラボレータで読破し、まとめ資料を全て共有するチャレンジです。CVPR（IEEE/CVF International Conference on Computer Vision and Pattern Recognition）はコンピュータビジョンやパターン認識の分野におけるトップ国際会議と位置付けられ、採択率は例年20%代と競争が熾烈であり必然的に論文のクオリティも高度になります。したがって、同会議に採択された全論文の要旨をまとめることは分野の現在を映し、世界中の研究者のアイディアや研究手法を知識として捉えることができると考えます。同国際会議に提案された技術を、新規性が保たれているうちに皆さんにお知らせするため、締め切りを2018年8月中（～2018/8/31）と設定して全979論文の完全読破に取り組みます。今回はただ論文をまとめるのみならず「論文を読む」という行為自体の統計解析や将来実施すべき研究課題のピックアップにも積極的に取り組みます。チャレンジにて得た知見はGitHub（まとめスライドはこちら）やTwitter、SlideShareなどにて公開予定です。

http://hirokatsukataoka.net/project/cc/cvpr2018survey.html

論文サマリの例



質問，コメント，誤植報告，ディスカッション等あれば
GitHub"CVPR2018_Survey"のIssueまで

cvpaper.challengeのCVPR採択論文

本会議2本，WS2本，コンペティション2件

- T. Suzuki*, H. Kataoka*, Y. Aoki, Y. Satoh, “Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident DB”（本会議論文）
- K. Hara, H. Kataoka, Y. Satoh, “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?”（本会議論文）
  - GitHubも宜しくお願いします https://github.com/kenshohara/3D-ResNets-PyTorch
- Y. Qiu, H. Kataoka, “Image generation associated with music data”（WS論文）
- K. Yoshida, M. Minoguchi, K. Wani, A. Nakamura, H. Kataoka, “Neural Joking Machine: An image captioning for a humor”（WS論文）
- T. Wakamiya, T. Ikeya, A. Nakamura, K. Hara, H. Kataoka, “TDU&AIST Submission for ActivityNet Challenge 2018 in Video Caption Task”（ActivityNet Challenge）
- K. Hara, H. Kataoka, Y. Satoh, “AIST Submission to ActivityNet Challenge 2018”（ActiivtyNet Challenge）
- 論文/プレゼン資料等のダウンロードこちら http://hirokatsukataoka.net/

CVPR 2018 まとめのまとめ

関連リンク

- @jellied_unagi氏
  - Tracking: https://www.dropbox.com/s/lzhekcmtfgl2woi/CVPR2018%20Tracking.pdf?dl=0
  - Forecasting: https://www.dropbox.com/s/rwzraw2rhv7i5gy/CVPR2018%20Forecasting.pdf?dl=0
  - Human Action: https://www.dropbox.com/s/n57hhpl08hcv2by/CVPR2018%20Human%20Action.pdf?dl=0
- 橋本敦史氏 https://atsushihashimoto.github.io/cv/
  - CVPR2018 参加報告(速報版)初日 https://www.slideshare.net/atsushihasimoto/cvpr2018-102697489
  - CVPR2018 参加報告(速報版)2日目 https://www.slideshare.net/atsushihasimoto/cvpr2018-2
  - Cvpr2018 参加報告（速報版）3日目 https://www.slideshare.net/atsushihasimoto/cvpr2018-3
- Acroquest Techonlogy株式会社ブログ
  - 1日目 http://acro-engineer.hatenablog.com/entry/2018/06/19/140042
  - 2日目 http://acro-engineer.hatenablog.com/entry/2018/06/20/145859
  - 3日目 http://acro-engineer.hatenablog.com/entry/2018/06/21/130625
  - 4日目 http://acro-engineer.hatenablog.com/entry/2018/06/22/125831

- 他にもCVPR 2018 まとめありましたらお知らせ下さい！

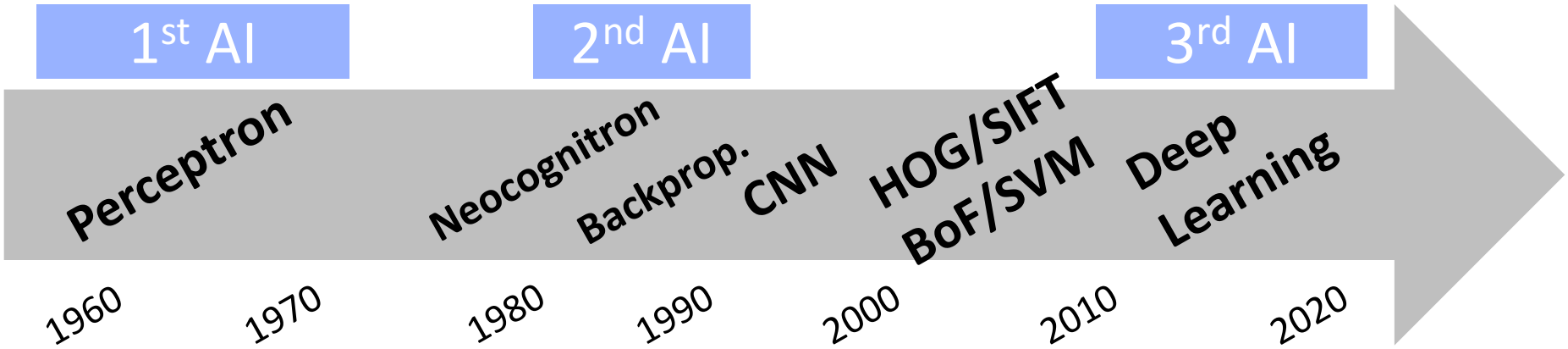
DNNの概要

DNNの動向（1/12）

DNN時代以前の動向

– Perceptron, MLP, Neocognitron, BackProp, CNN

– DNNが流行る前の画像認識では局所特徴が使用

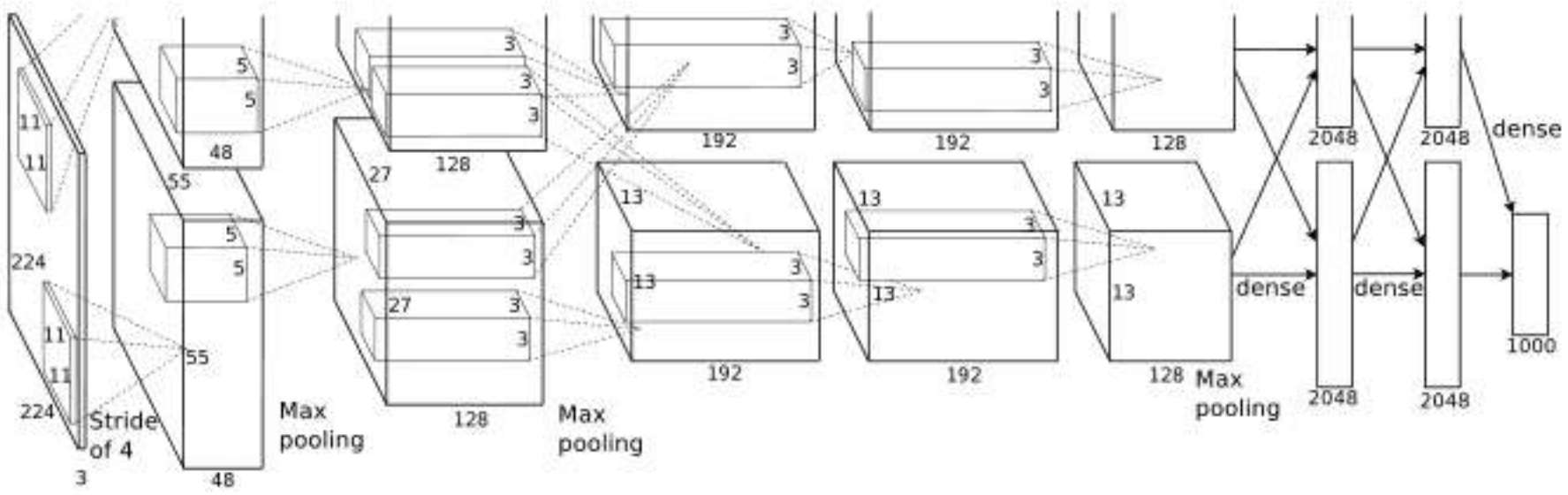


F. Rosenblatt et al. “Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms” in 1961.
Rumelhart et al. “Learning representations by back-propagating errors” in Nature 1986.
K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”, in 1980
Y. LeCun et al. “Gradient-based learning applied to document recognition” in IEEE 1998.

DNNの動向（2/12）

ILSVRCを発端とする画像識別タスクへの応用

- AlexNet @画像認識コンペILSVRC2012
  - 第一著者Alexさんのネットワーク（仕掛け人のHintonNetになってたかも？）
- 背景にはBelief Propagation, ReLU, SGD, Dropoutなど構造をDEEPにする技術が揃ってきた

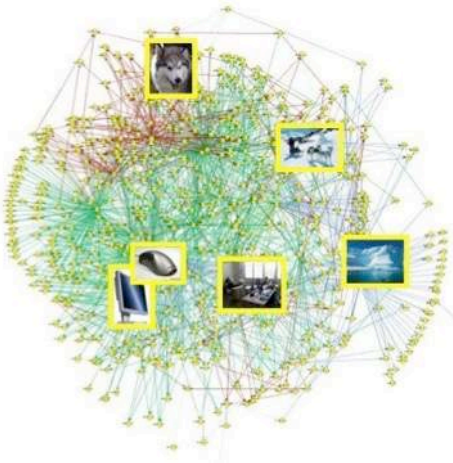


DNNの動向（3/12）

DNNが勝てた背景

– ImageNet!（データが最も重要）

– NVIDIA!（圧倒的な計算力）



IMAGENET

http://www.image-net.org/

×

NVIDIA®

http://cvpr2017.thecvf.com/

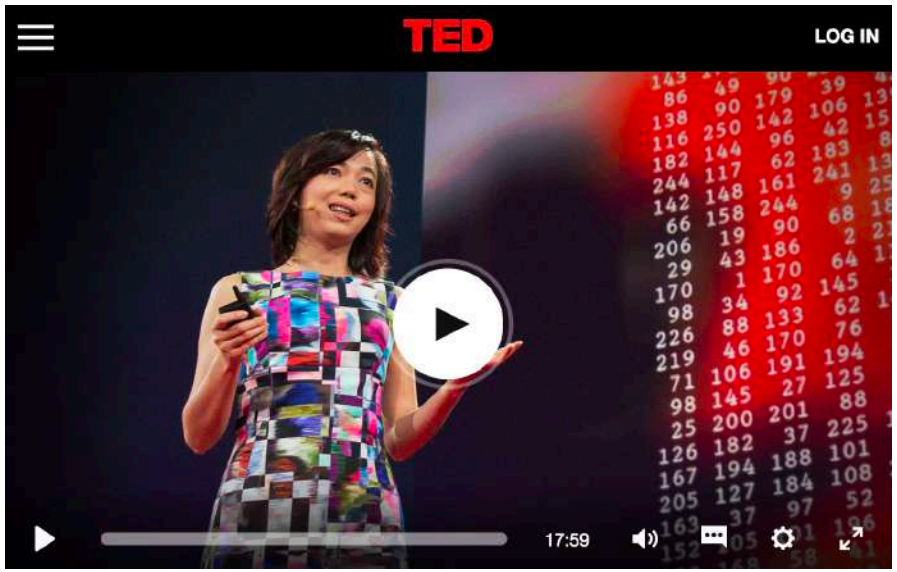
DNNの動向（4/12）

ImageNetの収集について

- 14,000,000+ imgs / 20,000+ categories
- 2007年からデータを収集，2009年CVPR発表
- その後もデータ収集は継続して，現在は上記の規模に

http://fungai.org/images/blog/imagenet-logo.png

ImageNetのロゴ，右側はStanfordの赤，左は前所属のPrinceton，そして上の緑はWorldPeace－世界平和－を示す（らしい）



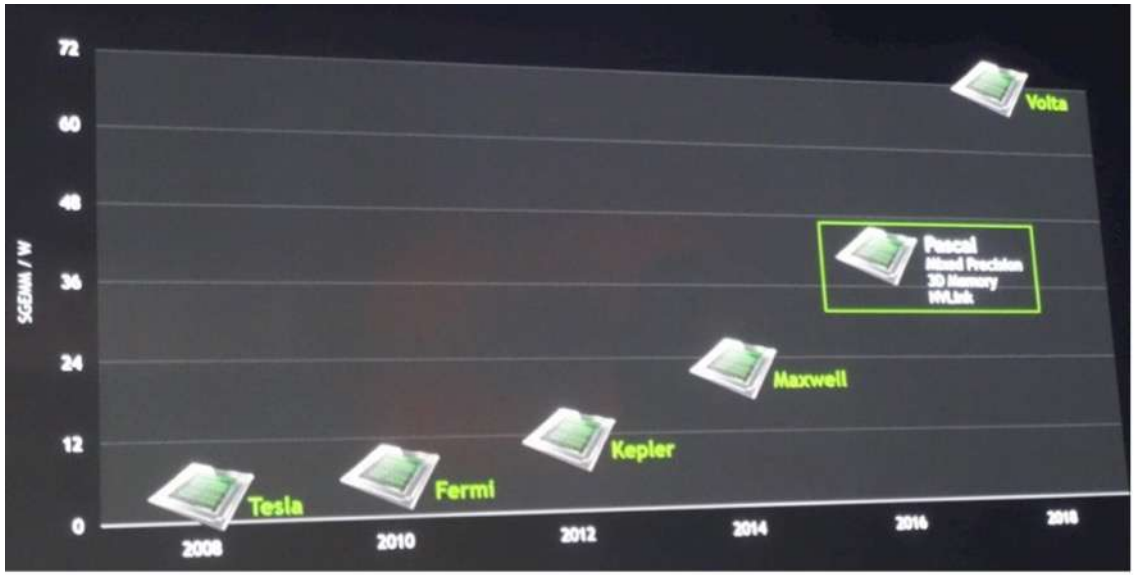
https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures/up-next?language=ja

Fei-Fei氏のTED動画（右）資金繰りの苦労や，2000年代当時はアルゴリズム至上主義でデータを収集することが理解されなかった

DNNの動向（5/12）

計算機環境（主にGPU）の発展

- 特に3rd AIブームからはNVIDIAの隆盛ぶりがすごい
- NVIDIA，最初はゲーム用グラフィックボードを売っていたらしいが，深層学習に会社の命運を託すと明言
- 結果，下記の性能向上と世界的な提携/資金獲得である



Tesla（2008年）からVolta（2018年）世代までの性能向上

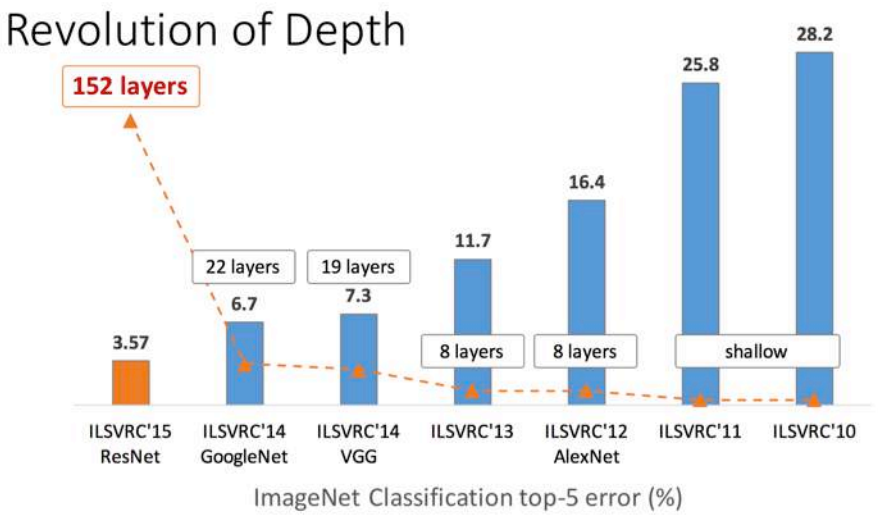
https://www.nextplatform.com/2015/03/18/nvidia-tweaks-pascal-gpus-for-deep-learning-push/

DNNの動向（6/12）

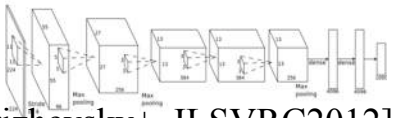
構造の深化

– 2014年頃から「構造をより深くする」ための知見が整う

– 現在（主に画像識別で）主流なのはResidual Network



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.

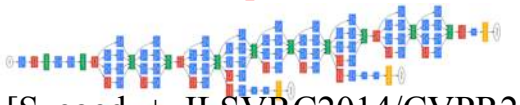


AlexNet [Krizhevsky+, ILSVRC2012]

ILSVRC2012 winner，DLの火付け役

VGGNet [Simonyan+, ILSVRC2014]

16/19層ネット，deeperモデルの知識



GoogLeNet [Szegedy+, ILSVRC2014/CVPR2015]

ILSVRC2014 winner，22層モデル

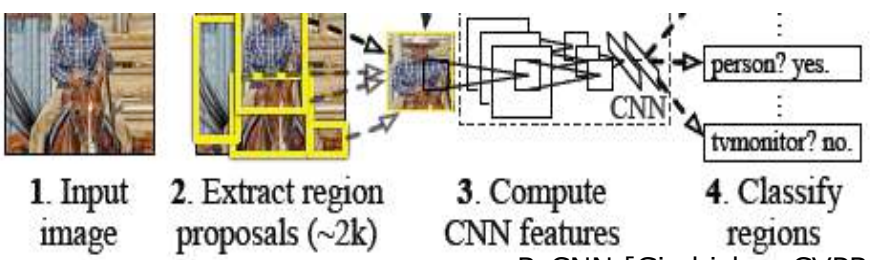
ResNet [He+, ILSVRC2015/CVPR2016]

ILSVRC2015 winner， 152層！(実験では10^3+層も)

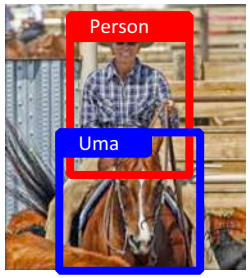
DNNの動向（7/12）

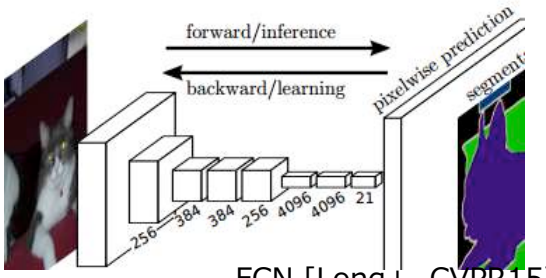
他タスクへの応用（画像認識・動画認識）

- R-CNN: 物体検出
- FCN: セマンティックセグメンテーション
- CNN+LSTM: 画像説明文
- Two-Stream CNN: 動画認識

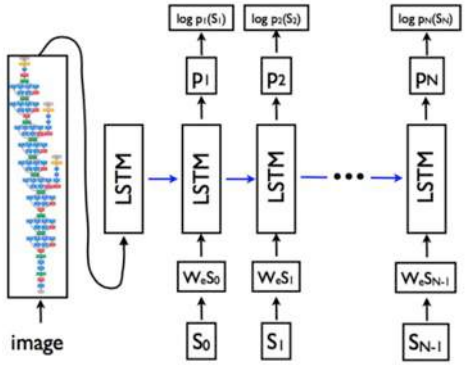


R-CNN [Girshick+, CVPR14]

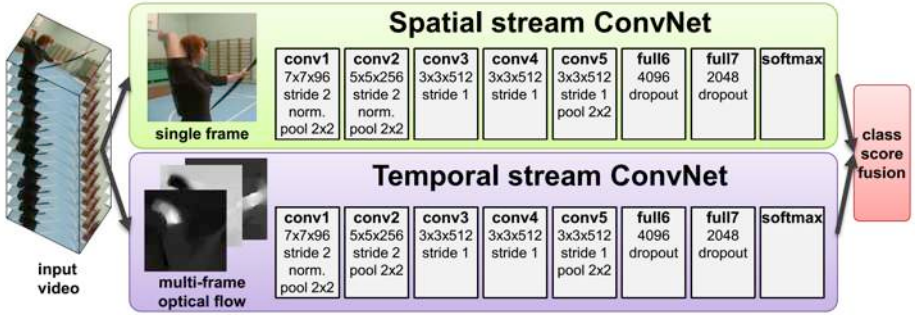




FCN [Long+, CVPR15]



Show and Tell [Vinyals+, CVPR15]



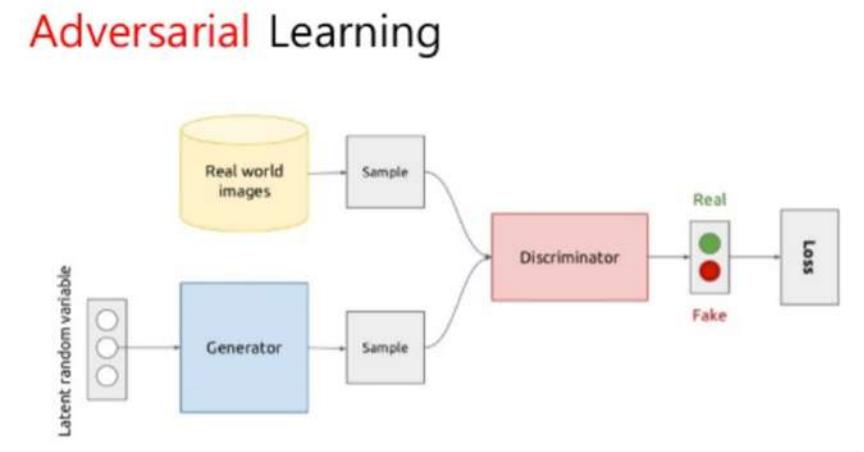
Two-Stream CNN [Simonyan+, NIPS14]

DNNの動向（8/12）

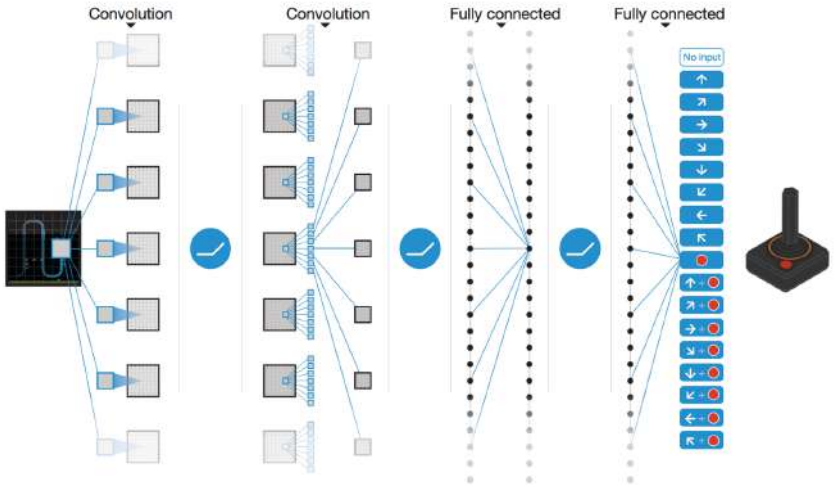
画像生成・強化学習への移行

– GAN（左図）：敵対的学習，Generator（G）とDiscriminater（D）が競い合いリアルな画像を生成

– DQN（右図）：強化学習モデル



https://www.slideshare.net/nmhkahn/generative-adversarial-network-laplacian-pyramid-gan



https://cdn-images-1.medium.com/max/1600/1*T54Ngd-b_CKcP3N6hyXLVg.png

DNNの動向（9/12）

教師なし学習/少量の教師あり学習への拡がり

- キーワード
  - {Un-, Weak-, Semi-, Self-} supervision
  - {Zero-, One-, Few-} shot learning
  - Domain Adaptation
- 教師がない/間接的に教師を与える，場合でも学習できるような仕組みに対する競争も激化
- 巨大IT企業のようにデータを持っていなくても学習を成功させる（アルゴリズム至上主義への回帰？）

DNNの動向（9’/12）

学習法の簡単な整理

- {Un-, Semi-, Weak-, Self-} supervision
  - Un-supervision
    アノテーションが一切ないデータで学習
  - Semi-supervision
    アノテーションを持つデータと持たないデータで学習
  - Weak-supervision
    出力として必要な情報よりも拘束力の弱いデータを用いて学習
    - ex)物体検出を行う際に画像ラベルのみを用いて学習
  - Self-supervision
    人手によるアノテーションではなく、数値表現のみを学習データとして使用。
    - ex)レンダリング画像から得られるデプスを使用

DNNの動向（10/12）

学習データ生成

- キーワード
  - Synthetic Data
  - Adversarial Learning
  - Data Augmentation
- 前ページと同様，大規模データ収集問題への対策
- CGなど合成（Synthetic）でデータを作成する
- 敵対的学習（Adversarial Learning）
  - 少量のサンプルから画像生成
  - 合成をリアルに近づける
- データ拡張（Data Augmentation）
  - データの水増しをあらゆる方法（e.g. 反転，回転，統合）で実現

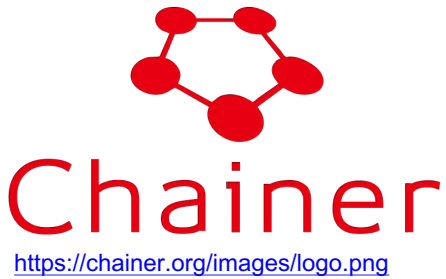
DNNの動向（11/12）

DNNのフレームワークが次々にリリース

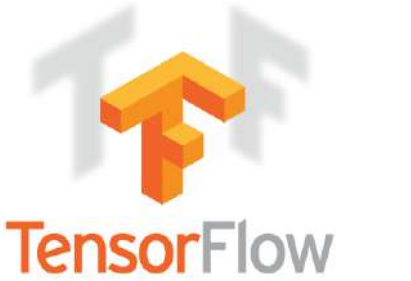
– Caffe/Caffe2, Theano, Chainer, TensorFlow, Keras, Torch/PyTorch, MatConvNet, Deeplearning4j, CNTK, MxNet, Lasagne

（順不同，その他多数）

– 特に，Caffeが出てきてからCVにおけるDNNの研究は爆発的に広がった



https://chainer.org/images/logo.png



https://www.tensorflow.org/_static/images/tensorflow/logo.png

http://pytorch.org/docs/master/_static/pytorch-logo-dark.svg

日本ではChainer？ EuroではMatConvNet？ 世界的にはTensorFlow？ コード量の少なさからKerasもよく聞く
当グループではPyTorchでほぼ統一して共有（Facebookでも研究はPyTorch，プロダクトはCaffe2）

DNNの動向（12/12）

現在も進化の一途を辿り，社会実装が進む

- 自動運転/ADAS
- ロボティクス
- ファッション
- 画像/動画検索
- 物流（ピッキング等）
- 等

研究者としては「こんなこともできる」を世に出したい

CVPRで議論されているサブタスクとホットな領域

- サブタスク（CFPより）
- CVPRの三大分類？
  - アルゴリズム考案，データ問題，新規問題設定

サブタスク一覧

CVPRのCFPより (1/2)

– Computer Vision Theory, 3D Vision, Action Recognition, Biometrics, Big Data, Large Scale Methods, Computational Photography, Sensing and Display, Deep Learning, Document Analysis, RGBD sensors and analytics, Face and Gesture, Low-level Vision, Image Processing, Medical/Biological and Cell Microsopy, Image Analysis, Motion and Tracking

サブタスク一覧

CVPRのCFPより (2/2)

– Optimization Methods, Performance Evaluation and Benchmark Datasets, Physics-based Vision and Shape-from-X, Recognition: Detection, Categorization, Indexing Segmentation, Grouping and Shape Representation Statistical Learning, Video Analytics, Vision for Graphics, Vision for Robotics, Vision for Web, Applications

CVPRの三大分類

アルゴリズム考案

- 従来の問題設定の延長だが，精度向上/タスク解決に対して効果的な手法を提案（データに関する項目は次ページ）
- 昔からホットな領域
  - 物体検出（含:人物），人物行動認識，手部領域追跡，セマンティックセグメンテーション，ビューポイント変換，Shape-from-X，SLAM，Computational Photography，自動運転，ロボット応用，，，
- 最近ホットになった領域
  - 画像説明文，視覚的質問回答（VQA），GAN，ファッション，ピッキング，画風変換，，，
- 今後ホットになりそうな領域？
  - 超多タスク学習（Best Paperを受けて），強化学習, 3D×GAN，，，
  - あとはみんなで予想しましょう！

CVPRの三大分類

データ問題

- 大規模データ収集/アノテーション問題を解決および緩和
  - DNNの動向と重なりますが，深層学習の1st waveがアーキテクチャ改善だとすると，データ作成/少量データ学習が2nd wave
- {Un-, Weak-, Semi-, Self-} Supervised Learning, {Zero-, One-, Few-} shot learning, Domain Adaptation, Synthetic Data, Adversarial Learning, Data Augmentation

上記の改善とともに，野心的な新しい学習法などが出る

（少なくとも世界的に取り組まれている）

CVPRの三大分類

新規問題設定

- 問題設定と同時にデータセット作成/アルゴリズム考案
- ベンチマーキングにより新しい視点を与える
- 立ち止まって網羅的解析により問題を見直す
  - などの論文が通っている（し増えている）

研究に哲学がある方が論文は面白くなる！

新しい問題を考えよう

CVPR 2018の動向・気付き

- 今回どんな研究が流行っていた？
- 海外の研究者は何をしている？
- 「動向」や「気付き」をまとめました

CVPR2018の動向・気付き（1/61）

– 重要論文はどこの研究機関からでも出るようになった

- 昔（少なくとも5年前くらい）は目立つCV分野の論文は大体において研究チームが限られていた
- DNNの提案により裾野が広がった
- フレームワーク（e.g. Caffe/Caffe2, TensorFlow, PyTorch, Chainer）/arXiv, Openでダウンロード可能な論文やコード等の充実

まさに今，誰もが当事者になるチャンス！

努力次第ではトップ会議採択/産業応用展開も可能！！

CVPR2018の動向・気付き（2/61）

- DNN時代になってから、人物/物体/シーン認識/3DVisionなど関係なく良い手法が出てくる
  - DNNという共通の枠組みができた
  - ひとまず動かすだけなら比較的簡単（DNNフレームワークも充実）
  - 応用範囲が広い上に組み合わせ可能（分野間がシームレス化して来た）
  - さらに相互利用ができる枠組みになってきた

CVPR2018の動向・気付き（3/61）

– タイトルに頻出の単語を調べてみた

- Learning 217 →Deep Learningに関連した単語強し
- Network(s) 205
- Image 158
- Deep 122
- 3D 85 →3Dメッシュを扱う論文が多い
- Multi 82
- Object 74
- Video 70
- Detection 69 →物体検出はまだまだ健在？
- Visual 66
- Adversarial 61 →GANを使った論文がかなり存在

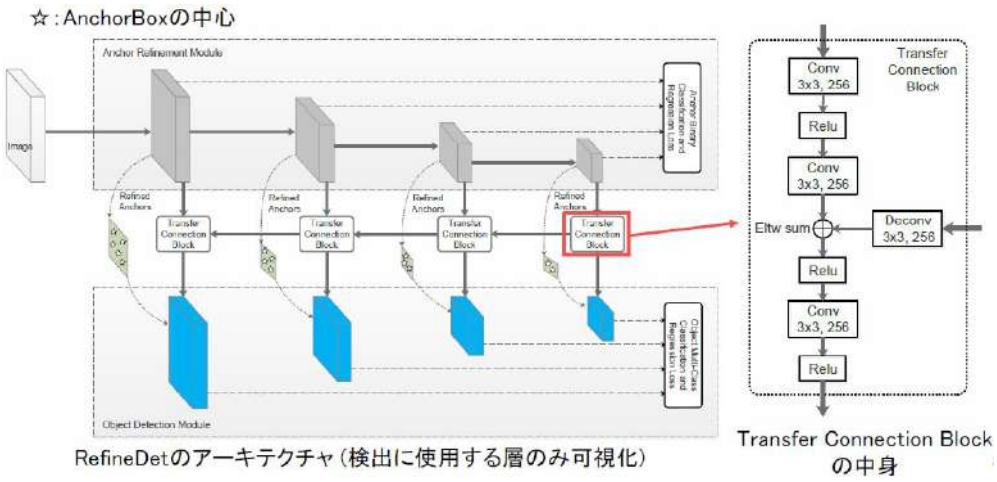
CVPR2018の動向・気付き（4/61）

- 「こんな新しいことができた」ということ自体でさえ形骸化してしまった？
  - 研究者ですらここまで早いか，という流れの中にいる
  - 大規模にデータを集め，入力と出力の対応関係さえ教示するラベルが揃っていればDNNでなんとかしてくれる
  - アーキテクチャ探索のように，未解決問題の空間を探索している（いい意味です）

CVPR2018の動向・気付き（5/61）

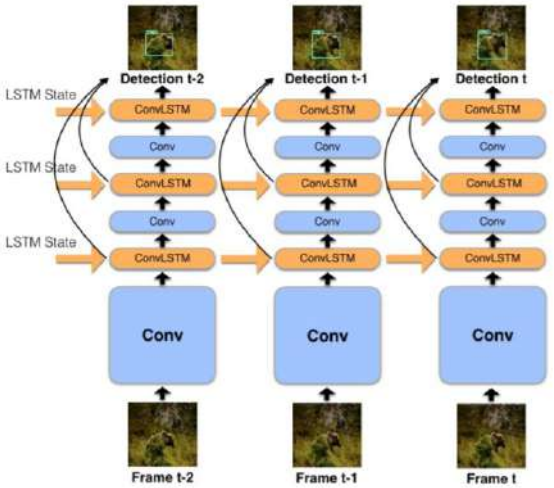
– 物体検出は従来法 + αの手法が多い

- Single-shot（SSDなど）ベース
  - RefineDet：**畳み込み後の特徴マップ**から検出boxの高精度化
  - Bottleneck-LSTM：SSDのリアルタイム検出に**LSTM**導入で時系列考慮



RefineDetのアーキテクチャ（検出に使用する層のみ可視化）　Transfer Connection Blockの中身

RefineDet



Bottleneck-LSTM

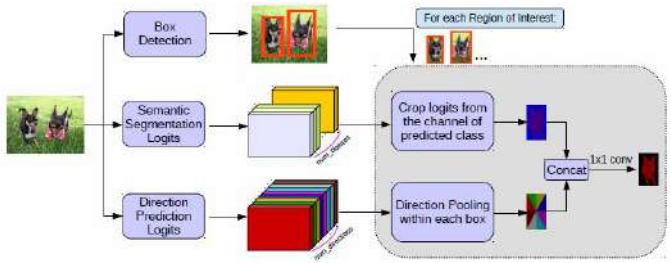
CVPR2018の動向・気付き（6/61）

– 物体検出は従来法 + αの手法が多い

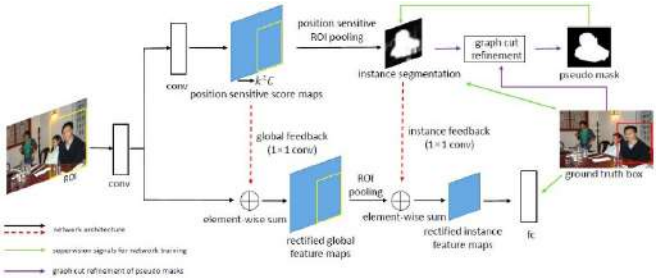
- RPN + classificationの2-shot（Faster R-CNNなど）ベース
  - MLKP：**マルチスケール**の検出手法をR-CNNに応用
  - MaskLab：Mask R-CNN＋**Direction prediction**
  - Pseudo-mask augmented：**グラフカット**でMask高精度化
  - Future selective networks：**Attention**を用いてRoI補正
  - drl-RPN：RPNに**強化学習**を用いてRoIの選択

Det + Seg

RoI refinement

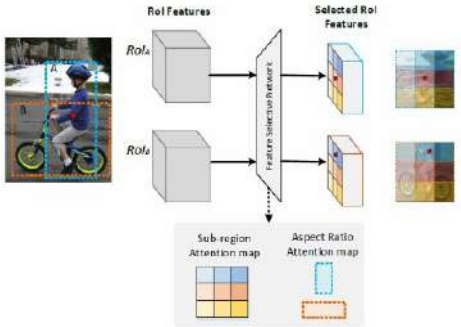


MaskLab

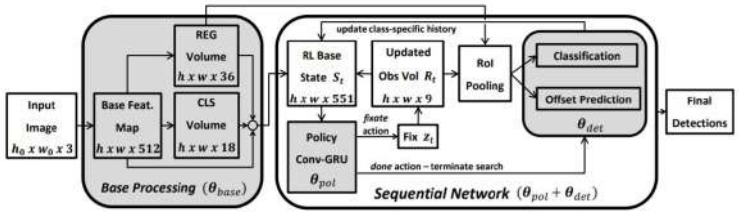


Pseudo-mask augmented

Detection + Segmentation



Future selective networks



drl-RPN

RoI refinement

CVPR2018の動向・気付き（7/61）

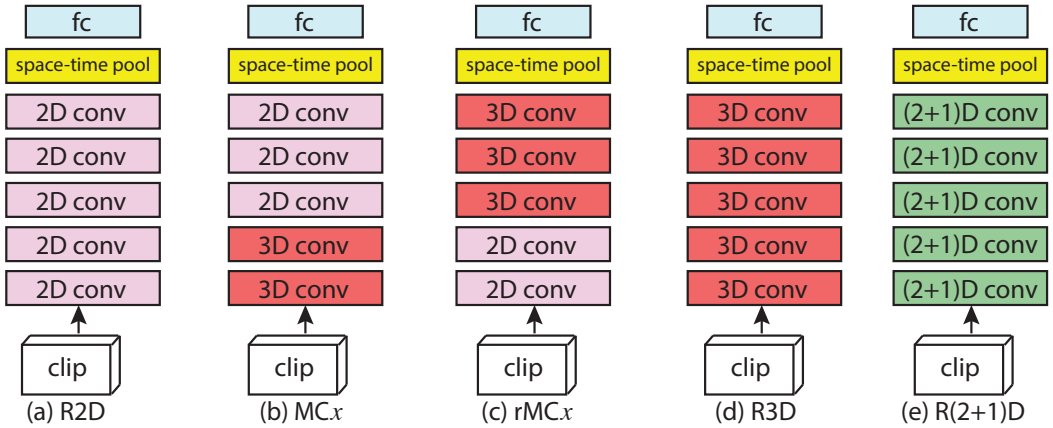
– 物体検出問題，未だに根強くFaster R-CNNも残る

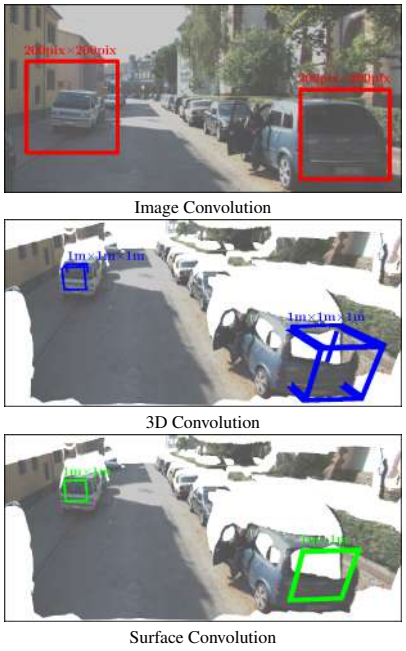
- YOLO/SSDなど早い手法が提案されているにも関わらず，アンドMask R-CNNのようにインスタンスセグメンテーションができる手法ができても，である
- これは，(1)候補領域抽出 (2)物体識別という2-stageの構造から？
  - (1)の部分があることで，より理解しやすい手法となっている
  - ただし，候補領域で抽出されていないものは検出されない
  - 多少の解釈性を保有していることが実利的
- 実験的には，YOLO/SSDよりもFaster R-CNNの方が精度がよい場合も多い（あくまでも実験的）

物体検出の覇権争いはここまでで落ち着いた？

CVPR2018の動向・気付き（8/61）

- 3D(xyt, xyz) の畳み込みfilterの再考
  - 動画解析における、(3D conv) VS (2D conv + α)
    - 3D ResNet：大規模DBによる学習で3D convの有効性を解析
    - R(2+1)D：同一param数では(2+1)D convの精度が良いことを主張
    - MiCT：3D, 2D convの組み合わせが良いと主張
  - RGB-D 画像解析における3D convの冗長性を問題視
    - SurfConv：画像内の深度情報を考慮した2D convを用いて解決





CVPR2018の動向・気付き（9/61）

– Semantic Segmentationの動向・気付き

- Semantic Segmentationの覇権争いが激化
  - 単純な精度比較ではなく、問題をさらに細分化して解かれている印象
    1. global contextの取得
    2. シーン中の物体のスケール変動への対応
    3. データ不足への対応
- Semantic Segmentation + ***X***
  - Detection
    - » MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features など
  - Depth
    - » PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing
  - Deburring
    - » Deep Semantic Face Deblurring

CVPR2018の動向・気付き（10/61）

- Semantic Segmentationの流行り
  - ピラミッド表現の利用
    - 空間情報を保持しつつglobal contextを抽出し、シーン中の物体のスケール変動へ対応するための、multi-scaleな特徴抽出を行う
      複数のdilated rateによるdilated convが利用されている。
    - **PSPNetとDeeplabが与えた影響は大きい！**
  - データ不足の対応
    - Semantic Segmentationは特にアノテーションコストが高い！
    - 弱教師あり学習も昨年に引き続きホットであるが
      今年は特に、**Domain Adaptation**、**Interactive Segmentation**も盛ん。
  - Video & 3D
    - 去年から静止画から動画像や3Dデータへ、より難しいドメインへと変遷。
    - これまでのSemantic Segmentationの問題を時系列やボリュームデータから解決できるか？ -> **3D Convも利用できそう？！**

CVPR2018の動向・気付き（11/61）

– Semantic Segmentation周辺分野の未来予想

1. Semantic Segmentation + ***Detection***
   – Semantic SegmentationはDetectionを内包したタスク。
   Mask RCNNをはじめとして統合が進む？
2. Sementic Segmentation + ***Instance Segmentation***
   – 1と相まって今後統合されるのは確実
   – Panoptic Segmentation https://arxiv.org/abs/1801.00868
3. Stuffクラスの認識
   – Stuffクラスはシーンや出現物体の理解の手がかりとなる。
   – COCO-Stuff: Thing and Stuff Classes in Context
4. Domain Adaptation
   – Semantic Segmentationとアノテーションコストは切り離せない。
   – 今回のDomain Adaptationの研究の数をみても必然。
5. Semantic Segmentationのさらなる高精度化
   – CityscapesやPASCAL VOCでもまだまだ完璧ではない。
   CNNができないことを追求して海外と戦うべき？(逆にやる人がいない？)

CVPR2018の動向・気付き（12/61）

– Semantic Layoutからの画像生成

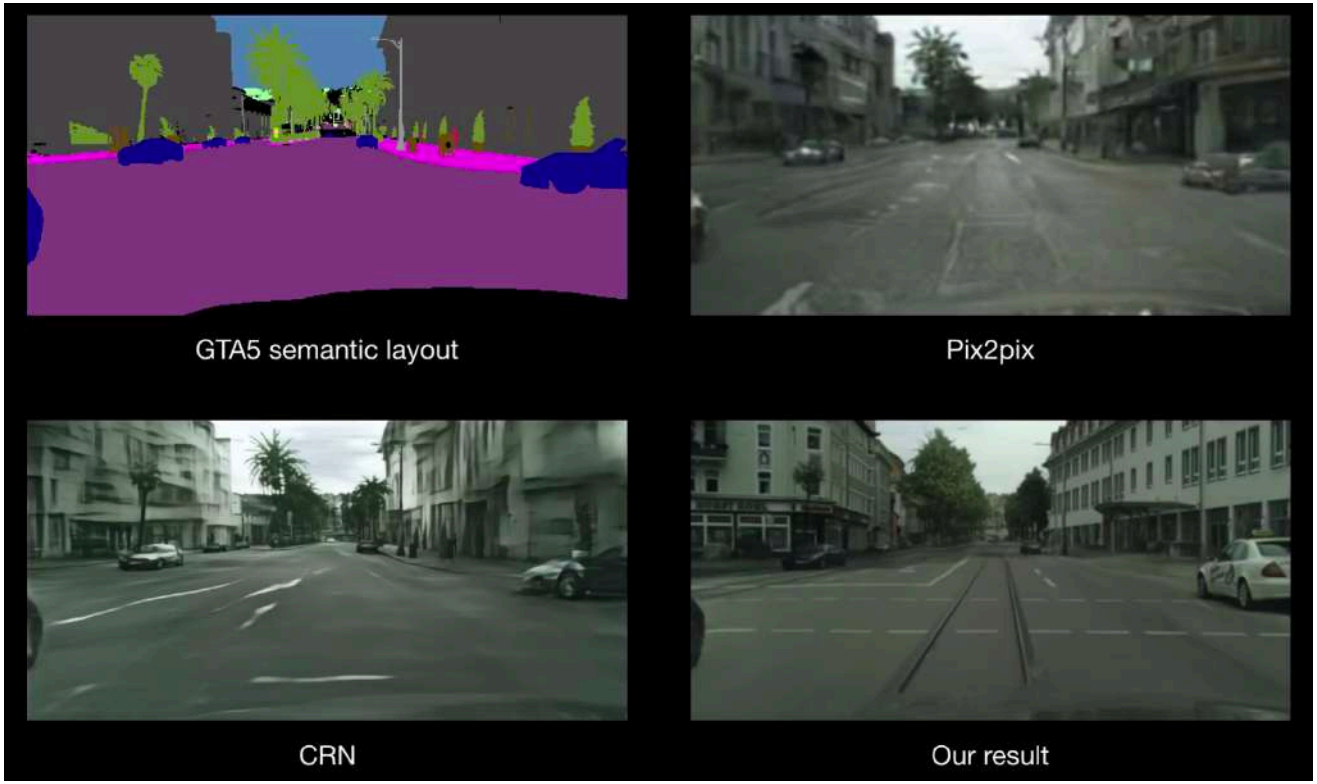
- ラベルデータがあれば、高解像度高品質なデータが作成できる。
  - High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs
  - Semi-Parametric Image Synthesis
- **データ不足への新たな対応策**
  - 今後は出力の**多様性**が重要になる。

High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs

CVPR2018の動向・気付き（13/61）

– Semantic Layoutからの復元，ホントにキレイになる

- これはもう，各方面で学習データに使うしかない
- （１）シミュレータをセマンティックラベルに復元，（２）この手法でセマンティックラベルを任意のリアル画像に復元
- （１）と（２）を対応付ければ学習の入出力画像が完成！



CVPR2018の動向・気付き（14/61）

– Object Tracking w/ CNN

- CNN + Correlation Filterが主流
  - CNNから得られた特徴量にCorrelation Filterを適用させる
- 物体追跡にDNNを適用するために、リアルタイム性と追跡精度を両立させるオンライン学習の実現が必要
- 特にネットワークとしては、Siamese Networkが使用される傾向がある
- Attentionを利用した手法も散見された
- 3DCNNがフローなどの時間的特徴を取得できるようになれば、物体追跡にも適用されていくと期待される

CVPR2018の動向・気付き（15/61）

– Object Tracking w/ GAN

- Positive Sample数を著しく少ないため、Generatorで生成する手法がいくつか提案されている
- 今後も同様の方向性の手法が提案されてくると期待できるが、リアルタイムにオンライン学習できるかがボトルネックとなる

CVPR2018の動向・気付き（16/61）

- 群衆データの追跡を扱う研究が少ない
  - 群衆データの追跡はまだ早い？

- 理由？
  - 2D画像を用いた人の判別はまだ精度があがる
    - 判別精度を高めてから追跡に入りたい
  - 3D画像は1人1人の細かい動きを把握したい
    - 運動をしている時の手の動き、体の動きが多い（姿勢推定）
    - ダンス、野球、サッカー規模の人数が限界？

CVPR2018の動向・気付き（17/61）

– 群衆データの追跡研究の未来について

- 追跡の研究は人の判別精度が頭打ちになってから始まりそう
  – 人の判別は2Dでも3Dでもまだまだ精度が向上している

- カメラの使用度と判別の精度で研究の方針が変わりそう
  – 2D画像はカメラのコストパフォーマンスが良いが計算処理が大変
  – 3D画像は人の判別は単純そうであるがコストが高い

CVPR2018の動向・気付き（18/61）

– GAN系は大体ベースラインが決まってきた？

- WGAN, WGAN-GP
  - 通常のGANのネットワーク構造のままWGANを適用可能で，より学習が安定.
  - WGANを試さない理由がむしろない.

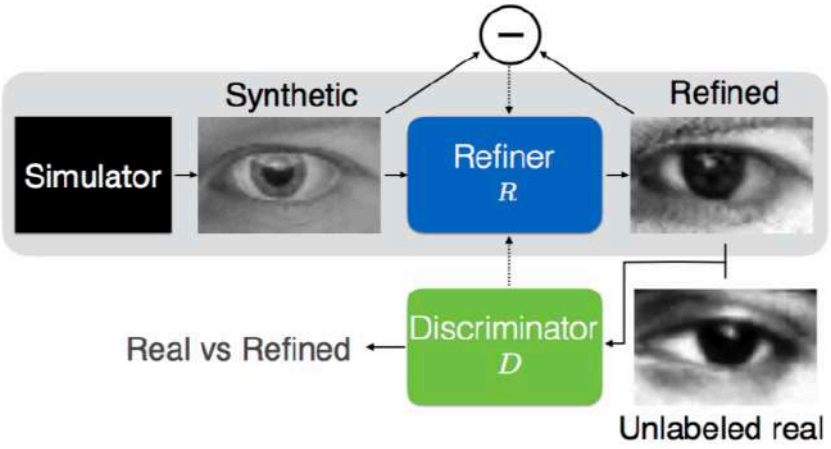
- CycleGAN, Cycle-Consistency
  - ペアを作らずに２つのドメイン間の変換を学習
    ペア不要によりデータセットの作成コストが削減
  - 様々なタスクで利用可能.
    画像間の変換(生成)，Domain Adaptationなど

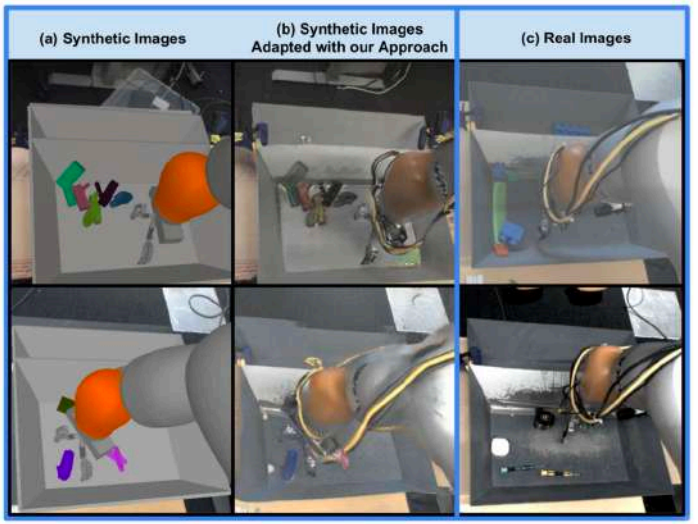
- これら二つは，取り入れ易くかつ強力であるため，**ひとまず**試されている様子

CVPR2018の動向・気付き（19/61）

– GANにより変換/生成した画像も学習に使えるように

- 左図：SimGAN
  - CVPR2017 Best Paper
  - CGをリアルに近付けるRefiner（R）と識別器（D）
- 右図：GraspGAN
  - ICRA2018
  - 上記論文を元ネタとしてロボットシミュレータ画像をリアルに近づけて，マニピューレーションを実行
- 今後も出てくる雰囲気がある





CVPR2018の動向・気付き（20/61）

– 画像の詳細な部分まで復元可能に

- 画像生成の分野では，U-Netのような入力画像の高次な特徴量を残す構造が一般的になってきている
  – 入力画像から形状が変わるような画像生成の際にも，U-Net構造が使えるような工夫が多くなされている

- ２段階で生成する手法が増加，ネットワークを組み合わせることで複雑なタスクにも対応可能
  – 一段階目で入力と形状が変わった解像度の荒い画像を生成
  – 二段階目で入力の詳細な情報を持った高精細な画像を生成

CVPR2018の動向・気付き（21/61）

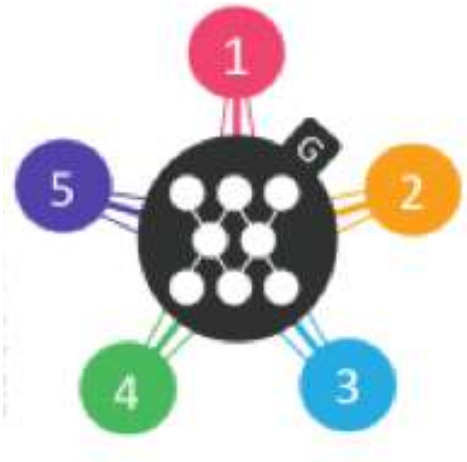
- 弱教師付き/半教師あり学習が高いレベルで実現された？
  - ものによっては教師あり学習を超える場面も
    - 教師あり学習よりもデータ量を確保することで精度が向上しているパターン
  - 画像生成/ドメイン変換なども使用して精度向上
    - やはりラベルは多少曖昧でもデータ量で精度を上げる方が良い？
    - 弱いとはいえ，数年前より適切なラベルを与えられるようになってきた？
  - 両方の枠組み，どちらでも学習可能なモデルも登場

弱いラベルも実利用に耐えうるレベルになる？

CVPR2018の動向・気付き（22/61）

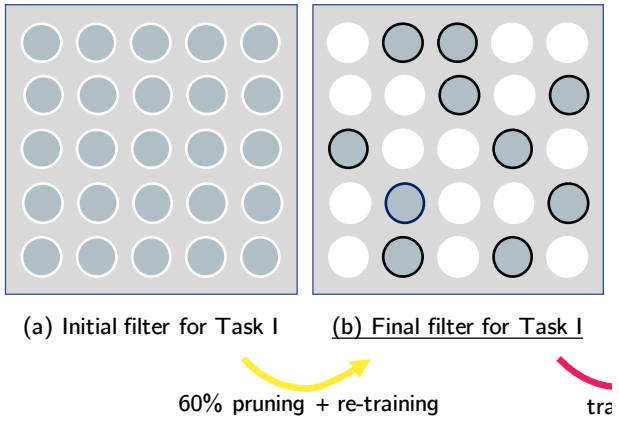
– 学習を効率化するような手法が目立つように

- ラベルなしデータの活用
  - 被っていたので略
- 他ドメイン(データセット)の活用
  - Domain Adaptation関係の手法
  - 弱教師あり学習の活用
  - PackNetのような複数のデータセットの情報を残していく学習方法



Star GAN

Domain Adaptive Faster R-CNN



PackNet

CVPR2018の動向・気付き（23/61）

- ドメインアダプテーション（ドメイン変換）
  - Source Domain（大量にラベル付データがある）からTarget Domain（少量のラベル付データしかない）への特徴転換問題
  - CG/WebDataをSourceにして，実空間というTargetに転換
  - Fine-grained Categorizationまで来ている（車両認識）
  - Object-level（ImageNet）画像からUnlabeled/Labeled な人物行動動画へのドメイン変換
  - Domain Adaptation Challenge: VISDA at ECCV2018
    - シミュレーションから実空間へのドメイン変換

CVPR2018の動向・気付き（24/61）

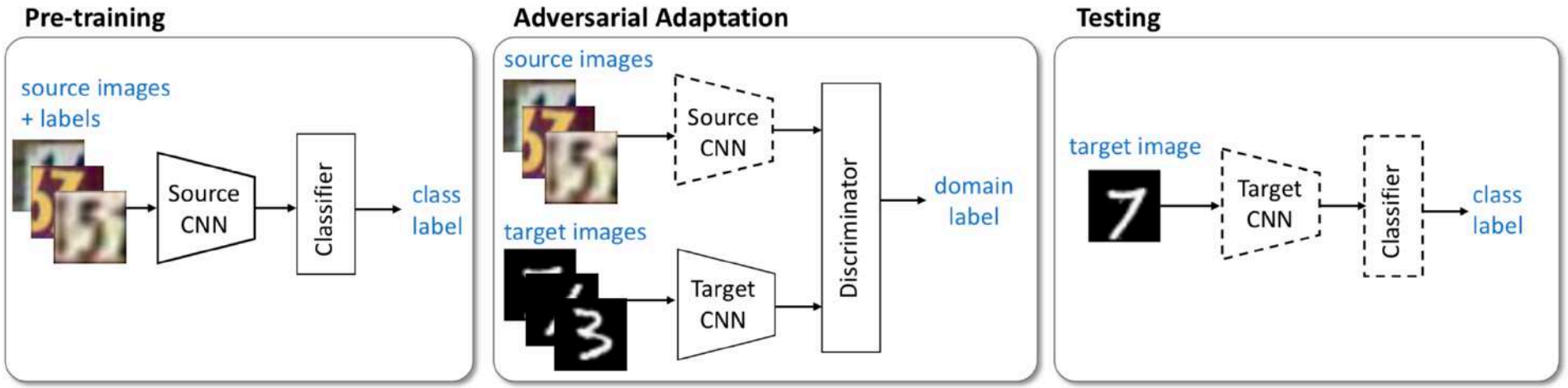
- Domain Adaptation (DA)激増？
  - DNNの進化に伴いDBが増えたが，ドメインを超えた学習や収集が困難なDBが存在
    - ex) 実画像に対する完璧なデプス、同一照明環境における顔画像 etc.
  - そんな時は教師なしDA (UDA; Unsupervised Domain Adaptation)で解決！
    - アノテーションが豊富なドメインでタスクを学習，アノテーションが{ない, 少ない}ドメイン学習に利用
    - ex) レンダリングされた3D顔合成データを用いてアルベド，シェイプなどを学習→アノテーションが一切ない実顔画像のアルベド，3Dシェイプ推定 etc.

CVPR2018の動向・気付き（25/61）

– UDAの代表格ADDA

- ADDA (Adversarial Discriminative Domain Adaptation)
  - GANを用いてドメインに不変な特徴量を取得
  - https://arxiv.org/abs/1702.05464
- ADDAを皮切りにGANを用いたUDAが流行！

ドメイン変換激増の一因か？



CVPR2018の動向・気付き（26/61）

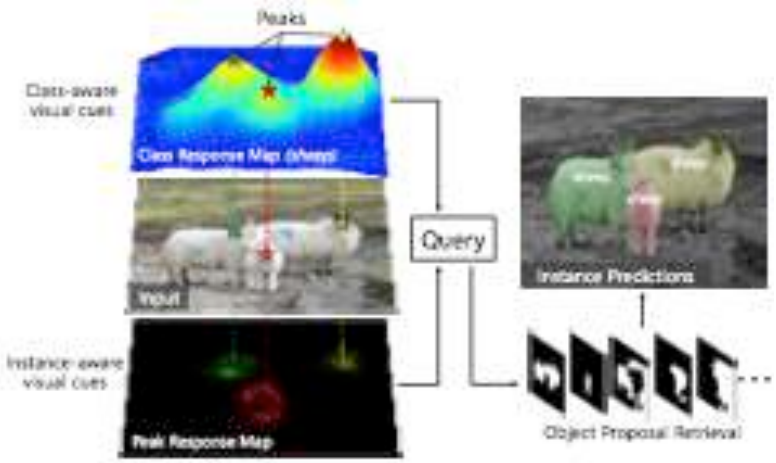
– 今回のアップデート版ADDA

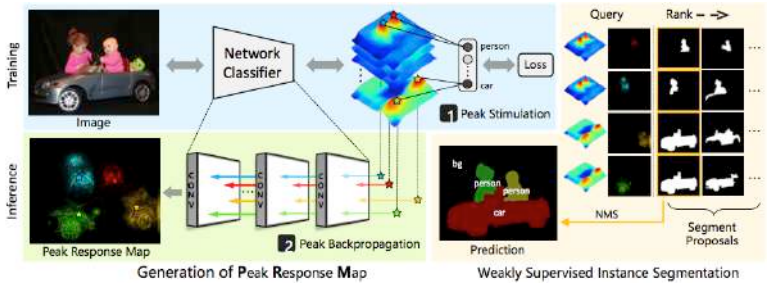
- 汎用的なタスクに適用可能な手法
  - CNNの浅い層ではドメイン固有な特徴量を、深い層ではドメインに不変な特徴量をもつことを利用
  - Cycle GANでconsistency lossによって再び同じドメインへ帰ってくることを強制
  - クラス特徴量の重心とCNNによる特徴量の相関を学習
- タスク特化な手法
  - 画像をハッシュ化するタスクにおいてハッシュ不変性をロスに取る
- タスク特化なUDAも提案されているが、汎用的なタスクに適用可能な手法が多く提案されているため、より広がりを見せて行くことが期待される！

CVPR2018の動向・気付き（27/61）

– 弱教師あり学習の発展

- Classificationのラベルのみを使って物体をローカライズする研究
  - セグメンテーション・グラフカットでローカライズの精度を上げる例が多い
- CNNの認識に関する根拠を知るために使われることが多い
  - 今年は単純に物体領域を求めるために使われる例も多い
- 去年くらいから流行っていたけど，今年も勢いがすごい



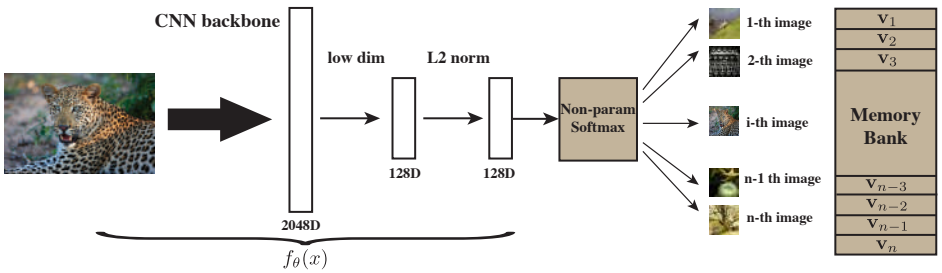


CVPR2018の動向・気付き（28/61）

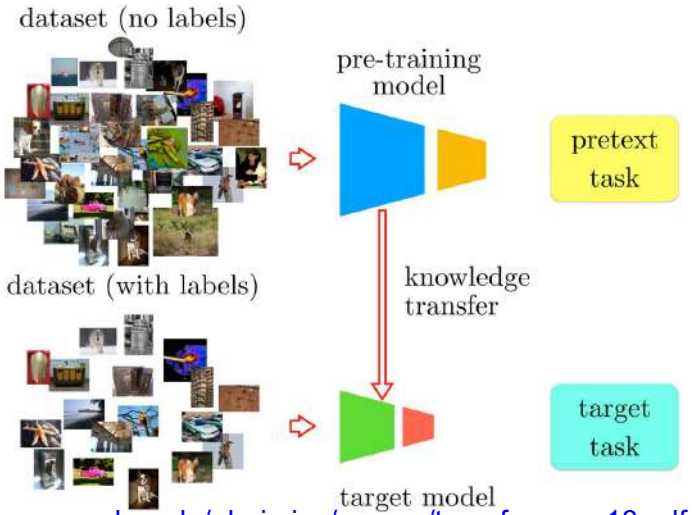
- Self-supervised学習の知見が整ってきた
  - Self-supervised学習：Fine-tuningの前に”教師なし”で特徴表現学習を行う
    - 例：動きを学習するために動画の順番を当てる問題を解いてから，人物行動DBにより学習
  - 2018時点では，簡単に思いつく擬似タスクはほぼ出た
  - 今回は精度以外の点での貢献にフォーカス
  - 特定のタスクに特化した表現獲得も議論

CVPR2018の動向・気付き（29/61）

- Self-/Un-supervisedな表現学習
  - 単純なアイデアは出尽くされてきた（e.g. 回転，カウント）
  - より洗練されたアイデアならば通っている
    - 画像インスタンスレベルの識別による表現学習 (1)
    - Artifact検出による敵対的表現学習
  - 既存の枠組みの改善手法が提案
    - コンテキスト（文脈）ベース学習の改善
    - 蒸留を利用したアーキテクチャに捉われない特徴表現の利用 (2)



(1) https://arxiv.org/abs/1805.01978



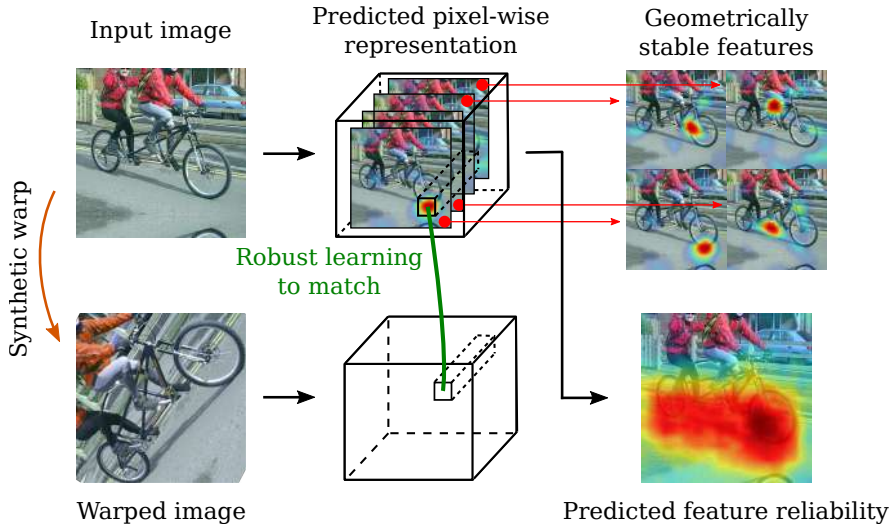
(2) https://www.csee.umbc.edu/~hpirsiav/papers/transfer_cvpr18.pdf

CVPR2018の動向・気付き（30/61）

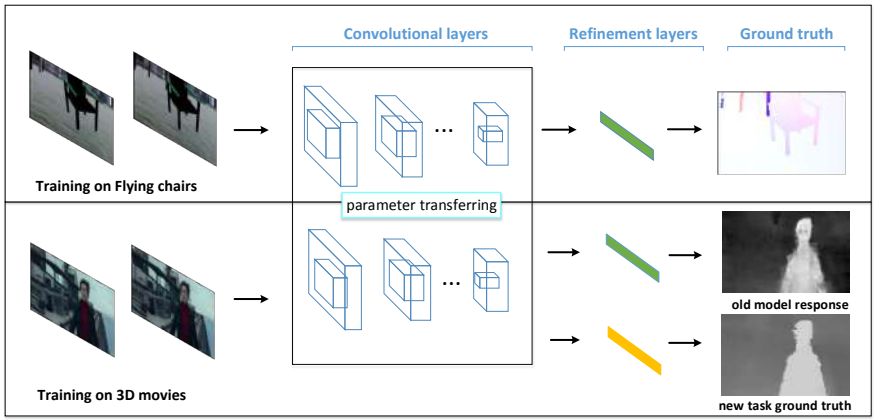
– Self-/Un-supervisedな表現学習

- Target taskをより具体化したものが今後注目？
  – キーポイントマッチングのための表現学習 (3)
  – シーン認識、行動認識のための表現学習 (4)

ImageNet学習済みモデルを超える挑戦にも引き続き注目！



(3) https://arxiv.org/abs/1804.01552

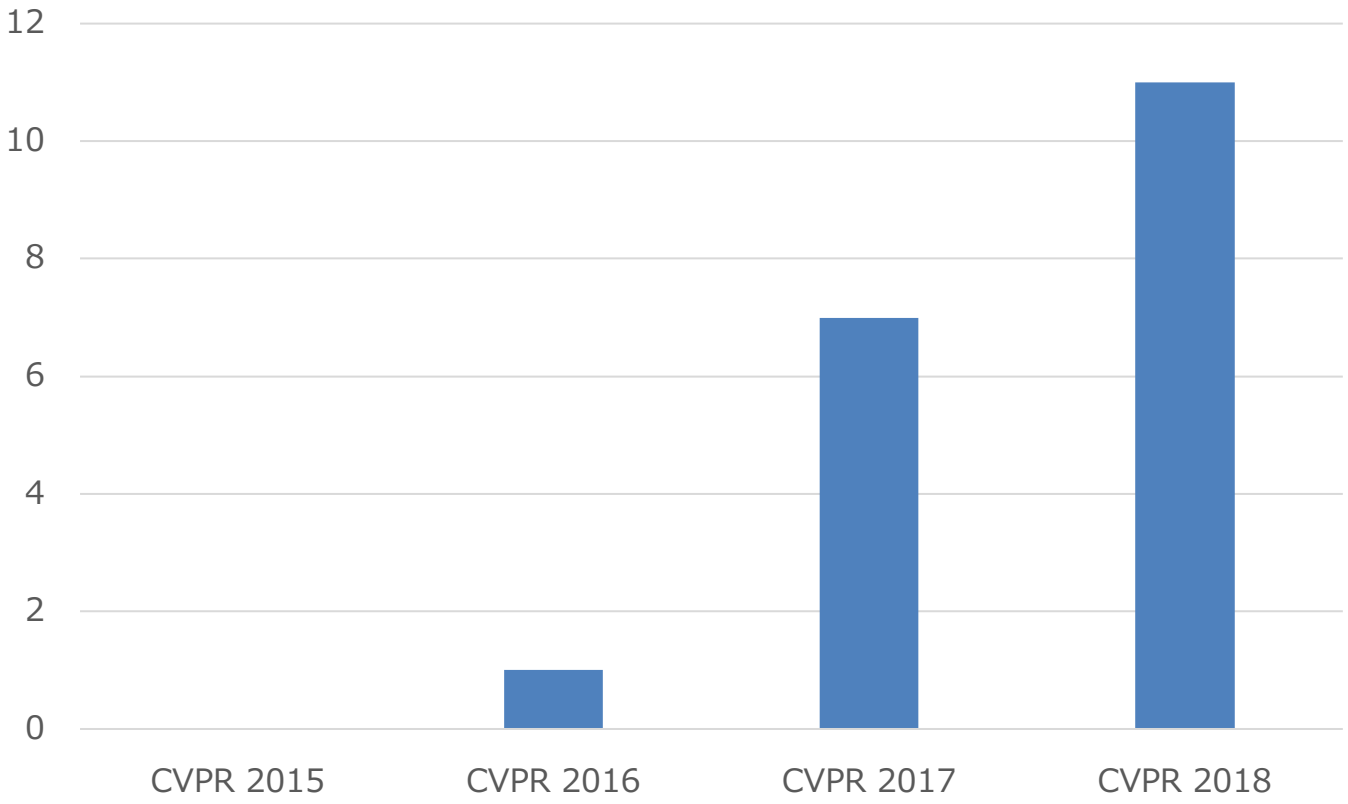


(4) http://cseweb.ucsd.edu/~haosu/papers/cvpr18_geometry_predictive_learning.pdf

CVPR2018の動向・気付き（31/61）

– 強化学習をCVのタスクに適用した論文も増加

題名にReinforcementという単語を含む論文



CVPR2018 は１１本に増加

CVPR2018の動向・気付き（32/61）

– 強化学習をCVのタスクに適用した論文も増加

特定のタスクへの適用

- Video Fast-Forwarding http://vcg.engr.ucr.edu/publications/Shuyue_CVPR.pdf
- Video Captioning https://arxiv.org/abs/1711.11135
- Image Cropping https://arxiv.org/abs/1709.04595
- Color Enhancement https://arxiv.org/abs/1804.04450

既存手法の精度向上

- Joint Optimization:
  http://openaccess.thecvf.com/content_cvpr_2018/papers/Xie_Environment_Upgrade_Reinforcement_CVPR_2018_paper.pdf
- Hyperparameter Optimization
  http://openaccess.thecvf.com/content_cvpr_2018/papers/Dong_Hyperparameter_Optimization_for_CVPR_2018_paper.pdf

CVPR2018の動向・気付き（33/61）

– 強化学習をCVのタスクに適用した論文も増加 続き

- いずれの論文もタスクをマルコフ過程として定式化
  - 代表的な手法としてDQN/A3Cを使用
- Action/Rewardの定義のみ与えれば教師データが少なくてよい
  - そうは言ってもこれが難しい！
  - 実空間の問題を扱いたいが（特徴という意味でも）探索範囲が広い
- 複雑ネットワークでなくてもよい
  - シンプルでないとパラメータを最適化できない？

今後も強化学習が適用可能なタスクの探索研究は増加する！

Semantic Segmentation, Object Localization,
Saliency Estimation, 3D Shape Learning などは早い者勝ち？

CVPR2018の動向・気付き（34/61）

– Blur周りでどのようなことがされているか

- Blur除去
  - GAN，U-Netで除去してるのが今回発表されている
  - セマンティックセグメンテーションを利用しているのもある
- Blurが望ましいか否かと，それが写真のクオリティに良い影響を与えているかどうかを判定
- 1枚のBlur画像から時系列フレームを推定して動画像を生成

一般的には不要な情報を有効活用する考えが面白い

CVPR2018の動向・気付き（35/61）

- Meshを生成する方法自体が徐々に精度を上げている
  - 数年前はノイズがあり散々な結果だった
    - それでもコンセプトが認められ論文自体は通っていた
  - すでに3Dプリンタ生成に耐えうる手法も
    - AtlasNet: https://arxiv.org/pdf/1802.05384.pdf
  - 3D MeshのレンダリングについてBackProp.ができる手法
    - 原田・牛久研の加藤氏

メッシュ生成，
前はこんな感じ

今はこんな感じ！

正解はこれ

日本発のNeural 3D Mesh RendererもCool!

http://hiroharu-kato.com/projects/neural_renderer.html

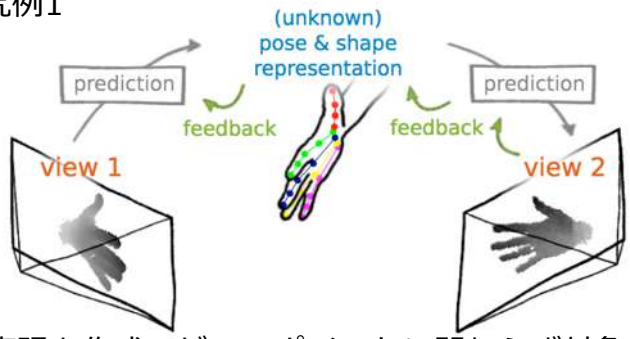
https://arxiv.org/pdf/1802.05384.pdf

CVPR2018の動向・気付き（36/61）

– 間接的視点情報を弱教師付き学習に利用

- 研究例1
  - ビューポイント変換を相殺するような中間表現を生成
  - 中間表現から変換して誤差を計算，フィードバック
- 研究例2
  - 入出力1: ある視点の2次元画像から3次元形状を復元
  - 入出力2: 入力1とは異なる視点の2次元画像からカメラ角度推定
  - 正解値: 入力2のシルエット画像
  - 誤差計算: 出力1の3次元形状を，入出力2のカメラ角により回転，入力2のシルエットと変換したシルエットを比較

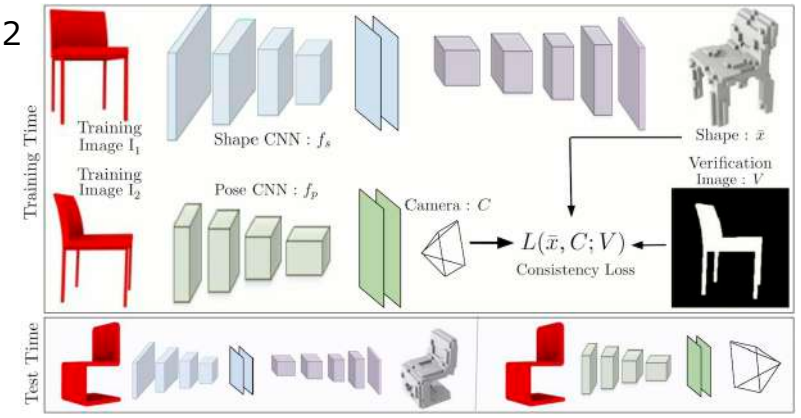
研究例1



中間表現を作成，ビューポイントに関わらず対象の姿勢を計算できるように（視点によらず姿勢を計算）

http://openaccess.thecvf.com/content_cvpr_2018/papers/Poier_Learning_Pose_Specific_CVPR_2018_paper.pdf

研究例2

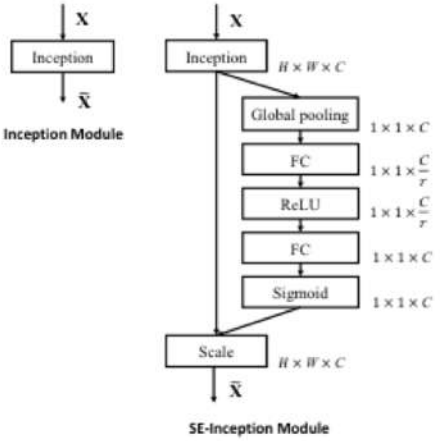


https://shubhtuls.github.io/mvcSnP/

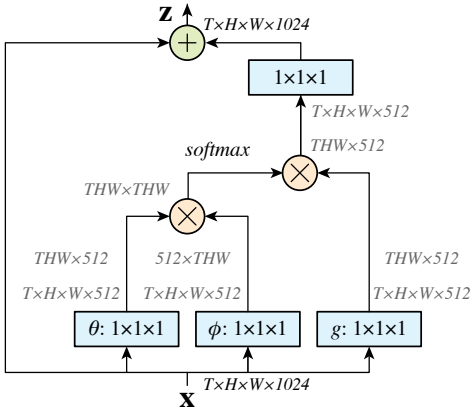
CVPR2018の動向・気付き（37/61）

– 他分野の技術が積極的に取り入れられる

- Attention：あまり無かった1枚の画像のAttentionも出てきた
  - Non-local Neural Network：NLPっぽいAttention機構の使い方
  - Squeeze-and-Estuation Network：CNNに特化したAttention機構
- Embedding：画像＋αの組み合わせ
  - 言語から画像中の物体の位置を推定
  - 音から画像中の物体の位置を推定



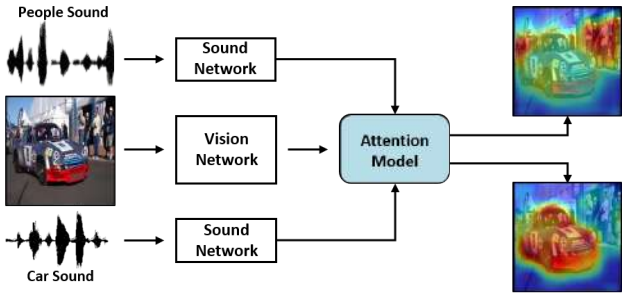
SE block



Non-local block

Attention

与えられたワードから物の位置を推定



与えられた音から物の位置を推定

Embedding

CVPR2018の動向・気付き（38/61）

– Language & Visionの研究が加速

- 言語情報をナビゲーションに使用する動き

  例1）言語情報を用いてロボットに目的地へ向かわせる

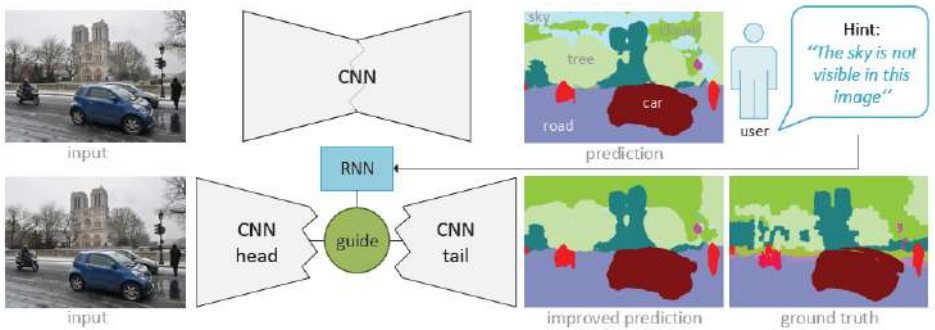
  例2）言語情報によってCNNの予測結果を修正する

- VQAの研究も健在(Questionをタイトルに含む論文22本)

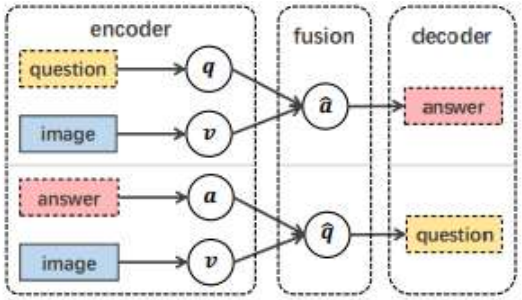
  例3）質問への回答(Answer)と質問の作成(Generation)を同時に学習

  例4）VQAにおけるAdversarial Attackの検証

例1



例2



例3

例4

CVPR2018の動向・気付き（39/61）

– Language & Vision続き

- 会場でポスターを見た感じ，大まかに：Image/Video Captioning, VQA(TQA, VDなど含む), Phrase localization (言語の画像中の領域を抽出)，Visual Reasoning (広義にはL&V) などのサブ分野がある
- CVPR本会議のスレッドのうちの1つ ”Object Recognition & Scene Understanding”のなかに“Language＆Vision”の論文が増加傾向
- 全体的に，CVPR2018においてL＆Vの研究が2017年より注目を浴びている傾向が見られる (論文数，オーラル発表数，会場でポスターを見る人数)
- Language＆Visionにおける“VQA”の研究の勢い ＞ “Image/Video Captioning”, 今後は更に難しい問題設定に移行する傾向 (例:Embodied Question Answering, Visual Dialog, TextQA, IQAなど)
- Language(かSound)＆Visionの研究が他分野の知識と融合し，応用分野の広がり・更にシーンなどに対する深層理解などができ，DNNやmulti-modal技術の成熟により今後は更に期待されると感じる

CVPR2018の動向・気付き（40/61）

- Visual Question Answering (& Visual Dialog)
  - VQA(VD)の勢いがすごい！(Oral:6/70,Spotlight:約20/224)
  - VQAの重要スポット時間軸：
    - 2014: 問題提出 ->2015: データセットVQA1.0
    - > 2016: VQA Challenge -> 2017: VQA2.0, VQA Challenge 2nd
    - > 2018: VQA Challenge 3rd (40teams) -> Who's the next?
  - VQAの技術遷移：LSTM-Q Norm-I -> MCB -> NMN -> Attention
  - 現在に至るVQAの技術的改善の議論は主にattentionの改善が中心
    例: Oralでは2本attentionの改善(下)、posterも数本
    - Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering(faster-rcnn　ベースなattention領域検出を導入)
    - Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering(stack可能なDense Co-Attentionモジュールの提案)

CVPR2018の動向・気付き（41/61）

– Visual Question Answering (& Visual Dialog)

- VQAの従来問題点“Vが重視されない”解決への挑戦も待った残ってる
  - 代表例①②：逆問題(回答画像から質問を予測)を導入し，Vを重要視
    - » Visual Question Generation as Dual Task of Visual Question Answering
    - » IVQA: Inverse Visual Question Answering
  - 代表例③：データセットをバランスよくする
    - » VQA1.0－＞VQA2.0（2016年の研究で，VQA1.0をもっとバランスよく調整する）
    - » VQA2.0 →C-VQA : Don’t just assume:　Look and Answer: Overcoming Priors for Visual Question Answering
- ホットな領域の方法(例:Domain adaption/GAN/Unsupervised)などをVQAに応用する研究も多い (ので，この方向で通りやすい？)
- 今後は精度向上よりは，新規な問題設定・更なる複雑な問題設定(TQA，VD，IQA，Embodied QA)に向けて発展していく

CVPR2018の動向・気付き（42/61）

– Visual Question Answering (& Visual Dialog)

- Visual Dialogは提案されたばかりだが，2018のCVPRではVisual Dialogの論文数がVQAと同レベルくらい
- 来年はVD＞VQAになる？(CVの人もNLPに適応し始めた？)
- TQA(Text QA)も今後増える傾向？今回以下のTQAがある
  - Focal Visual-Text Attention for Visual Question Answering
  - Textbook Question Answering under Instructor Guidance with Memory Networks
- Language-Vision Navigate, Interactive環境でのVQAが今後の動向？
  - Embodied Question Answering
  - Vision-and Language Navigation:Interpreting visually-grounded navigation instructions in real environments*
  - IQA: Visual Question Answering in Interactive Environments

* （第一作者がBottom-up,Top-downの著者，第二作者がVisual Dialogのオーラル発表の人なので，今後やっぱりVision-and Language Navigationが正解？更にVisual-Dialogも融合？）（しかも，この論文が主にデータセット提案なので，今後このデータセットに対して強い研究が出されそう！！）

CVPR2018の動向・気付き（43/61）

– CNNの解釈はやはり注目されているトピック

- Interpretable Convolutional Neural Networksなど
  ポスター聴講者が多い
- What Do Deep Networks Like to See? とかも
- What Makes a Video a Video や What Have We Learned From Deep Representations for Action Recognition?
  など動画でも分析系が複数通っている
- ML系とは違い，理論的な解析とか解釈というよりは
  可視化などを通して感覚的に理解しようとしているものが多い印象

CVPR2018の動向・気付き（44/61）

– リアルタイム化，省メモリ化が加速した？
  - 現実に適用可能な技術の確立を目指す

– 理由？
  - 現実指向派がCV研究の速い流れに追いついた？
  - 精度指向が一定の満足を得た？
  - 研究成果を生み出す制約の一つとしてみなされた
  - 低予算系研究グループの反乱！？？？
    – かつてはCVといえば低予算でできる研究分野だった

CVPR2018の動向・気付き（45/61）

– DNNの更なる横の広がりの加速

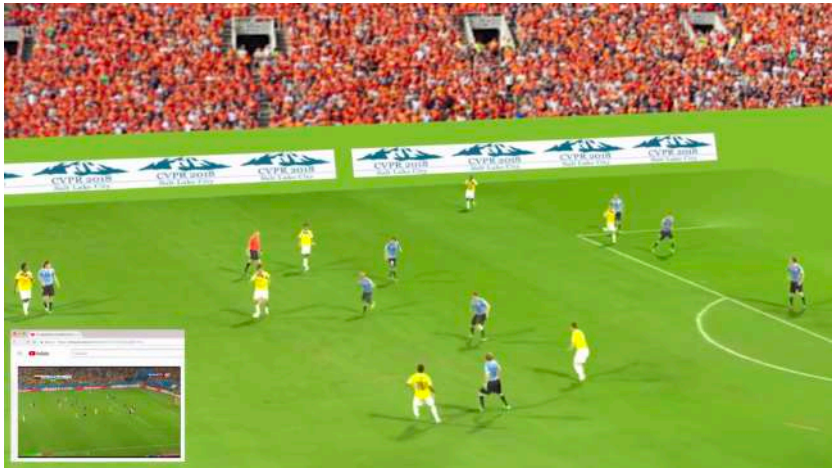
- 「微分不可能な問題を微分可能にした」系がちらほら
- 数学勢の参入？
- パターン認識系における，かつての現実問題のモデリング→精度向上のやり方の雰囲気を感じる

- AI(DNN）が世の中のあらゆる問題に対応するには，まだまだ人の手を借りないとムリ？

CVPR2018の動向・気付き（46/61）

– 新しいCVPR論文の通し方？

- Soccer on Your Tabletop
  - http://grail.cs.washington.edu/projects/soccer/
- すごい実装力！スポーツの面白さを倍増させる！！
- 査読者だったら動画見ただけで採択を決めたくなる？
  - 数値的な実験結果があまりないが，デモや動画に命を吹き込んだ
- 日本でもプロジェクトが行われていた自由視点３次元映像の発展版
  - http://www2.coara.or.jp/cgi-bin/demo/read2.pl/02/110010/0375



YouTubeなどの動画像からサッカーシーンを復元，Hololens等でVR映像を確認

CVPR2018の動向・気付き（47/61）

– Software Engineering in Computer Vision Systems

- MS/AWS/Fyusion/Kitware/SIEMENS/Uber/HERE各社の中心エンジニアの開発成果を知る機会
  - https://www.here.com/en/secvs-cvpr-2018
- 論文にはならない開発について語るチュートリアル
- トピック: オープンソース, データ作成, モデル作成, テスト, ストレージ, 製品開発, 多言語開発

研究・開発を手助けするサポート体制が整っている

（深層学習で言うと，おそらくCaffeから始まっている）

CVPR2018の動向・気付き（48/61）

- IT企業にはパラメータチューニング屋さんがいるのでは？
  - ベースラインの精度が異常に高い論文があった（が，実現不可能ではない）
  - 会場にていろんな議論した結果，ベースの精度を異常に高くする専門家がお抱えでいるのでは？となった
  - 論文の通しやすさにダイレクトに関わる！
- 誰にも作れないラベルを作り出すアノテータもいる？
  - ADE20K Dataset
    - Our dataset was annotated by a single expert annotator, providing extremely detailed and exhaustive image annotations（2万枚もあるSem.Segment.のラベル付を一人のエキスパートが行なっている）となり話題になった http://people.csail.mit.edu/bzhou/publication/scene-parse-camera-ready.pdf
  - トップアノテータ（Top Annotator）というAIにラベルを提供する先生役がいてもよい！

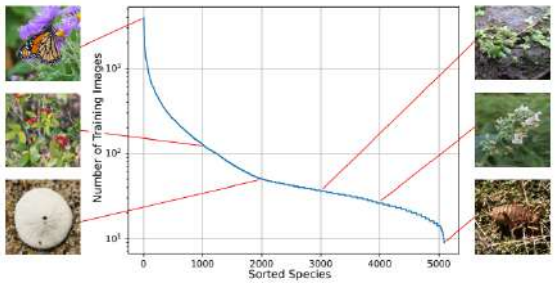
    日本も，誰もが欲しがるデータを作る人/モデルの精度を上げる人に対してお金を出すのはいかがでしょう？（但し年収数千万です）

CVPR2018の動向・気付き（49/61）

– チャレンジングなデータセットの提案

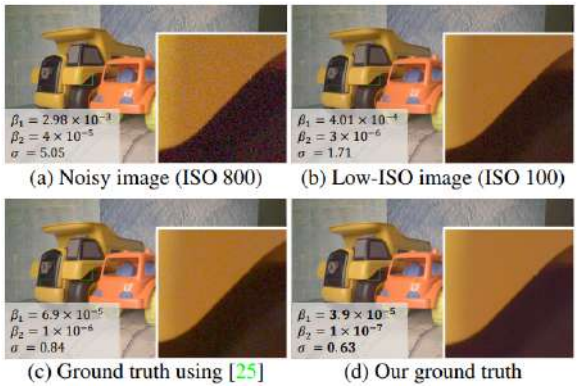
- 画像の不均衡、アノテーションや画像のノイズを残したDBの提案
- また、それらのDBを使って学習問題を解くことで今日までのDB問題を解決しようという試みも
- 研究例1：画像数が極端に偏ったDB
- 研究例2：ノイズが乗ったスマホ画像DB
- 研究例3：ラベルノイズが乗ったDBで反復学習

研究例1



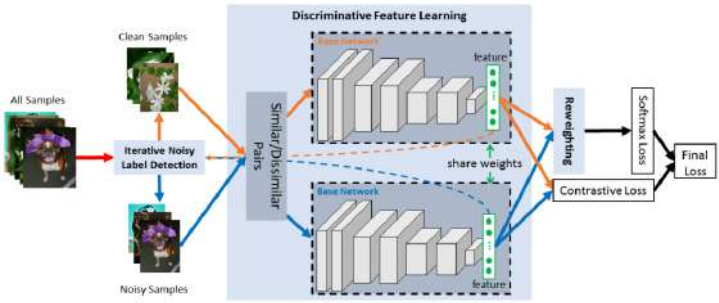
グラフがクラスごとの写真の枚数を表す
https://arxiv.org/abs/1707.06642

研究例2



スマホ特有のノイズが乗っている
http://www.cse.yorku.ca/~mbrown/pdf/sidd_cvpr2018.pdf

研究例3



ラベルノイズを検出しつつ学習
https://arxiv.org/pdf/1804.00092.pdf

CVPR2018の動向・気付き（50/61）

- データセットを人様が作らず，自動的に作る動き？
  - Webでの収集画像を学習する際，間違ったラベル（ノイズ）付き画像を除去する研究を複数確認
- 日々増加するWebデータを正しいラベルに仕分け
  - 学習量増加により精度の向上につながる?

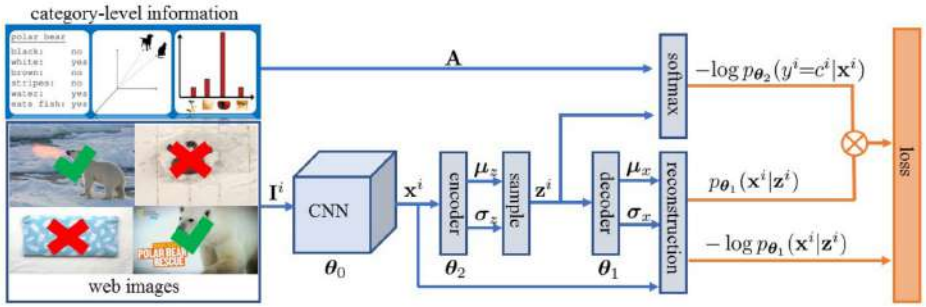
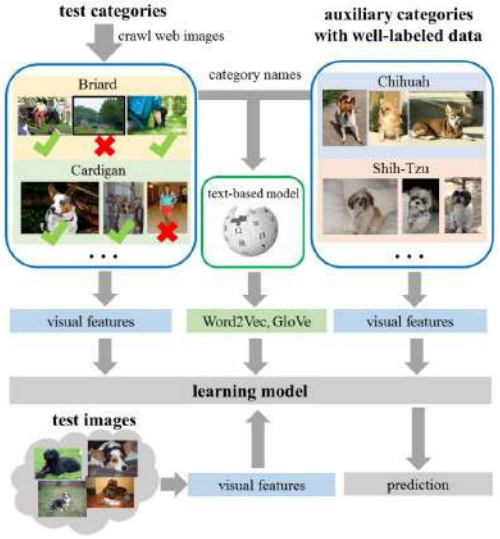


Fig. 1: Flowchart of learning from noisy web data with category-level semantic information. The top flow is classification network and the bottom flow is variational autoencoder. Two flows share the common model parameters θ_0 and θ_2.

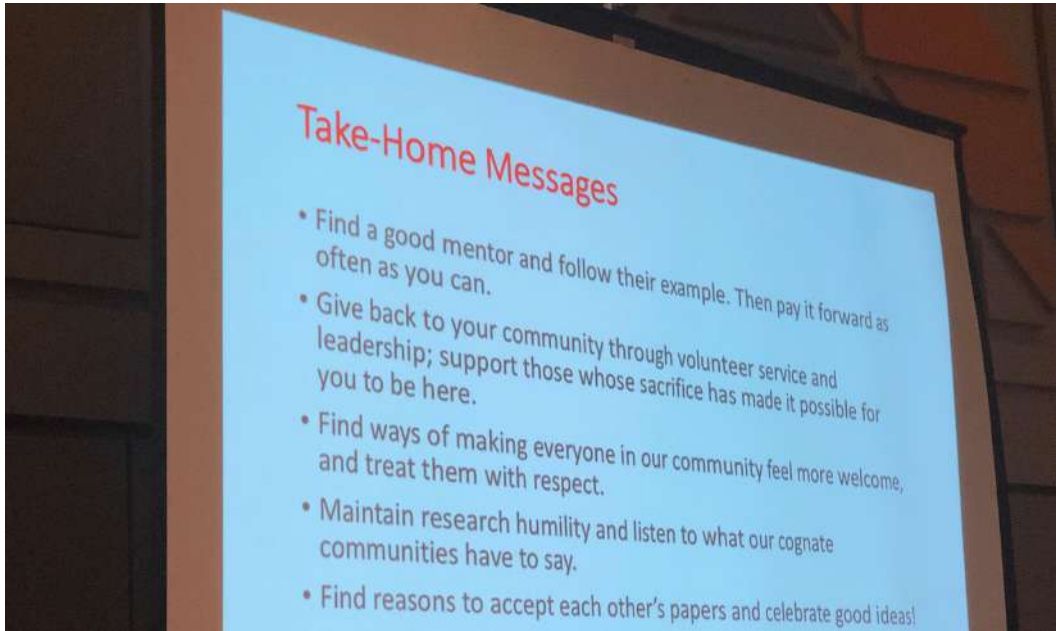
Learning from Noisy Web Data with Category-level Supervision
https://arxiv.org/pdf/1803.03857.pdf



Webly Supervised Learning Meets Zero-shot Learning: A Hybrid Approach for Fine-grained Classification
https://arxiv.org/pdf/1711.11585.pdf

CVPR2018の動向・気付き（51/61）

- How to be a good citizen of the CVPR community
- Sven Dickinson https://www.cs.toronto.edu/~sven/
  - CVPRコミュニティが多様化して来て，強くなって来た
  - 壁を作らない！オープンマインドで行こう！
  - Rejectする理由を書こう！環境が研究者を育てる
  - 良いメンターを見つけ，ついていこう！
  - 良いアイディアをたたえよう！アクセプトする理由を見つけよう！



https://pbs.twimg.com/media/DgTxwoMUwAEcdNc.jpg:large

CVPR2018の動向・気付き（52/61）

– Vladlen Koltun http://vladlen.info/

- SoTAを試してみよう
  - 自分で実装/あればコードを落として試す
  - パラメータをいじったり，異なる文脈で試したり．．．
- Quality over quantity
  - コミュニティにノイズはいらない（979本もいらない）
  - 必要なのは意味のある貢献だけだ（強い）
- Research over time
  - 大きなゴールを設定して何度も立ち返ろう
  - たくさん読む，アイディアを書き下そう
  - コミュニティが知らないことをやる，そのために現状を知る
  - じっくり読む/書く/考える時間を確保しよう
- メンターが読むべき論文を選ぶ，導く
- ネガティブな結果も面白い
- どんな時もアイディアを考えよう

CVPR2018の動向・気付き（53/61）

- Adriana Kovashka http://people.cs.pitt.edu/~kovashka/
  - Research in Context
  - Organizing workshop（ワークショップを企画する）
    - むずかしい，苦しいではなくて楽しい！
    - 良い人脈を築く
    - コミュニティにとって有益
  - 女性をエンカレッジするコミュニティ
    - Women in CV https://wicvworkshop.github.io/
    - Olga & Fei-Fei’s AI4ALL http://ai-4-all.org/
  - 学部生を育てよう
    - 面白い視点を持っている
    - 良い学生は必ずいて，良い視点を持っている/いずれ良い研究をする

CVPR2018の動向・気付き（54/61）

- 裾野が広がり，基礎的な内容/個別化研究に別れた？
  - 基礎的な内容：アーキテクチャ/学習戦略/誤差計算などすべての領域に対して使用できるような学習
  - 個別化研究：ドメインに合わせて深化している，CVのみならず周辺分野/関連分野に対する知見が必要

CVPR2018の動向・気付き（55/61）

– AI分野の中心はどこにあるのか？

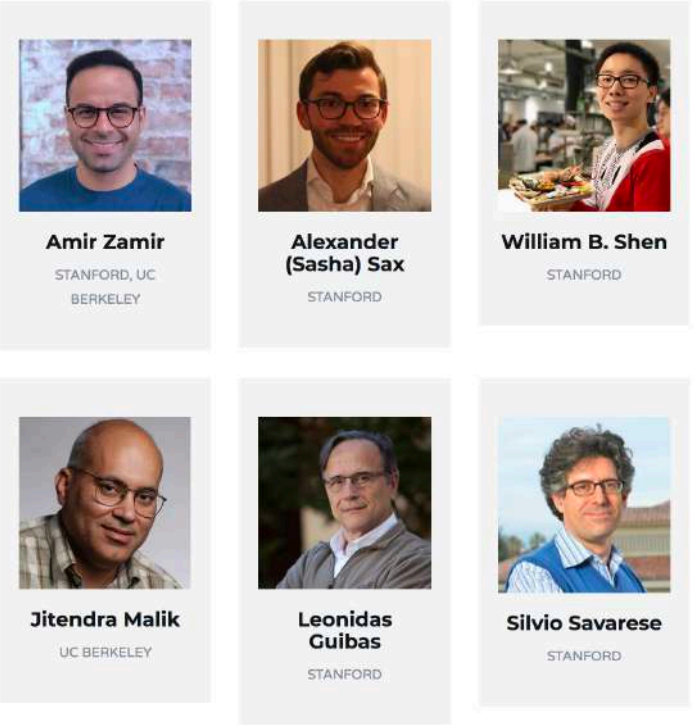
- 米国：「中国に抜かれる」，中国：「米国にはまだ勝てぬ」と言う
- 一方，米国の大学：「巨大IT企業の流れについていくのは大変」
- 巨大IT企業：「まだまだ人材が足りぬ」

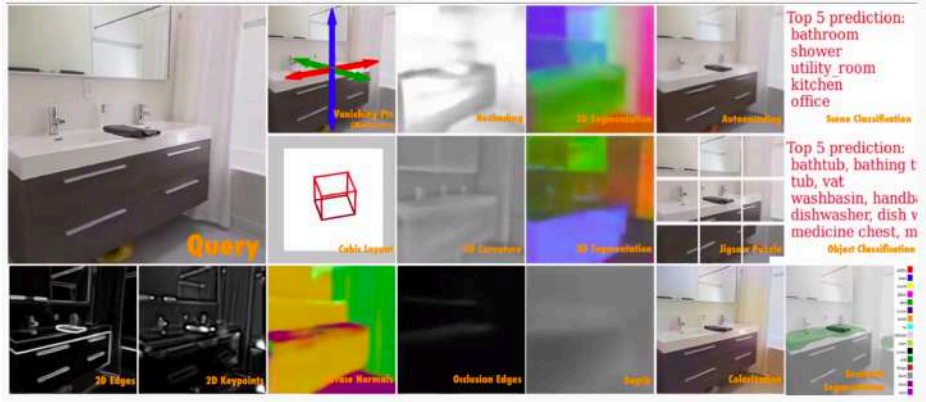
中心なんて無くて，世界的にAIを取り扱う大学・企業など研究機関が巨大な生き物のように見えている？

CVPR2018の動向・気付き（56/61）

– 研究機関の枠を超えて良い研究が生まれてくる

- Ph.D./Masterの大学院生が他大学/企業でインターン
  - 国際会議で履歴書を持って直接話しかける姿も見かける
  - 一方で，元々の知り合いが共同研究をやっている例も多い
  - その研究をちゃんと国際会議に通している人も多い
  - 何の気なしに有名研究者にメール送っても（世界からメール来すぎて）返ってこない
- Best PaperのTaskonomy，3研究室が合同で問題を解いている





データセットも大規模に収集

Best Paperの著者リスト

ベイエリアオールスターズという感じ

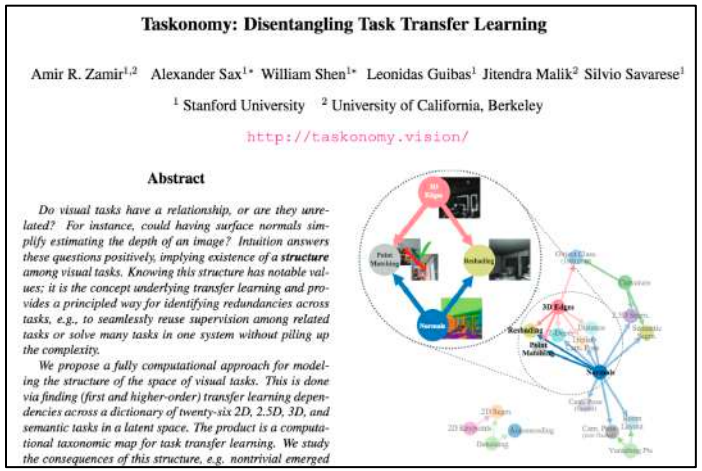
http://taskonomy.stanford.edu/

CVPR2018の動向・気付き（57/61）

– Best Paperについて

- 26種のタスク間の関連性を調べる
  - CVの歴史の中で別々に議論されたいたサブタスクを繋げる
  - しかし，膨大な時間の学習（47,886 GPU hours. . . ）
- 思いついたとしても，誰もやらなかったアイディアに真っ向勝負！
- 失敗したら膨大な課金に対し成果ゼロ
  - だが，リスクを承知で結果的に最大の業績（3,300論文中の1位）を獲得した！

小さくまとまるよりも，リスクをとって分野に最大の貢献を！



Taskonomy: Disentangling Task Transfer Learning

Amir R. Zamir[1,2] Alexander Sax[1*] William Shen[1*] Leonidas Guibas[1] Jitendra Malik[2] Silvio Savarese[1]

[1] Stanford University [2] University of California, Berkeley

http://taskonomy.vision/

Abstract

Do visual tasks have a relationship, or are they unrelated? For instance, could having surface normals simplify estimating the depth of an image? Intuition answers these questions positively, implying existence of a ***structure*** *among visual tasks. Knowing this structure has notable values; it is the concept underlying transfer learning and provides a principled way for identifying redundancies across tasks, e.g., to seamlessly reuse supervision among related tasks or solve many tasks in one system without piling up the complexity.*

We propose a fully computational approach for modeling the structure of the space of visual tasks. This is done via finding (first and higher-order) transfer learning dependencies across a dictionary of twenty-six 2D, 2.5D, 3D, and semantic tasks in a latent space. The product is a computational taxonomic map for task transfer learning. We study the consequences of this structure, e.g. nontrivial emerged

これぞStanford！という感じの論文

CVPR2018の動向・気付き（58/61）

– IT企業を中心に目立っている

- SenseTimeは44本論文を通す，が話題になった！ https://www.sensetime.jp/single-post/2018/05/15/CVPR-2018%E3%81%AB44%E6%9C%AC%E3%81%AE%E8%AB%96%E6%96%87%E3%81%8C%E6%8E%A1%E6%8A%9E
- Google 54本 https://ai.googleblog.com/2018/06/google-at-cvpr-2018.html
- Facebook 35本 https://research.fb.com/facebook-research-at-cvpr-2018/
- NVIDIA 19本 http://www.nvidia.com/object/cvpr-2018.html
- Amazon 11本 https://www.amazon.jobs/jp/landing_pages/CVPR
- 内情を見ると大学/企業の両方に肩書きがある人も

本数が全てではないが，数の暴力はある

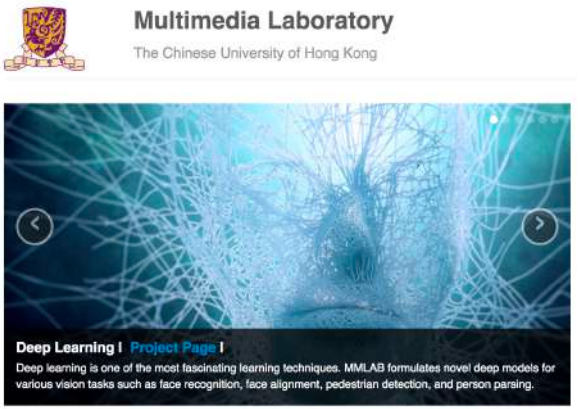
CVPR2018の動向・気付き（59/61）

- SenseTimeがなぜこんなに論文を通せるか？
  - （下記は推測も含みます）
  - トップ研究者であるProf. Xiaoou Tangが会社を設立
  - 大学研究室CUHK MMLab./SenseTimeを両輪で成長
    - 実際にOBがSenseTime入りするケースもあり，内外部からインターンの受け入れもあり？で成長する仕組みが整う
  - 潤沢な資金（資金調達で6億USD(630億円程度)を獲得 https://glotechtrends.com/sensetime-alibaba-funding-180416/ ）/豊富な研究設備

研究が進む/人材が成長するエコシステムが整っている

2001年7月	Xiaoou Tang教授 MMLabを創立
2011年	世界に先駆けてコンピュータビジョンにディープラーニング技術を応用
2014年	顔認識技術で人間を超える性能を実現（Facebook社の性能も超える）
	ImageNet物体認識コンテストにて，Googleに次ぐ世界第2位
	SenseTime Group Limited 設立
2015年	ILSVRC2015の動画物体認識コンテストにて，世界第1位を獲得
2016年	（株）センスタイムジャパン設立
	ILSVRC2016で、Object Detection, Object Detection and Tracking, Scene Parsing の３部門で１位を獲得

https://www.sensetime.jp/history



http://mmlab.ie.cuhk.edu.hk/

CUHK MMLab.は教員7名，ポスドク/Ph.D.学生は合計50+名とも聞く

CVPR2018の動向・気付き（60/61）

– NVIDIAの隆盛ぶりが著しい

- NVIDIA主催のGTC (GPU Technology Conference; 右下図) は世界的に行われる
- 当然のようにCVPRでも勢力は強い
  - NVIDIA AT CVPR 2018（左下）
  - 採択論文は19本，受賞論文含む
- ワークショップでも賞景品としてGPUを提供

ゴールドラッシュでつるはしを売る！

（AI時代にGPUを売るNVIDIAは強いの意）

NVIDIA AT CVPR 2018 http://www.nvidia.com/object/cvpr-2018.html?ncid=so-ele-cr28-42860

SILICON VALLEY
March 26-29, 2018

TAIWAN
May 30-31, 2018

JAPAN
Sept. 13-14, 2018

EUROPE
October 9-11, 2018

ISRAEL
October 17-18, 2018

WASHINGTON, D.C.
October 22-24, 2018

https://www.nvidia.com/en-us/gtc/schedule/

CVPR2018の動向・気付き（61/61）

- 論文を投稿しよう！
- 出さないと通らないし落ちても次への経験値
  - 海外ではPh.D.学生/ポスドクが中心となり論文投稿
  - 日本は修士学生が中心（もちろん先生方のサポートもある）
  - Ph.D.学生が増えないからしょうがないとネガティブになるのは早い！（修士の研究プロジェクトでも論文は通っている; 下記もし他にもあればお知らせください）
    - Kuniaki Saito, et al. “Maximum Classifier Discrepancy for Unsupervised Domain Adaptation”（M2の研究成果; Oral ）
    - Yuki Fujimura, et al. “Photometric Stereo in Participating Media Considering Shape-Dependent Forward Scatter”（M2の研究成果; Oral）
    - Naoto Inoue, et al. “Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation”（M2の研究成果）
    - Takuma Yagi, et al. “Future Person Localization in First-Person Videos”（M1の研究成果; Spotlight Oral）
    - Tomoyuki Suzuki, et al. “Anticipating Traffic Accidents With Adaptive Loss and Large-Scale Incident DB”（M1の研究成果）
    - Daiki Tanaka, et al. “Joint Optimization Framework for Learning with Noisy Labels”（なんとB4の研究成果）

学生からするとメジャー会議への投稿0を1にするのは大きい！

これから引用されそう（流行りそう）な論文

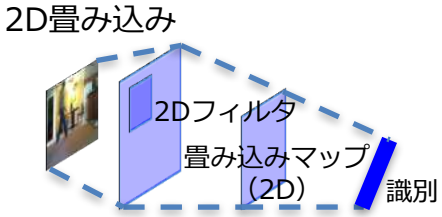
- すでに引用されている論文も含みます
- （取捨選択のためにもう少し時間が欲しかったですね）

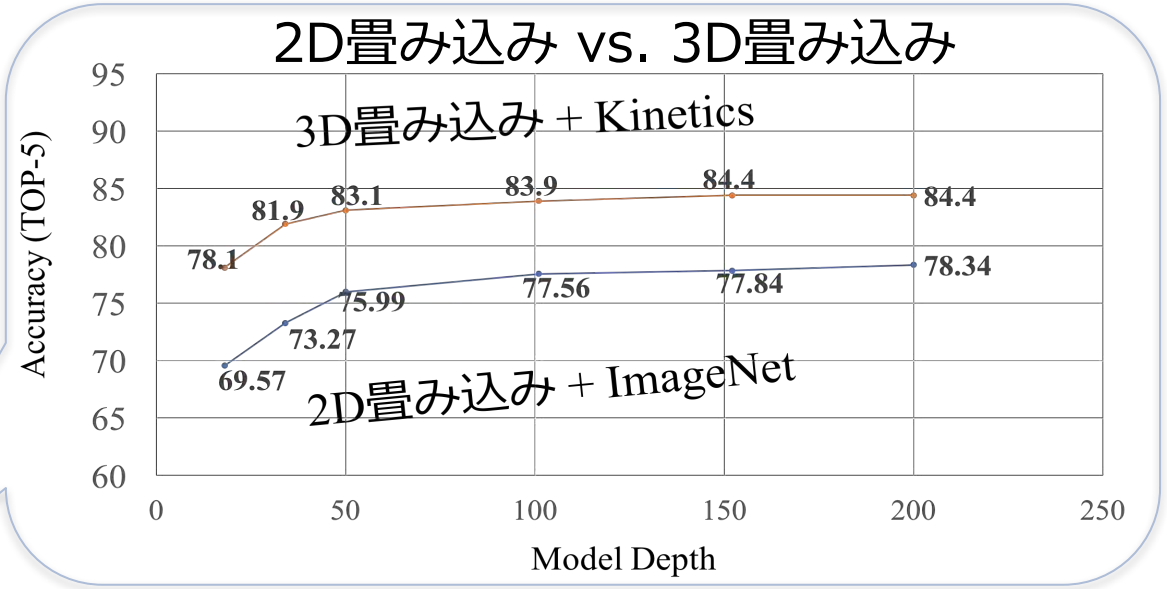
引用されそうな論文（1/26）

動画認識は画像認識(ImageNet)の歴史を辿るのか？

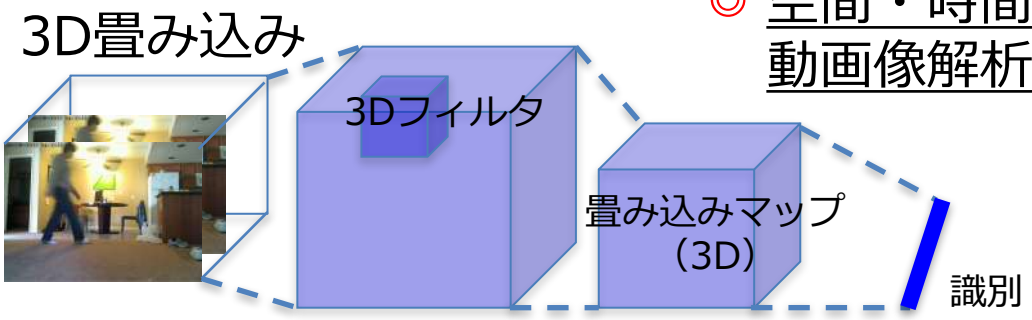
3D Convの解析

Kinetics datasetが動画像のImageNetに相当する可能性を明らかに





◎ 空間・時間を同時処理する3次元畳み込み法は動画像解析のコア技術となる



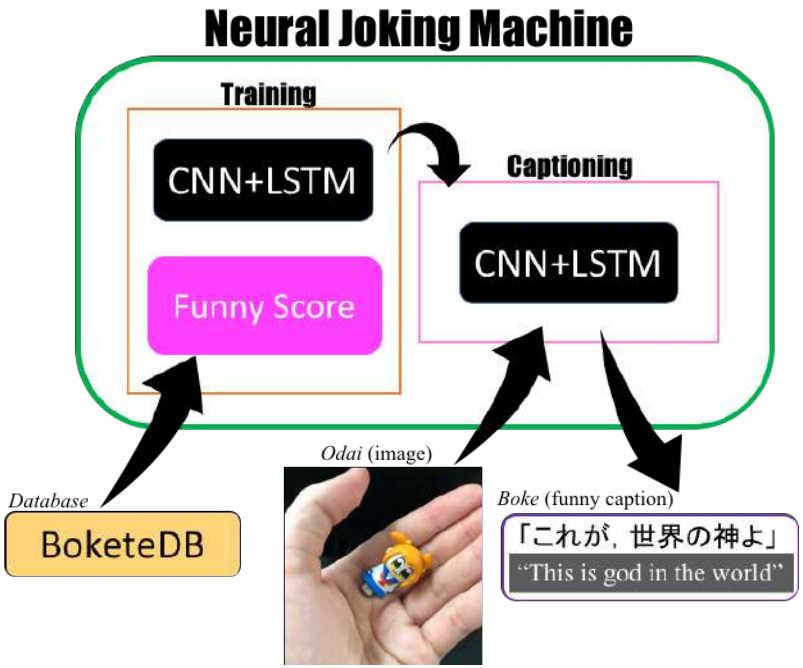
K. Hara, H. Kataoka and Y. Satoh “Can spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?”
https://github.com/kenshohara/3D-ResNets-PyTorch

~~引用され~~面白そうな論文（2/26）

CVPR2018 Language & Vision Workshop

Neural Joking Machine(NJM)

- ボケをかますAI
- Webサイト「ボケて」からデータ収集しBoketeDBを構築
- 学習時星の数によって重みを与えるFunny Scoreを定義

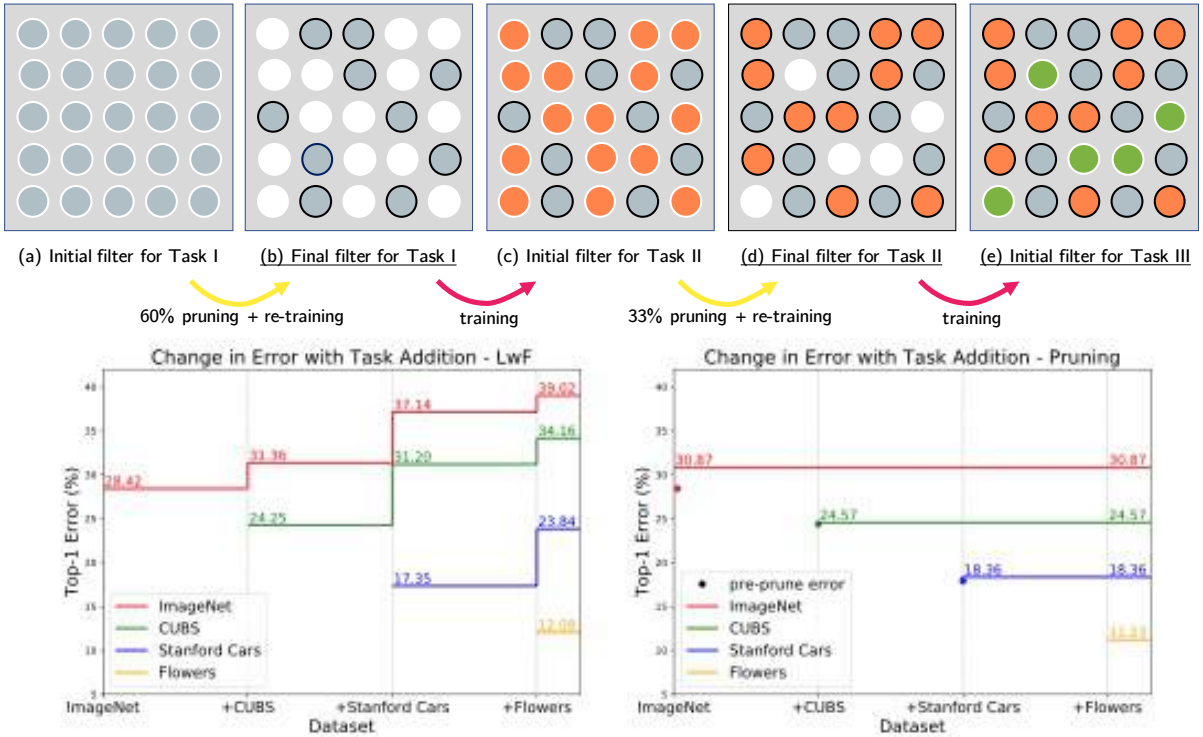


K.Yoshida, M.Minoguchi, K.Wani, A.Nakamura and H.Kataoka,"Neural Joking Machine : Humorous Image Captioning".
https://arxiv.org/abs/1805.11850

引用されそうな論文（3/26）

PackNet

- 過去のデータセットの記憶を残しつつ複数のデータを学習
- プルーニングとファインチューニングの繰り返しで学習
  - ネットワークの各パラメータがそれぞれのデータセットに割り振られるような学習
  - 学習をサボってるパラメータを違うデータの学習で有効活用しながら学習

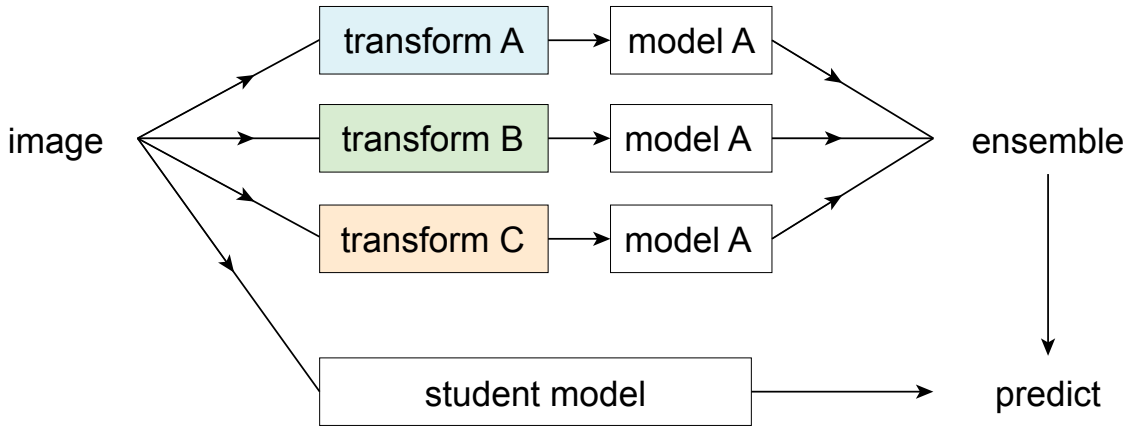


引用されそうな論文（4/26）

Data Distillation

– セルフトレーニングを導入したラベルなしデータの効率的な学習

- 変化を与えた複数のサンプルをモデルに与えて結果を受け取りstudent modelを学習
- ラベルなしデータを使ってCOCOの物体検出・Keypoint検出で高精度化



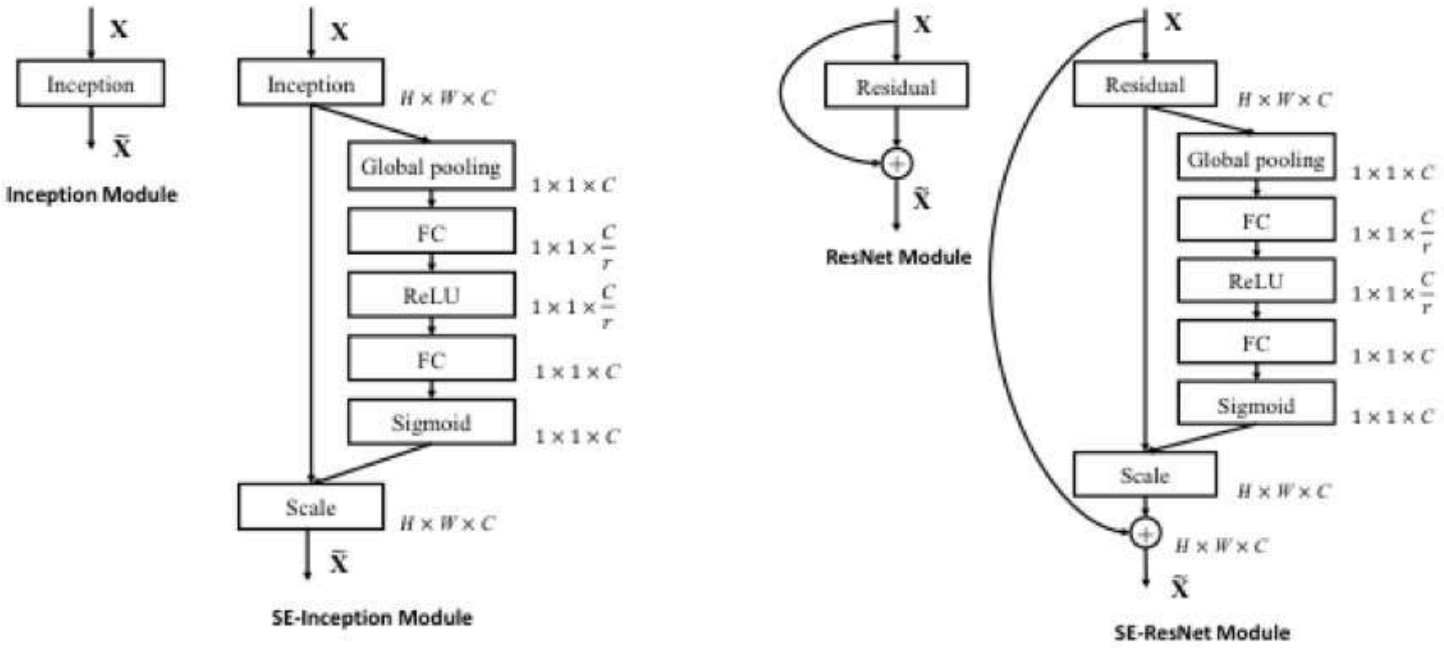
Data Distillation

引用されそうな論文（5/26）

Squeeze-and-Excitation Networks

– CNNのチャンネルにAttention機構を導入

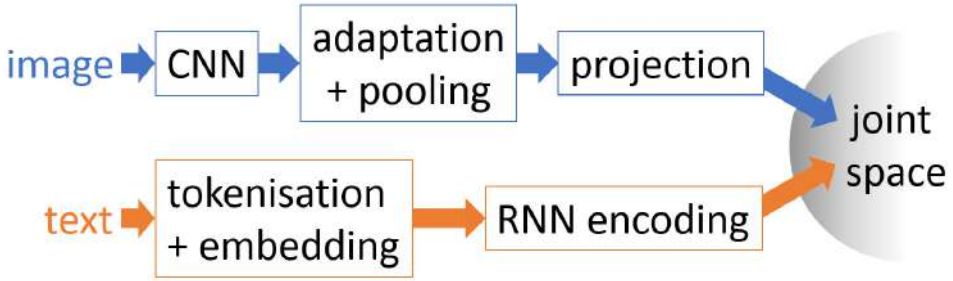
- 畳み込み層のチャンネルに対してAttentionを与えるSE Blockを導入
- ImageNet2017で1位のネットワーク
- すでにSENetを使った手法がNIPSの論文でちらほら見かける



引用されそうな論文（6/26）

Finding beans in burgers

- 画像中から与えられた単語の物体を見つける研究
- 単語&画像ベクトルを同一特徴空間で学習
  - 推定時はClass Activation Mappingに似た方法でヒートマップを出力してローカライズ

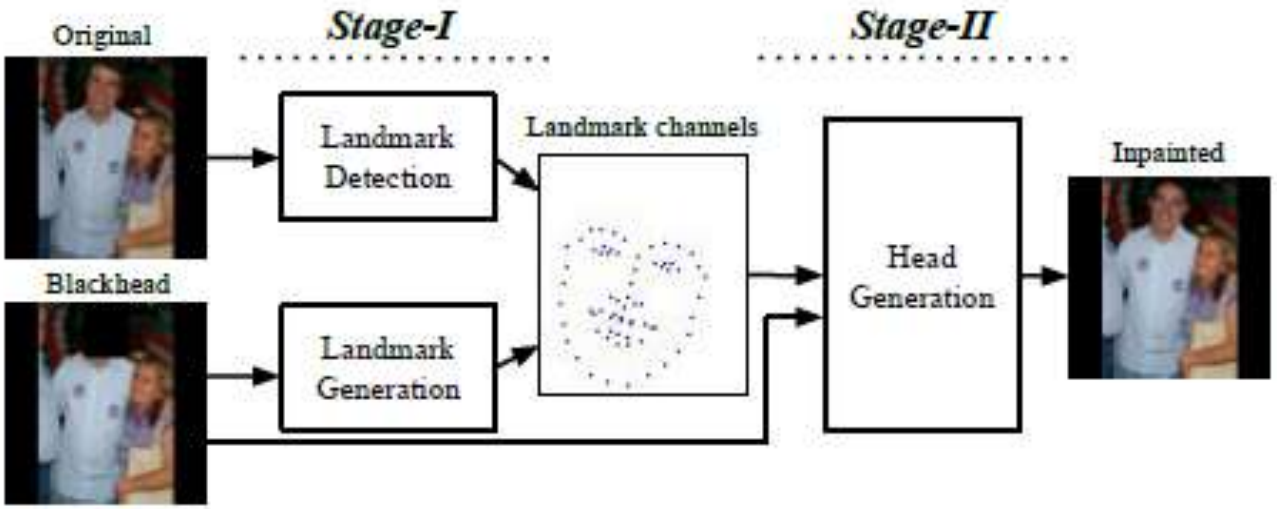


引用されそうな論文（7/26）

Natural and Effective Obfuscation by Head Inpainting

- プライバシー保護のために黒塗りするが，不自然さが残るので他人の顔（この世に存在しない顔？）で置き換える
- が，ホントに存在しないか不明

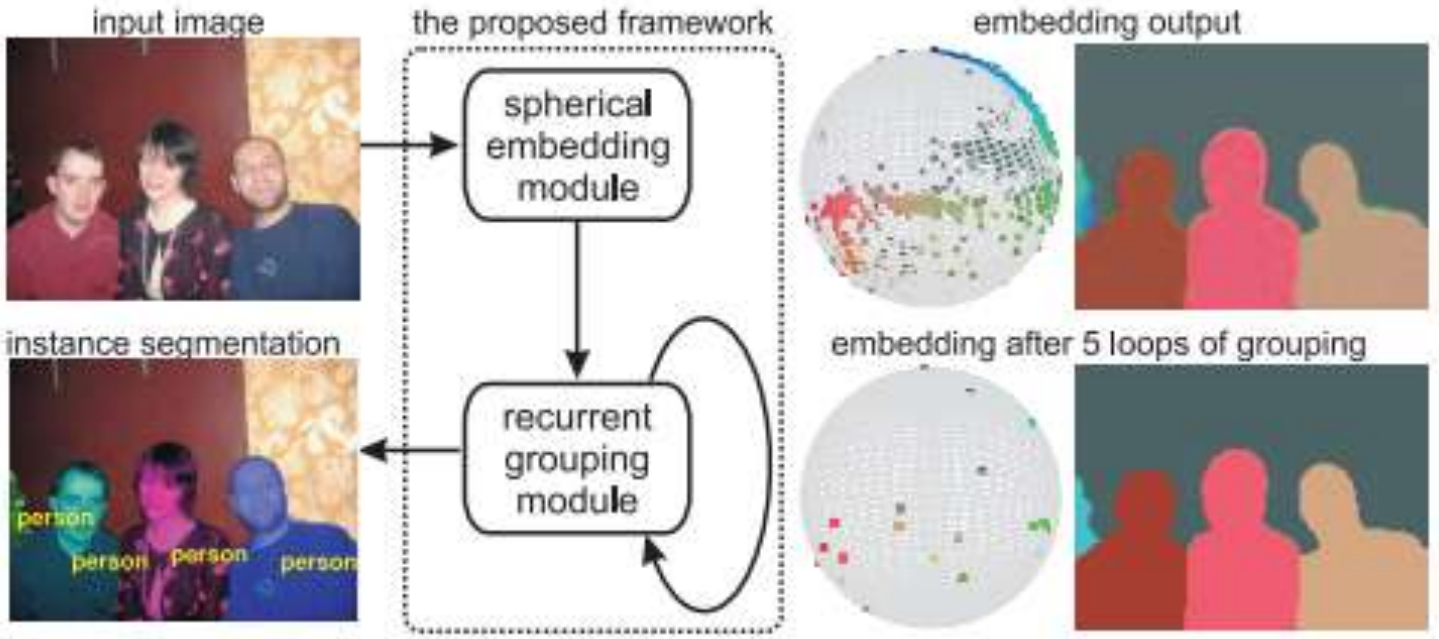
SNSでも賛否両論が起こった研究



引用されそうな論文（8/26）

Recurrent Pixel Embedding for Instance Grouping

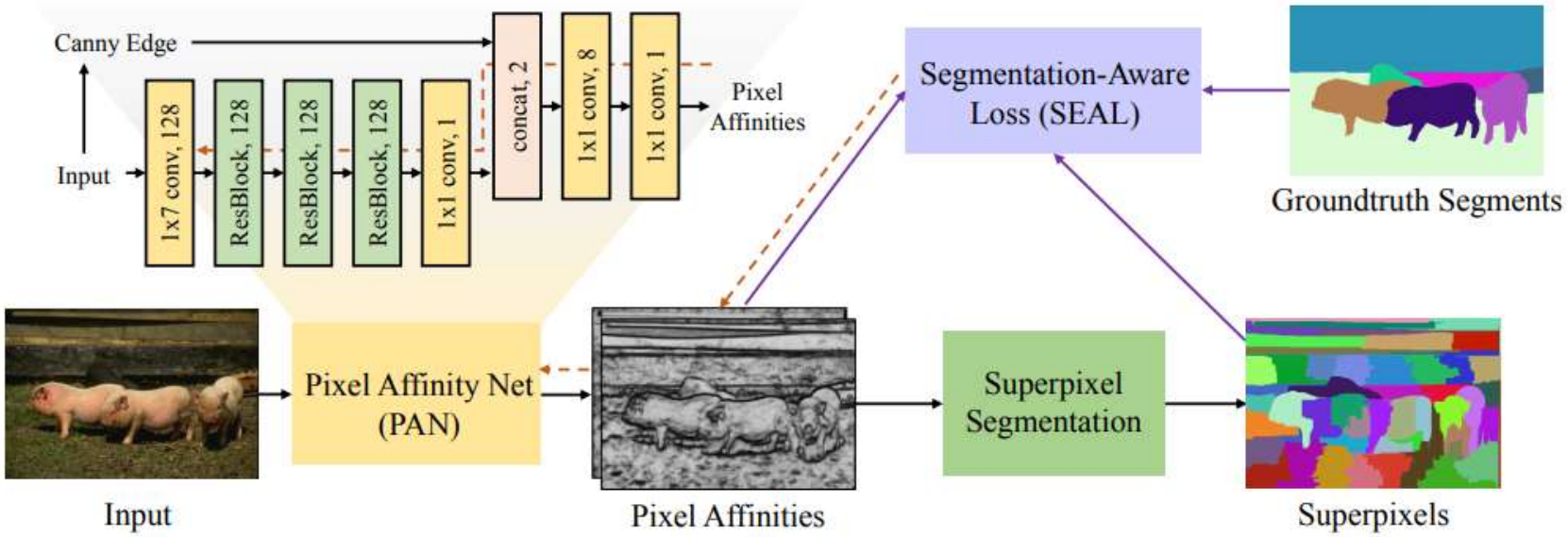
- End-to-Endインスタンスセグメンテーション
- 同領域の画素はCosSimが高くなるよう，異領域の画素はMargin以下の画素になるよう超球面状に回帰



引用されそうな論文（9/26）

Learning Superpixels with Segmentation-Aware Affinity Loss

- スーパーピクセル（SP）のためのピクセル類似度学習法
- セグメント計算の誤差から類似性を評価するPart Affinity Net (PAN)を提案，SPと深層学習を統合



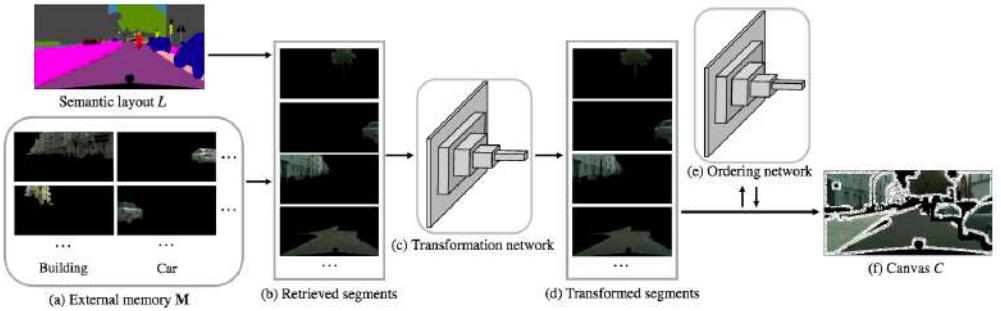
引用されそうな論文（10/26）

Semi-parametric Image Synthesis

– 意味ラベルから写真のようにリアルな画像を生成

– ノンパラとパラメトリックの両者を統合したセミパラメトリック法を提案，Conv-Deconv構造で最適化

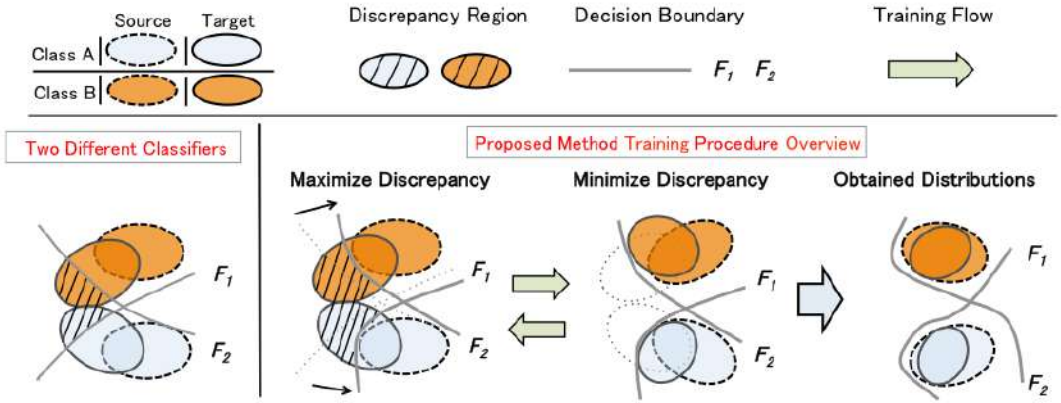
自動運転の学習画像はシミュレータ => セマンティックラベル => リアルで生成可能？



引用されそうな論文（11/26）

Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

- 目的タスクに特化した２つの分離境界を利用したドメイン適応
- 従来の埋め込み空間においてドメイン間の分布を単に近づける方法に対して、あるタスクと解くための分離境界を考慮して適応を行う



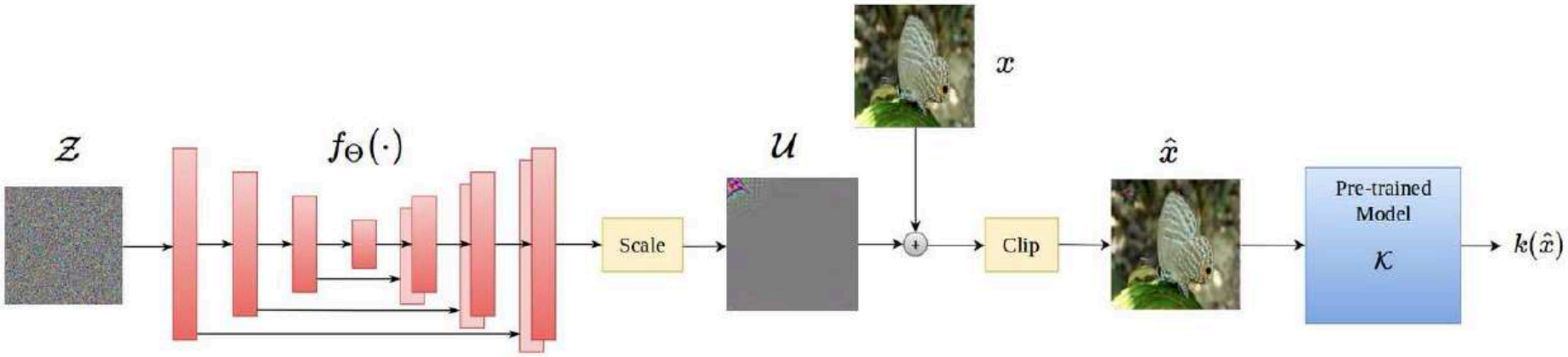
アイデアの面白さと同時に論文が非常にわかりやすかった

引用されそうな論文（12/26）

Generative Adversarial Perturbations

– 汎用的かつ画像依存性ありの摂動ノイズを，画像識別/セマンティックセグメンテーションに有効

– Universal Perturbationsの枠組みを生成モデルにより実装



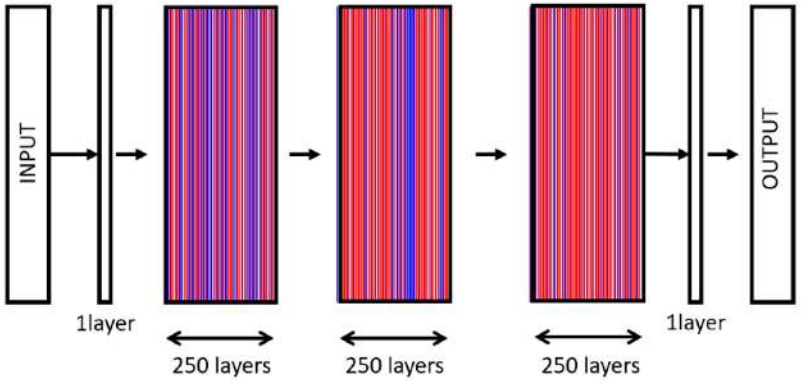
(a) Perturbation　(b) Perturbed image　(c) Prediction for perturbed image　(d) Target

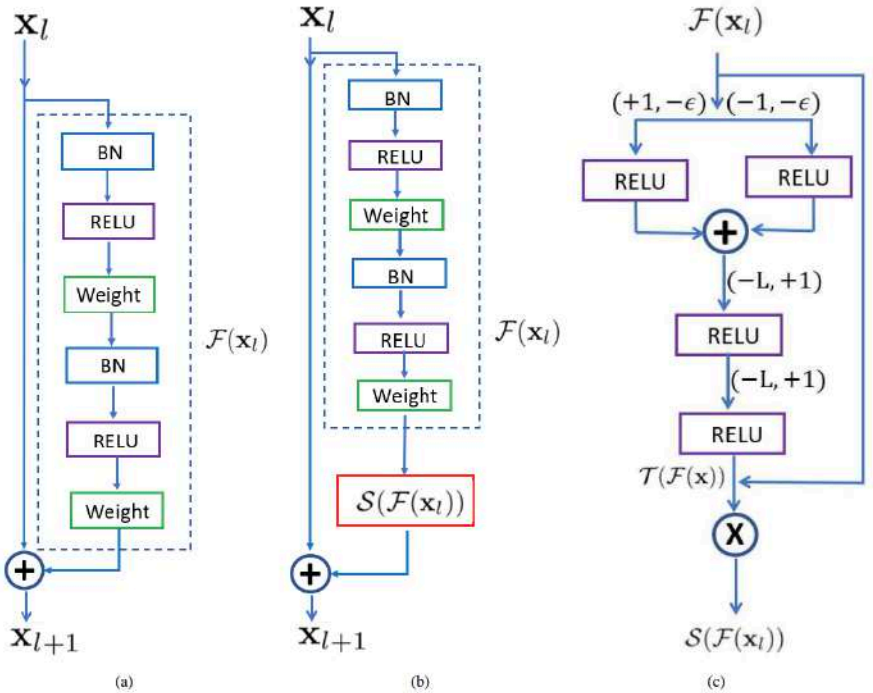
引用されそうな論文（13/26）

ε-ResNet

– 冗長な層を自動で除去するResNetで，超ディープなモデルに対して有効（左図は752層）

– ユニットは4つのReLUと閾値カット処理を増やすだけ



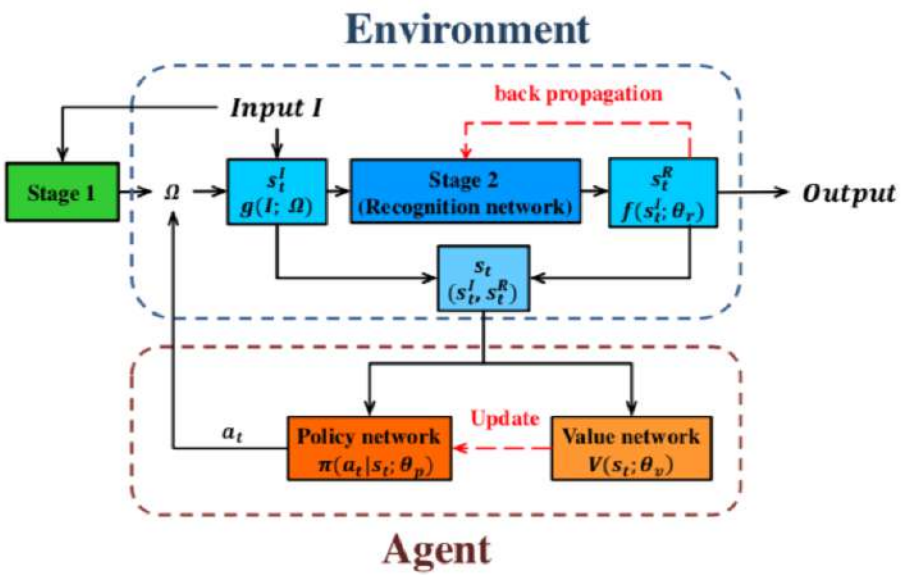


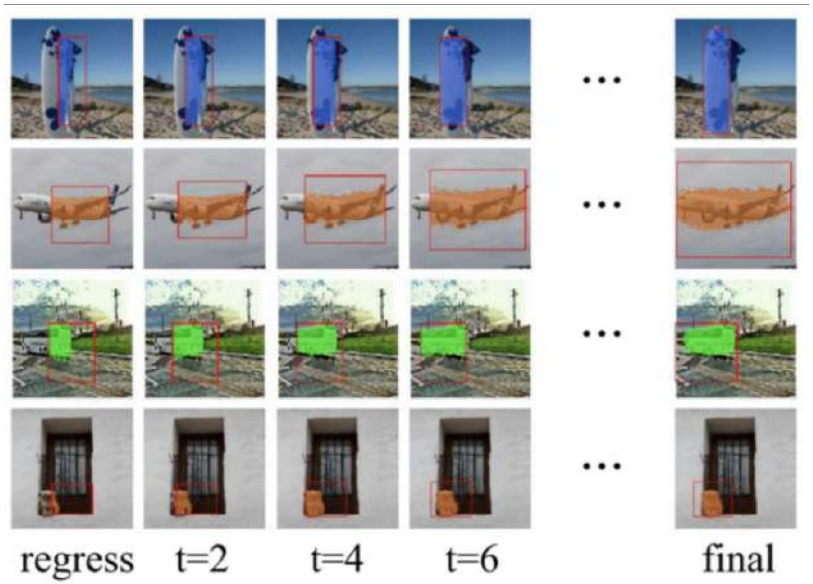
引用されそうな論文（14/26）

Environment Upgrade Reinforcement Learning for Non-differentiable Multi-stage Pipelines

- End-to-End学習が出来ないmulti-stage pipelineに置いてjoint optimizationを行う方法を提案

- 強化学習の Agent が下流の出力を受けて上流の出力に変更を与える



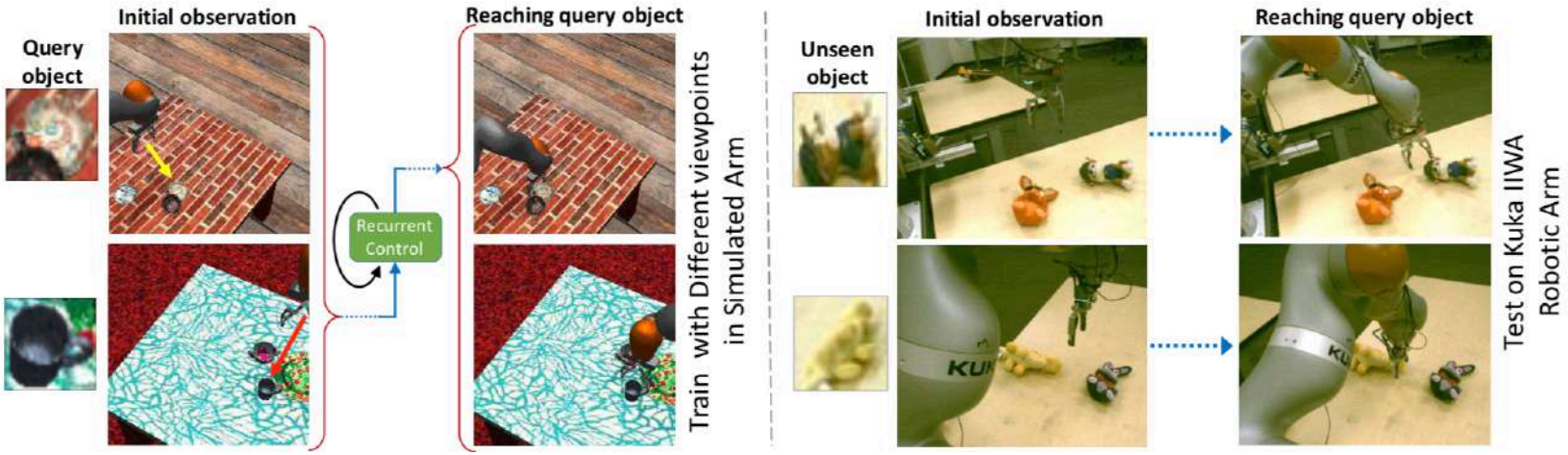


引用されそうな論文（15/26）

Sim2Real Viewpoint Invariant Visual Servoing by Recurrent Control

- DNNを用いた視点に依存しないビジュアルサーボの能力を学習する手法を提案

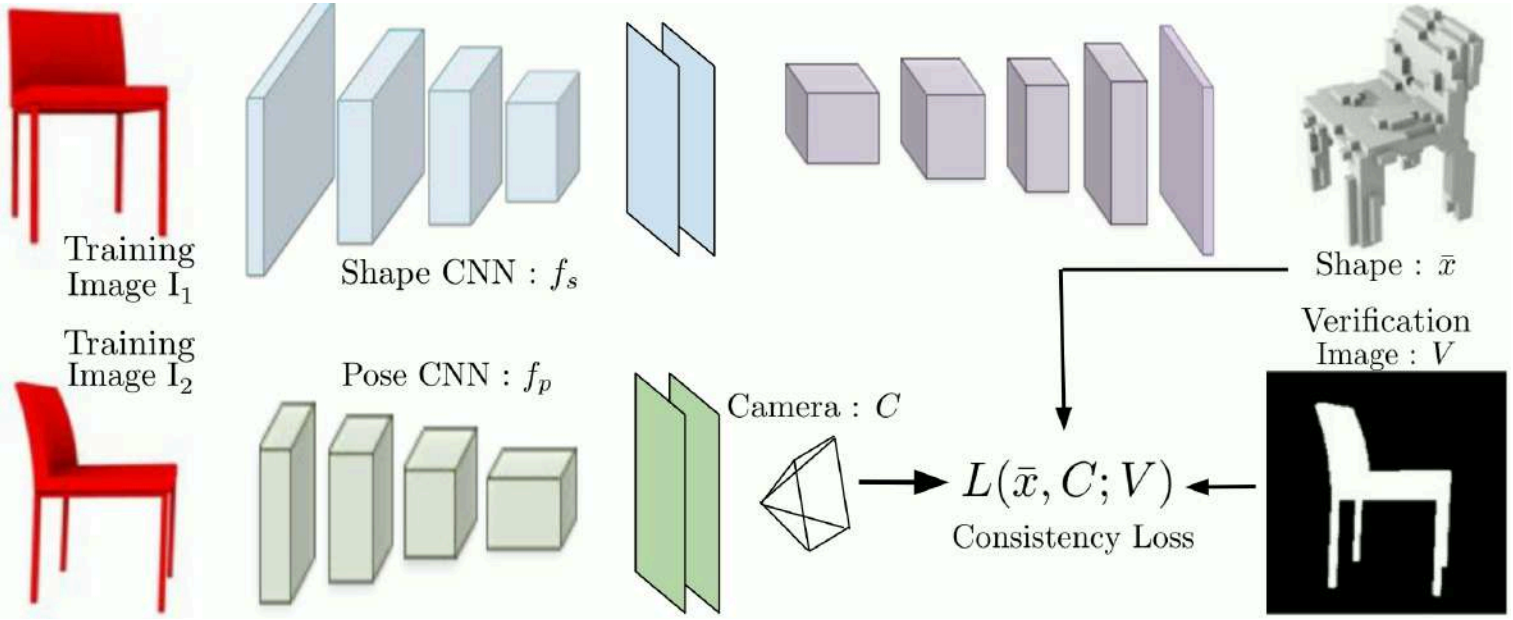
- 様々な視点,光源環境,物体の種類や位置に置けるタスクをシミュレーション上で学習することで, 未知の視点において自動でキャリブレーションを行うことが可能



引用されそうな論文（16/26）

Multi-view Consistency as Supervisory Signal for Learning Shape and Pose Prediction

- 1枚のRGB画像から物体形状とカメラ姿勢の両方を推定する
- 異なる視点から見たときの一貫性(物体の輪郭または深度情報の一貫性)を教師情報として用いるため, 学習時に物体の３次元形状と姿勢のいずれについても教師データを必要としない

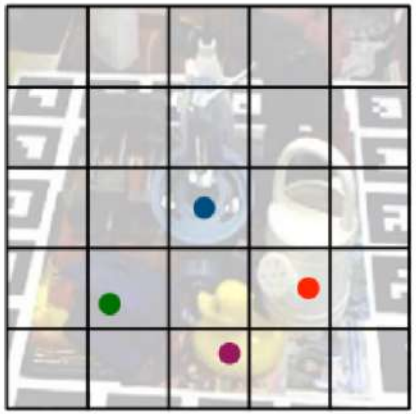


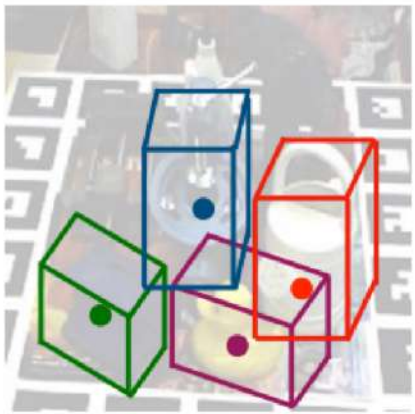
引用されそうな論文（17/26）

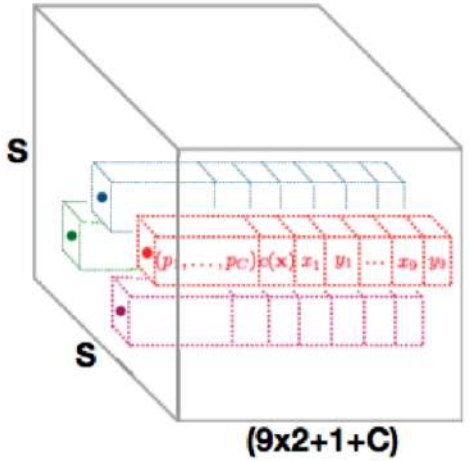
Real-Time Seamless Single Shot 6D Object Pose Prediction

- CNNを用いた単一のネットワーク (YOLOv2 ベース)で1枚のRGB画像から物体の6次元姿勢 (3D bbox)を直接推定

- Post-process 無しで高精度な姿勢推定が可能なため, 従来手法(BB8やSSD-6D)と同程度の推定精度を50fpsで達成した







引用されそうな論文（18/26）

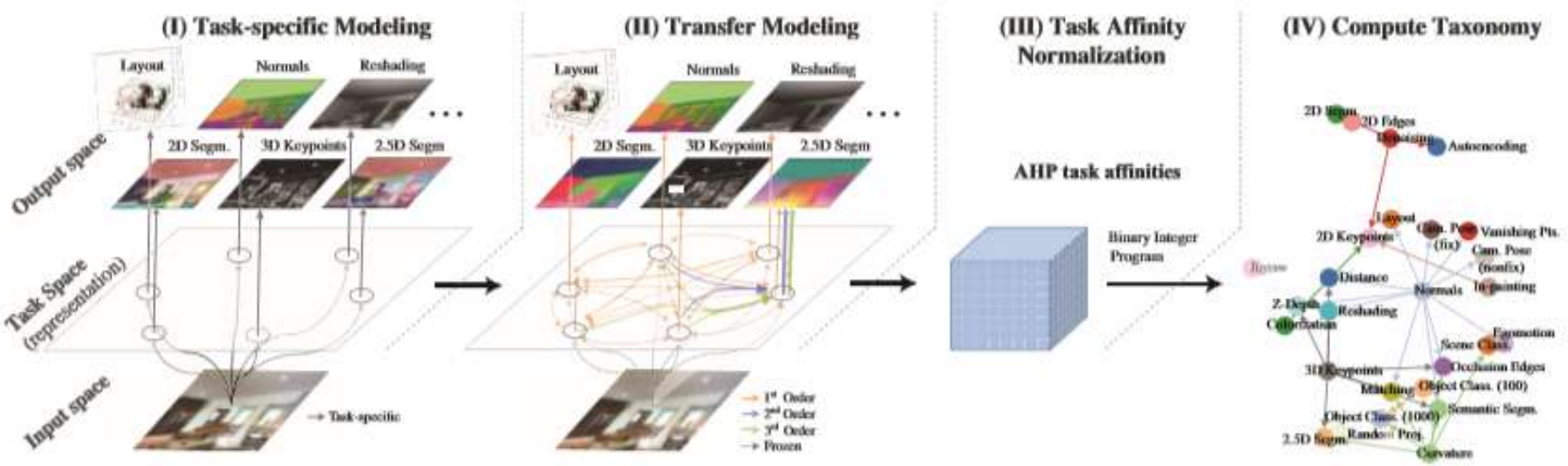
CVPR2018 Best Paper

Taskonomy: Disentangling Task Transfer Learning

- 26種のタスク遷移における潜在的依存関係の探索

  ビジョンタスクの空間構造をモデリングする手法を提案

- Source taskで得られた良い結果を他のtaskに用いることを可能に，また新しいタスクに対して転移学習により学習データの削減を可能にした

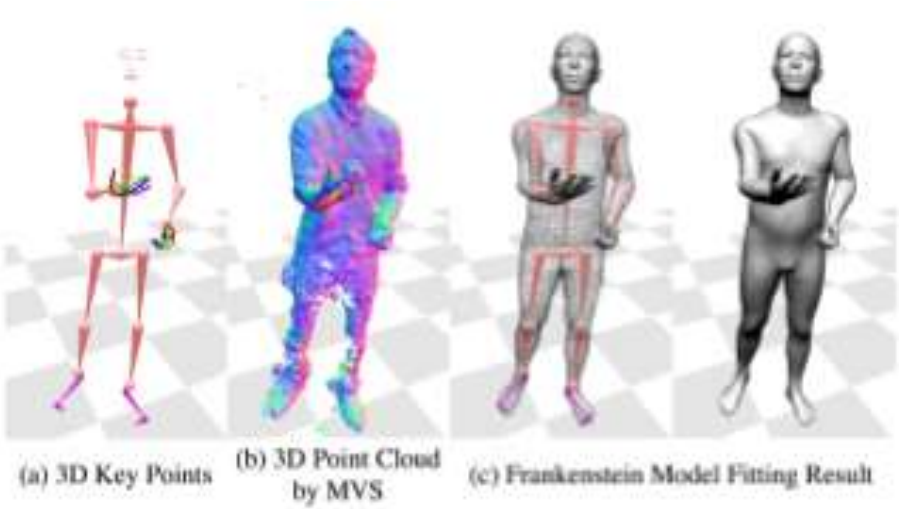


引用されそうな論文（19/26）Best Student Paper

Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies

- 顔表情，ボディモーション及び手の姿勢などのヒューマン行動のマルチスケールなマークレスなunified変形モデルの提案.

-大規模なボディモーション及び繊細な顔・手などのモデリングを同時に行える．従来手法SMPLよりリアルである同時に，顔・手などの表現がより詳細.



引用されそうな論文（20/26）Honorable Mention

Deep Learning of Graph Matching

- unaryおよびpairwiseノード近傍などを含む伝統的なgraph matchingのパラメータを学習可能なend-to-endなモデルの提案．一致性に基づいたジョイント最適化によりgraph matching及びfeature learningをリファイン．

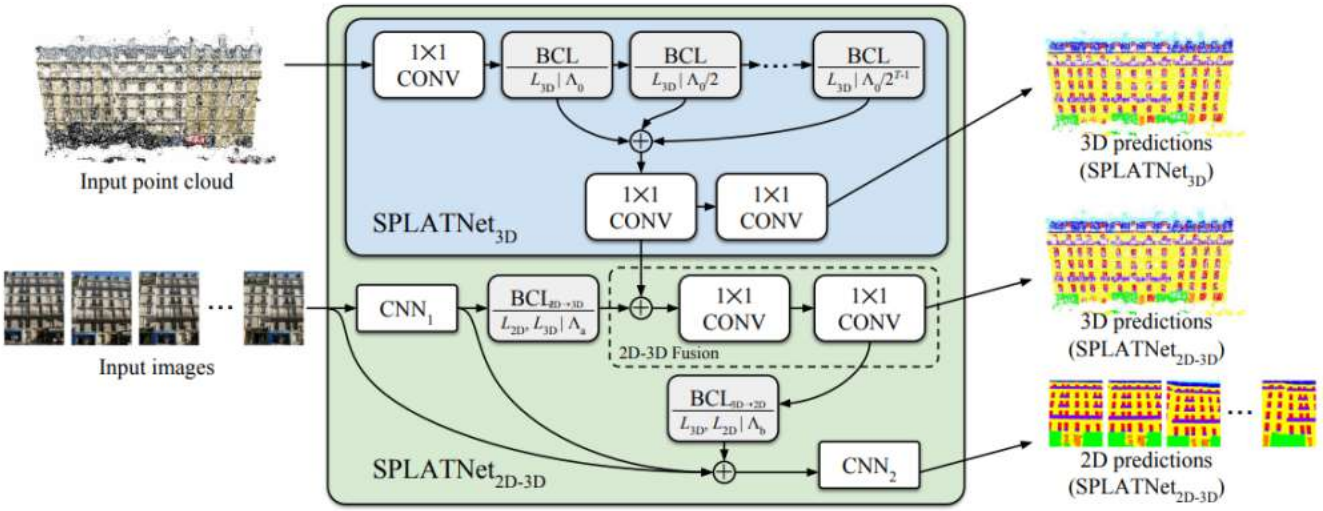
- PASCAL VOCキーポイント検出，Sintel, CUBなどにおいてSoTAな結果を達成．

引用されそうな論文（21/26）Honorable Mention

SPLATNet：Sparse Lattice Networks for Point Cloud Processing

-BCL(Bilatgeral Convolution Layer)を点群処理に応用し，点群の直接畳み込み(SPLATNet 3D)・ジョイント2D-3D学習(SPLATNet 2D-3D)を可能にした.

-点群処理をflexibleにした．Part-segmentation, Façade Semantic SegmentationにおいてSoTA.

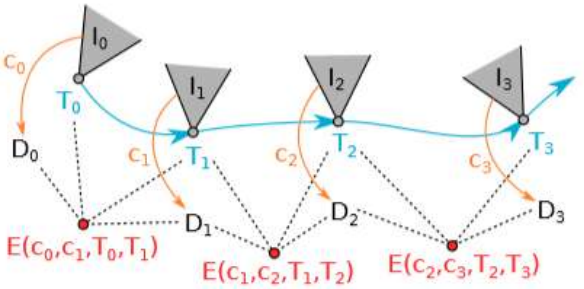


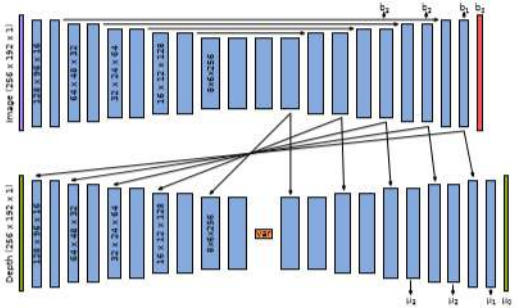
引用されそうな論文（22/26）Honorable Mention

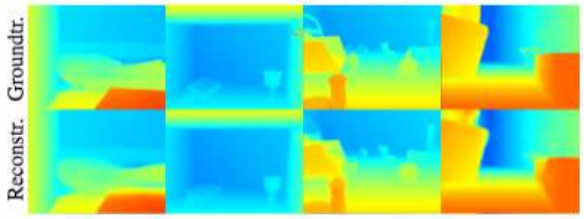
CodeSLAM: Learning a Compact, Optimisable Representation for Dense Visual SLAM

- リアルタイムなデンスなキーフレームベースな幾何情報およびphotometricのジョイントコストのSLAMシステムの提案.

- auto-encodingにより生成できるdepth mapsのコンパクトrepresentationを提案した．これによりSparse SLAMのコンパクト性およびDense SLAM情報性を同時に実現できる.





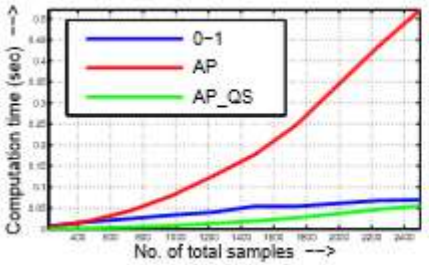


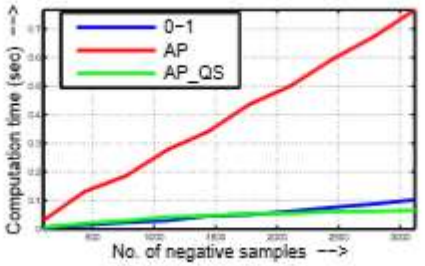
引用されそうな論文（23/26）Honorable Mention

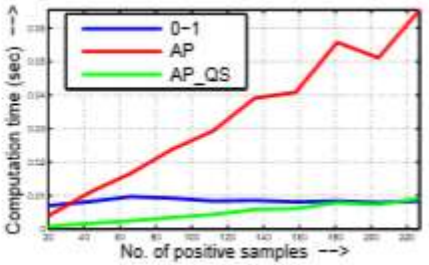
Efficient Optimization for Rank-Based Loss Functions

-分解不可損失関数(例: AP及びNDCG)用の新しいクイックソート(QS)法アルゴリズムの提案，QSを用いてロス関数のloss-augmented inferenceの複雑度を減少できる.

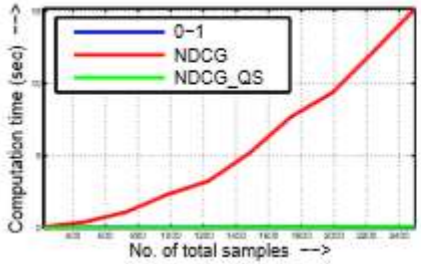
- QS操作を適応できるランキングベースなロス関数に関して徹底的な実験により特徴付けを行った.

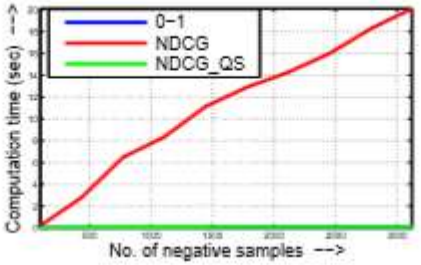


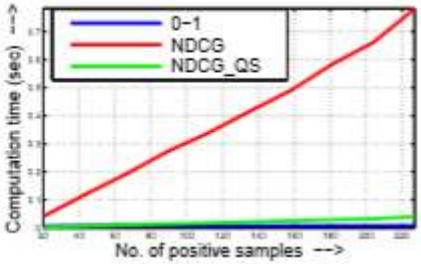




$$\Delta(\mathbf{R}^*, \mathbf{R}) = \sum_{\mathbf{x} \in \mathcal{N}} \delta_{ind-(\mathbf{x})}(rank(\mathbf{x})).$$





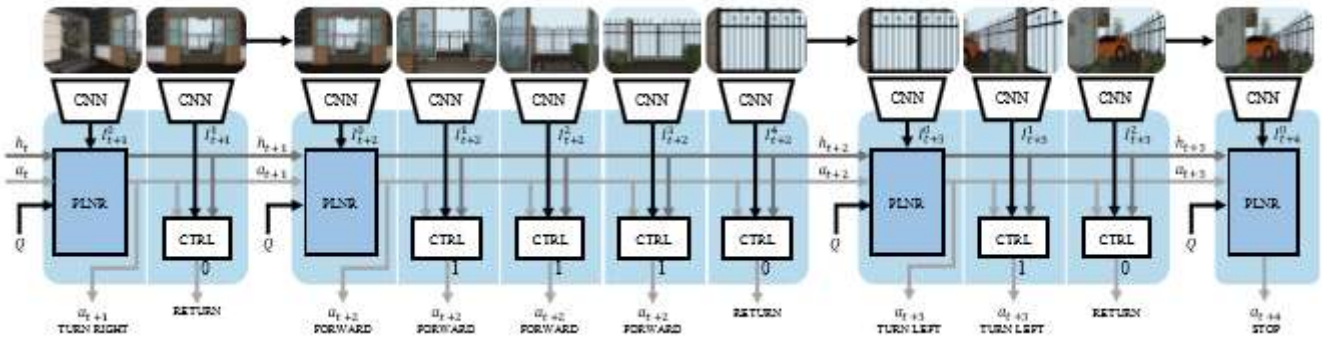


引用されそうな論文（24/26）

Embodied Question Answering

- エージェントがシミュレーション環境で自己ナビゲート及び視覚質問応答を同時に行う新たなタスクEmbodied QA及びデータセットを提案.

-RNNベースなHierarchical Planner-Controller Navigation Policyを提案し，画像の抽出特徴・現在状態などにより次の行動の計画若しくは行動のステップを決められる.

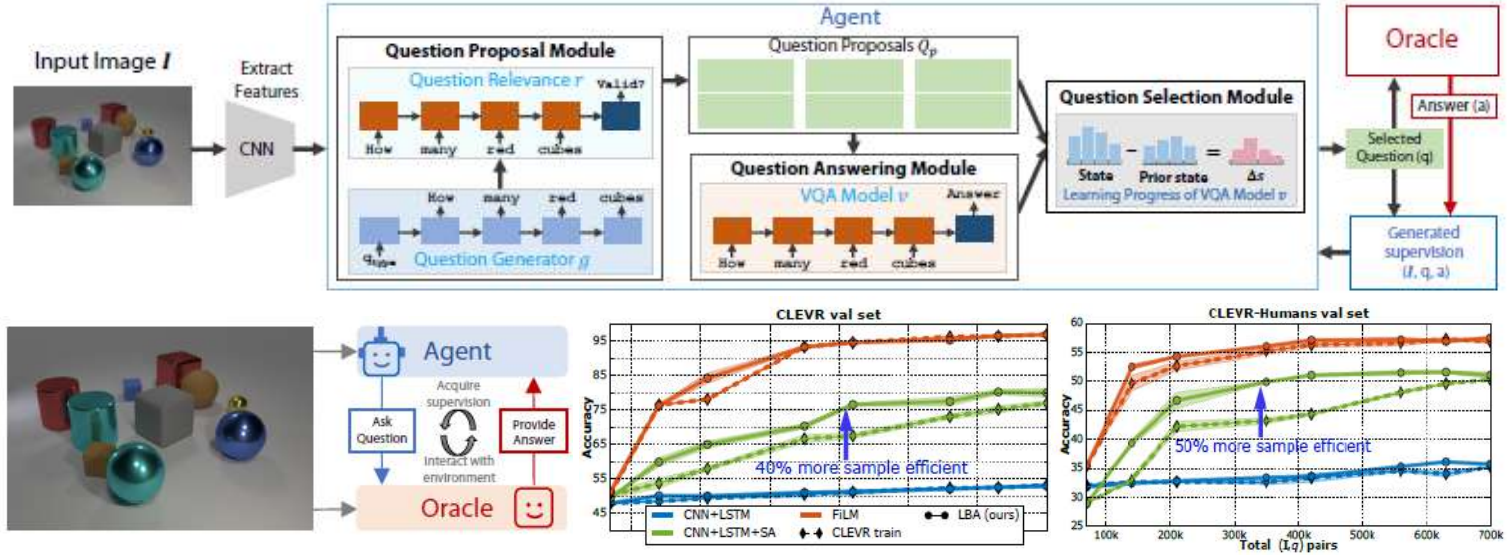


引用されそうな論文（25/26）

Learning by asking questions

- VQAタスクに用いられる新たなインターアクティブ学習フレームワークを提案した.

- 質問文の自動生成できる及び質問を選択する構造を導入し，自動的でインターアクティブで環境から情報を獲得することを可能にした．学習によりimprovementをベースに，カリキュラムで学習の難易度により学習を計画する.



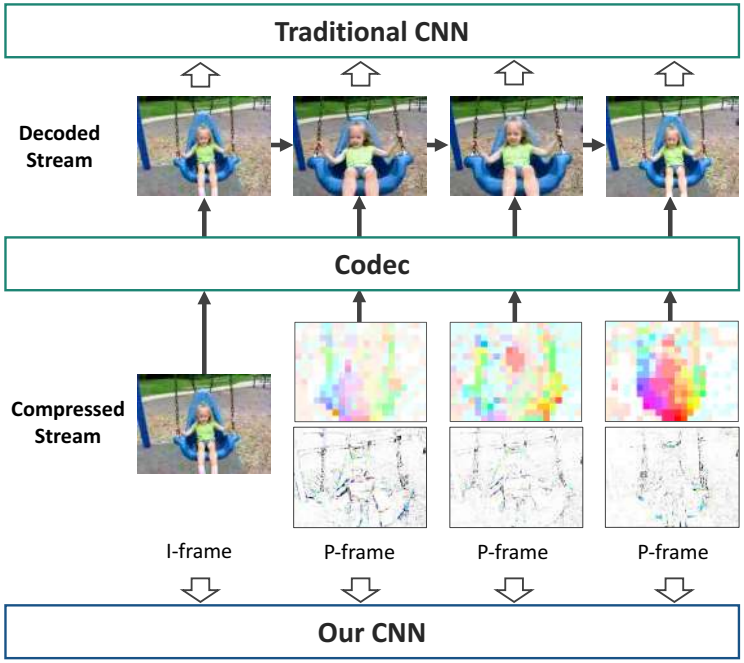
引用されそうな論文（26/26）

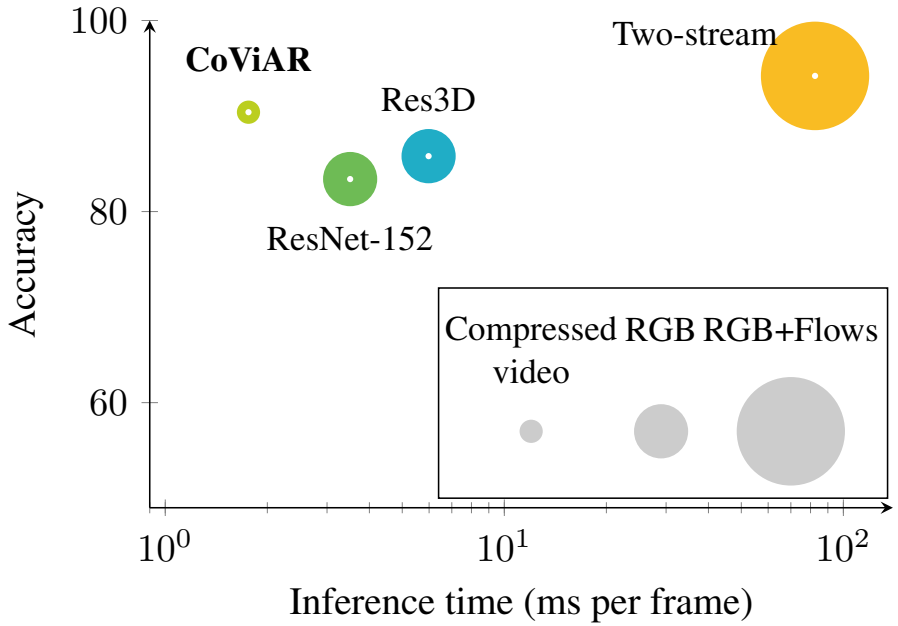
Compressed Video Action Recognition

– ビデオをフレームに分解せず、圧縮された情報を直接3DCNNで処理することで行動認識を行う

– 学習・推論の高速化が可能

– データサイズという動画認識の大きな悩みに着目





研究アイディア

- CVPR 論文を読んで考案したアイディア
- 実際に行っても結果が出るとは限りません
- その他も自己責任でお願いします

研究アイディア（1/39）

画像識別/物体検出の問題拡張？

- Webベースの今の問題 (e.g. ImageNet/PascalVOC) は精度が収束しつつある
- 新規の実世界問題だと新ベースラインが誕生？
  - 実世界問題例：Amazon Picking(Robotics) Challenge，自動運転，ドローン，サービスロボット。。。
  - 今のベースライン：Faster R-CNN/YOLO/SSD/Mask R-CNN
  - 違うデータ構成ではResNetとは異なる仕組みのConvがSOTA？
- 何が新ベースラインに？
  - ドメイン変換/転移学習を考慮？
  - より深い構造？より大規模なデータ？

新しい問題を立ち上げよう，強いアルゴリズムを考えよう！

研究アイディア（2/39）

SLAM + 物体検出（3D bbox/6D Est.）

- 単一の枠組みで実装し，End-to-Endで学習したい
- 従来の画像/動画像に対するマルチタスク学習を利活用しつつ新しい知見を創出する
- 特に6D物体姿勢推定は未だ萌芽期にあり未解決問題？

研究アイディア（3/39）

Multi-view Semantic Segmentation

- Multi-view対応のセマンティックセグメンテーション
- {FCN, SegNet, U-Net, PSPNet}とMVCNN（Multi-View CNN）を単一のフレームワークとする
- 撮影/学習時にはマルチビュー，テスト時にはシングルビューで推定可能

研究アイディア（4/39）

Human vs. AI in Computer Vision

- 強化学習ではゲームAI/AlphaGo@DeepMindが有名
- アテンション対決
  - 人間視線 vs. 顕著性
- 人物検出（既出だが再戦あり）
  - How Far are We Solving Pedestrian Detection? https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/people-detection-pose-estimation-and-tracking/how-far-are-we-from-solving-pedestrian-detection/
  - プロドライバー vs. {Faster R-CNN, SSD, YOLO, Mask-RCNN}

研究アイディア（5/39）

伝統的な分野を取り扱う，組み合わせる

- Symmetry：対称性
- Saliency：顕著性

研究アイディア（6/39）

戦いの中で強くなる識別器！

- GANは画像生成器Gに目が行きがちだが，識別器Dにも注目すべき
- Gは学習画像を生成して，Dをアシストするような仕組みになるとよい？
- 推定した姿勢が自然かどうか？で強くする研究は登場している
  - End-to-End Recovery of Human Shape and Pose
    https://akanazawa.github.io/hmr/

研究アイディア（7/39）

強化版GAN：自然界の法則により多種多様な機能が磨かれていくエコシステムを提案

- 識別器D・生成器Gもひとつじゃなくてもよい
  - 識別器が環境を理解する，生成器が環境を作る，としたら？
- G1（シーン全体を作る），G2（シーン内の人物/物体を作る），G3（シーン内の生合成を整える）
- D1（人物/物体/シーンを認識），D2（人物/物体/シーンの状態を推定），G3（シーン全体の関係性を捉える）
- 全体がホントっぽいかを判断するものがあっても良い？

GANはまだまだ進化しそう！

研究アイディア（8/39）

目にいい画像の自動生成

- コンピュータは人間を鍛える時代？
- 視線等から目に良い領域学習/個人の効き目を判断して左右均等に使うような画像生成
- 人間の行動履歴から報酬を得て学習する画像生成器

研究アイディア（9/39）

AffordanceNetなどをうまく利用すれば，個別化をより効果的に支援できる？

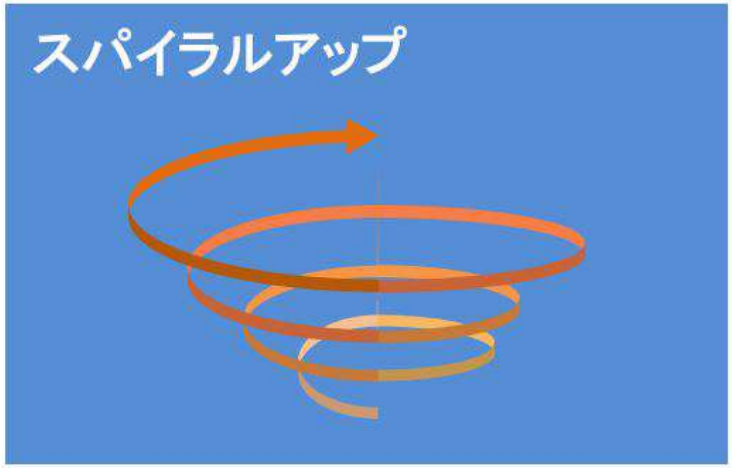
- 昨日のセグメンテーションを実装するアフォーダンスネット
- 個人を観察してアフォーダンスを決定していく
- AffordanceNet (ICRA 2018): https://github.com/nqanh/affordance-net
- Action Print: http://jglobal.jst.go.jp/public/20090422/201202294758884091

研究アイディア（10/39）

ドメイン適応を繰り返すことで特徴を強化

– ドメインをまたぐと特徴を強化するという発表あり
– では，ドメインを何度もまたぐ/ループするのは？
– エンジニアリング的なスパイラルアップができる？

適用環境が増えるほど良い特徴になっていくモデル



Spiral-Up Domain Adaptation
のイメージ図

http://te28.net/pdca/

研究アイディア（11/39）

自然言語処理の知見を動画解析に活かしたい

- Doc2Vecの知見を動画の表現学習に使用
- 時間長が異なる動画を単一のベクトルで扱えたら。。
  - 検索に有利
  - 高速な識別処理
- 何とか2Vec自体を新しく考案

研究アイディア（12/39）

反面教師学習：失敗例のみから正解を導く

- 教師あり学習と強化学習の中間のような感じ？
- 正常以外が異常というCHLAC的な問題設定が必要
- 反面教師あり学習、アイディアは出ていた？だが誰もやっていないのでやる価値はあり。 http://d.hatena.ne.jp/repose/20091122/1258827365
- 反面教師あり学習、Google翻訳すると「On the other hand teacher learning」（どうでもいい）

研究アイディア（13/39）

Adversarial Examplesを埋め込んで，何に見間違える傾向にあるかを調査

- 何を見ても特定の物体にしかならないように現実を改善？
- ゲシュタルト崩壊は関連するかも？
- 犬見すぎると何の物体かわからなくなる（ゲシュタルト崩壊はDNNでも起きるか？違う意味での過学習？）
- ゲシュタルト崩壊が起きやすい漢字があるらしい，パターン解析できないか

研究アイディア（14/39）

GANの学習安定化

- 識別器ではなく，生成安定器にするとか
- ロスを返すだけでなく分布自体にフィードバックする
- Goodfellow氏のGAN10選
  - https://gist.github.com/zafartahirov/8fcd00f703b27c4426e7c74f2900f2dd

研究アイディア（15/39）

モダリティ間のパスを通す

- RGB => 距離画像
- グレースケール => カラー画像
- 画像 <=> 音声
- 画像 <=> 文章
- ができるようになった，のであらゆる物理量をひとつの空間で扱えるようなものができないか？

研究アイディア（16/39）

画像キャプショニングで使用されている情報は画像の一部に過ぎない

- あらゆる可視な情報/背景に隠れている情報を顕在化
- 画像に対して発火する領域をスパースかつ詳細化できると画像を網羅的に理解できる？

研究アイディア（17/39）

リアル画像を一枚も含まないDB

- 部分的にはできている（人検出, 目検出, 特定物体…）
- 主にCGが使われているが，これに対してあらゆる法則や人間の知見のようなものをモデル化
- 学習ベースではなくモデルベースのデータ生成の隆盛があってもよい？

研究アイディア（18/39）

写真を膨大に集めると自然発生的にデータセットを構成してモデルを自動生成

- CNNのモデルに関する知見（こういう時にはこのネットワークが良い）も揃ってきたので，データから大体のモデル構造やタスクを決定可能？
- アルゴリズムと人間の強調，ボトムアップなアルゴリズムからの提案とトップダウンな人間の感覚のコラボ

研究アイディア（19/39）

VQAの再帰的繰り返し

– 質問回答したらそれに関する次の質問と回答を自動生成

– 再帰的に自問自答できるようなVQAが次のフェーズか

研究アイディア（20/39）

データセットで取得される時間幅が限定的

- 静止画は一瞬，動画DBでも長くても数分
- Super Long-Termなデータが必要？
  Day/Month/Year/Decadeレベルの解析
- 長く映像が残っている有名人などを追いかける？

研究アイディア（21/39）

CVアート分野を確立

- 新規分野創出ができるのではないか
- 正解値がない問題に対する提言，CVは高精度/アルゴリズムが評価されるが，そうではない部分があっても良い

研究アイディア（22/39）

データ量で圧倒するDBならどのくらい必要？

- 画像なら数十億？
- 動画なら数千万？
- 3Dモデルなら数千万インスタンス？
- しかも体系化されたデータ収集

研究アイディア（23/39）

超大規模データから，タスクに合わせてデータセットを再構成する

- ポジティブはもちろん，ネガティブサンプルも最適なものを選択できるか
- 学習の中でも最適なデータを取捨選択できるように

研究アイディア（24/39）

人間にも機械にとっても優しい文字生成

- 人間にとって優しい=書きやすい，読みやすい
- 機械にとって優しい=認識しやすい，理解しやすい

研究アイディア（25/39）

Multimodal Knowledge Distillationは可能か

- 音声認識で学習したパラメータ => 画像認識で使用するパラメータ
- 感覚をまたいで学習することができるかどうかを検証

研究アイディア（26/39）

簡単な/難しいデータを使わないでちょうど良いデータのみを使うと学習が成功しやすい

- Active Bias: https://papers.nips.cc/paper/6701-active-bias-training-a-more-accurate-neural-network-by-emphasizing-high-variance-samples
- 学習時に現在のモデルの状態から動的に次の最適なバッチをとってこれると良さそう
- それをデータ収集の段階からできると面白い

研究アイディア（27/39）

CNNにも符号化を導入するとよいか？

- すでにLBP-CNNやBinaryNetのようなものもある
- さらに熟慮された手法も検討可能か

研究アイディア（28/39）

適応的物体検出

- 物体検出で9000クラスなど対応できる（e.g. YOLO9000）みたいだけど，環境によって認識するクラス数を適応的に変更させるのできないか？
- メタ学習？の枠組みか

研究アイディア（29/39）

2D（xy）から3D（xyz）の復元はあるけど，2Dから4D（xyzt）の復元はできるか？

- 動きと奥行きの情報をプラス
- 意味空間（セマンティックラベル）にん落とし込んで，実空間形状と想定される動きを同時推定することは可能？

研究アイディア（30/39）

多クラス物体検出とYouTube/SNS

- YOLO9000物体検出プラスYouTube/SNSデータを用いれば，概念追加や概念強化が可能か？
- 概念追加：全く見たことない概念を新しくカテゴリとして追加
- 概念強化：今まで見たことあるカテゴリの特徴量を強化

研究アイディア（31/39）

現在作り得る「最強の学習済みモデル」とは？

- Googleは3億枚/Facebookは35億枚の学習で精度向上
- HybridNet（ImageNet+Places205/365）で精度向上
- マルチタスク問題を大量画像で学習？
- 生成画像/教師なし学習で強い学習済みモデルはできるか？
- 巨大IT企業でもない限り億単位のラベル付きデータは手に入らないので，ImageNet/ Places/ COCO/ OpenImages/ YFCC100Mあたりのデータを組み合わせる+マルチタスク学習+弱教師付き学習を実行か

研究アイディア（32/39）

視線計測器でアノテーション

- 矩形にキリトリ
- 詳細なカテゴリ付け（HCIのアイディアが必要？）
- 視線でセグメンテーションは目が疲れそう？

研究アイディア（33/39）

データアノテーションによる収入生活は成り立つか，を真面目に考察すると良いのでは？

- クラウドでの仕事が発達しつつある
- トップアノテータという年収数千万円もらえる職業がある
  といわゆるAI研究が捗るかも？

研究アイディア（34/39）

かっこいい論文の図検索

- 論文がアートになる
- 見た目だけで通る論文を生成
- トップ会議の査読オン仕組みから論文を生成・変換
- 拘束条件はペーパーフォーマット

研究アイディア（35/39）

最適化された形状はなぜ美しいかを検証

- 大規模データの美しさと自然の法則の類似性
- 学習済みパラメータを可視化した際の動きについて考える

研究アイディア（36/39）

CV x セキュリティ

- 公開鍵暗号を２つの学習済みネットワーク間で行う
- 近年ではセキュリティとビジョンの話が萌芽期に位置付けられている
- Privacy and Security Workshop: http://vision.soic.indiana.edu/bright-and-dark-workshop-2018/

研究アイディア（37/39）

音楽とCV研究の垣根を橋渡し

- 例えば音楽ジャケットと音楽スタイルとの関係
- 指揮者の動きから音楽性を定量的に見つけ出したい
- 演奏者の細かな動きと音楽の音との関係
- Sight and Sound というワークショップが開かれた
  http://sightsound.org/

研究アイディア（38/39）

質感を保持したままのSLAM

- 材質推定とSLAMの同時処理
- SLAMは付加情報をさらに付け加えてもよい？

研究アイディア（39/39）

動物の動画DBを作りたい

- 子供よりもペットの数の方が多い時代
- ペットのサポートもできるようになるとよい
- 静止画だとすでに詳細画像分類問題のDBが存在
- 人間だと個人情報保護が必要だが，ペットだと同様に法律関係が絡んでくるか？

今後の方針

- では，どうすればよいか？

今後の方針（1/5）2017からの再掲

他の追随を許さぬ強い手法を作る！

- 受賞論文や注目論文にあるように「極めて高速かつ高精度」を実現，コードをリリースして分野に貢献が理想
- 問題設定に対しての強い手法でも構わない
- 作れたらいいですね！=> e.g. DenseNets, YOLO9000, PAF, PSPNet, Taskonomy[new!], R-FCN-3000-30fps[new!]
  …

今後の方針（2/5）2017からの再掲

高品質論文でないと記録・記憶に残らない

- トップ会議の論文とて例外でない（CVPRは今年979本）
- 中途半端に分割した複数論文よりもパーフェクトな１本
- 動画やスライド公開・コード共有・DBリリースなども（できる限り）徹底して揃える

今後の方針（3/5）2017からの再掲

今まで以上にチームの力が重要

- 高品質論文には１人のパワーでは不十分？
- cvpaper.challengeでは仕組みを再考
  - 通常の学生：１人１テーマ３年間継続（学部~修士を想定）
  - cvpaper.challenge：２~４人１テーマ0.5~１年でテーマ拡張/変更

今後の方針（4/5）

論文を投稿しよう！

- まずは量，次に質を伴わせる
- 学部/修士の学生だって通せるポテンシャルは持っている
  - もちろん，先生方の支援/労力が大きいことも忘れてはいけない！
  - Rejectされたとしても投稿と改善で論文は磨かれる，経験値を蓄積していく
- cvpaper.challengeは2018年中にトップ会議に20本投稿する，が目標の一つ（もう一つはCVPR完全読破）

今後の方針（5/5）

研究連携する

- 研究機関の垣根を越えて論文を書く/産業に結びつける
- 良いアイディア，良いテーマ，ひいては良い研究というのはディスカッションの中で磨かれる
- cvpaper.challengeでは，研究テーマは全員で考案
  - 2018は約30名，4ヶ月（1~4月），全15回の全体集合でブレスト
  - 研究テーマのためのアイディアを「436」蓄積
  - 上澄み「20」テーマのうち，学生が楽しいと思う&学術的に意味のありそうなテーマを研究グループに対して割り振り

チームとしての研究が自然に進むエコシステムを作る

今後の方針（Bonus Slide）

研究，楽しもう！

– Enjoy!

以下、まとめ論文集

- 500本弱あります
- 詳細には下記をご覧ください
  https://cvpaperchallenge.github.io/CVPR2018_Survey/#/
- Web版ではブラウザの横幅によって
  文字サイズを変更できるので,
  下にはみ出している場合には横幅を狭くしてください
- 誤字脱字などの修正，ディスカッションなどはIssueまでお願いします（小さいミスでも構いません）
  https://github.com/cvpaperchallenge/CVPR2018_Survey/issues

[#1]

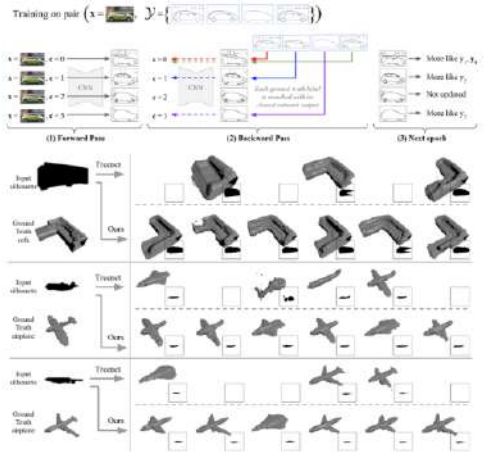
Yoshihiro Fukuhara

DiverseNet: When One Right Answer is not Enough

Michael Firman et al. CVPR 2018

概要

教師あり学習において, test 時に同じ入力から異なる結果を出力可能にする Loss と学習方法 (DiverseNet) を提案. 提案手法はあらゆる教師あり学習の手法に対して適用が可能であり, 提案された Loss は GAN などで報告されている mode-collapse を起こしにくい. 複数のタスクに対して評価実験を行い有効性を確認した.



新規性・結果・なぜ通ったか？

- 学習の画像と一緒に制御変数（整数）を入力する, 制御変数を変更することで test 時に同じ画像から異なる結果を得られる
- 複数の正解ラベルについて Loss の和をとると mode-collapse を起こしやすいため, 提案された Loss では各ラベルについてそれぞれ Loss を計算し, 最小の値を取ったものを Loss として使用
- 提案手法はあらゆる教師あり学習の手法に対して適用が可能. また, 正解ラベルが 1 つしか無いタスクにおいても, 最もらしい結果を複数生成可能
- 評価実験では提案手法を 2D image completion, 3D volume estimation, flow prediction などの複数のタスクにおける手法に適用し, 特に小さなネットワークのモデルに対して良い結果となった

コメント・リンク集

- [論文] Optimizing Video Object Detection via a Scale-Time Lattice
- [Project page] Optimizing Video Object Detection via a Scale-Time Lattice

TakumuIkeya

Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification

Kiang Long et al. CVPR2018 1711.09550

概要

- 動画のクラス分類タスクにおいて時系列の情報，特に長期間のパターンは必要な情報ではないことを示し，純粋にattentionに基づいた局所特徴の統合フレームワークを提案をした研究である．
- 提案したフレームワークを用いて動画分類タスクを実行することで評価した．

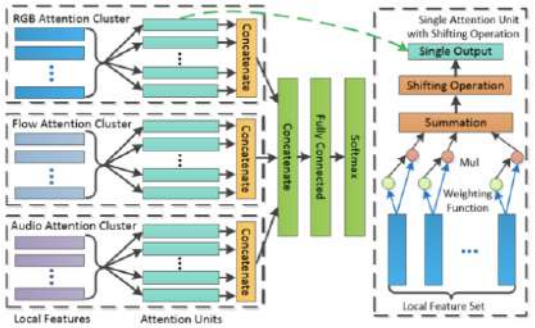


Figure 2. Multimodal Attention Clusters with Shifting Operation: The overall architecture for video classification. Separate attention clusters are applied for different feature sets and then the outputs are concatenated for classification.

新規性・結果・なぜ通ったか？

- 提案したフレームワークはKineticsデータセットにおいてtop-1で79.4%,top-5で94.0%の精度を達成した．
- 提案したフレームワークではシフト操作を伴うMultimodal Attention Clustersを導入することでフレームの類似性が高い動画に対しても良好な結果が得られる

コメント・リンク集

- 論文

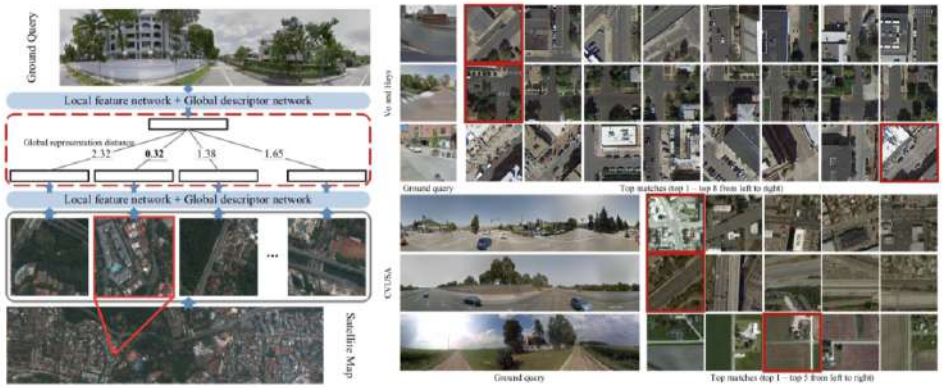
Yoshihiro Fukuhara

CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization

Sixing Hu et al. CVPR 2018

概要

Ground-to-Aerial Geolocalization の研究. CNNを用いて局所特徴量を抽出した後, NetVLAD によって局所特徴量から大域特徴量を生成してマッチングを行う. また, 新しい Loss を提案し学習時間を短縮した. CVUSA dataset 等を用いて行った評価実験では既存手法に大差で優位な結果を達成した.



新規性・結果・なぜ通ったか？

- 地上で撮影された写真から, 衛星写真上のどの位置で撮影されたかを推定する（Ground-to-Aerial Geolocalization）
- 両方の写真からCNNを用いて局所特徴量を抽出した後, NetVLAD によって局所特徴量から大域特徴量を生成, 後述の weighted soft margin ranking loss を用いて学習を行う
- 新しく提案した weighted soft margin ranking loss は従来の soft-margin triplet loss よりも学習の収束の速度を早めると共に, ネットワークの精度を向上させた
- CVUSA dataset と Vo and Hays dataset を用いて行った評価実験では既存手法に大差で優位な結果を示した（評価基準は上位 1% の recall）. 特にパノラマ写真を入力とした場合は90%以上の精度を達成

コメント・リンク集

- [論文] CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization
- [Code] GitHub

Kazuki Inoue

Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery

Zhongzheng Ren and Yong Jae Lee CVPR 2018 Poster

概要

人手によるアノテーションを使用しない本当の意味での自己教師学習を行うために、合成画像の法線マップ、デプス、物体輪郭と実画像とのadversarial trainingを行う手法を提案。実画像に対して汎用的な特徴量が取得できたことを主張している。 輪郭線はキャニーフィルタによるエッジだが、これによって人がつける曖昧なアノテーションを緩和することができる。 デプスを推定することで高次元のセマンティックな情報やオブジェクトの相対的な位置を得ることが可能。 既存研究により法線マップとデプスのそれぞれの推定が良い影響を与えることがわかっているため、法線マップの推定も行う。GANの学習において、ディスクリミネータの更新は実画像、合成画像に対するGANのロス、ジェネレータの更新は合成画像に対するGANロス、 3つのタスクの推定におけるロスを使用している。ドメインに不変な特徴料を得るために実画像を用いたジェネレータの学習も行ったが、 精度が良くなかった。

新規性・結果・なぜ通ったか？

- 人手によるアノテーションを使用せずに自己教師学習を行うために合成画像の法線マップ、デプス、オブジェクトの輪郭を推定するネットワークを構築し、さらに実画像に対して汎用的な特徴量を得るために実画像とのadversarial trainingを行う。
- PASCAL VOCを用いた最近傍によるリトリーバルを行った。トレーニングデータにはバスや車などの区別しづらい画像が含まれているにも関わらず、車を入力した際には車のりトリーバルに成功。

コメント・リンク集

- 論文
- Project page

[#5]

Kazuki Inoue

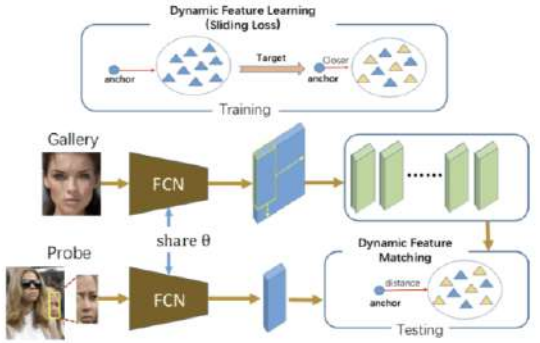
Dynamic Feature Learning for Partial Face Recognition

Lingxiao He, Haiqing Li, Qi Zhang, Zhenan Sun CVPR 2018 Poster

概要

マスクなどから見えている顔領域のみを検出するPartial face recognition(PFR)をFCNで高速かつ高精度に行う手法を提案。トレーニング時には顔全体と顔が見えているパッチのそれぞれに対してパラメタを共有したFCNをで特徴量マップを適用し、 パッチ領域から得られる特徴量マップと同サイズのマップを顔全体からえられた特徴量マップからスライディングウィンドウによって複数個切り出し、 パッチから得られた特徴量マップとの比較を行う。 この比較のことをDynamic Feature Matching(DFM)と読んでいる。 DFMを行う際の工夫として、パッチから得られた特徴量マップを顔全体から得られた特徴量ウィンドウの線形和で表す際の重み、 パッチから得られた特徴量マップと特に類似している特徴量ウィンドウに対する重みの学習を行っている。

Figure 1. Partial face images are produced in unconstrained environments. A face may be 1) occluded by sunglasses, a hat and a scarf; 2) captured in various poses; 3) positioned partially out of cameras filed of view.



新規性・結果・なぜ通ったか？

- PFMを行う際に顔全体から得られた特徴量マップを切り出した複数の特徴量ウィンドウと顔パッチ部分から得られた、特徴量ウィンドウと同サイズの特徴量マップを比較するDFMを行う手法を提案。
- 既存手法であるMR-CNNの20倍の速度で実行可能。
- CASIA-WebFace 1万枚を用いて学習。LFWなどのデータセットでテストを行う。face recognition, verificationにおいてSoTA。
- 切り取るサイズや、パラメタに対する考察も行っている。

コメント・リンク集

- FCNを用いることで任意のサイズの入力を扱えることに着目したことが根幹となるアイディア。
- 論文

[#6]

Kazuki Inoue

Mean-Variance Loss for Deep Age Estimation from a Face

Hongyu Pan, Hu Han, Shiguang Shan, Xilin Chen CVPR 2018 Poster

概要

顔画像から年齢を推定する際に正確に年齢を推定するのではなく、ガウス分布を用いてある程度幅のある推定を行う手法を提案。大きなコントリビューションはロス関数としてガウス分布の平均値と分散に関するロスをとったことであり、 平均値はGTの年齢との差分をとり、分散は分布がよりシャープになるようにロス関数を設計する。 学習の際には上記2つのロス関数の他に1歳刻みの年齢をそれぞれクラスと見立てソフトマックスロスを取る。 分布を学習する既存手法と異なる点は、提案手法ではGTの平均値、分散を使用しない点である。

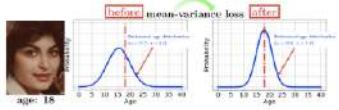


Figure 1. An example of age distribution learning using the proposed mean-variance loss. Our mean-variance loss aims to learn an age distribution which has not only a mean value close to the ground-truth age (red dotted line), but also a concentrated shape.

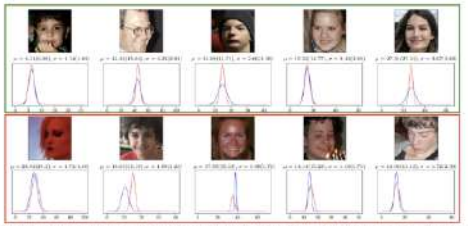


Figure 6. Examples of distributions estimated by our approach on the CLAP2016. The top two rows show good distribution results, and the bottom two rows show poor distribution results. The red and blue curves are the estimated and ground-truth distributions, respectively. The numbers below each image show the age and std of ground-truth and estimated of the subject, i.e., μ = ground-truth age (estimated age), σ = ground-truth std (estimated std).

新規性・結果・なぜ通ったか？

- 人間の年齢は正確に推定することは難しいが、ある程度の範囲内であれば推定は容易、という観察に基づいてロス関数を設計。
- FG-NET, MORPH Album Ⅱ, CLAP2016, AADBデータセットにおいてMAE、CSを評価指標として使用し多くのテストプロトコルにおいてSoTA。
- 照明環境に依存し、顔が赤い光で照らされているなどの特殊な照明環境では推定誤差が大きい。

コメント・リンク集

- 年齢推定だけでなく、同様の性質を持つタスクならば適用可能。
- 論文

[#7]

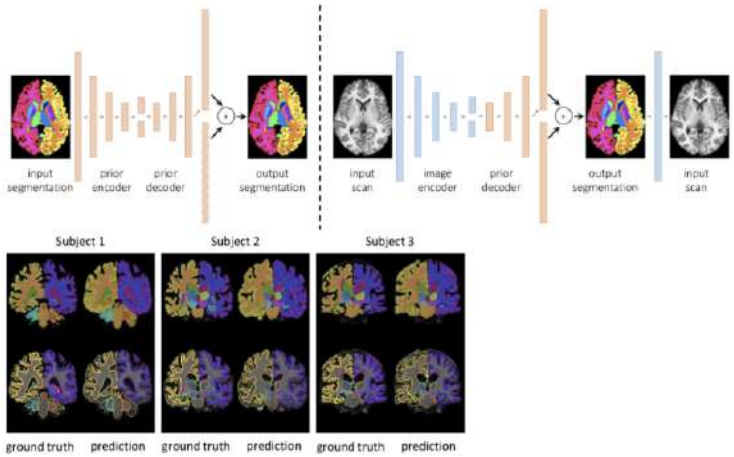
Kazuki Inoue

Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation

Adrian V. Dalca, John Guttag, Mert R. Sabuncu CVPR 2018 Poster

概要

MRIのスキャンデータに対するセグメンテーションを、MRIのソース画像とセグメント画像のペアを使用せずに行う手法を提案。はじめにセグメント画像のみを用いてVAEを学習。 次に教師無しでセグメンテーションを行うためにdecoderの重みを固定してソース画像に対するセグメンテーションの推定を行う。



新規性・結果・なぜ通ったか？

- 医療用画像に対する教師無しのセグメンテーション手法を初めて提案。
- T1w scanデータセットのうち、5000枚のセグメンテーション画像を使用してauto-encoderをプリトレーニング。残りの9000枚のスキャンデータを用いて教師無し学習。
- T1wデータセットよりも解像度が低く、スライス間隔も広いT2-FLAIR scanデータセットでもテストを実行。ただしアノテーションが存在しないのでセグメンテーションの見た目で良し悪しを判断。
- 評価尺度はGTとの領域の重なりを評価するDice。Dice、セグメンテーションの結果の見た目として良好な結果が得られていると主

コメント・リンク集

- Diceを使って定量的に評価しているため、境界線の引き方などの細かい部分のセグメンテーション結果を詳細に評価していないが、実用上は問題無いのだろうか？
- 論文
- Supplementary material
- GitHub

Kazuki Inoue

GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose

Author CVPR 2018 Poster

概要

単視点動画に映っている物体を静的物体と動的物体に分離することで教師なしでデプス、オプティカルフロー、カメラ向きを推定する手法を提案。フレームワークは二段階で構成されており、 まずはじめにデプスとカメラ向きをそれぞれ独立に推定することで道路や街路樹などの静的物体のモーション情報を得る。 続いて静的物体との差分情報を使用することで歩行者などの動的物体のモーション情報を得る。教師無しの推定を行うため、 参照フレームから推定されたモーション情報の逆変換をターゲットフレームに適用し参照フレームを推定することで consistency lossをとることで精度が向上。

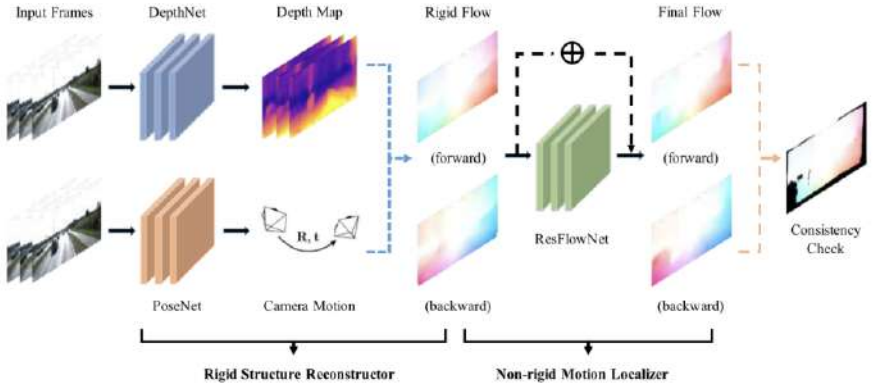


Figure 2. Overview of GeoNet. It consists of rigid structure reconstructor for estimating static scene geometry and non-rigid motion localizer for capturing dynamic objects. Consistency check within any pair of bidirectional flow predictions is adopted for taking care of occlusions and non-Lambertian surfaces.

新規性・結果・なぜ通ったか？

- consistency lossによってオクルージョンに対する精度の向上も確認。
- 同じネットワークを持つ既存研究に対して、ロス関数の優位性を確認

コメント・リンク集

- 論文
- GitHub

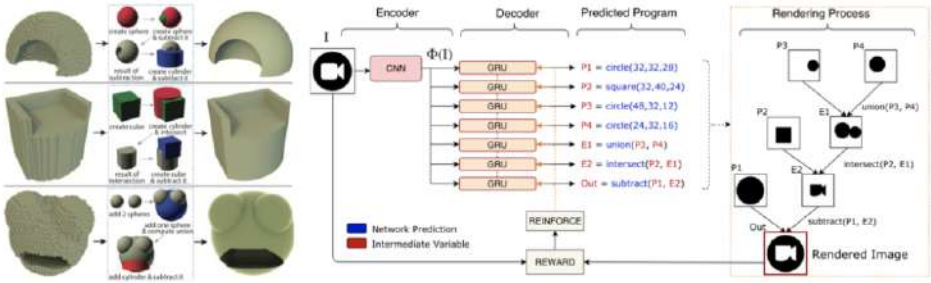
Yoshihiro Fukuhara

CSGNet: Neural Shape Parser for Constructive Solid Geometry

Gopal Sharma et al. CVPR 2018

概要

Shape Parsing の研究. ２次元画像, ３次元ボクセルから同じ形状を生成するプログラムを推定する. 学習のための2次元や３次元のLogoやCADモデルなどを含む synthetic dataset を作成・公開した. また, 教師データが無い場合でも強化学習を用いた学習が可能.



新規性・結果・なぜ通ったか？

- 入力された形状からCNNで特徴量を抽出し, RNN（GRUs） によって形状を生成する一連のプログラムを生成
- Ground Truth が無い場合は強化学習(Policy Gradient)で学習可能（評価実験では教師ありと強化学習を組み合わせたものが一番高精度）
- 2次元や３次元の形状とそれを生成するプログラムのデータセット(2D and 3D synthetic dataset)を作成・公開
- 評価実験では, 2次元と３次元のいずれの場合も Nearest Neighbor を用いた手法よりも高精度を達成
- また, Primitive detection のタスクにおいては Faster R-CNN よりも高い Mean Average Precision を達成

コメント・リンク集

- [論文] CSGNet: Neural Shape Parser for Constructive Solid Geometry
- [Code] GitHub

[#10]

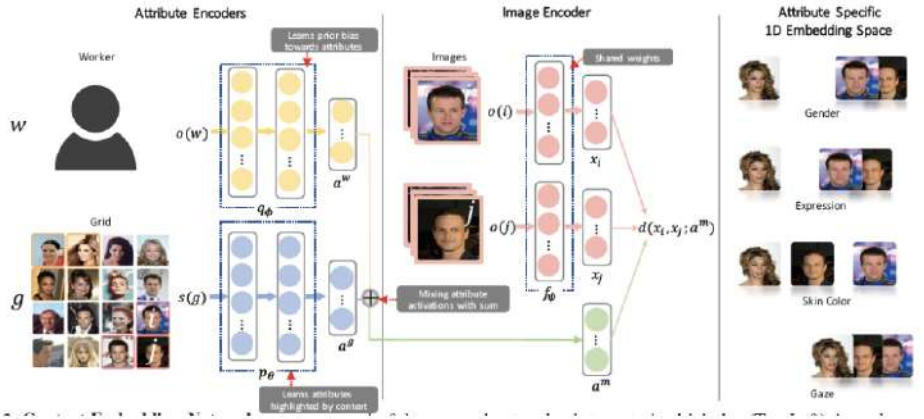
Shintaro Yamamoto

Context Embedding Networks

Kun Ho Kim, Oisin Mac Aodha and Pietro Perona CVPR2018

概要

ラベル付けする人の評価尺度やcontextを考慮して画像の類似度を求めるContext Embedding Networksを提案した。クラウドワーカーによるアノテーションは、個人独自の評価尺度やコンテキストに影響される。 例えば、人物顔画像をクラスタリングする際にはある人は性別によってクラスタリングするが、別の人は表情によってクラスタリングしてしまうと考えられる。 そこで、workerと見せた画像(context)それぞれから、画像のどのような点に注目するかを表すattributeをAttribute Encoderにより求める。 画像の類似度は、2枚の画像それぞれに対してImage Encoderから得られる画像特徴を、attributeによる重みつきの類似度によって求める。



新規性・結果・なぜ通ったか？

クラウドワーカーに応じた類似度の算出が可能になった。各クラウドワーカーがどのattributeに基づいて画像クラスタリングをしているかを予測することに成功した。

コメント・リンク集

クラウドソーシングによるアノテーションにおいて、クラウドワーカーの個人差は避けては通れないので重要な問題になりそう。

- 論文

[#11]

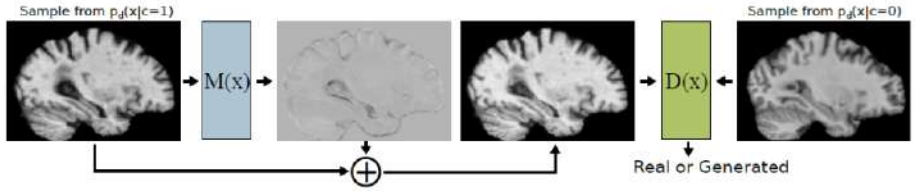
Shintaro Yamamoto

Visual Feature Attribution using Wasserstein Gans

Christian F. Baumgartner, Lisa M. Koch, Kerem Can Tezcan and Jia Xi Ang CVPR2018

概要

画像中のどの箇所がクラス分類に寄与するかを可視化する手法を提案。多くの手法は、クラス分類のタスクを学習することで重要な特徴を調べている。 しかし、識別への寄与が強い特徴が存在する場合ネットワークは強い特徴のみに注目してしまい、他の特徴は無視されてしまう。 医療画像からの病気の診断では、病気のステージを見極める、複数の要因が絡む病気を発見するなど無視されてしまう特徴を探すことは極めて重要である。 本研究では、Wasserstein GANを用いてある病気を発見する上で重要な領域を示したマップMを生成する。 病気のラベルがついた入力画像xに対して、x+Mが病気でないと判定されるMを生成するGeneratorを学習する。 その際、患者の個人性による画像の違いを考慮するためにL1正則化項をロスに加える。



新規性・結果・なぜ通ったか？

合成画像と実際の医療画像の2種類により評価した。従来の特徴を可視化する手法は、病気の際に見られる特徴のうち一部しか取れない、エッジなどの高周波情報が取れないという結果に対して、提案手法はこれら2つを改善した。 Normalized Cross Correlation(NCC)による数値評価では、ベースラインと比べ提案手法が最も良い数値を記録した。

コメント・リンク集

- コード
- 論文

[#12]

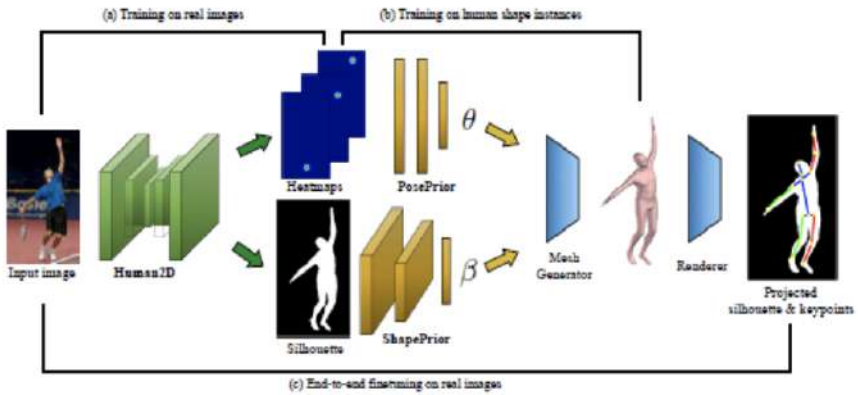
Shintaro Yamamoto

Learning to Estimate 3D Human Pose and Shape from a Single Color Image

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou and Kostas Daniilidis CVPR2018

概要

1枚のRGB画像から人間の全身の3次元モデルを推定するEnd-to-Endのネットワークを提案した。DNNを用いた3次元モデルの推定は、膨大なアノテーションが必要となり現実的ではない。 そこで、画像からの2次元特徴の抽出と2次元特徴から3次元モデルの推定の2段階に分けることによりDNNベースの手法を実現する。 始めに、Human2DというRGB画像から2次元の特徴点及び人物のシルエットを推定する。 2次元特徴点及びシルエットから3次元モデルの推定には、SMPLという統計モデルを用いて作成した学習データにより学習を行う。 加えて、得られた三次元モデルから2次元特徴点とシルエットを取得し、画像から得られた情報と一致するかをロスに加える。



新規性・結果・なぜ通ったか？

推定した3次元モデルの誤差を評価したところ、提案手法が最もground truthに近づいたことを確認した。1枚の画像に対して50msという従来研究と比べ大幅に高速化することができた。

コメント・リンク集

データ作成の問題をCGを駆使して解決しており、同様のアイデアを活用できないだろうか？

- 論文

[#13]

Zero shot Kernel Learning

Hongguang Zhang and Piotr Koniusz CVPR2018

概要

ゼロショット学習のオープンな問題に取り組む上で，カーネルを利用したゼロショット学習の手法を提案する．

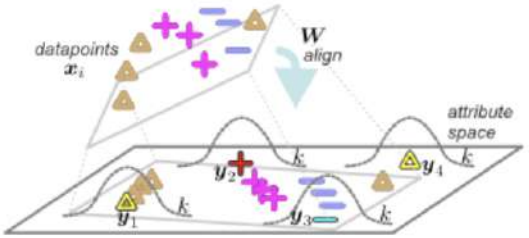


Figure 1: Zero-shot kernel alignment. Datapoints x are projected to the attribute space via the rotation and scaling matrix W which we learn. Kernels k are centered at y which are attribute vectors.

Method		AWA1			AWA2			SUN			CUB			aPY			Better than SOA
		ts	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	*ts*	*tr*	*H*	
DAP	[20]	0.0	88.7	0.0	0.0	84.7	0.0	4.2	25.1	7.2	1.7	67.9	3.3	4.8	78.3	8.0	
IAP	[20]	2.1	78.2	4.1	0.9	87.6	1.8	1.0	37.8	1.8	0.2	72.8	0.4	5.7	65.6	10.4	
CONSE	[24]	0.4	88.6	0.8	0.5	90.6	1.0	6.8	39.9	11.6	1.6	72.2	3.1	0.0	91.2	0.0	
CMT	[33]	0.9	87.6	1.8	0.5	90.0	1.0	8.1	21.8	11.8	7.2	60.1	8.7	1.4	85.2	2.8	
*CMT**	[33]	8.4	86.9	15.3	8.7	89.0	15.9	8.7	28.0	13.3	4.7	60.1	8.7	10.9	74.2	19.0	
SSE	[41]	7.0	80.5	12.9	8.1	82.5	14.8	2.1	36.4	4.0	8.5	46.9	14.4	0.2	78.9	0.4	
LATEM	[38]	7.3	71.7	13.3	11.5	77.3	20.0	14.7	28.8	19.5	15.2	57.3	24.0	0.1	73.0	0.2	
ALE	[1]	16.8	76.1	27.5	14.0	81.8	23.9	21.8	33.1	26.3	23.7	62.8	34.4	4.6	73.7	8.7	
DEVISE	[13]	13.4	68.7	22.4	17.1	74.7	27.8	16.9	27.4	20.9	23.8	53.0	32.8	4.9	76.9	9.2	
SJE	[2]	11.3	74.6	19.6	8.0	73.9	14.4	14.7	30.5	19.8	23.5	59.2	33.6	3.7	55.7	6.9	
ESZSL	[29]	6.6	75.6	12.1	5.9	77.8	11.0	11.0	27.9	15.8	12.6	63.8	21.0	2.4	70.1	4.6	
SYNC	[7]	8.9	87.3	16.2	10.0	90.5	18.0	7.9	43.3	13.4	11.5	70.9	19.8	7.4	66.3	13.3	
SAE	[17]	1.8	77.1	3.5	1.1	82.2	2.2	8.8	18.0	11.8	7.8	54.0	13.6	0.4	80.9	0.9	
Polynomial, $r=2$		5.8	77.3	10.7	6.4	78.8	11.8	20.6	31.5	24.9	16.7	61.3	26.2	4.8	77.5	9.0	0/5
Polynomial, $r=4$		5.7	78.7	10.6	7.0	83.0	13.0	20.0	31.7	24.5	**24.2**	63.9	**35.1**	5.7	79.2	10.6	1/5
Polynomial, $r=6$		8.3	78.1	15.0	8.7	81.6	15.7	21.0	31.0	25.1	23.8	58.6	33.8	4.9	78.3	9.2	0/5
Cauchy		6.0	79.9	11.1	6.2	82.7	11.5	16.1	29.7	20.9	18.2	49.6	26.6	1.0	84.9	2.0	0/5
Cauchy-Ort		**18.3**	79.3	**29.8**	17.6	80.9	29.0	19.8	29.1	23.6	19.9	52.5	28.9	11.9	76.3	**20.5**	**3/5**
Gaussian		6.1	81.3	11.4	7.3	79.1	13.3	18.2	33.2	23.5	17.5	59.9	27.1	3.0	82.3	5.8	0/5
Gaussian-Ort		17.9	82.2	29.4	**18.9**	82.7	**30.8**	20.1	31.4	24.5	21.6	52.8	30.6	10.5	76.2	18.5	2/5

Table 3: Evaluations on the generalized zero-shot learning protocol and the newly proposed datasplits. We indicate the mean top-1 accuracy on (*tr*) train+test classes and (*ts*) test classes only. Moreover, (*Better than SOA*) indicates the number of datasets on which our methods outperform the other state-of-the-art methods (the upper part of the table) according to the harmonized score (*H*).

新規性・結果・なぜ通ったか？

提案する手法は，回転とスケーリングが組み込まれているため，制約のないモデルでは，より自由度が高いために過学習を防止することができる．1枚目の画像はゼロショットカーネルの配置． 2枚目の画像は一般化ゼロショット学習プロトコルと新たに提案されたデータ集合についての評価． （tr）はtrain + testクラス，（ts）はテストクラスの平均トップ1精度，(H)はハーモナイズされたスコア，

コメント・リンク集

- link1

[#14]

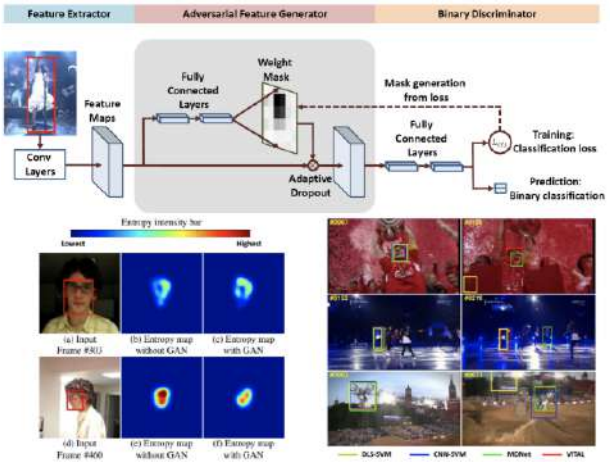
Takahiro Itazuri

VITAL: VIsual Tracking via Adversarial Learning

Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson Lau, Ming-Hsuan Yang

CVPR 2018

概要



tracking-by-detectionベースの手法は、(1)各フレームにおけるpositive sampleが空間的に重なった領域を取りやすいため、十分な見た目のばらつきを学習できない点と(2)positive sampleとnegative sampleの不均等さ（class imbalance）が顕著に出てしまうという点が問題である。本論文では、positive sampleのデータ拡張を行うため、GANを用いて長い時間のスパンで頑健な特徴を学習可能なVITALアルゴリズムを提案した。またclass imbalanceを解決するため、識別が容易なnegative sampleを取り除くためのhigh-order cost sensitive lossを提案した。

新規性・結果

提案手法はCNNで抽出した特徴量に適用するマスクを複数（論文では9個）用意し、マスクを通じて重み付けられた特徴量に対して識別器Dが対象物体か背景かの二値分類を行う。学習時には識別器Dに最も悪い識別性能を出させたマスクを学習させる。テスト時には生成器Gは取り除いておく。また識別が簡単すぎる大量のnegative sampleのロスが合計されて大きくなってしまう現象であるclass imbalanceを、あまり学習に寄与しないようにする。

リンク集

- 論文
- デモ動画
- プロジェクト
- GitHub

[#15]

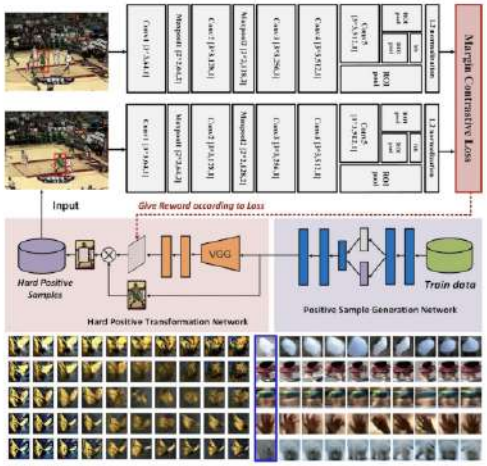
Takahiro Itazuri

SINT++: Robust Visual Tracking via Adversarial Positive Instance Generation

Xiao Wang, Chenglong Li, Bin Luo, Jin Tang CVPR 2018

概要

物体追跡タスクでは追跡対象の画像を１フレーム目においてのみ与えられるため、トレーニングデータの多様性が不足していることがDNNを適用する際の障壁となっている。そこで変形や遮蔽といった困難な環境下における正解サンプルを生成する手法（SINT++）を提案した。提案手法は他の物体追跡手法に取り入れることが可能である点も非常に重要である。



新規性・結果

VAEを用いて追跡対象の多様体を生成し、その多様体局面上を移動させることで正解サンプルを増やすネットワーク（PSGN）と識別器の認識性能にクリティカルな領域を探すように遮蔽領域を決定する強化学習ネットワーク（HPTN）を用いて、正解サンプルの多様性を増幅させる。追跡器はSINTを用いているため、与えられた追跡対象の画像に対するオフライン学習も、追跡中のオンライン学習も行わない。

リンク集

- 論文
- プロジェクト

[#16]

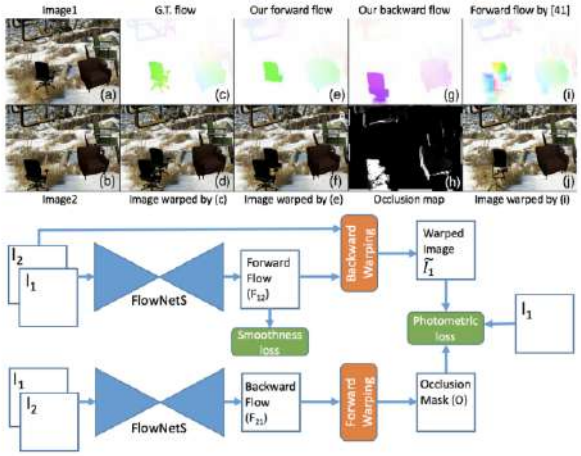
Takahiro Itazuri

Occlusion Aware Unsupervised Learning of Optical Flow

Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, Wei Xu CVPR 2018

概要

オプティカルフローのアノテーションが困難であることから、教師なし学習ベースのオプティカルフロー推定手法が提案されているが、十分な精度が出ていない。そこで問題とされている遮蔽と大きな動きに対応したネットワークを提案。教師なし学習ベースの手法では最も良い精度を出し、教師あり学習ベースの手法とのギャップを埋めた。



新規性・結果

２枚の画像に対して、１枚目から２枚目へのオプティカルフローと、２枚目から１枚目のオプティカルフローを推定する。２枚目の画像と前者のオプティカルフローを用いて、１枚目の画像を復元する。復元した１枚目の画像のうち遮蔽が発生していない部分に対して、本物の１枚目の画像との差を損失として用いる。

リンク集

- 論文

[#17]

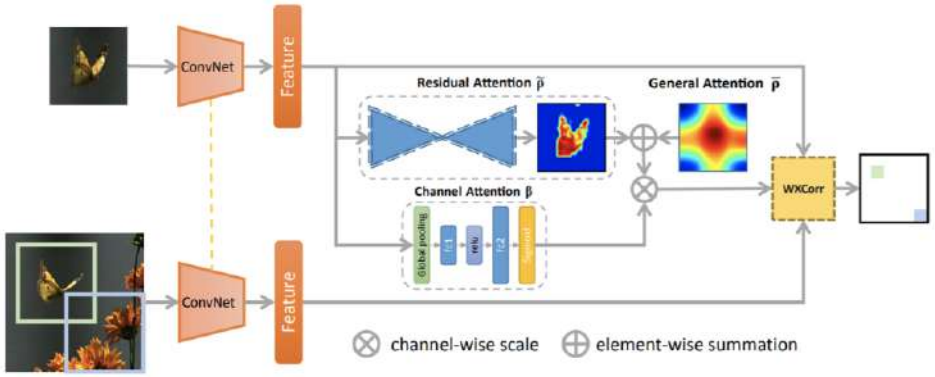
Takahiro Itazuri

Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking

Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, Steve Maybank CVPR 2018

概要

物体追跡のためのオフライン学習ベースの手法は精度とスピードにおいて高いポテンシャルがあるが、追跡対象に適応させることは困難である。一方で、オンライン学習ベースの手法は計算コストとオーバーフィッティングが問題になっている。本論文では、Siamese NetworkにおけるCross CorrelationをAttentionで重み付けしたRASNet（Residual Attentional Siamese Network）を提案し、リアルタイムを超える速度（83fps）とSOTAを実現した。



新規性・結果

Siamese NetworkにAttention Mechanismを導入した。Attention MechanismにはResidual AttentionとGeneral Attentionを含むDual Attentionと、Channel Attentionを導入した。Resiual Attentionは追跡対象に特化させるようにオンライン学習をし、Channel Attentionはチャンネルごとの特徴量の質を示している。

リンク集

- 論文
- GitHub

[#18]

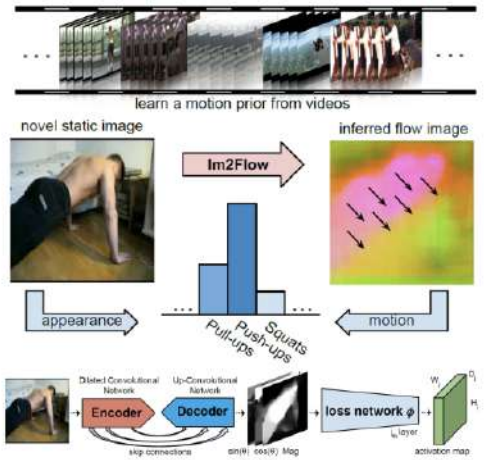
Takahiro Itazuri

Im2Flow: Motion Hallucination from Static Images for Action Recognition

Ruohan Gao, Bo Xiong, Kristen Grauman CVPR 2018

概要

人間が一枚の静止画から動き情報を推定可能であることを受け、一枚の静止画から動き情報（フロー）の事前知識を得る手法を提案。具体的には動き情報の表現方法とU-Netの構造を変形させたエンコーダ・デコーダネットワークを提案。提案手法で得たフロー情報を利用することで、行動認識の精度が向上した。



新規性・結果

動き情報を動きの大きさと角度（角度はコサインとサインに分解）の計３チャンネルで表現する。角度は周期的な構造であるが、三角関数を用いることでこれを避けることができる。損失関数は(1)フロー自体の損失と(2)動き情報のコンテンツの損失の和で構成される。動き情報のコンテンツは、ResNetをUCF-101データセット上で行動認識にfine-tuningさせたものから取得し、推定したフローと正解のフローから得られたコンテンツの差から損失を得る。

リンク集

- 論文

[#19]

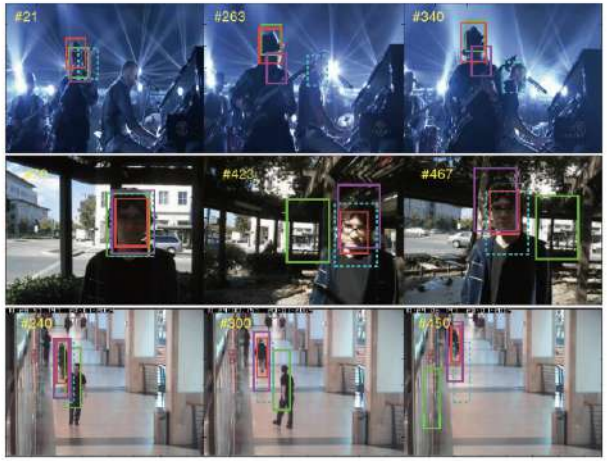
Takahiro Itazuri

High-Speed Tracking With Multi-Kernel Correlation Filters

Ming Tang, Bin Yu, Fan Zhang, Jinqiao Wang CVPR 2018

概要

物体追跡タスクにおいて、Multi-Kernel Correlation Filter (MKCF)はKernelized Correlation Filter (KCF)のカーネルを複数にすることで識別性能を向上させているが、計算量がボトルネックとなっていた。そこで目的関数の上界を目的関数として再設定し、上から押さえるように最適化問題を解くことで、MKCFより高速（150fps）かつ高識別性能な物体追跡手法 (MKCFup)を提案した。



新規性・結果

MKCFupは従来のMKCFの最適化問題における目的関数の上界を最適化する。上界を最適化する問題に再定式化することで高速かつ高精度な追跡を実現しており、DNNを使っていない数少ない論文の1つ。Correlation FilterがDNNベースの物体追跡に利用されているように、今後DNNベースの物体追跡手法が使用する可能性がある。

リンク集

- 論文

[#20]

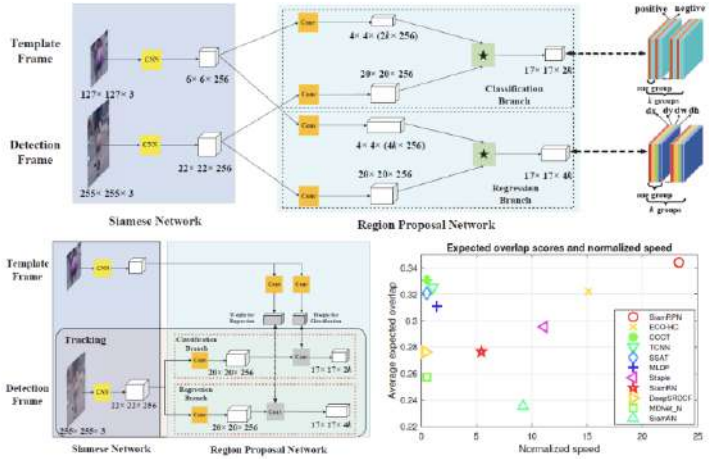
Takahiro Itazuri

High Performance Visual Tracking With Siamese Region Proposal Network

Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, Xiaolin Hu CVPR 2018

概要

オフラインで学習させたDNNで得た特徴量を使用した物体追跡手法は、ターゲットの動画に特有の情報を使用していないことから、相関フィルタベースの手法より良い精度が出ていなかった。提案手法は大規模な画像ペアデータを用いて学習し、同じ特徴量抽出器を２つの入力に適応させて得た特徴量の類似度を比較するSiamese NetworkとFaster R-CNNで提案されているRegion Proposal Network（RPN）を組み合わせた上で、物体追跡をlocal one-shot detectionとして定式化することで、高速かつ高精度な追跡を実現した。



新規性・結果

従来のSiamese Networkを利用した手法とは異なり、RPNを用いることで物体の変形に合わせた矩形領域を提示することによって高い精度を出すことが可能である。また物体追跡をlocal one-shot detectionとして定式化する。

リンク集

- 論文

[#21]

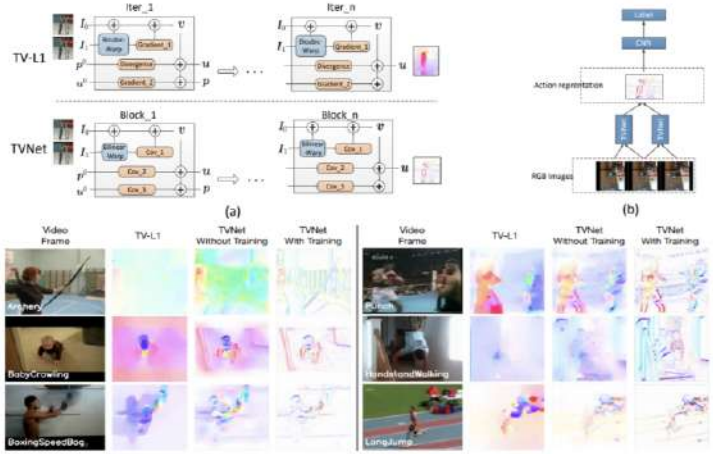
Takahiro Itazuri

End-to-End Learning of Motion Representation for Video Understanding

Lijie Fan, Wenbing Huang, Chuang Gan, Stefano Ermon, Boqing Gong, Junzhou Huang CVPR 2018

概要

深層学習の成功に反して映像解析では未だに手作りのオプティカルフローが使用されている。通常のオプティカルフローは、それを利用したCNNと独立してしまっている点と時間的・空間的計算コストが非常に大きい点が問題である。本論文では、オプティカルフローに代わる特徴をEnd-to-Endに学習可能なネットワーク（TVNet）を提案した。End-to-Endに学習可能になることで、特定のタスクに特化した動き特徴量を学習できる。



新規性・結果

オプティカルフロー抽出手法の1つであるTV-L1をDNNにカスタマイズさせた。End-to-Endのネットワークにすることで、フロー抽出後のタスクから得られた誤差を伝搬することができるため、特定のタスクに特化した動き情報の抽出が可能となっている。

リンク集

- 論文GitHub

[#22]

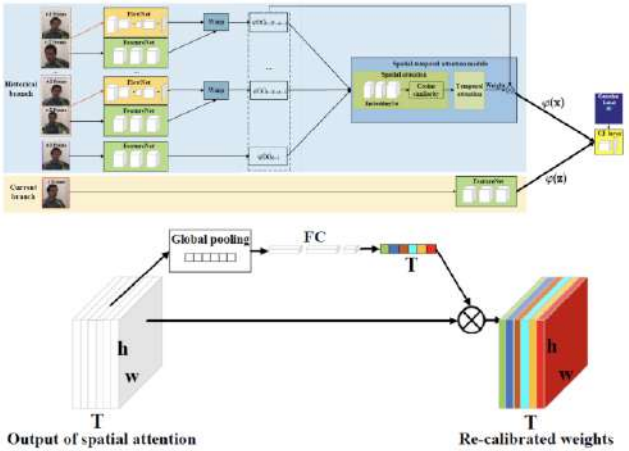
Takahiro Itazuri

End-to-End Flow Correlation Tracking with Spatial-temporal Attention

Zheng Zhu, Wei Wu, Wei Zou, Junjie Yan CVPR 2018

概要

従来のCorrelation Filterベースの物体追跡手法は現在のフレームの見た目しか考慮できておらず、フレーム間の情報や動きの情報を考慮していなかった。本論文ではフロー情報を直接的に考慮することで時間変化に関する情報を考慮することが可能な物体追跡手法を提案した。



新規性・結果

通常のネットワークに対してフロー情報を追加しただけではなく、Spatial AttentionとTemporal Attentionも提案した。これにより空間情報と時間情報を効率的に考慮することが可能となった。

リンク集

- 論文

[#23]

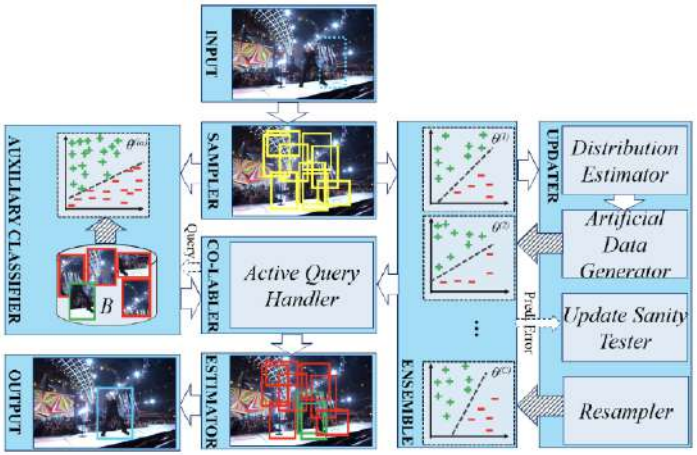
Takahiro Itazuri

Efficient Diverse Ensemble for Discriminative Co-Tracking

Kourosh Meshgi, Shigeyuki Oba, Shin Ishii CVPR 2018

概要

tracking-by-detectionベースの物体追跡手法は識別器の不完全性からオンライン自己学習するため、自己学習のループでドリフト問題が発生する。そこで学習する識別器に対する教師が必要であるという発想から、相補的に教師になるアンサンブル学習ベースの手法が提案されている。しかし、アンサンブル学習ベースの手法は、各識別器が互いに重複した領域を対象にする冗長性が発生する。本論文ではその冗長性を軽減することが可能なリアルタイム物体追跡手法（DEDT: Diversified Ensemble Discriminative Tracker）を提案する。



新規性・結果

DEDTは高い適応性と多様性を持つ識別器群であるCommitteeモデルと長期記憶を持つAuxiliaryモデルからなり、Committeeモデルが不明確な回答を出した入力に対しては、Auxiliaryモデルが代わりに回答する。Committeeモデルは自身が不明確な回答をしたデータを用いて学習する。またこれまでのデータから不明確な回答になるようなデータを人工的に生成し、そのデータにおけるエラー率が、推定時に冗長な結果が得られたデータのエラー率より小さくなるまで繰り返し、更新することで、冗長性を回避する。一方でAuxiliaryモデルはCommitteeモデルより更新頻度が低くすることで長記憶性を持つ。

リンク集

- 論文

[#24]

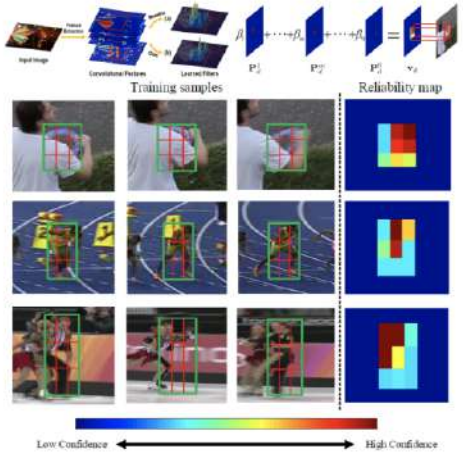
Takahiro Itazuri

Correlation Tracking via Joint Discrimination and Realiability Learning

Chong Sun, Dong Wang, Huchuan Lu, Ming-Hsuan Yang CVPR 2018

概要

Correlation Filterベースの物体追跡手法は識別性と信頼性を学習するべきであるが、従来手法は識別性に着目したものが多く、Bounding Box内の予期されない顕著な領域に影響を受ける可能性がある。本論文では信頼性の高い領域に特に着目して物体追跡を行う手法（DRT）を提案した。



新規性・結果

提案手法は識別性を保持するbase filterと信頼性を保持するreliability termのアダマール積を取ることで、より信頼性の高い領域に着目する。目的関数には学習サンプルの分類誤差に関する項と、局所応答に一貫性を持たせる制約項、L2ノルム正則化項からなる。

リンク集

- 論文

[#25]

Takahiro Itazuri

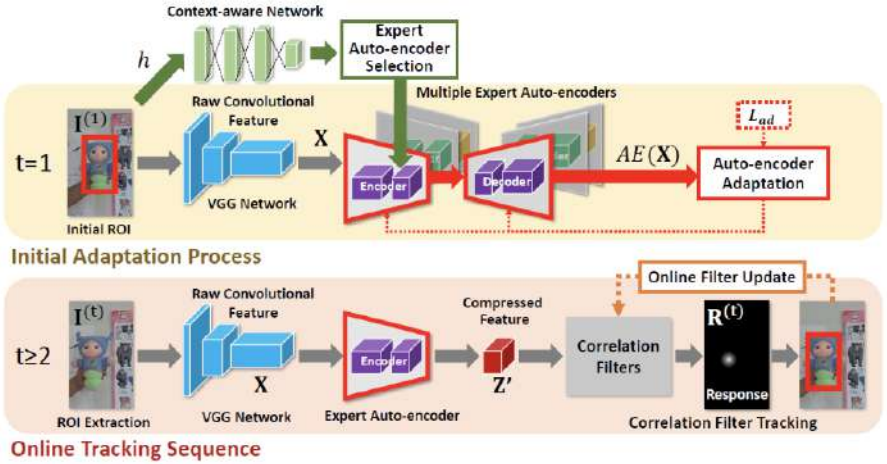
Context-aware Deep Feature Compression for High-speed Visual Tracking

Jongwon Choi, Hyung Jin Chang, Tobias Fischer, Sangdoo Yun, Kyuewang Lee, Jiyeoup Jeong, Yiannis Demiris, Jin Young Choi

CVPR 2018

概要

コンテキストを考慮したCorrelation Filterによる物体追跡手法を提案した。カテゴリごとに事前学習したオートエンコーダーのエキスパートを複数用意し、その中からコンテキストネットワークが1つ選択する。



新規性・結果

リアルタイム性が重要である物体追跡タスクでは、リアルタイムにDNNを学習することは困難である。本論文では事前に各物体のカテゴリ別に学習したオートエンコーダーを用意し、その中から1つを選択することで、ある程度既に特定の物体に特化したネットワークを使用できるため、再学習の必要性を軽減することができる。

リンク集

- 論文

[#26]

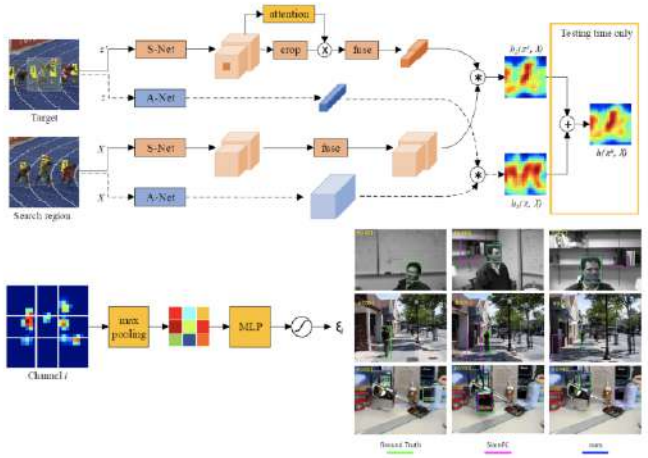
Takahiro Itazuri

A Twofold Siamese Network for Real-Time Object Tracking

Anfeng He, Chong Luo, Xinmei Tian, Wenjun Zeng CVPR 2018

概要

物体追跡手法の１つであるSiamFCは効率的なオフライン学習を行うことで、非常に高い識別性能を持つが、追跡対象の見た目の変化に弱かった。そこで、見た目特徴量とセマンティックな情報を別々に抽出する２つのSiamese Networkを利用することで、追跡対象の見た目変化にも強い物体追跡手法を提案した。セマンティックな情報を抽出するネットワークは画像分類タスクで学習させることで、見た目の変化に頑健な特徴量を抽出することが可能となる。



新規性・結果

推論フェーズでは、それぞれのネットワークで別々に追跡対象画像と探索画像の類似度を計算し、それを統合する。セマンティックな情報を抽出するネットワークは、見た目変化には頑健ではあるが、識別性能は不十分であるため、与えれた追跡対象に反応するチャンネルの重要度を増やすChennel Attentionを追加する。これによって追跡対象に適応する最低限の機能を追加している。

リンク集

- 論文

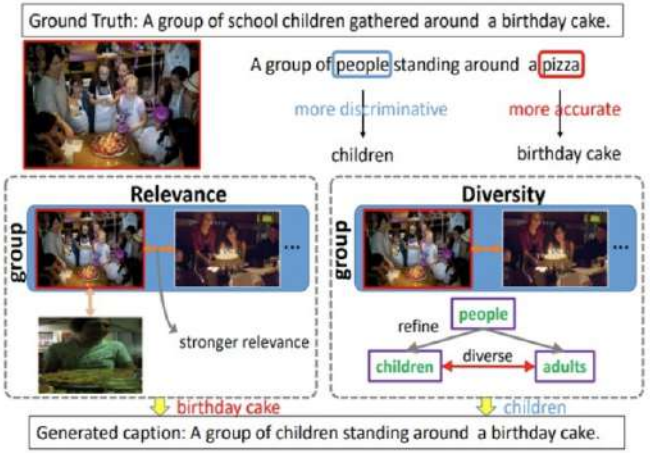
Munetaka Minoguchi

GroupCap: Group-based Image Captioning with Structured Relevance and Diversity Constraints

Fuhai Chen, et al. CVPR 2018

概要

画像グループ内での関連性や相関関係などを考慮し、キャプションを出力するGroupCapの提案。まず、個々の画像でvisual tree parser(VP-Tree)を構成し、文字ベースで意味の相関を構築。次にツリーの関係から、画像間での関連性と多様性をモデル化。この制約関係をもとにLSTMでキャプション生成。これらをトリプレットロスとしてend-to-endで学習する。



新規性

従来のイメージキャプショニングでは、単一画像に対して説明文を生成している場合がほとんど。これらはオフラインで学習し、画像間での視覚的構造関係を無視して推定している。本手法のグループベースの手法によって、グループ画像内での構造的関連性や多様性を協調して学習することでキャプションの正確性を向上させる。

結果・リンク集

MSCOCOをもとに作成した2グループキャプションデータセットを使用して評価し、優れていることを示唆。

- 論文

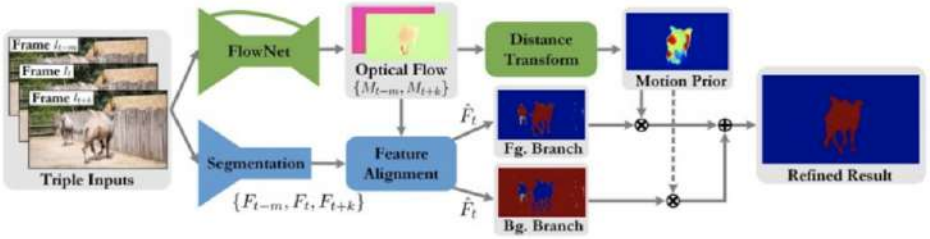
Munetaka Minoguchi

MoNet: Deep Motion Exploitation for Video Object Segmentation

Huaxin Xiao, et al. CVPR 2018

概要

動画中の物体にセグメンテーションを行うタスクにおいて、フレーム間処理をモーションキューによって改善するMoNetの提案。オプティカルフローを利用し、その近傍の表現を統合することにより、ターゲットフレームでの表現を強化する。これにより、時間変化におけるコンテキスト情報を活用することができ、外観変動やモーションブラー、物体の変形に頑健となる。また、動作の一致性を考慮することで、ノイズの大きいモーションキューを前景または背景に変換し、精度を向上させている。



新規性

セグメンテーションの改良と、フレームごとの学習を行うという観点からモーションキュー(オプティカルフロー)を利用している。これによって、前景と背景の分離する制度を向上。 また、distance transform layerを提案し、動作が一致しないインスタンスと領域をフィルタリングすることができる。

結果・リンク集

実験において、モーションキュー利用の有効性と、 distance transform layerの有効性を示している。

- 論文

[#29]

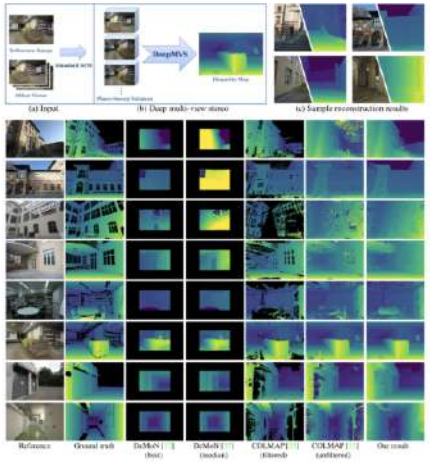
Yoshihiro Fukuhara

DeepMVS: Learning Multi-view Stereopsis

Po-Han Huang et al. CVPR 2018

概要

Learning-based Multi-View Stereo の研究. 任意の枚数の画像から, 視差 Map の推定を行う（推定結果は入力の順番に依存しない）. また, ネットワークの学習のため, 新しい synthetic datasets (MVS-SYNTH dataset) を作成・公開した. ETH3D を用いた評価実験では DeMoN を上回り, COLMAP と同等の結果を達成した.



新規性・結果・なぜ通ったか？

- 複数枚の画像（1枚の参照画像と複数枚の近傍画像）を入力とする, Learning-based Multi-View Stereo（MVS）の手法を提案
- 入力画像に対して通常の SfM(COLMAP) を用いてポーズの推定を行った後 , D段階の離散的な視差の大きさ毎に近傍画像を参照画像に Warp した画像群 (plane-sweep volume) を生成
- 参照画像と各 plane-sweep volume に対して Patch matching を行って抽出された特徴量を encoder-decoder 型のネットワークで統合した特徴量を用いて視差 Map を推定
- ネットワークを上手く学習させるためには real と synthetic の両方のデータセットが重要であるとし, 新しい synthetic datasets (MVS-SYNTH dataset) を作成・公開した
- ETH3D を用いた評価実験で COLMAP[Schonberger+16] と

コメント・リンク集

- [論文] DeepMVS: Learning Multi-view Stereopsis
- [Project page] DeepMVS: Learning Multi-view Stereopsis
- [Code] GitHub

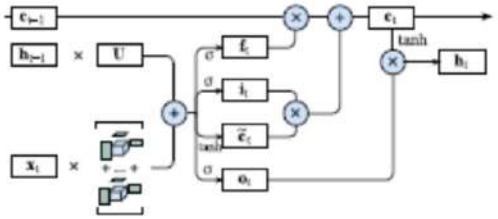
Kazuma Matsui

Learning Compact Recurrent Neural Networks with Block-Term Tensor Decomposition

Jinmian Ye et al. CVPR 2018

概要

RNNは強力なシーケンスモデリングツールであるが，高次元の入力を扱う場合，RNNのトレーニングはモデルパラメータが大きくなるため計算に時間がかかるという問題がある．これは，RNNがビデオや画像キャプションのアクションレコグニションなど，多くの重要なコンピュータビジョンのタスクを行うことを妨げる．この問題を解決するためにRNNのパラメータを大幅削減し，トレーニング効率を向上させるコンパクトで柔軟な構造「Block-Termテンソル分解(BTD)」を提案し，これをBlock-Term RNN (BT-RNN)と名付ける．テンポトレインRNN (TT-RNN)のような他の低ランク近似とBT-RNNを比較すると，同じランクを使用する場合，より簡潔でより良い近似が可能であり，より少ないパラメータで元のRNNに戻すことが可能である．ビデオ，画像キャプション，画像生成のアクションレコグニションを含む3つの困難なタスクに対し，BT-RNNは予測精度と収束速度の両方でTT-RNNや標準のRNNより優れていると言える．この研究において，BT-LSTMはUCF11データセットのアクションレコグニションのタスクで15.6%以上の精度向上を達成するために，標準LSTMより17,388回少ないパラメータを使用した．



新規性・結果・なぜ通ったか？

BTDは最適なTT-rankの設定を見つけることを困難にする代わりに次のような利点がある．・Tucker分解は異なる次元間の相関関係を表し，より良い重み分担を達成するためにコアテンソルを導入している。 ・コアテンソルのランクを等しくすることができ，異なる次元での不均衡な重みの共有を避けることができ，かつ入力データの異なる順列に対して頑強なモデルを導くことができる． ・BTDは，

コメント・リンク集

- 論文

[#3]

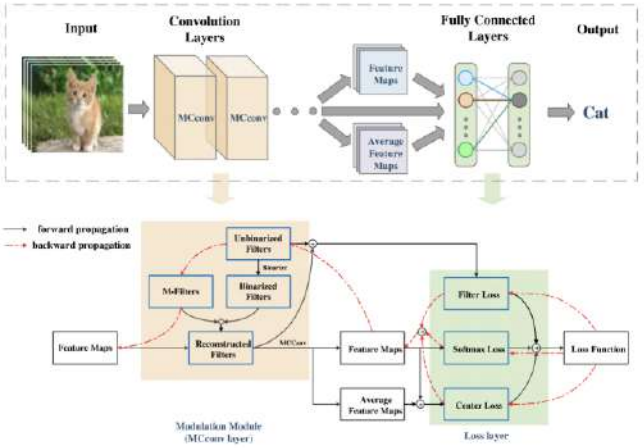
Kazuki Tsubura

Modulated Convolutional Networks

Xiaodi wang , Baochang Zhang

概要

・CNNは画像処理の様々なタスクをこなすうえでとても有効だが，ネットワークのストレージにかなりのコストを要求するため，展開が制限される．2値化フィルタを用いたCNNの移植性向上のための新しい変調畳み込みネットワーク(MCNs)を提案する．MCNでは，end-to-endフレームワークにおけるフィルタ損失，中心損失，ソフトマックス損失を考慮した新しい損失関数であるM-フィルタを提案する．



新規性・差分

・非二項フィルタを復元するために，M-フィルタを導入しネットワークモデルを計算するための新しいアーキテクチャを導出する．MCNは完全精度モデルとは対照的に，畳み込みフィルタの必要な記憶スペースのサイズを32倍に縮小することができ，最先端の2値化モデルよりもはるかに優れた性能を達成した．また，MCNは完全精度のResentsおよびWideResentsと同等のパフォーマンスを達成した．

- 論文

End-to-End Dense Video Captioning with Masked Transformer

Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, Caiming Xiong CVPR 2018

概要

動画内のいつ行動が行われたかのTemporal Action Proposals(TAP)とどのような行動が行われたかのキャプションを行うタスクにおいて，self-attentionを用いて既存手法を改善する．

img(src=`${figpath}End-to-End_Dense_Video_Captioning_with_Masked_Transformer_1.png`,al
to-End_Dense_Video_Captioning_with_Masked_Transformer_1)

新規性・結果・なぜ通ったか？

ActivityNet CaptionsとYouCookIIでキャプションの評価を行い，METEORスコアが10.12と6.58であった．

SoTAではないが，時間的なイベントの検出とイベントのキャプショニングをEnd-to-Endに行う手法であること．また，このようなタスクで初めてのRNN-basedでは無い手法を提案したこというところが新規性．

コメント・リンク集

- 論文
- arxiv

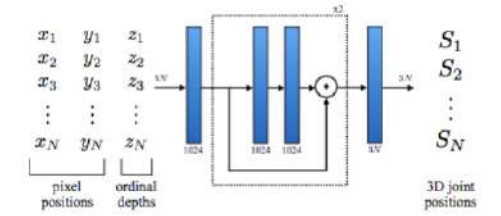
Shusuke Shigenaka

Ordinal Depth Supervision for 3D Human Pose Estimation

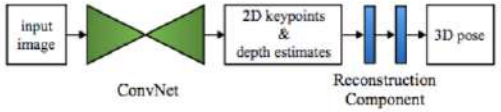
Georgios Pavlakos, Xiaowei Zhou, Kostas Daniilidis CVPR 2018

概要

3D ground truthの存在しないデータに対し人間の関節の奥行きデータの監視信号を使用することを提案。人体関節の奥行きを用いて3Dの姿勢推定をConvNetsで学習すると正確な関節座標で学習結果を得ることができる。 通常の深さ注釈をもつ2Dポーズデータセット(LSPとMPII)はConvNetsの学習に容易に組み込むことができるため、 ポーズデータセットを拡張させることにより3Dの姿勢に対する序数の深さ正確なものにし、 標準のベンチマークでstate-of-the-artを達成した。



(a) The reconstruction component.



(b) Integration of the reconstruction component.

新規性・結果・なぜ通ったか？

- 3D ground truthを必要としない
- 2Dポーズデータセットを使うことで、スタジオ以外の条件での3Dポーズ推定でも高い精度を得ることができる
- Human3.6Mのデータセットではこれまで誤差が47.7だったのに対し41.8を達成しており、HumanEva-Iデータセットにおいてはこれまで誤差が24.6だったのに対し18.3と大幅に更新をしている

リンク集

- 論文

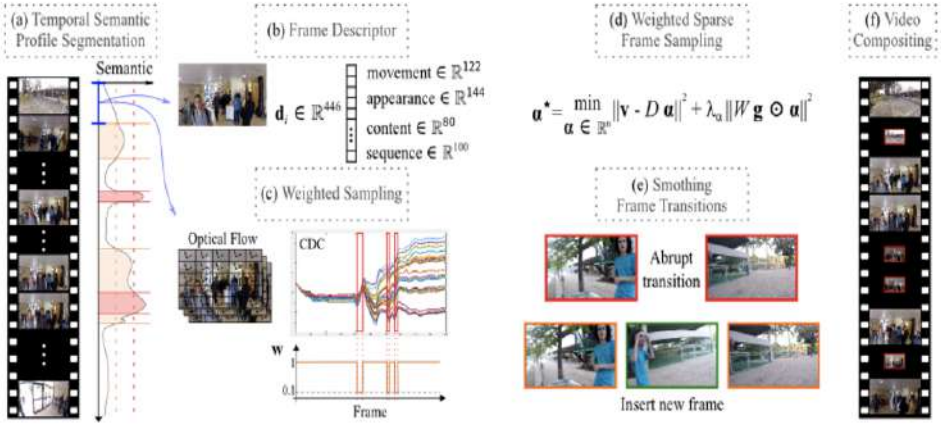
Ryota Suzuki

A Weighted Sparse Sampling and Smoothing Frame Transition Approach for Semantic Fast-Forward First-Person Videos

M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos CVPR2018

なめらかに**早送り**するという，ビデオ要約の新たな形を提案．

新しい適応的なフレーム選択手法を提案．重み付き最小値再構築問題として定式化． そこに，スムーズなフレーム遷移の手法を組み合わせる． 通しで見るとなめらかに見えるようにフレームを落とす．



新規性・結果・なぜ通ったか？

問題設定が面白い．流行りのビデオ要約の流れを汲みつつ，意識的に新しい枠組みを提案している． しかも十分実行可能と思われる問題である．想定される成果の見栄えもよい． 解き方もちゃんとしている．

コメント・リンク集

- 論文
- プロジェクトページ
- ソースコード

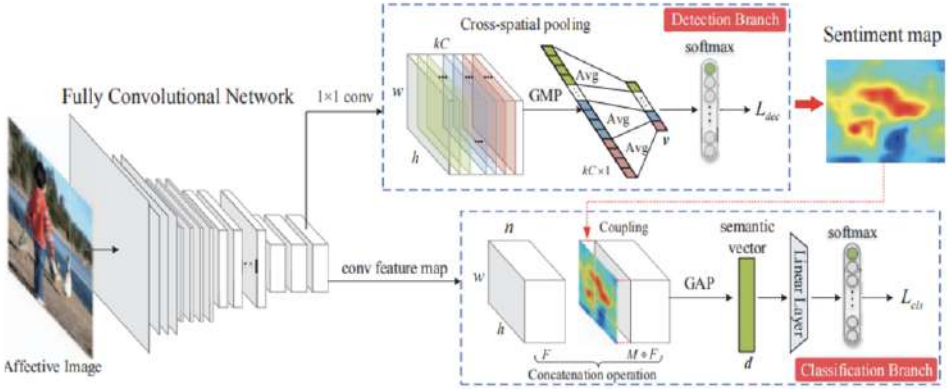
Ryota Suzuki

Weakly Supervised Coupled Networks for Visual Sentiment Analysis

J. Yang, D. She, Y. Lai, P.L. Rosin, M. Yang CVPR2018

画像で感情分析を行う研究．従来法は全体的な画像特徴からセンチメント表現を学習していたが，本研究では局所特徴もとらえるようにした．

弱教師付き二つ組CNNによる．(1)感情に特定的にソフトマップを検出するFCNN．画像レベルのラベルだけ必要にしたので，画素レベルアノテーションのようなアノテーション負荷が低くて済む．(2)ロバストなクラス分類のために，深層特徴を使い，感情マップを2つ組することによって，全体・局所情報の両方を活用．そして，これら2つを統合してEnd-to-Endで最適化できるようにする．



新規性・結果・なぜ通ったか？

より詳細に画像を見るように設計した．その結果，6つのベンチマークで評価を行い，SOTA性能を達成．

コメント・リンク集

- 論文

[#7]

Ryota Suzuki

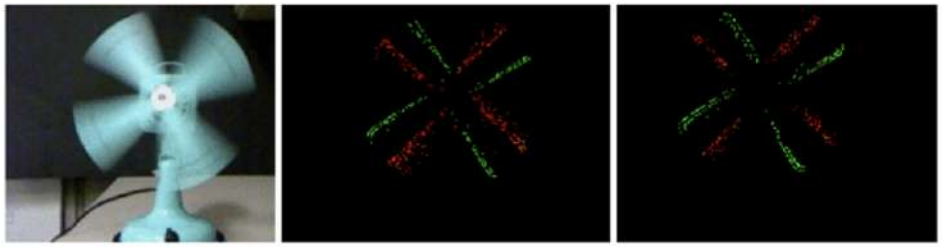
A Low Power, High Throughput, Fully Event-Based Stereo System

A. Andreopoulos, H.J. Kashyap, T.K. Nayak, A. Amir and M.D. Flickner CVPR2018

著者らIBMが開発した100万個のノードが伝達しあうニューラルネットワークを模倣したプロセッサ「TrueNorth」を使った，新しいカメラ「Dynamic Vision Sensor」を使ってステレオしてみた論文．

Dynamic Vision Sensorは，通常カメラのフレーム撮影方式ではなく，イベントベースに，各画素が非同期で撮影するという新たな撮影方式のセンサである． これにTrueNorthを組み合わせれば，完全にグラフベースで，配列などのあらゆるデータ構造無しに フォン・ノイマン型計算モデルの計算が可能である．

これにより，2000fpsの視差マップ生成を達成．通常のカメラではとらえられない急激な変化をとらえることが可能． しかも200倍省エネ．



新規性・結果・なぜ通ったか？

上記参照．

コメント・リンク集

新製品の宣伝的論文っぽい．確かに面白いカメラシステムなので，今後これを軸に新たな枠組みが発生するかもしれない？

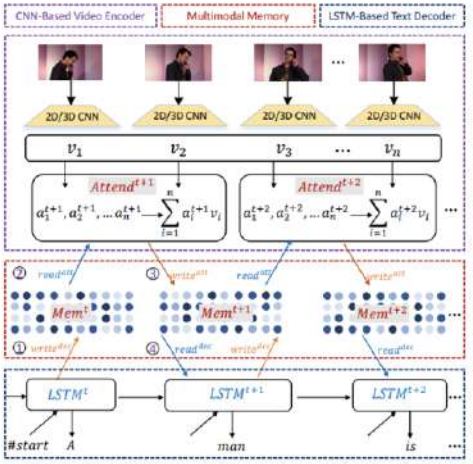
- 論文

Ryota Suzuki

M3: Multimodal Memory Modelling for Video Captioning

J. Wang, W. Wang, Y. Huang, L. Wang, T. Tan CVPR2018

ビデオキャプショニングの話題．Long-Termのマルチモーダルな依存性のモデリングと 文脈的ミスアラインメントがあるのに対し，(1)メモリモデリングするのは Long-Term系列的問題に対して 潜在的な利点がある （なにそれ），(2)視覚的アテンションにおいてワーキングメモリは主要素，という二点の事実を考慮した，Multimodal Memory Modelling（M3）を提案． LSTMの外部に視覚-テキスト間共有メモリを持ち，Long-Termな視覚-テキスト間依存性をモデル化する．



新規性・結果・なぜ通ったか？

MSVD，MSR-VTTで評価し，BLEU，METEORにおいてSOTA性能．

コメント・リンク集

HMMのように見える．

- 論文

Ryota Suzuki

Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push

S. Gorji and J.J. Clark CVPR2018

画像における静的なSaliency Modelを，動的なビデオのSaliencyの予測に使う手法．この著者らは，前回に写真内に写っている人の注視（Attention）をCNNのAttentionと組み合わせるというShared Attentionに関する論文を出していたが，今度は写真を撮る人・シーンに映っている人のShared Attentionについて取り組んだ．

マルチストリームCNN-LSTM構造を提案．これはSoTAなSaliencyをDynamic Attentional Pushに拡張する．

4つのステージからなる．Saliencyステージと，3つのAttentional Pushステージ．この複数ステージ構造は，Augmenting ConvNetに従っている． ConvLSTMの補足（complementary）と時間変化出力組み合わせで学習． 拡張したSaliencyと，ビデオにおける「見ている人」修正パターンの間のRelative Entropyの最小化を行う．

新規性・結果・なぜ通ったか？

動画データセットHOLLYWOOD2，UCF-Sport，DIEMにおいて，SoTAな時空間Saliency推定性能を達成．

コメント・リンク集

発展ネタを自分で出して，しかもCVPR連続当選．

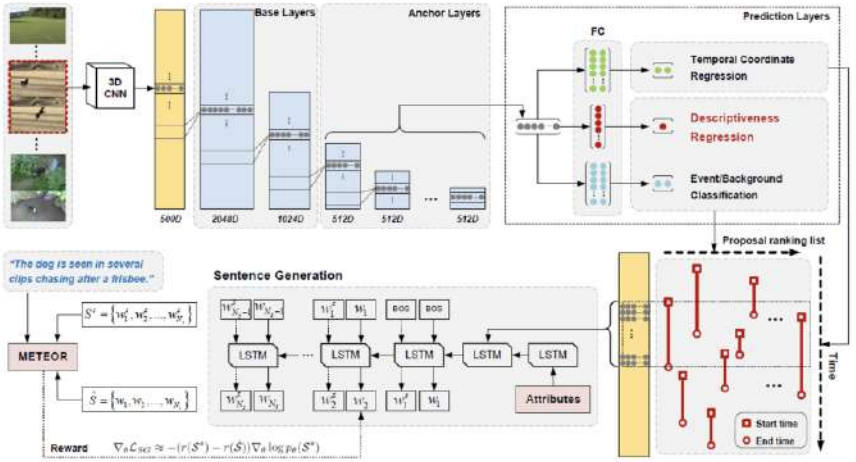
- 論文

Ryota Suzuki

Jointly Localizing and Describing Events for Dense Video Captioning

Y. Li, T. Yao, Y. Pan, H. Chao and T. Mei CVPR2018

Dense Video Captioningの話．イベントの発生時間のプロポーザルと，それぞれのイベントにおける文章生成の両者を結合的にEnd-to-Endで学習する， Descriptiveness Regressionを提案． シングルショット検出に組み込む．これは文章生成を経由したプロポーザル時間ごとの説明的複雑性を推論する． これが時間定位の調節につながるらしい． キャプショニングと検出の結合・汎用最適化をするところが他手法と異なるらしい．



新規性・結果・なぜ通ったか？

動画データセットActivityNetにおいてSoTAを達成．著者らはMETEORで12.96%出たのがすごいと言っている．

コメント・リンク集

Dense Video Captioning: イベントの時間的定位と説明文を付けるタスク．

- 論文

[#11]

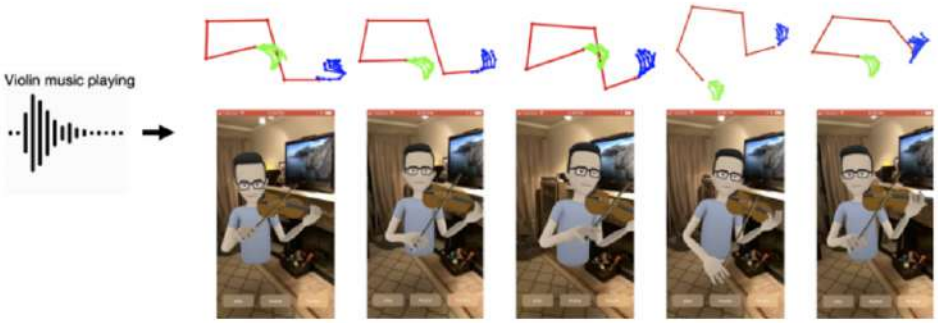
Ryota Suzuki

Audio to Body Dynamics

E. Shlizerman, L. Dery, H. Schoen and I. Kemelmacher-Shlizerman CVPR2018

「音から手の動きは生成可能か？」バイオリンやピアノ演奏の音声を入力すると，アバターが演奏しているかのようにアニメーションするようなスケルトンの推定を行う手法を提案． 結論：できる．

実際ちゃんとやるにはいくつかアドホックな工夫が必要なようで，詳細はおのおの論文を確認してもらいたい．学習時に使うスケルトンデータはYouTubeのリサイタル動画からOpenPoseやMaskRCNNを駆使して生成する． 入力音声からこの手法で 13次元ベクトルに変換し，さらにその時間差分や音量エネルギーを足した28次元ベクトルにする． これから上半身のスケルトンの時系列を生成するLSTMを作り， スケルトンにアバターを着せてアニメーションを作成する．



新規性・結果・なぜ通ったか？

アプリケーション枠らしく，見た目の良さがあり，また実装上の困難と解決についてちゃんと書いているのが評価されたものと思われる． アプリケーションとして利用するに当たって，どれだけうまくいけるのかが窺い知れる資料として 貴重に思われる．

コメント・リンク集

1ページ目が既に他の論文と一線を画そうとしている．Fun to read という点で参考になるので，一度読んでみることを勧める．

- 論文
- 動画など

Ryota Suzuki

Separating Self-Expression and Visual Content in Hashtag Supervision

A. Veit, M. Nickel, S. Belongie, L. Maaten CVPR2018

Facebookでの研究．ユーザのこれまでのハッシュタグから，一意に同定できない意味の単語のハッシュタグでもユーザが意図した画像検索ができるようにした． 画像のDeCAFを取り，ユーザの履歴特徴，ハッシュタグ特徴を埋め込んだ３次テンソルを構成，多クラスロジスティック関数などで評価する．

img(src=`${figpath}Separating_Self-Expression_and_Visual_Content_in_Hashtag_Supervision.png`,alt="

新規性・結果・なぜ通ったか？

MLPによる手法よりこちらの方が良い性能を示した．Top1で43.7%，Top10で72.12%のAccuracy．

コメント・リンク集

- 論文

Ryota Suzuki

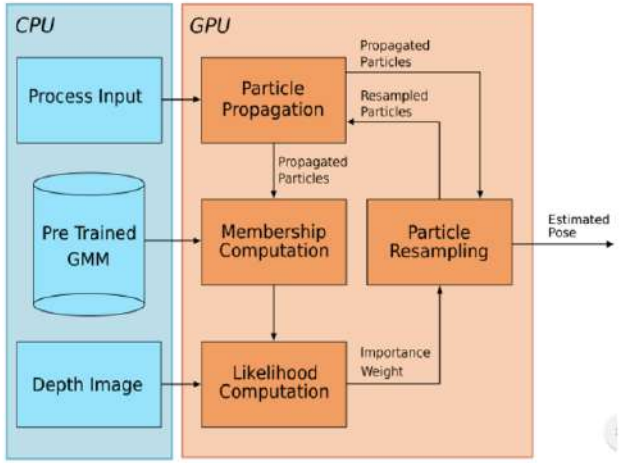
Human-centric Indoor Scene Synthesis Using Stochastic Grammar

S. Qi, Y. Zhu, S. Huang, C. Jiang, S. Zhu CVPR2018

3D部屋レイアウトとその2D画像との合成の話題．

Spatial And-Or Graph (S-AOG) ※ で屋内シーンを表現する．終端ノードは物体エンティティ（部屋とか家具とかその他）．

終端ノードに対し，マルコフランダム場（MRF）を用い，人間の文脈で関係性をエンコードする．屋内シーンデータセットから分布を学習し，モンテカルロマルコフ連鎖（MCMC）を使って新しいレイアウトをサンプルする．



新規性・結果・なぜ通ったか？

3つの視点で有効性を確認．

- SOTAな部屋アレンジ手法と比較しての，視覚的リアルさ
- GTに対する，アフォーダンスマップの精度
- 合成部屋の機能性，自然っぽさを人間の被験者で評価

コメント・リンク集

※S-AOGは確率的文法モデルの一つ．

- 論文

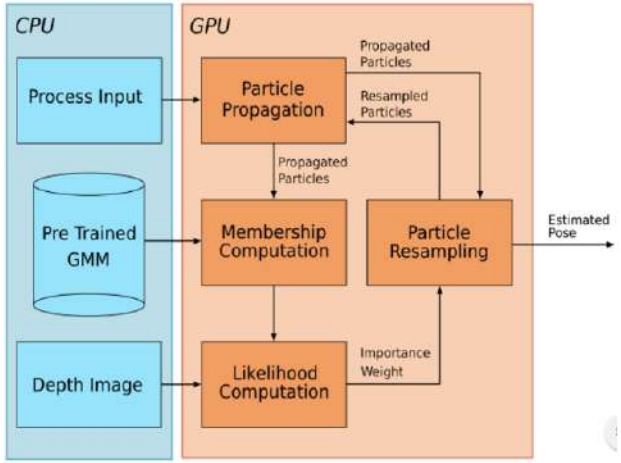
Ryota Suzuki

Fast Monte-Carlo Localization on Aerial Vehicles using Approximate Continuous Belief Representations

A. Dhawale, K.S, Shankar, N. Michael CVPR2018

ドローンのようなサイズ，重さ，力が制約されたプラットフォームでも，3D自己位置同定を高速に行えるフレームワークを提案．点群データの混合ガウス分布（GMM）表現による圧縮をキーアイデアとしている．

デプスセンサのデータと，オンボード姿勢参照システムからピッチとロールを得る．データをGMMで表現した尤度を使って，複数仮説パーティクルフィルタにより定位．



新規性・結果・なぜ通ったか？

CVPRでは，高速性・省メモリに関するトピックに興味があるかもしれない．SLAM系はICRAでは大変多く議論されている話題だが，逆にCVPRだとアプリケーション枠で 通る可能性があるかもしれない．

コメント・リンク集

- 論文

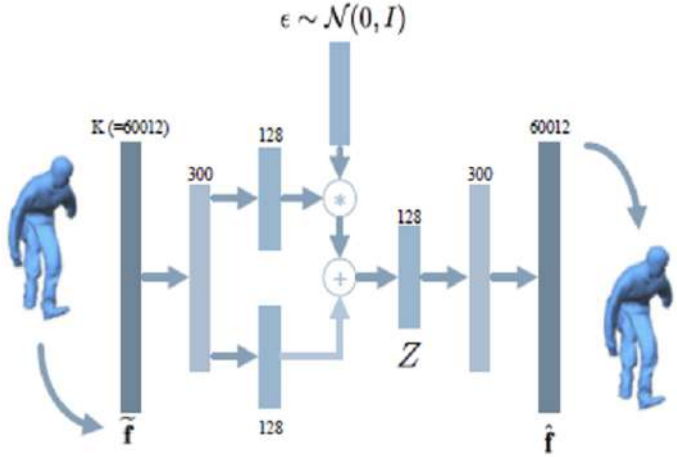
[#15]

Ryota Suzuki

Variational Autoencoders for Deforming 3D Mesh Models

Q. Tan, L. Gao, Y. Lai and S. Xia CVPR2018

3Dメッシュの変形に関して，Variational AutoeEcoder(VAE)を使ってみたという研究．可能な変形の確率的潜在空間の探索を行う． 学習は簡単で，学習データも少なくて済む（どれくらい？） 事前分布を代替することで，異なる潜在変数の顕著性（Significance）を柔軟に調節可能な拡張モデルも提案．



新規性・結果・なぜ通ったか？

形状生成，形状補完，形状空間埋め込み，形状探索においてSoTA越え．

コメント・リンク集

- 論文

Kodai Nakashima

Density-aware Single Image De-raining using a Multi-stream Dense Network

He Zhang and Vishal M. Patel CVPR 2018

概要

DID-MDN (density-aware multi-stream densely connected convolutional neural network-based algorithm) と呼ばれる、画像内の雨量密度推定と雨除去を行うアルゴリズムを提案。雨のストロークをより良く特徴づけるため、multi-stream densely connected de-raining networkでは異なるスケールの特徴量を効率的に活用する。また、雨密度ラベル付き画像を含むデータセットを新たに作成した。このデータセットを学習に使うことにより、state-of-the-artな手法を超えることができた。

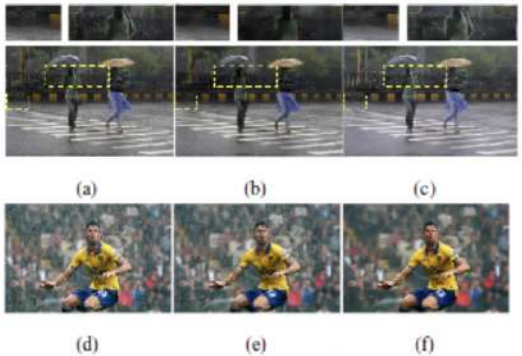


Figure 1: Image de-raining results. (a) Input rainy image. (b) Result from Fu *et al.* [6]. (c) DID-MDN. (d) Input rainy image. (e) Result from Li *et al.* [33]. (f) DID-MDN. Note that [6] tends to over de-rain the image while [33] tends to under de-rain the image.

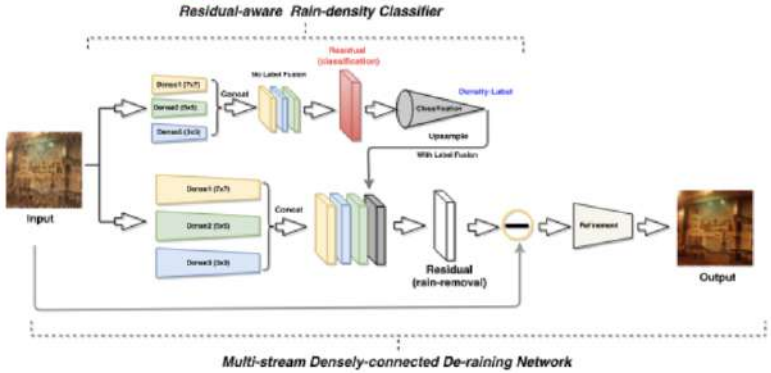


Figure 2: An overview of the proposed DID-MDN method. The proposed network contains two modules: (a) residual-aware rain-density classifier, and (b) multi-stream densely-connected de-raining network. The goal of the residual-aware rain-density classifier is to determine the rain-density level given a rainy image. On the other hand, the multi-stream densely-connected de-raining network is designed to efficiently remove the rain streaks from the rainy images guided by the estimated rain-density information.

Table 1: Quantitative results evaluated in terms of average SSIM and PSNR (dB) (SSIM/PSNR).

	Input	DSC [29] (ICCV'13)	GMM [17] (CVPR'16)	CNN [5] (TIP'17)	JORDER [33] (CVPR'17)	DDN [6] (CVPR'17)	JBO [41] (ICCV'17)	DID-MDN
Test1	0.7781/21.15	0.7896/21.44	0.8352/22.75	0.8422/22.97	0.8622/24.32	0.8978/ 27.35	0.8522/23.05	**0.9087/ 27.95**
Test2	0.7695/19.31	0.7825/20.08	0.8305/20.66	0.8289/19.73	0.8405/22.26	0.8851/25.60	0.8356/22.45	**0.9092/ 26.0745**

新規性・結果・なぜ通ったか？

コメント・リンク集

- 論文URL

[#17]

Kodai Nakashima

SeGAN: Segmenting and Generating the Invisible

Kiana Ehsani, Roozbeh Mottaghi and Ali Farhadi CVPR 2018

概要

オクルードされている物体の全体像を推定するため、SeGANを提案。SeGANは物体の見えていない領域のセグメントを生成することができる。また、occluderとoccludeeの関係も推定することができる。さらにSeNetはcategory-agnosticでありカテゴリー情報を必要としない。データセットにはDYCEを使用。

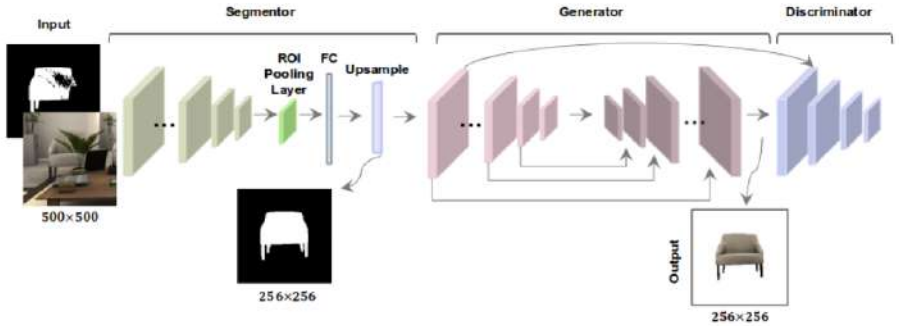


Figure 2. **Model architecture.** Our network has three parts: *segmentor*, *generator*, and *discriminator*. The input to our model is an RGB image and a mask for the visible region of an object which is obtained automatically by [51]. The output is an RGB image that shows the appearance and segmentation for the full object (visible and reconstructed invisible regions). The segmentor part outputs an intermediate mask (the mask shown in the middle) that represents the full object, which is passed to the generator part of the network.

Model	Training Data	Input Mask	Loss Mask	Visible ↑	Invisible ↑	Visible ↑	Invisible ↑
[BBE] [7]	natural and synthetic	SV	SF	31.0	25.2	5.1	
Multipath [51]	natural and synthetic	SV	SF	36.0	34.8	12.5	
Multipath [51]	natural	SV	SU	47.51	48.58	9.01	
Pix2Pix [17]	natural and synthetic	SV	SF	52.3	49.6	11.9	
SeGAN (ours) w/ predicted SV	natural and synthetic	SV	SF	**68.78**	**64.76**	**15.39**	
SeGAN (ours) w/ GT SV	synthetic	SV	SF	75.71	68.05	23.26	

Table 1. **Segmentation evaluation.** We compare our method with [7], [17], and [51] on the synthetic test data. SV and SF refer to the mask for the regions of objects and the full object, respectively. We evaluate how well we predict 'visible' regions, 'Invisible' regions and their combination. The bottom row is not comparable with other rows since it uses groundtruth information.

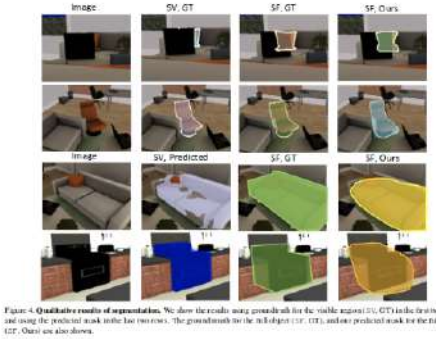


Figure 4. **Qualitative results of segmentation.** We show the results using groundtruth for the visible regions (SV, GT) in the first two rows, and using the predicted mask in the last two rows. The groundtruth for the full object (SF, GT), and our predicted mask for the full object (SF, Ours) are also shown.

新規性・結果・なぜ通ったか？

右図に示すように、他のセグメントベースラインと比べ、SeGANが見える領域、見えない領域、それらの組み合わせの全てにおいて最も良い結果を出した。ここで、SUは見える領域のセグメント、SIは見えない領域のセグメント、SFは全体像のセグメントを表してい

コメント・リンク集

- 論文URL
- github

[#18]

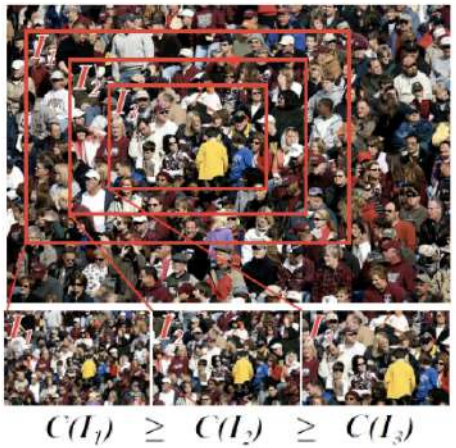
Shusuke Shigenaka

Leveraging Unlabeled Data for Crowd Counting by Learning to Rank

Xialei Liu, Joost van de Weijer, Andrew D. Bagdanov CVPR 2018

概要

群衆の画像データにおいて、ネットワークの訓練を改善するためのself-supervisedタスクを提案。タスクは集計情報とランキング情報の両方を組み合わせたマルチタスクフレームワークであり、群衆カウントのためにend-to-endで訓練できる。 群衆画像をだんだん小さくするように切り取って人数をランク付けおり、提案されたself-supervisedタスクはラベル付けのされていない群衆画像のCNNに大きく貢献した。 提案手法は群衆計測の困難なデータセットShanghaiTechとUCF CC 50においてstate-of-the-artを得ている。



新規性・結果・なぜ通ったか？

- 困難とされている2つのデータセットでstate-of-the-artを得たこと
- 大人数のデータはその人数のデータより少ない数で観察というルールに基づいて計測を行っているため、大規模なトレーニングデータセットの欠如に対処することができている

リンク集

- 論文
- github

[#19]

Kodai Nakashima

Conditional Image-to-Image Translation

Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen and Tie-Yan Liu CVPR 2018

概要

image-to-image translationタスクで用いられるモデルは、ターゲットドメインの翻訳結果をコントールする機構がなく、出力結果が多様性に乏しい。この研究では、1. conditional image-to-image translationをいう問題を新たに設定し、2. この問題を解くためにconditional dual-GAN (cd-GAN) を提案する。 1では、複数の画像を組み合わせたtarget domainが入力されたsorce domainを変換する問題を扱う。複数の画像をどのようにして組み合わせるかで多様性に富んだ変換結果が得られる。

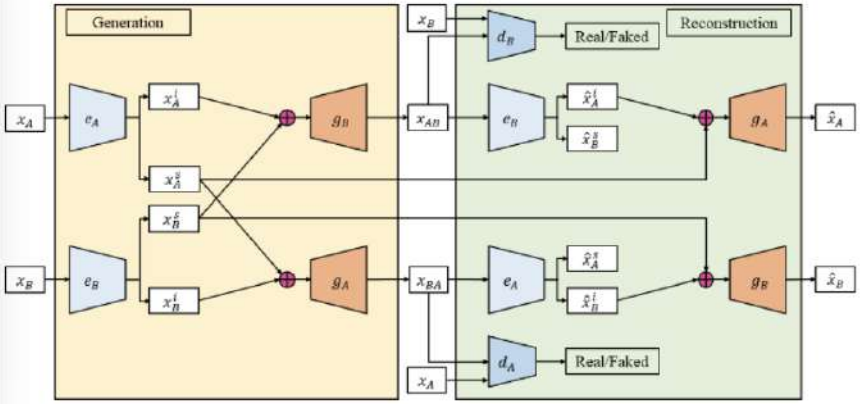


Figure 2. Architecture of the proposed conditional dual GAN (cd-GAN).

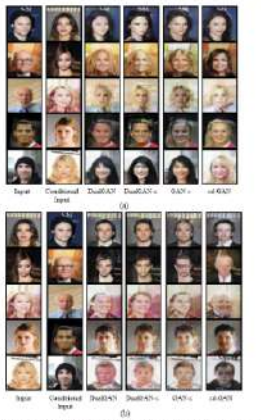


Figure 3. Conditional face-to-face translation. (a) Results of conditional men→women translation. (b) Results of conditional women→men translation.

新規性・結果・なぜ通ったか？

入力は64x64とする。eA, eBは3つの畳み込み層で構成されており、各畳込み層の活性化関数にLReLUを用いる。デコレーターネットワークであるgAとgBは4つのデコンボリューション層から構成されて

コメント・リンク集

- link11. link3
- link22. link3
- link33. link3

[#20]

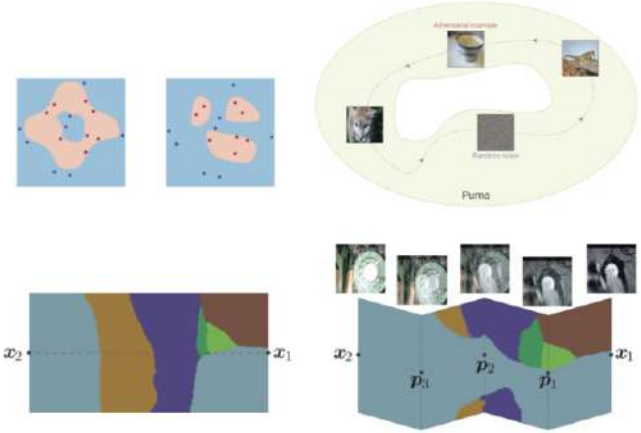
Yoshihiro Fukuhara

Empirical study of the topology and geometry of deep networks

Alhussein Fawzi et al. CVPR 2018

概要

DNN 画像クラス分類器の入力空間における位相的・幾何学的性質を実験的に分析した研究. DNN が学習している各クラスの領域は接続されたものであり, その境界は少数の大きな曲率をもつ方向と, 平坦な大多数の方向があることが確認された. また, 大きな曲率をもつ方向はデータ間で共有されており, これらの方向とネットワークの摂動に対する感度に関係性があることを確認した.



新規性・結果・なぜ通ったか？

- 理論のみを用いた解析は困難なため, 実験を行って性質の分析を行った
- DNN が学習している同じクラスの領域は接続されたものであり , その領域はほぼ凸集合になっている（凸集合に近いが実際には違う）
- クラスの境界の主曲率は多数の方向で 0 であったが, 大きな値をもつ方向が少数存在
- 主曲率の値は非対称で大きな負の値を持つ方向が多い (この結果はネットワークの構造やデータセットなどを変えても共通して確認された)
- 主曲率の大きな値をもつ方向はデータ間で共有されていることを確認

コメント・リンク集

- [論文] Empirical study of the topology and geometry of deep networks
- 本研究で確認された入力空間における位相的性質と同様の性質が, weight の空間でも報告 [Freeman+ 16] されており, 2 つの空間の関連性を調べることは今後の課題とされている.

[#21]

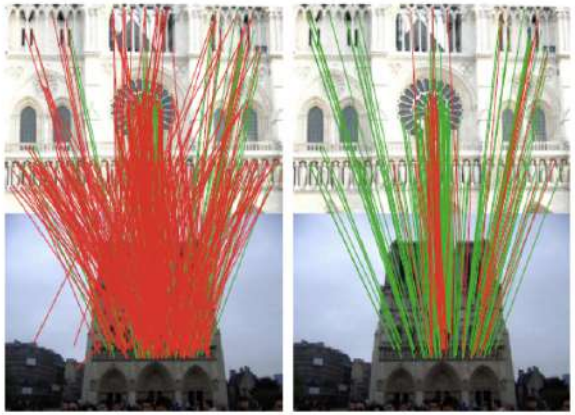
Shintaro Yamamoto

Learning to Find Good Correspondences

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann and Pascal Fua CVPR 2018

概要

2枚の画像間の対応点探索を学習ベースで行う方法を提案。従来のhandcrafted特徴(SIFTなど)による手法は、特徴量により候補を決めた上でRANSACなどのアルゴリズムで対応点かそうでないかを決定する。 本研究では同様に、候補となる対応点の中から実際に対応しているペアをMulti Layer Perceptrons(MLPs)により決定する。 対応点の数は画像によって異なるので、ネットワークには対応点のペア(4変数)毎に実際に対応しているかの判定を行う。 一方で、中間層出力を全ペアの平均と分散により正規化することでglobal contextを考慮する。(Context Normalization) 学習は、ペアの判定が正しいか、判定結果を用いてessential matrixが正しく求められるかによって行う。 その際、学習データに対して対応点のアノテーションを手動で与えるのは非常に時間がかかってしまう。 そこでepipolar distanceを用いた閾値処理により対応点を取得する。



(a) RANSAC (b) Our approach

新規性・結果・なぜ通ったか？

ベースラインと比較して、学習したシーン、学習していないシーンどちらにおいても高い精度ないし同等の精度を出すことに成功。59枚の学習データのみで学習した場合であっても、ベースラインと比べ高い精度を出すことに成功。 RANSACのみで対応点を決定する場合より、提案手法により候補を絞った上でRANSACにより更に候補を削るほうが17倍計算時間が早い。

コメント・リンク集

- 論文

[#22]

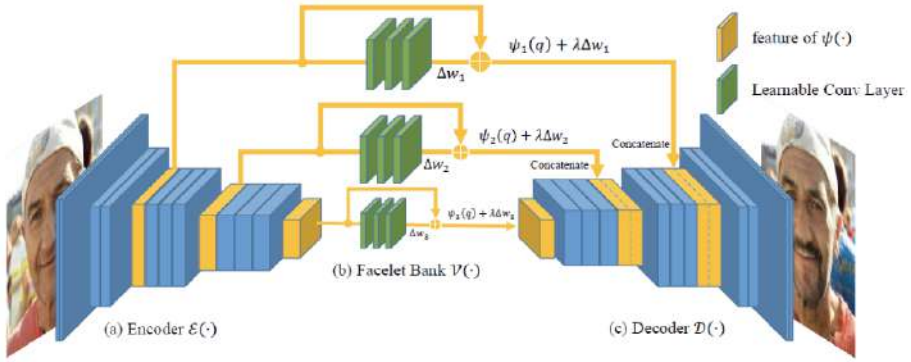
Shintaro Yamamoto

Facelet-Bank for Fast Portrait Manipulation

Ying-Cong Chen, Huaijia Lin, Michelle Shu, Ruiyu Li, Xin Tao, Xiaoyong Shen, Yangang Ye and Jiaya Jia CVPR 2018

概要

顔のattributeを編集するEnd-to-Endのネットワークを提案した。ドメイン間の変換を考えるのではなく、Encoderにより得られた特徴のドメイン間の差分を考えることにより特徴の付与を実現する。 ドメイン毎の特徴は、全ての学習データの平均ではなく入力画像の最近傍K枚の平均を考える。 Encoderにより入力画像から得られた特徴から、Facelet Bankというネットワークによりドメイン間の差分を求める。



新規性・結果・なぜ通ったか？

従来手法と比較して、artifactが少なく高解像度の画像を出力することが可能になった。女性に髭を付与するなど学習データには存在しないようなものの場合、従来法では男女の違いが付与されて髭以外の変化が加わってしまう。 しかし、編集に重要な領域(髭→口周り)のみに変化を施すため従来手法よりも自然な変化が実現可能である。

コメント・リンク集

比較的関連研究が多そうな研究だったが比較対象が2つと少なめ

- コード
- 論文

[#23]

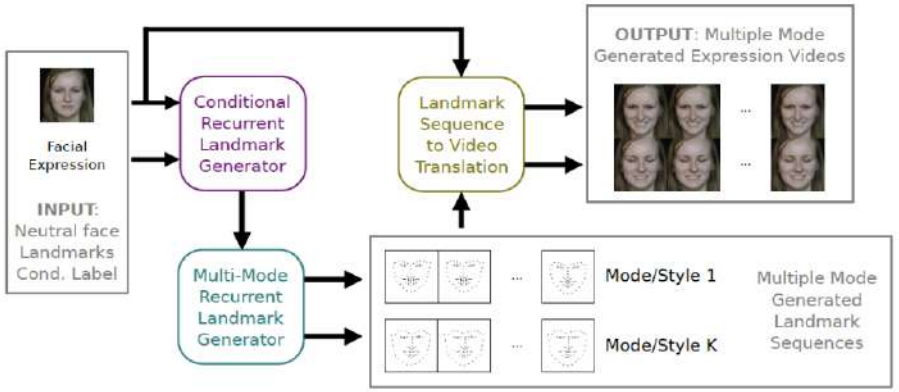
Shintaro Yamamoto

Every Smile is Unique: Landmark-Guided Diverse Smile Generation

Wei Wang,Xavier Alameda-Pineda, Dan Xu, Pasal Fua, Elisa Riccia and Nicu Sebe CVPR2018

概要

1枚の顔画像から、指定した表情に変化する動画を生成する手法を提案。たとえ同じ笑顔であっても、作り笑いとそうでない場合など目の動きなど顔の変化は異なる。 そこで、指定された表情に対して複数の動画を生成する手法を提案した。 入力画像とラベルから、指定されたラベルに対して適した顔特徴点の変化を複数のネットワークによって予測する。 その際、各ネットワークの予測がお互いに類似しないように最適化することで動画を複数用意することなく予測することを可能とする。 予測した顔特徴点から各フレームの顔画像を復元することにより、動画の生成を実現する。



新規性・結果・なぜ通ったか？

従来の動画生成に関する研究と比べ、artifactが少なく与えられた画像の人物の個人性を保った合成を実現した。ユーザースタディの結果、比較対象とした研究よりも提案手法により生成された動画のほうが圧倒的に好まれるということが分かった。 Action Unit(AU)の変化を調べたところ、提案手法により生成された動画は実際の動画に近い変化をすることが分かった。

コメント・リンク集

- 論文

[#24]

Shintaro Yamamoto

Creating Capsule Wardrobes from Fashion Images

Wei-Lin Hsiao and Kristen Grauman CVPR2018

概要

Capsule Wardrobesという、良い組み合わせが多数存在するファッションアイテムのセットを自動で作る手法を提案。ファッションアイテムのセットに対して、それで実現可能なファッションの親和性と多様性を最大化することによりセットを決定する。 注目レイヤー以外を固定して最適化することを繰り返すことでファッションアイテムの選択を行う。 ファッションの親和性を決定するために、トピックモデルをベースとした教師なし学習による全身画像からのファッションの評価方法を構築した。

新規性・結果・なぜ通ったか？

ファッションサイトに掲載されているCapsule Wardobesと作成したものに含まれるファッションアイテムの類似度を測った結果、ベースラインと比べ提案手法により選ばれたものの方が類似度が高いという結果が得られた。提案手法である繰り返しの最適化と貪欲法による最適化結果をユーザースタディで比べたところ、提案手法のほうが好ましいと答えた人が59%いた。 また、個人の好みに応じたCapsule Wardrobesの作成が可能である。

コメント・リンク集

- 論文

Kodai Nakashima

Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident DB

Hirokatsu Kataoka, Tomoyuki Suzuki, Yoshimitsu Aoki and Yutaka Satoh CVPR 2018

概要

交通事故予測のため, 1. loss関数としてAdaptive Loss for Earlay Anticipation (AdaLEA)と2. 予測のためのNear-miss Incident DataBase (NIDB) の提案を行った. AdaLEAにより, モデルが学習過程において, 徐々に早く危険を予測できるように学習される. モデルが交通事故を予測する速さでペナルティを与えることにより, これを実現する. NIDBは, 多くの交通ニアミス動画を含んでおり, 危険と危険要素予測の評価用アノテーションが付けられている.

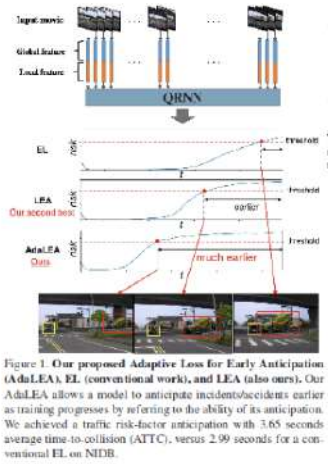


Figure 1. **Our proposed Adaptive Loss for Early Anticipation (AdaLEA), EL (conventional work), and LEA (also ours).** Our AdaLEA allows a model to anticipate incidents/accidents earlier as training progresses by referring to the ability of its anticipation. We achieved a traffic risk-factor anticipation with 3.65 seconds average time-to-collision (ATTC), versus 2.99 seconds for a conventional EL on NIDB.

	Chan16 [4]	Zeng17 [24]	Chan16 +AdaLEA	QRNN +EL	Ours1 LEA	Ours2 AdaLEA	Ours3 +NIDB
DSA	✓	–	✓	✓	✓	✓	✓
LSTM	✓	✓	✓				
QRNN				✓	✓	✓	✓
EL	✓	✓		✓			
LEA			✓		✓		
AdaLEA						✓	✓
NIDB							✓
mAP[%]	48.1	51.4	49.2	51.7	52.1	52.3	**53.1**
ATTC[s]	1.34	3.01	2.80	3.02	3.22	3.43	**3.44**

Table 2. Results of risk anticipation on DAD: NIDB in the table means NIDB-pretrain for global feature. The result of conventional methods are cited from [24].

	Chan16 [4]	Chan16 +AdaLEA	QRNN +EL	Ours1 LEA	Ours2 AdaLEA	Ours3 +NIDB
DSA	✓	✓	✓	✓	✓	✓
LSTM	✓	✓				
QRNN			✓	✓	✓	✓
EL	✓		✓			
LEA		✓		✓		
AdaLEA					✓	✓
NIDB						✓
mAP[%]	85.5	84.4	84.2	86.2	86.3	**90.1**
ATTC[s]	2.45	4.62	2.85	4.67	4.72	**4.81**

Table 3. Results of risk anticipation on NIDB: NIDB in the table means NIDB-pretrain for global feature.

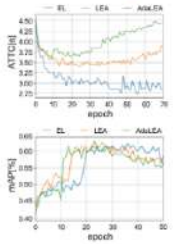


Figure 5. ATTC v.s. epoch (upper) and mAP v.s. epoch (lower) curves.

		Chan16 [4]	Chan16 +AdaLEA	QRNN +EL	Ours1 LEA	Ours2 AdaLEA
DSA		✓	✓	✓	✓	✓
LSTM		✓	✓			
QRNN				✓	✓	✓
EL		✓		✓		
LEA					✓	
Ada-LEA			✓			✓
mAP[%]	bicycle	57.3	57.7	60.0	**62.8**	56.8
	pedestrian	43.1	43.9	43.6	44.7	**47.9**
	vehicle	73.2	75.9	79.9	78.5	**81.4**
	Average	57.8	59.2	61.2	62.0	**62.1**
ATTC[s]	bicycle	2.94	3.38	3.51	3.22	**3.65**
	pedestrian	3.15	3.36	3.34	3.23	**3.56**
	vehicle	2.71	2.96	2.12	**3.99**	3.75
	Average	2.95	3.23	2.99	3.48	**3.65**

Table 4. Results of risk-factor anticipation on NIDB: we used NIDB-pretrain model to extract global feature in all methods.

新規性・結果・なぜ通ったか？

ベールモデルとしてDSA, LSTM, QRNN, loss関数としてEL, LEA, AdaLEAを用いて実験した.その結果, 危険予測では, mAPが6.6%上昇, ATTCが2.36sec速くなった. また, 危険要素予測では, mAPが4.3%上昇, ATTCが0.70sec速くなった.

コメント・リンク集

- 論文URL

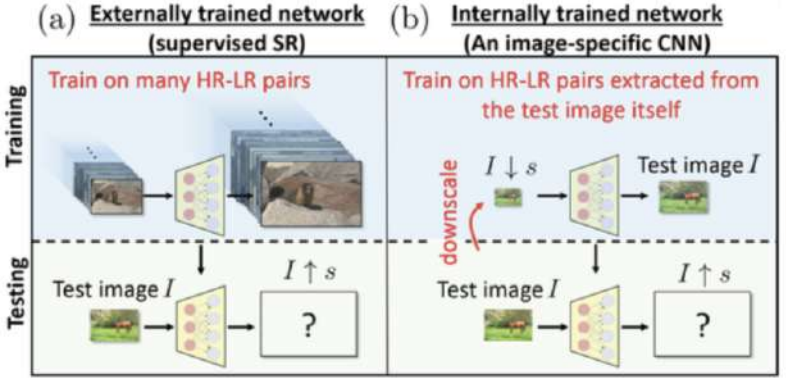
Shusuke Shigenaka

“Zero-Shot” Super-Resolution using Deep Internal Learning

Assaf Shocher, Nadav Cohen, Michal Irani CVPR 2018

概要

実際の古い写真,ノイズの多い画像,生物学的データ,取得プロセスが不明または非理想的な画像のSuper-Resolution(SR)を実行を行うことができるZero-Shot SR(ZSSR)を提案．過去の画像例や事前訓練に依存することなく,Low-Resolution(LR)とその縮小版から複雑な画像特有のHR-LR関係を推論するCNNを訓練を行うことにより, 実際のLRの画像において,State-of-the-artなCNNベースのSRおよび教師なしSRよりも優れている．



新規性・結果・なぜ通ったか？

SR-CNNは大規模な外部データベースの画像を事前に訓練しているのに対し,ZSSRは小さな画像から粗い解像度のテストデータを訓練．

ZSSRは同じ教師なしのSelfExSRにと比べ全てのDataSetにおいて優れている．教師あり学習でも通常のLRはあまり変わらない精度を出しており,未知LR画像で確認をするとかなり優れた精度を出している．

リンク集

- 論文
- Zero-Shot Super-Resolution

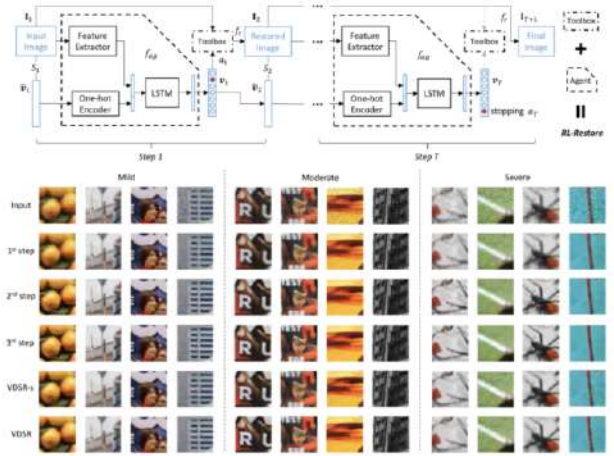
Yoshihiro Fukuhara

Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning

Ke Yu et al. CVPR 2018

概要

強化学習（Deep Q-learning）を用いた画像復元の研究. 単一の大きなネットワークを用いる手法とは対照的に, 特定の distortion に対する復元に特化した小さなネットワークを複数集めて toolbox とし, agent が各ステップにおいて最適な tool を選択することで段階的な復元を行う. 評価実験では従来の大きな単一のCNNを用いた手法と同程度の精度を20%程度の計算量で実現した.



新規性・結果・なぜ通ったか？

- 強化学習を用いて段階的に画像復元を行うフレームワークを提案
- agent は action として, 各ステップにおいて特定の distortion に対する復元に特化した小さなネットワークを複数集めた toolbox の中から最適なものを選択
- 段階的な復元を行うと中間のステップにおいて生じる複雑な atifact を扱うため agent と tool の joint training アルゴリズムを提案
- DIV2K dataset を用いて行った評価実験では, PSNR 尺度において単一の大きなCNNを用いた場合と同程度の精度を約20%計算量で実現

コメント・リンク集

- [論文] Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning
- [Code] GitHub
- どのネットワークを使うべきかという高次の意思決定を強化学習で学習するという方針が面白い. (Hierarchical Reinforcement Learning と類似の考え方)

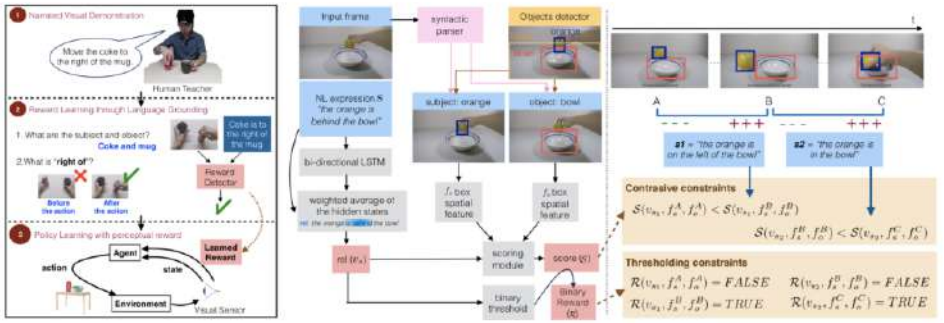
Yoshihiro Fukuhara

Reward Learning from Narrated Demonstrations

Hsiao-Yu Tung et al. CVPR 2018

概要

動画による教示と言語による説明を組み合わせて Reward の学習を行う研究. 言語情報によって与えられた目標の達成の可否を, 画像情報から判断する Instractable Perceptual Rewards を提案し, 学習用のデータセットを作成した. また, 評価実験では教師ありで静止画像のみから学習した場合と比較して, 優位な結果を達成した.



新規性・結果・なぜ通ったか？

- 言語情報によって与えられた目標の達成の可否を, 画像情報から判断する Instractable Perceptual Rewards を提案
- 上記の教師データとして, 動画による教示に言語による説明を付随した, Narrated Visual Demonstration (NVD) のデータセットを作成した
- 提案手法は hard negative mining によって少ない教師データからの効率的な学習が可能
- 評価実験では Visual Genome のみを用いて学習した手法 [Hu+16] と比較して優位な結果を達成

コメント・リンク集

- [論文] Reward Learning from Narrated Demonstrations

[#29]

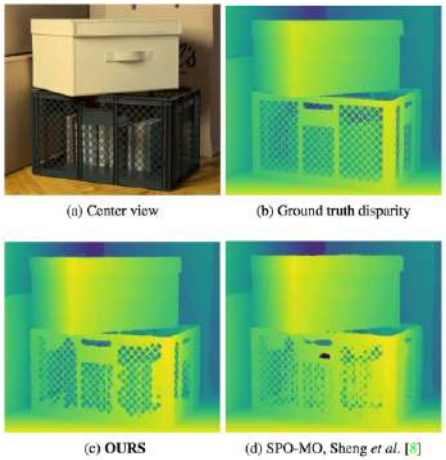
Hirokatsu Kataoka

Trust Your Model: Light Field Depth Estimation With Inline Occlusion Handling

Hendrik Schilling, Maximilian Diebold, Carsten Rother, Bernd Jähne CVPR 2018

概要

LightFieldカメラからの距離画像推定の問題を提案。オクルージョンに伴う物体境界の精度や質向上に対して操作を行なったことが貢献である。従来法とは異なり、PatchMatchをベースラインとして距離画像とオクルージョン領域を同時推定を直接的に行う。同時推定を行うことで、データを全て同時に学習に用いることができ、さらに前処理のステップが不要になる。結果的には、オクルージョン領域の推定を行い物体境界をケアしただけでなく滑らかな表面再構成に成功した。公開されているLightFieldデータセットにて評価した結果、12のうち9の指標においてState-of-the-artな数値を出した。



新規性・結果・なぜ通ったか？

ライトフィールドカメラを用いた距離画像推定においてオクルージョン対策を講じた。距離画像とオクルージョン領域を同時推定する手法では既存のライトフィールドカメラにおける評価指標においてState-of-the-art。さらに、平面推定においても高度な推定を実現した。

コメント・リンク集

同時推定のうまい手法を考案、副次的に平面が滑らかになるというのも面白い！

- 論文

[#30]

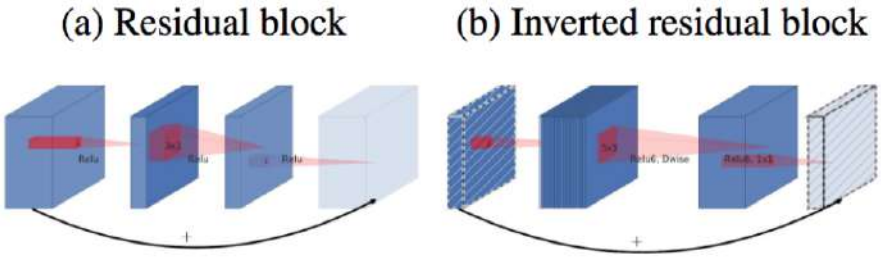
Hirokatsu Kataoka

MobileNetV2: Inverted Residuals and Linear Bottlenecks

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen CVPR 2018

概要

モバイルで動作する新規アーキテクチャMobileNetV2の提案論文、データセットを用いた複数タスクにてState-of-the-artな精度を達成した。物体検出のモデルであるSSDLiteやセマンティックセグメンテーションのモデルであるMobile DeepLabv3を考案した。これらはInverted Residual Structureと呼ばれる、ショートカットコネクションが小さなボトルネックレイヤに挟まれた構造を最小ユニットとして構成される。中間の拡張レイヤは非線形関数として軽量化されたdepthwiseの畳み込みとして実装される。右図に本論文の重要技術であるInverted Residual Blockについて示す。従来のResidual Block（左）は前後のdepthが広いが、提案のInverted Residual Blockは中ふたつがdepthが広く、前後は狭い。



新規性・結果・なぜ通ったか？

Inverted Residual Blockの提案等によりモバイルサイズのモデルにおいても良好な認識精度のモデルを提案することに成功。認識精度とパラメータ数のトレードオフについても良好で、さらにはCPUにおいても高速に動作することを示しCVPRに採択された。

コメント・リンク集

モバイルネットv2、応用範囲が広そう。

- 論文
- Google AI Blog
- GitHub
- GitHub2

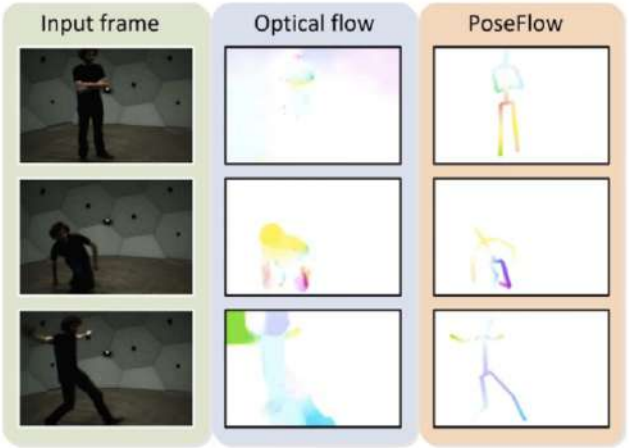
Munetaka Minoguchi

PoseFlow: A Deep Motion Representation for Understanding Human Behaviors in Videos

Dingwen Zhang, et al. CVPR 2018

概要

動画から人間の行動を理解するためのPoseFlowの提案。PoseFlowはオプティカルフローに代わる新しい動き表現であり、背景の動きによるノイズやオクルージョンに頑健。人間の骨格位置とマッチングの2つの問題を同時に解決するようなネットワークであるPoseFlow Net(PFN)を提案し、学習する。これにより、人体の部分のみに動きベクトルが付与された出力を得ることができる。



新規性

従来手法では、オプティカルフローを使ってモーションキューを探索している場合が多いが、背景の動きなども取ってしまうので“ノイズが多い動きの表現”であり、姿勢推定や行動認識のタスクにおいて支障をきたす。実験では、従来手法と比較して、姿勢推定や行動認識タスクにおいて高精度となっている。

結果・リンク集

図のように、オプティカルフローでは背景の動きも取ってしまい、ぼんやりとした出力になっているが、PoseFlowでは人間の骨格の動きのような情報を取得することができる。

- 論文

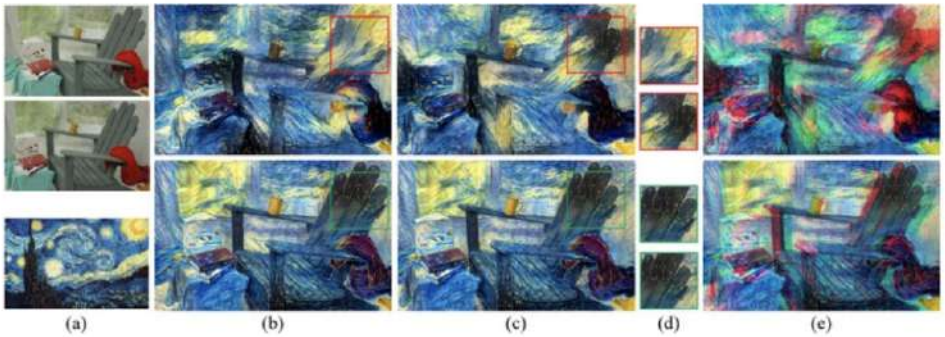
Munetaka Minoguchi

Stereoscopic Neural Style Transfer

Dongdong Chen, et al. CVPR 2018 1802.10591

概要

3D映画やAR / VRの需要に先駆けた、Stereoscopic Neural Style Transferの提案。スタイルトランスファーによって、左右視点での整合性を保持するために、style loss functionにdisparity lossを追加し、左右視点での視差制約を設けている。また、リアルタイム性を考慮したソリューションの開発に取り組み、stylization sub-networkとdisparity sub-networkの2つを共同してトレーニングできるモデルを提案。



新規性

ステレオカメラを使ったスタイルトランスファー手法。通常、図(a)のような左右視点の画像とスタイル画像を入力すると1行目のように，左視点(b)と右視点(c)のように左右の視点で差が生じる(d)。このような不一致性は、(e)のアナグリフ画像のようになり、視聴者へ左右視点での三次元的視覚疲労が生じさせる。提案手法ではこのような不一致性を抑制し、2行目のように整合性のとれたスタイルトランスファーを可能にする。

結果・リンク集

提案手法によって、時間的および視差の整合性を考慮しており、3D映像を拡張できる。定量的および定性的評価によって、従来手法よりも高精度であることを示唆。

- 論文

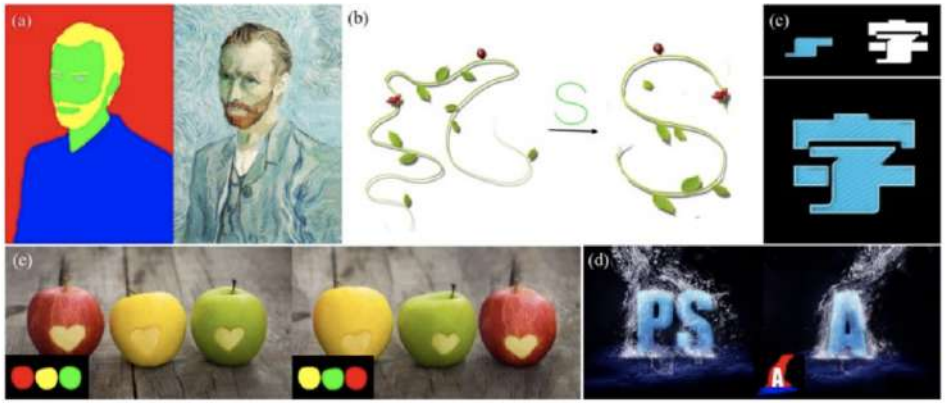
Munetaka Minoguchi

A Common Framework for Interactive Texture Transfer

Yifang Men, et al. CVPR 2018

概要

局所構造と視覚的豊かさの両方を保持できる、より汎用的なtexture transfer問題を解決するための提案。元画像と元画像のセマンティックマップ(aのようなセグメンテーション画像)と、変換後となるセマンティックマップの3つを入力とする。変換顔のセマンティックマップを元にスタイルトランスファーを実行する(ゴッホを痩せさせるなど)。contour key points match(CPD)やTPSアルゴリズムをベースとしたstructure propogation手法を提案している。



新規性

タスクの多様性と、ユーザガイダンスの簡潔さをテーマに取り組んでいる。図のように、(a)簡単な絵をアートワークに変更、(b)装飾パターンの編集、(c)テキストに特殊効果を付与、(d)テキスト画像における効果を制御、(e)テクスチャの交換、などユーザのガイダンスによってさまざまなテクスチャの変換を実現できる。

結果・リンク集

他の手法と比較して、人間の視覚的にもより自然な変換ができている。

- 論文

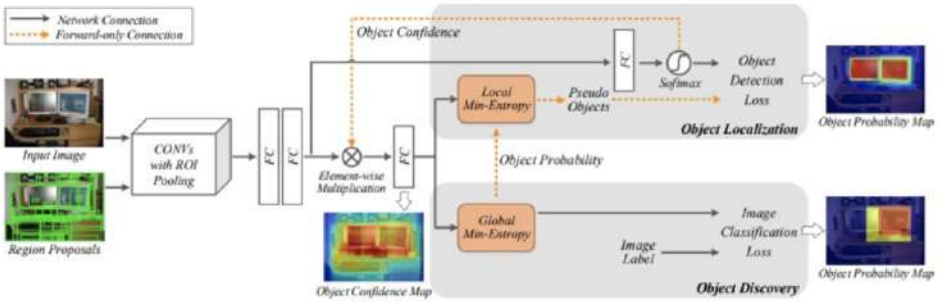
Munetaka Minoguchi

Min-Entropy Latent Model for Weakly Supervised Object Detection

Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han and Qixiang Ye

概要

弱教師付き学習で物体検出を行うmin-entropy latent model (MELM)の提案。MELMは、object discoveryとobject localizationの2つのサブモデルで構成され、end-to-endで学習可能。 object discoveryで、 global min-entropyと画像分類lossを最適化。local min-entropyとソフトマックスを最適化。グローバルとローカルそれぞれで物体を検出し、エントロピーを最小化し、グローバルからローカルへ物体確率を伝播。



新規性

弱教師付き学習による物体検出は、物体位置と検出を同時に学習するのが困難。弱教師と学習目標間に不一致が生じると物体位置にランダム性が生じ、検出器をうまく学習できない。min-entropyによって、学習中の物体位置のランダム性を計測し、物体位置を学習することができ、検出器のあいまいさを回避できる。

結果・リンク集

回帰的に学習することによって、弱教師であっても精度向上。

- 論文

[#35]

Yuta Matsuzaki

Avatar-Net: Multi-scale Zero-shot Style Transfer by Feature Decoration

Lu Sheng, Ziyi Lin, Jing Shao and Xiaogang Wang1 CVPR2018

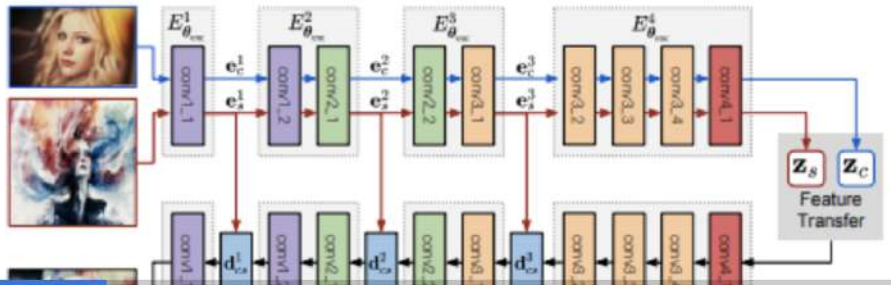
概要

既存手法のZero-shot style transferでは画像生成と効率のトレードオフによって，高品質な画像の生成とリアルタイムでの画像生成(style transfer)が困難．本稿ではこの問題を解決し，効率的かつ効果的な画像生成が可能なAvatar-Netを提案．提案手法では，高品質なstyle transferを可能にし，有効性および効率についても実証．さらに複数のスタイルの統合や動画のデザインを用いたアプリケーションも実装．

Method	Execution Time	
	256 × 256 (sec)	512 × 512 (sec)
Gatys *et al.* [11]	12.18	43.25
AdaIN [14]	0.053	0.11
WCT [21]	0.62	0.93
Style-Swap [6]	0.064	0.23
Ours-ZCA	0.26	0.47
Ours-ZCA-Sampling	0.24	0.32
Ours-AdaIN	0.071	0.28

Table 1. Execution time comparison.

新規性・結果・なぜ通ったか？



コメント・リンク集

- 論文

[#36]

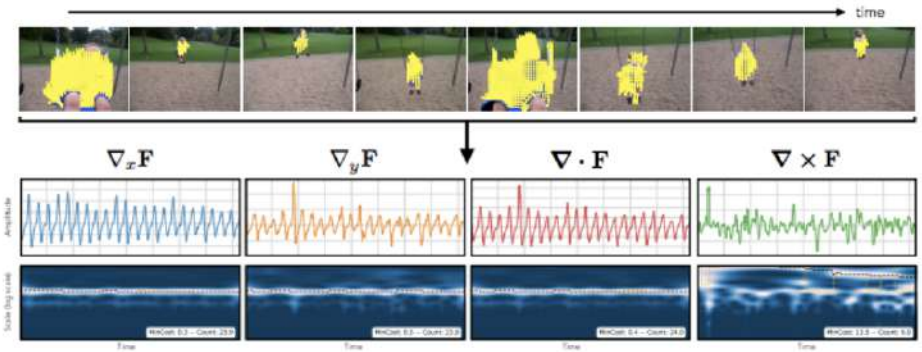
Yuta Matsuzaki

Real-World Repetition Estimation by Div, Grad and Curl

Tom F. H. Runia, Cees G. M. Snoek andArnold W. M. Smeulders CVPR2018

概要

動画中に存在する繰り返しの動作を推定する問題について考慮．既存の研究(フーリエベース)では静的および定常周期性という仮定のもとでは良好な精度であるが，現実的なシーンにおいては測定が困難．そこでウェーブレット変換を適用し，非静的かつ非定常な動画においても適切に処理できる手法を提案．また，非静的かつ非定常な動画で構成されるQUVA Repetition datasetを提案．動画内の繰り返し動作のカウント実験では深層学習による手法に比べ，良好な精度を実現．



新規性・結果・なぜ通ったか？

- 流動場とその微分から，3つの基本的な運動タイプと3次元内の固有周期性の3つの運動周期性を導出
- 3次元の周期性の2次元的な知覚は2つの極端な視点を考慮しており，18の基本的なケースを考慮
- 様々な繰り返し動作の出現に対応するために，セグメント化された前景の動きに対する時間変化量Ftおよびその差異∇Ft，∇・Ft

コメント・リンク集

- 論文

[#37]

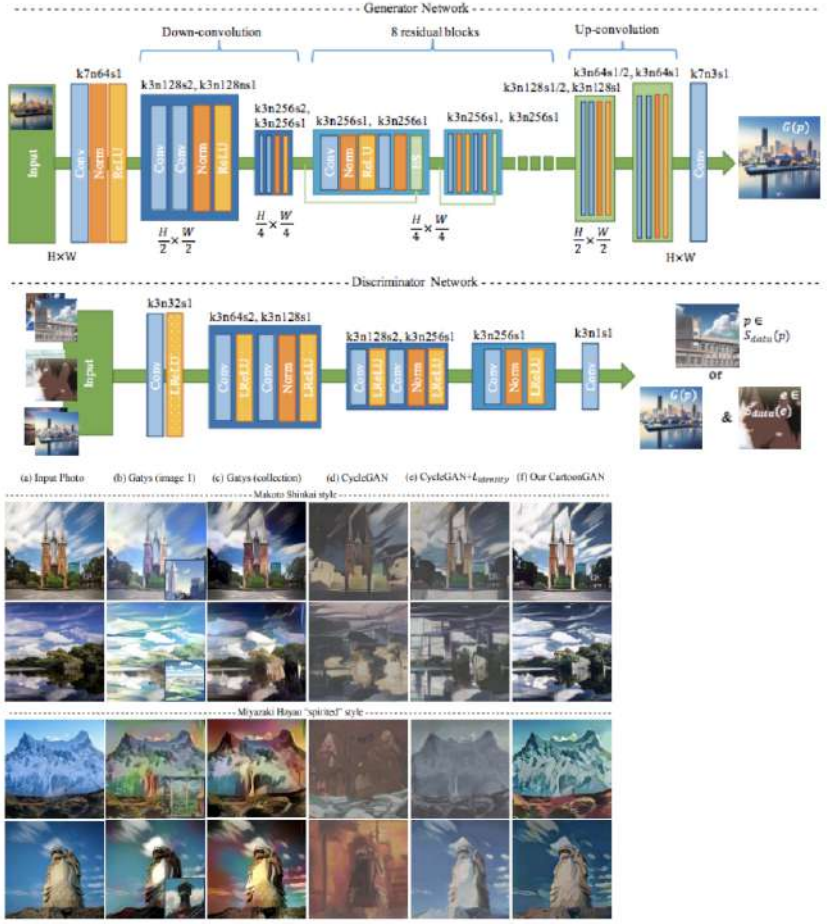
Yuta Matsuzaki

CartoonGAN: Generative Adversarial Networks for Photo Cartoonization

Yang Chen, Yu-Kun Lai and Yong-Jin Liu CVPR2018

概要

実世界の風景画(写真)を漫画スタイルの画像へ変換する手法の提案．漫画スタイル変換のためのGAN，CartoonGANを提案．ペアの画像を使用しない学習方法を採用し，そのための新規の損失関数を提案．実験では，写真のエッジや滑らかな陰影を保持したまま，アーティストのスタイルを表現することが可能であることを確認．



新規性・結果・なぜ通ったか？

画風変換には以下のような問題が存在，これにより既存の損失関数においては表現が困難

コメント・リンク集

- 論文

[#38]

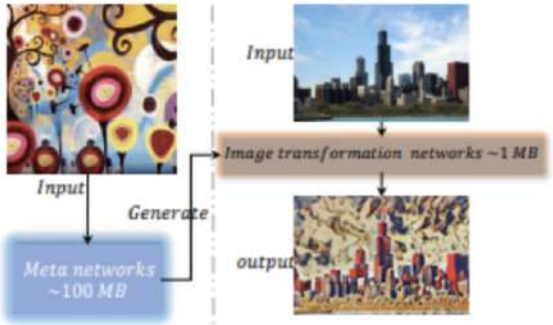
Yuta Matsuzaki

Neural Style Transfer via Meta Networks

Falong Shen, Shuicheng Yan and Gang Zeng CVPR2018

概要

本稿ではメタネットワークを用いた1つのフィードフォワードパスによる，(style transferのための)ニューラルネットワークパラメータを自動生成する手法を提案．最新のGPU 1つで19 ms以内に任意の新しいスタイルを表現することが可能．また，生成された画像変換ネットワークの容量はわずか449 KBでありモバイルデバイス上でリアルタイムでの実行が可能．



(a) Input (b) Style (c) Gatys *et al.* (d) John *et al.* (e) Ours (e) Ours-Fast

Method	*Encode*	*Transfer*	*Model*†
Gatys *et al.* [9]	N/A	9.52 *s*	N/A
Johnson *et al.* [17]	4 *h*	15 *ms*	7 MB
Chen and Schmidt [3]	0.4 *s*	0.17 *s*	10 MB
Huang and Belongie [14]	27 *ms*	18 *ms*	25 MB
Ours	19 *ms*	15 *ms*	7 MB
Ours-Fast	11 *ms*	8 *ms*	449 KB

新規性・結果・なぜ通ったか？

既存のstyle transferに関する研究の問題点

- スタイル毎にネットワークを学習する必要
- 推論の段階で確率的勾配降下による膨大な反復作業によって新規スタイルによる生成能力を欠く可能性

コメント・リンク集

- 論文

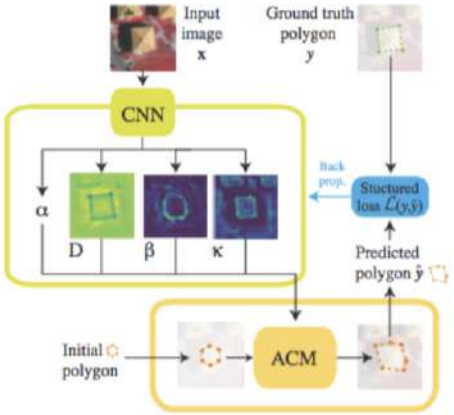
Shusuke Shigenaka

Learning deep structured active contours end-to-end

Diego Marcos, Benjamin Kellenberger, Lisa Zhang, Min Bai, Renjie Liao, Raquel Urtasun CVPR 2018

概要

この論文は,隣接する建物の境界線を幾何学的特性を利用して正確に描画するDeep Structured Active Contours (DSAC)の提案である．DSACは制約条件であるActive Contour Models(ACM)と従来のポリゴンモデルを使用している． 今回はCNNを用いてインスタンスごとのACMのパラメータを学習し, 構造化された出力モデルに全てのコンポーネントを組み込む方法を示し,DSACをend-to-endで学習可能にした． この論文は3つの困難なデータセット"building","instance","segmentation"をDSACで評価し, state-of-the-artと比較して優れた結果を残している．



新規性・結果・なぜ通ったか？

- CNNベースの方法に高度な幾何情報を利用可能にすることを目指している．
  - 明示的に多角形の出力を生成するCNNの作品はあまり行われていない
- CNNによる構造化学習はインスタンスレベルのセグメンテーションを扱う作業で認識されない．
  - 本手法は相互依存性をACMで調整することを学ぶため,損失をCNNで学習できる．
- IoUとエリア推定において従来のDSACより高い精度

リンク集

- 論文
- github

Takumu Ikeya

TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-ray

Xiaosong Wang et al. CVPR2018 1801.04334

概要

- 胸部のレントゲン写真から胸部疾病の分類及び報告を行うためのテキスト画像埋め込みネットワークの提案.
- 意味のあるテキストワードや画像領域を可視化するためのmultilevel attention modelsをend-to-endで学習可能なCNN-RNNアーキテクチャに統合.

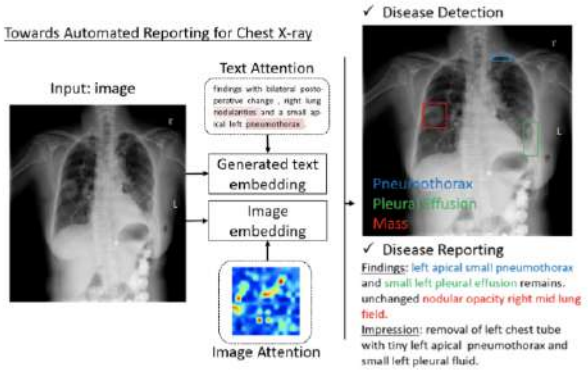


Figure 1. Overview of the proposed automated chest X-ray reporting framework. A multi-level attention model is introduced.

新規性・結果・なぜ通ったか？

- 分類精度を向上させるため，学習からattentionベースの画像と文字列内部表現の両方を組み合わせる手法が特徴.
- 提案したフレームワークは作成した評価用データセットの疾病ラベル割り当てタスクでAUCs平均0.9を達成.

コメント・リンク集

- 論文

[#41]

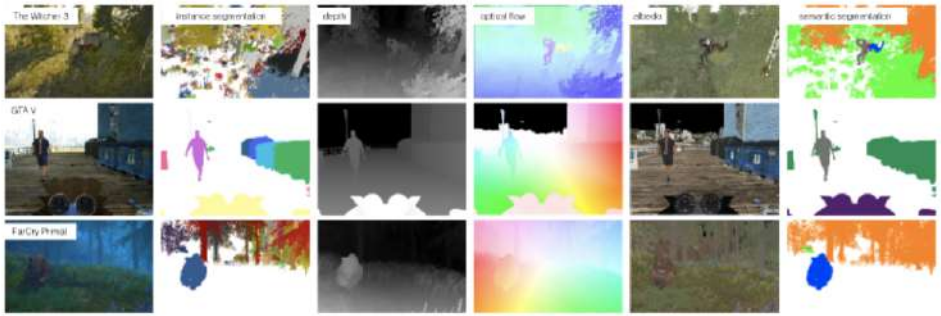
Masaki Miyamoto

Free supervision from video games

Philipp Krahenbuhl CVPR2018

概要

深層ネットワークでは大量のデータが必要で，ラベル付けされたデータはネットワークのデザイン同様深層ネットワークにとって重要である．しかし手作業の収集はお金と時間がかかる．そこでMicrosoftのDirectXレンダリングAPIを用いてゲームをやりながらリアルタイムでセグメンテーションやオプティカルフローなどのための正解ラベルを作成する手法を提案する．集めたデータセットは他の合成データセットより視覚的に現実世界と近いものになっている．



新規性・結果・なぜ通ったか？

このシステムはリアルタイムにすべてのラベルを計算するため直接ゲームのレンダリングパイプラインにコードを組み込んでいる．また人によるアノテーションが必要ない．さらに，様々なデザインの複数のゲームにおいてこの手法を用いることができる．

コメント・リンク集

- 論文

Tenga Wakamiya

Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh CVPR 2018 arXiv:1711.09577

概要

動画データセット上の比較的浅いものから非常に深いものまでの様々な3DCNNの構造を調べた．

Table 2: Accuracies on the Kinetics validation set. *Average* is averaged accuracy over *Top-1* and *Top-5*.

Method	Top-1	Top-5	Average
ResNet-18	54.2	78.1	66.1
ResNet-34	60.1	81.9	71.0
ResNet-50	61.3	83.1	72.2
ResNet-101	62.8	83.9	73.3
ResNet-152	63.0	**84.4**	**73.7**
ResNet-200	**63.1**	**84.4**	**73.7**
ResNet-200 (pre-act)	63.0	83.7	73.4
Wide ResNet-50	64.1	85.3	74.7
ResNeXt-101	**65.1**	**85.7**	**75.4**
DenseNet-121	59.7	81.9	70.8
DenseNet-201	61.3	83.3	72.3

新規性・結果・なぜ通ったか？

- ResNet-18の学習は，UCF-101，HMDB-51，およびActivityNetの過学習していて，Kineticsは過学習しなかった．
- Kineticsのデータセットは，深い層の3DCNNで学習するために十分なデータがあり，ImageNetの2D ResNetsと同様に，最大152のResNets層の学習を可能にし，ResNeXt-101は，Kineticsのテストセットで平均78.4%の精度がある．
- UCF-101およびHMDB-51上の複雑な2Dアーキテクチャよりも Kineticsの事前学習されたシンプルな3Dアーキテクチャが優れていて，UCF-101およびHMDB-51でそれぞれ94.5%および70.2%を達成した．

コメント・リンク集

- 論文

[#43]

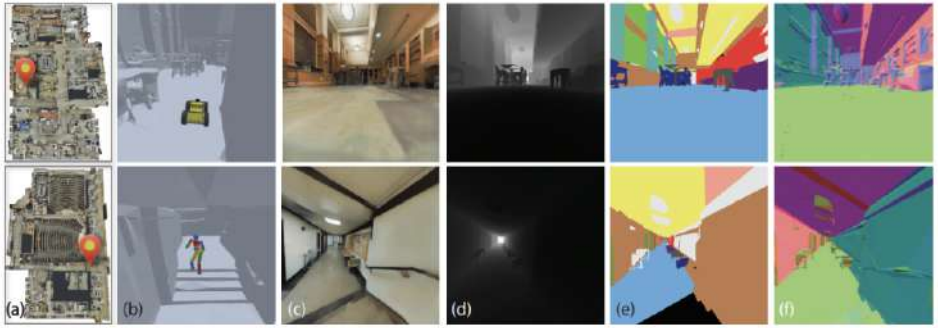
Shintaro Yamamoto

Gibson Env: Real-World Perception for Embodied Agents

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik and Silvio Savarese CVPR 2018

概要

ロボットなどのエージェントに知覚を身につけさせるためのGibsonという仮想環境を提案した。Gibsonは572の建物、1447のフロアから構築されている。 RGB-Dデータから、任意のカメラ位置でレンダリングする場合欠損が生じてしまう。 そこで、複数のカメラ位置でレンダリングした画像を組み合わせた上で、Neural Netにより欠損箇所を保管する。 得られた画像はリアルではないため、レンダリング画像とリアル画像間のドメイン変換手法Gogglesを提案した。 また、物理エンジンを組み込むことにより、実世界で起こる衝突などの判定を可能にした。



新規性・結果・なぜ通ったか？

目的地へ向かう、階段を上るといったエージェントのタスクに加え、depth推定、シーン認識によって有効性を検証した。実世界で撮影した画像によるテストでは、他のデータセットと比べ1番精度が良かった。

コメント・リンク集

- プロジェクトページ

[#44]

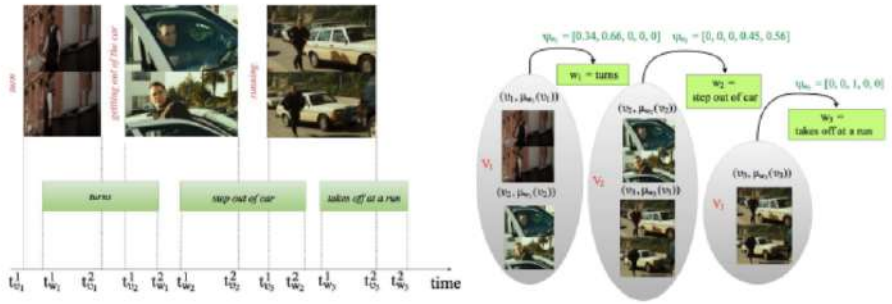
Shintaro Yamamoto

Multimodal Visual Concept Learning with Weakly Supervised Techniques

Giorgos Bouritsas, Petros Koutras, Athanasia Zlatintsi and Petros Maragos CVPR2018

概要

従来の動画認識に関する研究は、映像情報のみを用いているものが多く字幕のようなテキストや音などの情報は利用されていない。動画認識のタスクに、映像情報に加えテキスト情報を利用するための手法を提案した。 考慮すべきこととして、映像とテキストの情報が時系列的にどのように対応しているか、同じラベルに対してテキストでは複数の表現方法が存在している、という2つの点が挙げられる。 そこで、時系列的な対応付けを行うFuzzy Sets MIL(FSMIL)とテキストがどのラベルに対応しているかを推定するProbabilistic Labels MIL(PLMIL)の2つの学習方法を提案した。



新規性・結果・なぜ通ったか？

動画認識タスクとして、顔認識及びアクション認識の2つによりテストを行いベースラインと比べ精度が向上したことを確認した。

コメント・リンク集

- 論文

[#45]

Ryota Suzuki

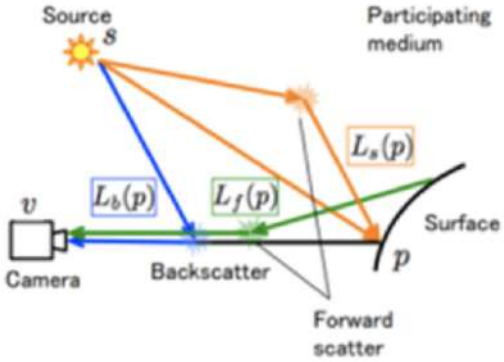
Photometric Stereo in Participating Media Considering Shape-Dependent Forward Scatter

Y. Fujimura, M. Iiyama, A. Hashimoto, M. Minoh CVPR2018

概要

濁った水や霧の中で撮影したような，散乱光により劣化したような画像に対して適用可能な3D復元手法の提案．

形状依存の前方散乱（forward scatter）を扱うモデルを考え，ルックアップテーブル使用で解析的に求める， それを空間的変化カーネルとして表現する． また，前方散乱の除去を可能にする，大規模密行列を疎行列に近似する手法を提案．



新規性・結果・なぜ通ったか？

厳密に形状依存の表面-カメラ間前方散乱をモデル化し，その解析的解法を提案したものは初めて．

実，合成データに対して改善的性能を示した．

コメント・リンク集

- 論文

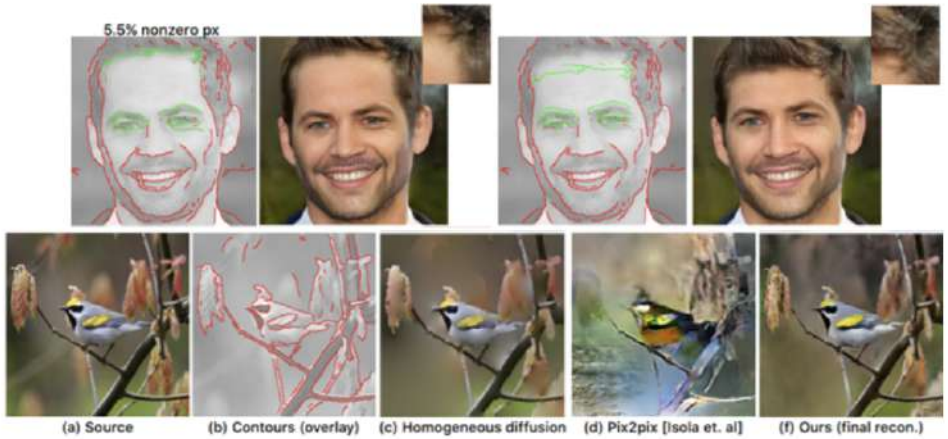
[#46]

Ryota Suzuki

Sparse, Smart Contours to Represent and Edit Images

T. Dekel, D. Krishnan, C. Gan, C. Liu, W. Freeman CVPR2018

概要



かなりスパースな輪郭線（元画像の4%程度のデータ量）から大変きれいな画像の復元ができ，更に輪郭線を調節すると大変きれいにパーツ位置を変えられる．参照画像も変更できるので，髪を生やせるし，（効果は薄いが）人の鼻を犬っぽくできる．

まず，入力の輪郭線を工夫する．この手法でスパースな輪郭線を取り，輪郭線の左右の画素の色(RGB)を色値（RGB×左右＝計6値）とする． また，画像の各色における勾配を取り，輪郭線の位置におけるRGB×XY成分＝計6値を勾配値とする． ここからN次元特徴マップを（GANを回している最中に）学習する． 構造はDeeplabを参考にしたDilated Conv.による簡素なネットワーク構造による．

この輪郭線特徴を入力として，2段階の復元用U-Netを生成器に，Dilated-Patch Discriminatorを判別器にしたGANを回す．

新規性・結果・なぜ通ったか？

アプリケーションとしてかなり使い出かあるように見える．

コメント・リンク集

実験的に見て，N=3がいいらしい．

- 論文
- プロジェクトページ

[#47]

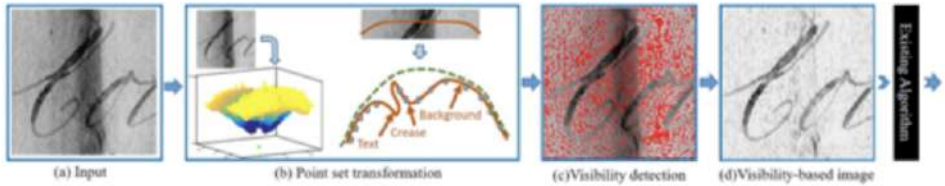
Ryota Suzuki

Document Enhancement using Visibility Detection

N. Kligler, S. Katz and A. Tal CVPR2018

概要

文書から二値化，陰影除去をするのに使えるDocument Enhancementの話．文書平面を三次元化し，文書面から凸凹を除去するという形で可視領域（Visibility）の検出をし，それをベースに鮮鋭化するというやり方．本手法を前処理として，二値化手法や陰影除去を適用するとSOTA性能を上回る．



新規性・結果・なぜ通ったか？

基本方針としては，識別性を高める高次元空間への変換のやり方を考えました，という非ディープなパタレコにおけるノリ．

論文の質としては他論文と比較して若干劣るように感じられるが，「平面だけど三次元点群にするとうまくいくとは，驚きだ！」と言っていて，それがウケたのだろうか．おそらく当初の発想も文書の凸凹を消すという発想だったと思われる．

コメント・リンク集

肝心の3次元空間への射影の具体的な実装（$(x, y) \rightarrow (\theta, \phi)$の部分）が読み取れませんでした．どなたか再現できたらご教授頂けますと幸いです．

- 論文

[#48]

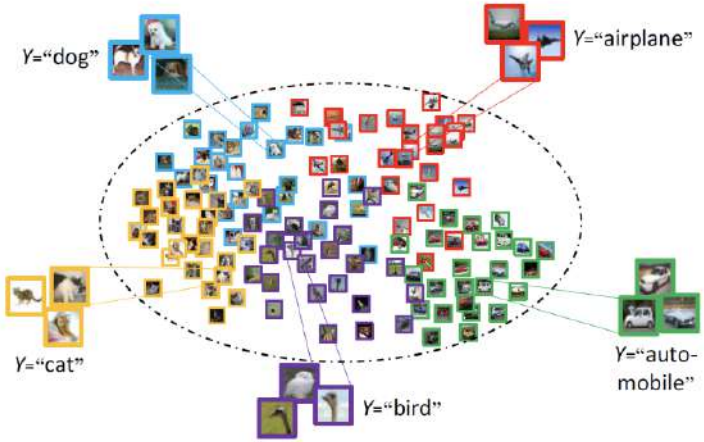
Hirokatsu Kataoka

An Efficient and Provable Approach for Mixture Proportion Estimation Using Linear Independence Assumption

Xiyu Yu, Tongliang Liu, Mingming Gong, Kayhan Batmanghelich, Dacheng Tao CVPR 2018

概要

混合分布内のラベルなしデータと少量のラベルありデータから正しく分布の重み（Weights of components）を推定し、画像分類を行う問題を提供。この問題自体をMixture Proportion Estimation（MPE）という。



新規性・結果・なぜ通ったか？

データに多数のノイズを含んでいても、少量のラベル付きデータから混合分布の割合を把握して正しく画像分類を行うことができるアルゴリズムを提案。Web画像に見られるラベルノイズが発生している学習/Semi-supervised学習、合成データ/実世界データの両者においてState-of-the-artな精度を達成した。

コメント・リンク集

ラベルノイズに関する新規の問題MPEを提供した。一見すると既存の問題と思われるようなものでもまだまだ重要で提案されていない問題は残っている？

- 論文

[#49]

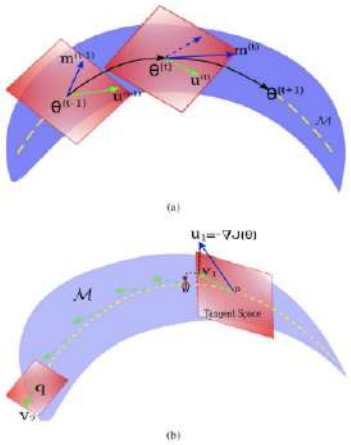
Hirokatsu Kataoka

Geometry Aware Constrained Optimization Techniques for Deep Learning

Soumava Kumar Roy, Zakaria Mhammedi, Mehrtash Harandi CVPR 2018

概要

勾配の最適化手法であるStochastic Gradient Descent（SGD）やRMSPropアルゴリズムをRiemannian Optimizationの設定にて一般化する手法を提案する。SGDはDNNでは一般的に用いられるが、勾配の最適化に大きな分散があり、一方でRMSPropやADAMがこの問題を解決するために提案されてきたが決定だとは言えなかった。本論文ではRiemannian Centroidsの計算や深層距離学習（Deep Metric Learning）を考慮して勾配最適化の不安定性に取り組む。詳細画像識別問題に取り組むことで提案手法の有効性を示した。右図は最適化のイメージ図であり、Riemannian多様体空間で勾配計算と誤差最適化を測ることで安定感のある最適化を実現。



新規性・結果・なぜ通ったか？

多様体空間で最適化を実現するcSGD-M/cRMSPropを提案、問題設定に対して拘束を強めてダイレクトに最適化ができる手法とした。機械学習の文脈において、PCA/DMLの拡張と位置付けられる手法を提案。同枠組みを詳細画像識別問題に適用したところ、Competitiveな結果を達成した。

コメント・リンク集

発想が数学の人、〜を＊＊の枠組みで最適化するというのは得意技？

- 論文

[#50]

Hirokatsu Kataoka

View Extrapolation of Human Body from a Single Image

Hao Zhu, Hao Su, Peng Wang, Xun Cao, Ruigang Yang CVPR 2018

概要

ある視点の人物画像からターゲットとなる視点（Novel View）の人物画像を復元するタスクを提案。従来法であるVSAP（参考文献40）では正確な視点変化に関するフローを推定することができなかったが、提案法ではまず距離画像を推定してからフロー推定することで精度を劇的に改善した。

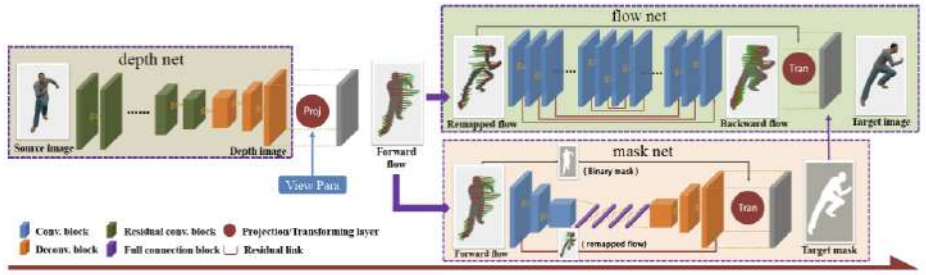


Figure 2. The full pipeline of our approach. The architecture of network is simplified, and the detailed parameters will be shown in supplement materials.

新規性・結果・なぜ通ったか？

距離画像の復元（予め形状を復元することに相当）することにより、ビューポイント変化に関するフローの推定精度を劇的に向上させ、さらにバックフローも組み合わせることでターゲット視点の人物画像復元を改善。距離画像の復元からオプティカルフローの推定を行うこのような枠組みをShape-from-Appearanceという？3次元的な情報があることで姿勢に関するバリエーションがあったとしてもロバストなビューポイント変化の人物画像推定が可能。合成データによる人物画像データセットも作成、2,000の姿勢に対して22のアピアランス変化を含む。

コメント・リンク集

以前は経由する情報をいかに少なくしてダイレクトに復元を行うか、が重要であったが、DNN時代になってから効果的な情報復元（この場合は距離画像による形状復元）を経由することにより推定精度が向上。

- 論文

[#51]

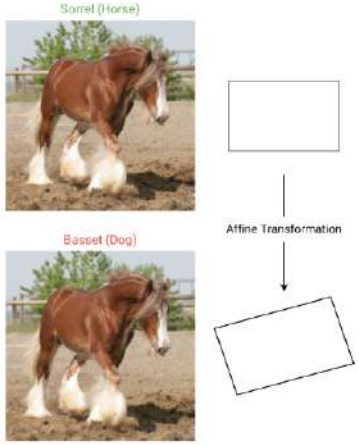
Hirokatsu Kataoka

Geometric robustness of deep networks: analysis and improvement

Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard CVPR 2018

概要

幾何学的な変換に頑健なDNNを考案。従来のDNNでは例えば右図のようなアフィン変換（ここでは主に回転）に対して脆弱であり、上図では馬の種類を答えていたものが、多少の回転を与えるだけで犬の種類を答えてしまう。本論文ではManiFoolというシンプルだがスケーラブル、多様体（Manifold）ベースのアルゴリズムManiFoolを提案、幾何学的な変化に対する不変性や複雑ネットワークに対する評価を行う。さらに、Adversarial Trainingにより幾何学的な変動に頑健なモデルとなるような学習法を実装した。



新規性・結果・なぜ通ったか？

最小の幾何学的変換により認識を誤ってしまう問題に対して不変性を計測するManiFoolを提案したことがもっとも大きな貢献である。ImageNet等の大規模データに対して幾何学的変換とそのロバスト性を評価した最初の論文である。ManiFoolアルゴリズムをAdversarial Trainingに応用して幾何学的変換に対してロバストな学習法を提案。

コメント・リンク集

実環境（撮影時のカメラのビューポイント）を多少回転させるのではなく画像をダイレクトなアフィン変換にて回転させるからエラーが生じる？もう少し解析して欲しいような気もする。

- 論文

[#52]

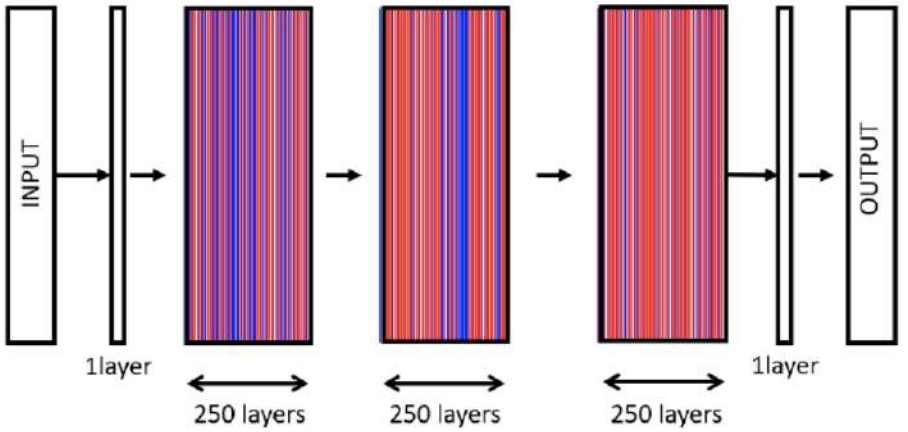
Hirokatsu Kataoka

Learning Strict Identity Mappings in Deep Residual Networks

Xin Yu, Zhiding Yu, Srikumar Ramalingam CVPR 2018

概要

自動的に冗長なレイヤを除外してくれるε-ResNetを提案し、よりコンパクトなサイズで最大限の認識パフォーマンスを実現する。ε-ResNetでは閾値εを設けて、これよりも小さい値を出力するレイヤに対して誤差を計算しないという方策を取る。提案法であるε-ResNetを実現するために、少量のReLUを加えることで実現した。CIFAR-10,-100,SVHN,ImageNetに対して単一のトレーニングプロセスで学習が成功し、なおかつ約80%ものパラメータ削減を実行した。右図は752層のε-ResNetを実装して最適化した例である。図中の赤ラインは除去されたレイヤ、青ラインは認識に対して必要と判断されたレイヤである。図の例では、CIFAR-100に対するオリジナル（ResNet-752）のエラー率が24.8%、提案法（ε-ResNet-752）のエラー率が23.8%であった。



新規性・結果・なぜ通ったか？

ResNetを対象として、レイヤを増加させることによる冗長性を自動的に除去してくれるε-ResNetを提案した。ε-ResNetは従来の枠組みに対して4つのReLUを組み合わせ、閾値カット処理だけで実装可能である。より深い層のモデルに対して有効であり、大体80%くらいの冗長生をカットする。パラメータ数を減らしつつも超ディープなモデルにおいて多少の精度向上が見込める。

コメント・リンク集

実装が非常に簡単そうであり、すでにDNNフレームワークにおいて実装されていれば、広く使ってもらえそう。また、各タスク（e.g. 物体検出、セグメンテーション、動画認識）において気軽に使用することができれば、広がりがありそう。

- 論文

[#53]

Hirokatsu Kataoka

Generative Adversarial Perturbations

Omid Poursaeed, Isay Katsman, Bicheng Gao, Serge Belongie CVPR 2018

概要

敵対的サンプル（Adversarial Examples）を生成的に作りだすモデルを考案し、自然画像に対して摂動ノイズを与えて学習済みモデルを効果的にだます手法（GAP; Generative Adversarial Perturbations）を提案する。提案のGAPは画像に依存する/しない摂動ノイズ、いずれも生成することが可能であり、画像識別やセマンティックセグメンテーションに対して有効。また、ImageNet/Cityscapesを用いたより高解像な画像においても効果的に識別器をだますことに成功した。さらに、従来の同様の枠組みよりもより速く推論を行うことができる。

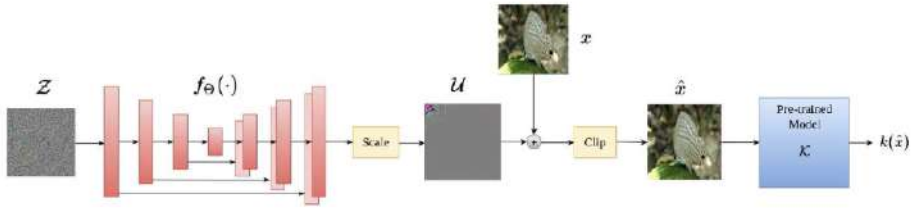


Figure 1: Training architecture for generating universal adversarial perturbations. A fixed pattern, sampled from a uniform distribution, is passed through the generator. The scaled result is the universal perturbation which, when added to natural images, can mislead the pre-trained model. We consider both U-Net (illustrated here) and ResNet Generator architectures.

新規性・結果・なぜ通ったか？

より汎用的かつ画像依存性のあり/なしに関わらない摂動ノイズを、画像識別/セマンティックセグメンテーションに対して行うことができる。それでいてUniversal Perturbationsの枠組みを生成モデルにより実装、より効果的にだますことに成功。

コメント・リンク集

この論文は引用されそう？だが、ホントの意味で騙せているのかは不明である。（Adversarial Examplesの論文は、会議の前に攻略法がarXivに載せられるなどまだまだ研究が必要である）

- 論文

[#54]

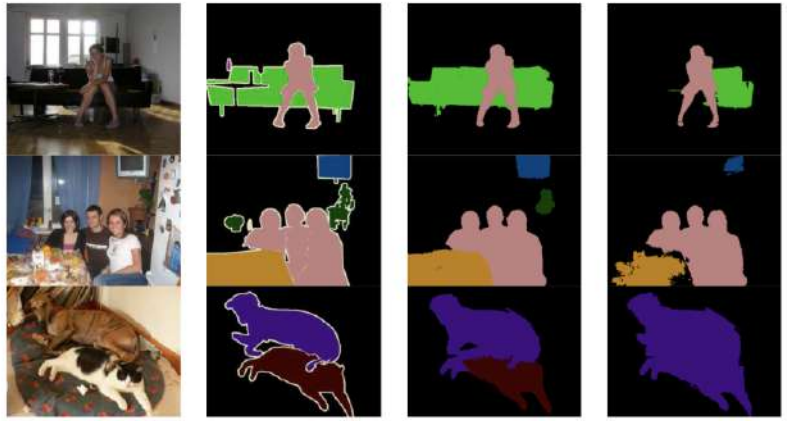
Hirokatsu Kataoka

The Lovasz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks

Maxim Berman, Amal Rannen Triki, Matthew B. Blaschko CVPR 2018

概要

セマンティックセグメンテーションにおいて、ピクセルごとの最適化ではなく領域（Intersection-over-Union）ごとの最適化を行うことで小領域を含む領域ベースのセグメンテーションを改良する。この問題に対して、サブモデュラ凸最適化手法Lovasz（参考文献26をベースとした）を用いることで誤差計算を行う。このLovasz-Softmax Lossは従来のCross-Entropy Lossよりも領域評価jに対して頑健であることを示した（右図）。位置付け的にはLovasz Hinge Lossのマルチカテゴリに対する一般化である。



(a) Input images (b) Ground truth masks (c) Lovász-Softmax + CRF (d) Cross-entropy + CRF

新規性・結果・なぜ通ったか？

セマンティックセグメンテーションにおいて特に小領域であったとしても適切に評価して誤差を計算できるLovasz-Softmax Lossを提案した。PascalVOCやCityscapesにおいてCross-Entropy Lossを用いた誤差計算よりも良好な性能を示すことが明らかとなった。

コメント・リンク集

IoUで最適化するとは？また、Jaccard indexとは何のことだろう？

- 論文

[#55]

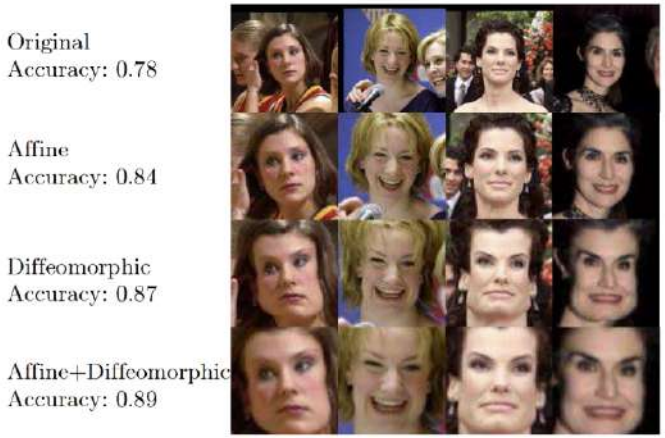
Hirokatsu Kataoka

Deep Diffeomorphic Transformer Networks

Nicki Skafte Detlefsen, Oren Freifeld, Søren Hauberg CVPR 2018

概要

顔認識において、本人認識率が向上するようにアフィン変換や形状変化（Diffeomorphic）を行うように変換を実装するネットワークDeep Diffeomorphic Transformer Networksを提案。直感的にはズームインだが、さらに形状変化を行うことが効果的であると判断してネットワークを構築した。



新規性・結果・なぜ通ったか？

顔認識においてアフィン変換によるズームインのみならず、認証率が向上するような形状変化方法であるDiffeomorphic Transferを提案した。同処理はCNN内に実装され、Deep Diffeomorphic Transformer Networksと呼ばれ、LFW/CelebA等でState-of-the-artであった。

コメント・リンク集

ネットワークに対して内的ではなく外的に変形させて精度向上するのは意外である。

- 論文

[#56]

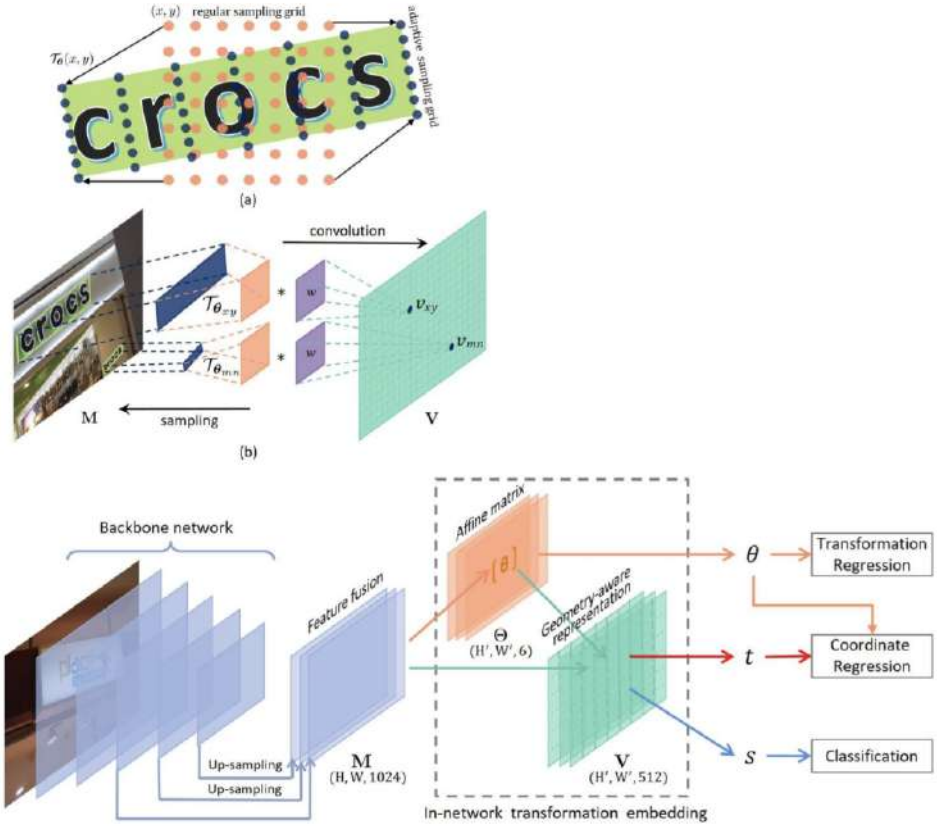
Ryosuke Araki

Geometry-Aware Scene Text Detection with Instance Transformation Network

Fangfang Wang, Liming Zhao, Xi Li, Xinchao Wang and Dacheng Tao CVPR2018 167

概要

幾何学的な表現を用いたEnd-to-endのシーンテキスト認識アプローチ．シーンテキストインスタンスの幾何学的構成をエンコーディングするため，幾何学的な表現を学習するInstance Transformation Network (ITN)を提案する．右図上部の（a）のように，いくつか並んだサンプルグリッド（橙色）をテキストにフィッティング（青色）する．また，（b）のように入力画像（の特徴マップ）からフィッティングのためのモデルを学習する．ネットワーク構成は，特徴抽出部，インスタンスレベルのアフィン変換を予測する部分，幾何学的表現部からなる．変換の回帰，座標の回帰，分類はマルチタスク学習となる．



新規性・結果・なぜ通ったか？

幾何学的表現で強いアフィン変換がかかっていても頑健なテキスト検出が可能である．データセットにはICDAR2015およびMSRA-TD500を用いて評価を行う．ベースネットワークにResNet50を用いた場合，MSRA-TD500のPrecisionは90.3，F値は80.3と非常に高精度な結果となった．ICDAR2015ではVGG16ベースの方が良い結果となり，Precisionは85.7，F値は79.5である．

コメント・リンク集

幾何学的なドット列をフィッティングする手法は他にも応用が効きそう．

- 論文

[#1]

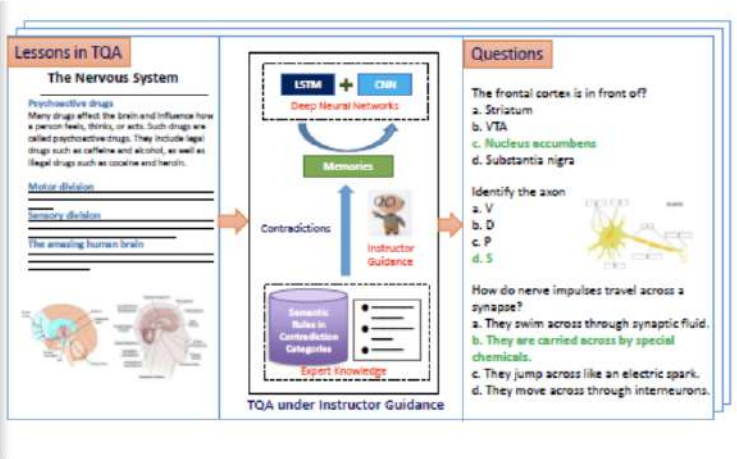
Shintaro Yamamoto

Textbook Question Answering under Instructor Guidance with Memory Networks

Juzheng Li, Hang Su, Jun Zhu, Siyu Wang and Bo Zhang CVPR 2018

概要

教科書(テキストデータ＋画像)に含まれている情報に関する質問に答える、Textbook Question Answering(TQA)に関する研究。質問の答えはテキストの局所的な部分に含まれていることが多く、テキストの要約によって答えを得ることが難しい場合が多い。 本研究では、テキストや画像から得られる因果関係や構造を表したContradiction Entity-Relationship Graph(CERG)を構築し、矛盾を探すための手がかり(Guidance)とすることで局所的な情報を使用して質問に答えることを可能とする。 CERGの構築には画像特徴とテキスト特徴を使用し、質問の答えには画像特徴とテキスト特徴に加えCERGから得られたGuidanceを用いることで出力を得る。



新規性・結果・なぜ通ったか？

Contextが多く要約することが難しい場合、得られる情報をグラフにして記憶することが効率的であるということを示した。ベースラインやランダムに選択する場合と比べて、あらゆる質問のタイプ(truth or falseやmultiple choise)において正解率が向上していることを確認した。

コメント・リンク集

一応画像情報を使用しているが、全体的にはNLP色が強いと感じた。手法としての完成度は非常に高く、評価は問題自体が新しいこともあり数値評価（従来法との比較、モデル設計の評価）及びqualitativeな比較であった。

- 論文

[#2]

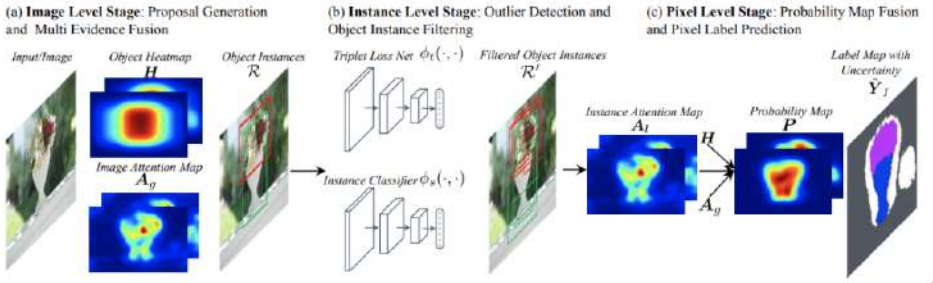
Kazuho Kito

Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning

Weifeng Ge, Sibei Yang and Yizhou Yu CVPR2018

概要

マルチレベルの物体認識，検出，セマンティックセグメンテーションのための弱教師カリキュラム付き学習のパイプラインを提案。このパイプラインは物体位置の中間点と訓練画像のピクセルのラベルの結果をを入手し、結果を用いて教師付きのやり方で特定のタスクの深層学習で訓練する。その全体のプロセスは4つのステージを含む、訓練画像の物体位置を含み、物体のインスタンスのフィルタリングと結合し、訓練画像のピクセルラベリングをし、特定のタスクのネットワークでトレーニングをする。訓練画像からキレイな物体のインスタンスを入手することで、物体のインスタンスのフィルタリング、結合、クラスファイリングのための新しいアルゴリズムを複数の解決策から集める。このアルゴリズムは、検出された物体のインスタンスをフィルタリングするため、metric learningと密度ベースのクラスタリングの両方を組み込んでいる。



新規性・結果・なぜ通ったか？

マルチレベルの画像の分類においてstate-of-the-artを達成.

コメント・リンク集

- 論文

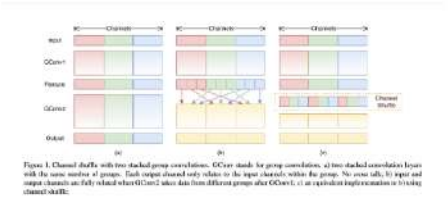
Takumu Ikeya

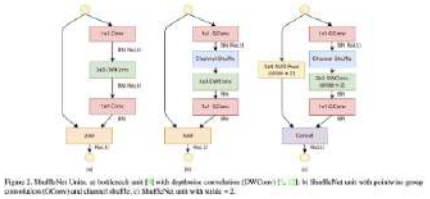
ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices

Xiangyu Zhang et al. CVPR2018 1707.01083

概要

- モバイルデバイス向けに特別に設計した非常に計算効率の良いCNNアーキテクチャである“ShufflNet”を開発した．このアーキテクチャではpointwise group convolutionとchannel shuffleという2つの新しい演算を使用し，精度を落とすことなく，計算コストを大幅に削減した．





新規性・結果・なぜ通ったか？

- ImageNetによる分類とMS COCOによる物体検出のタスクではほかのアーキテクチャよりも高い性能を示した．
- 40MFLOPの計算資源の制約のもと,ImageNet分類タスクで他のモバイルデバイス向けアーキテクチャよりもtop-1 エラーが7.8%低い結果が得られた．
- 既存のアーキテクチャよりも高精度で計算効率が非常に良い“ShufflNet”というアーキテクチャを提案した．

コメント・リンク集

- 論文

Takumu Ikeya

What have we learned from deep representations for action recognition?

Christoph Feichtenhofer et al. CVPR2018 1801.01415

概要!

- 動画中の行動を認識するためにtwo stream modelが学習したものを視覚化することで時空間表現がどのように働いているか調査した研究.
- 単純に形状特徴と動作特徴を分割するよりも，cross-stream fusionは正しい時空間特徴を学習することが可能.
- ネットワークはクラス特有の局所表現だけでなく，様々なクラスに対応できる汎用表現を学習することが可能.
- ネットワークの階層全体を通して，特徴はより抽象的になり，ある動作の区別にとって重要でないデータに対する不変性が増加.
- 視覚化は、学習された表現を確認するだけでなく，学習データの独自性を明らかにし，systemの失敗例の説目に利用可能.

(a) (b) (c) (d)

Figure 1. Studying a single filter at layer conv5_fusion: (a) and (b) show what maximizes the unit at the input: multiple coloured blobs in the appearance input (a) and moving circular objects at the motion input (b). (c) shows a sample clip from the test set, and (d) the corresponding optical flow (where the RGB channels correspond to the horizontal, vertical and magnitude flow components respectively). Note that (a) and (b) are optimized from white noise under regularized spatiotemporal variation.

新規性・結果・なぜ通ったか？

- ランダムに初期化されたノイズ画像とノイズ動画の入力から開始するモデルの時空間の入力を直接最適化する.

コメント・リンク集

- 論文

[#5]

Ryota Suzuki

A Perceptual Measure for Deep Single Image Camera Calibration

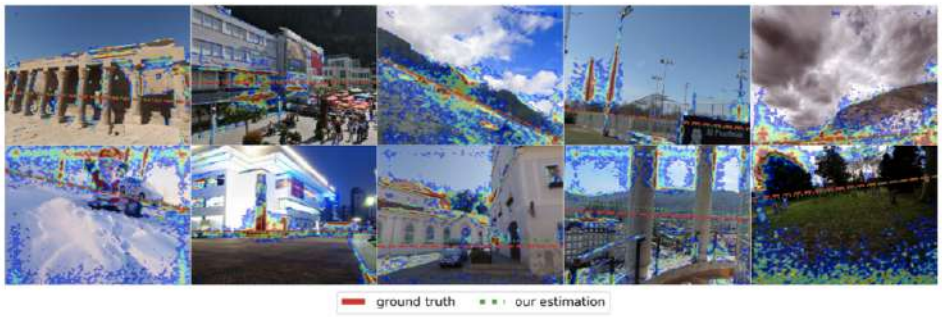
Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap and J.F. Lalonde CVPR2018

概要

単画像におけるカメラパラメータのキャリブレーションの話．事前知識なしに非コントロール環境でもちゃんと動くように，DCNNによるキャリブレーションパラメータの直接推測手法を提案する．

ImageNet学習済みDenseNetの最終層を３つの分離したヘッドに置き換え，それぞれ水平角度推定，水平線の中心からの距離，縦方向の場を表すように改造する． これを，大規模パノラマ画像データセットから自動生成したサンプルにより学習する．

評価については，実際人がおかしさを感じるかどうかによるので，AMTで聞いてみた結果から導いた人の誤差モデルをもとに語ってみる．



結果とアテンション

新規性・結果・なぜ通ったか？

結果はそれなりにできている．が，それなりっぽく見えてしまうので，人間の感じ方もちゃんと調べて載せた！ というのが評価されているように思う．

ネットワーク構造の簡単な調整で達成できたところが，DNNの手に掛かれば様々な問題が如何様にも解ける感じを醸し出していておもしろい．

アプリケーション枠狙いにするためか，アプリケーション例をいくつか掲載している．論文自体，他のアプリケーション系論文と比べて，読んでいて飽きない感じがする． 合わせ技一本，という感じがする．

コメント・リンク集

速読したからかもしれないが，不思議な構成の論文だった．論点が2つあるからだろうか．違和感は感じるが，なんとかうまく収めている感じもする．

NVidiaにGPUを寄付してもらったらしい．

- 論文

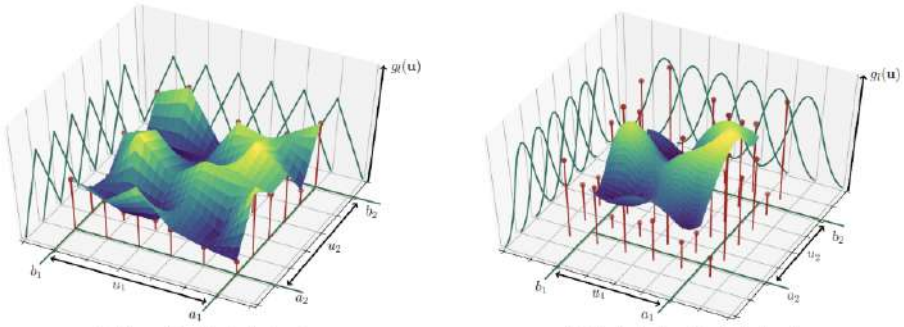
KazukiTsubura

SplineCNN: Fast Geometric Deep Learning with Continuous B-Spline Kernels

Matthias Fey, Jan Eric Lenssen, Frank Weichert, Heinrich M¨uller CVPR2018

概要

グラフなどの不規則な構造をした幾何学的入力のためのディープニューラルネットワークの変形であるスプラインベースの畳み込みニューラルネットワーク(SplineCNN)．スペクトル領域内でフィルタリングするのではなく，純粋に空間領域で特徴集計をする．
SplineCNNを使用することで，手作業による特徴記述子の代わりに入力として幾何学的構造を使用することで，深いアーキテクチャの完全なend-to-endの学習が可能になる．

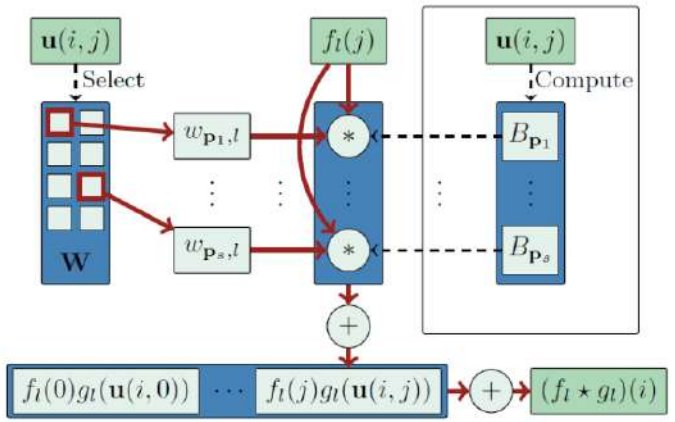


(a) Linear B-spline basis functions

(b) Quadratic B-spline basis functions

新規性・差分

グラフやmeshesのような不規則な構造をした様々な点で利用でき，空間上における入力の幾何学的関係を発見する．手作業による特徴記述子を使用せずにend-to-endの学習が可能になり，また，最先端の幾何学的な学習と同等である．



- 論文
- Github

Yoshihiro Fukuhara

Learning and Using the Arrow of Time

Donglai Wei et al. CVPR 2018

概要

DNN を用いて動画中の時間の流れている方向（Arrow of Time）を学習する研究. 人工的な信号を含むキューは Arrow of Time の学習に悪影響を及ぼすことを示し, それらの影響を取り除いた大規模 dataset を作成した. 評価実験では映画中の逆再生部分を検出するというタスクにおいて人間とほぼ同程度の精度を達成した.

新規性・結果・なぜ通ったか？

- Arrow of Time を学習する DNN アーキテクチャとして Temporal Class-Activation Map Network (T-CAM) を提案
- T-CAM は数フレーム分の optical flow を入力から Arrow of Time を推測
- 人工的な信号である camera Motion や black framing を含むキューは Arrow of Time の推定を容易にし, ネットワークの学習に悪影響を与えてしまうことを実験により示した
- 上記の人工的な信号を取り除いた Arrow of Time を学習するための大規模データセット, Flickr-AoT と Kinetics-AoT を作成
- 提案手法を用いて行った映画の逆再生部分を検出する実験では, 人間（80%）とほぼ同等（76%）の結果を達成
- また, Arrow of Time が flow-based の行動認識において self-supervised pre-training に有用であることを示した

コメント・リンク集

- [論文] Learning and Using the Arrow of Time
- [動画] YouTube

Ryota Suzuki

Missing Slice Recovery for Tensors Using a Low-rank Model in Embedded Space

T. Yokota, B. Erem, S. Guler, S.K. Warfield and H. Hontani CVPR2018

概要

テンソルがスライス方向に欠けてしまった場合の復元についての論文．このケースでは，よく行われる核ノルム利用やその他正則化手法ではムリ．遅れ／シフトに不変な構造を捉えることが重要になることから，「高次元空間への低ランクモデルの埋め込み」を行うことで解決する．時系列の遅延埋め込みを，テンソルにおける「複数方向遅延埋め込み変換」を行い，不完全なテンソルを高次不完全ハンケルテンソルへと変換する．その後，この高次テンソルをタッカー展開の枠組みで低ランク化することで 復元が行われる．

新規性・結果・なぜ通ったか？

伝統的に行われてきた行列・テンソル解析系の論文．情報学部出身の読者になるべく分かりやすいように丁寧に書いているように見受けられる．画像で言えば，伝送エラーなどで行の一部分や下半分が吹き飛んでしまった時などに使える復元手法．

コメント・リンク集

きちんと読み手への導入は行われているものの，読み下すには，テンソル分解程度の数学の知識が必要．ついでに，カオスのような時系列システムも知っているとわかりやすい（図中の説明での事例がそれ）．まとめ人にとっては数学の復習になったので，ぜひ論文を読んでみていただきたい．

- 論文

[#9]

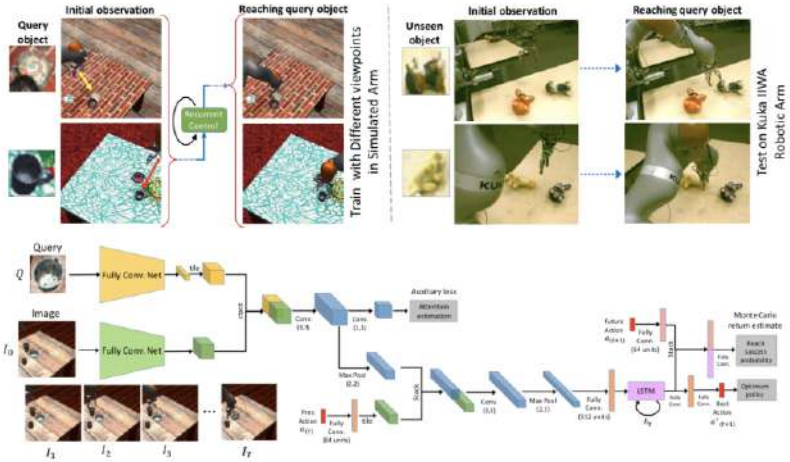
Yoshihiro Fukuhara

Sim2Real Viewpoint Invariant Visual Servoing by Recurrent Control

Fereshteh Sadeghi et al. CVPR 2018

概要

ロボットアームを用いたビジュアルサーボについての研究. DNN を用いた視点に依存しないビジュアルサーボの能力を学習する Recurrent Convolutional Neural Network Controller を提案. 様々な視点, 光源環境, 物体の種類や位置に置けるタスクをシミュレーション上で学習することで, 未知の視点において自動でキャリブレーションを行うことが可能.



新規性・結果・なぜ通ったか？

- コントローラーは目的物体のクエリ画像, 現在の観測画像, 1つ前の行動, 現在の内部状態から次の行動と内部状態を決定する
- LSTM を用いてネットワークが過去の行動の結果を参照できるようにすることで Jacobian (action と motion との関係) についての事前知識無しでの学習を可能とした
- ロス関数にはとった行動によって目的物体との距離がどのように変化したかと, 長期的な行動の価値を学習するための Q-関数 (行動状態価値関数) を用いる
- 少数のアノテーション付きシークエンスがあれば, シミュレーション上で学習結果を実際のロボットへ転移することが可能（追加で学習が必要なのは画像特徴の部分のみのため）
- 実際のロボットに学習結果を転移して行った評価実験では, 物体へ

コメント・リンク集

- [論文] Sim2Real Viewpoint Invariant Visual Servoing by Recurrent Control
- [動画] YouTube

[#10]

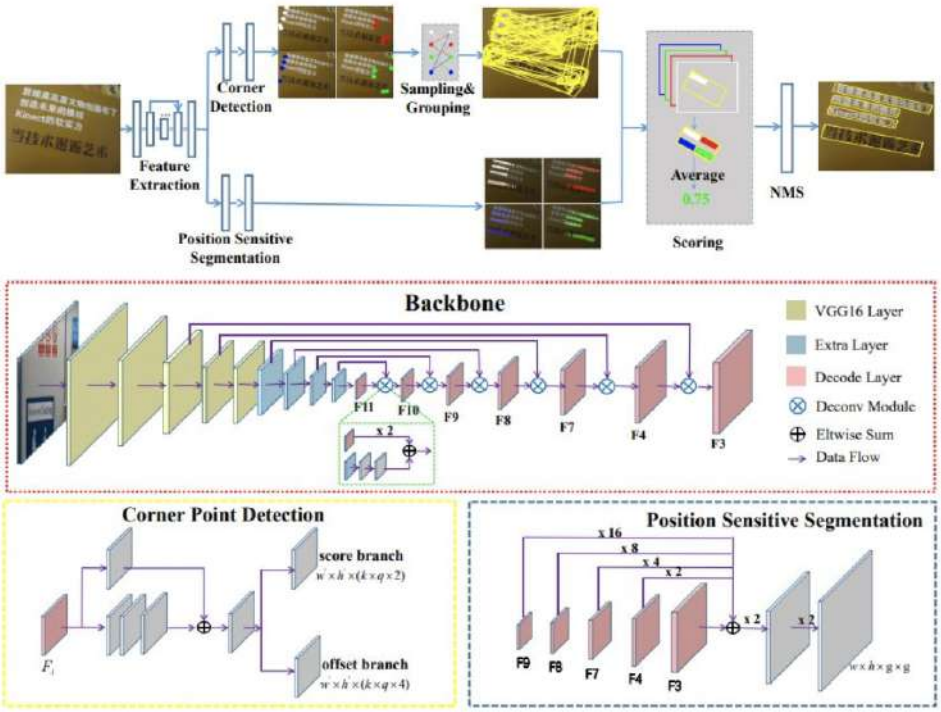
Ryosuke Araki

Multi-Oriented Scene Text Detection via Corner Localization and Region Segmentation

Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan and Xiang Bai CVPR2018 982

概要

コーナー検出とセグメンテーションを用いた高速かつ高精度なテキスト検出手法．テキスト検出時，ボックスのコーナー点を局所化し，テキスト領域を相対位置でセグメンテーションする．画像を入力すると，DSSDベースのNWで特徴抽出をし，コーナー点検出とコーナー位置に基づくセグメンテーションを出力する．コーナー点はサンプリングおよびグループ化され複数の候補ボックスとなる．セグメンテーション結果とあわせてスコア付けしてNMSする．長いテキストを自然に検出でき，複雑な後処理をする必要もない．



新規性・結果・なぜ通ったか？

Deepベースのテキスト検出は，テキストを物体の一種として扱いb-boxの回帰を行うか，テキスト部分を直接抽出する手法である．前者はアスペクト比によっては検出できず，後者は複雑な後処理を必要とする．本手法はその2つを組み合わせて，両者の欠点を補う．SynthText，ICDAR2015，2013，MSRA-TD500，MLTおよびCOCO-Textのデータセットで評価して，ほとんどがSOTAを達成した．とくに，ICDAR2015では84.3%（F-measure），MSRA-TD500では81.5%

コメント・リンク集

非常にシンプルながらも高精度なテキスト検出． DSSDのデコーダ部分の特徴マップからセグメンテーションを行う最近よくある手法をテキスト検出に応用している．

- 論文
- arXiv

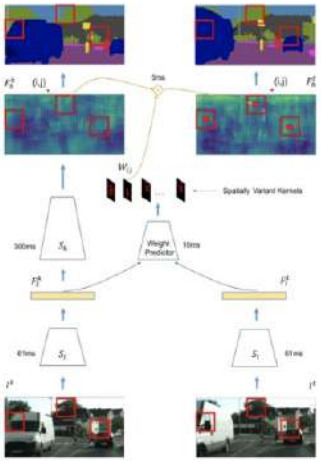
Munetaka Minoguchi

Low-Latency Video Semantic Segmentation

Yule Li, et al. 1804.00389

概要

動画によるセマンティックセグメンテーションにおいて、精度を向上させつつ、処理速度を上げる手法の提案。2つのコンポーネントを組み込んだフレームワークで構成している。1つ目は、時間変化に伴って空間的な畳み込み処理を変化させ、特徴を適応させる特徴伝播モジュール。2つ目は、精度予測に基づいて、計算を動的に割り当てるスケジューラ。



新規性

動画のセマンティックセグメンテーションには、高スループットやコスト、低遅延などの問題があり、 自律運転などにおいて重要となる。時間的変化に適応させた処理によって精度向上、処理速度向上を図る。

結果・リンク集

CityscapesとCamVidにおいて、最新の手法と競合する精度で、遅延を360msから119msに抑えられる結果に。

- 論文

Yoshihiro Fukuhara

VirtualHome: Simulating Household Activities via Programs

Xavier Puig et al. CVPR 2018

概要

家の中の環境をシミュレーションするための仮想環境 VirtualHome を作成した. また, 家の中で典型的に起こる様々な行動を自然言語とプログラムの形式で表現し, それらを仮想環境上でシミュレーションした動画を組みにした VirtualHome Activity Dataset を公開した. 加えて, LSTM を用いて動画やテキストからプログラム形式の表現を生成する手法を提案した.

新規性・結果・なぜ通ったか？

- VirtualHome には様々な種類の間取りや物体（平均357個）があり, Agent も複数の種類が用意されている
- dataset では家の中で行われる様々な行動に対して, 名前と自然言語形式での行動の説明と行動をプログラムの形式が与えられている
- VirtualHome 上でプログラムをシミュレーションすることで作成された動画には, Agent の姿勢やフロー, 物体のクラスなど様々な情報が与えられている
- LSTM を用いた encoder-decoder 型のネットワークに強化学習を適用し, 動画やテキストからプログラム形式の表現を生成する手法を提案

コメント・リンク集

- [論文] VirtualHome: Simulating Household Activities via Programs
- [Project page] VirtualHome: Simulating Household Activities via Programs

[#13]

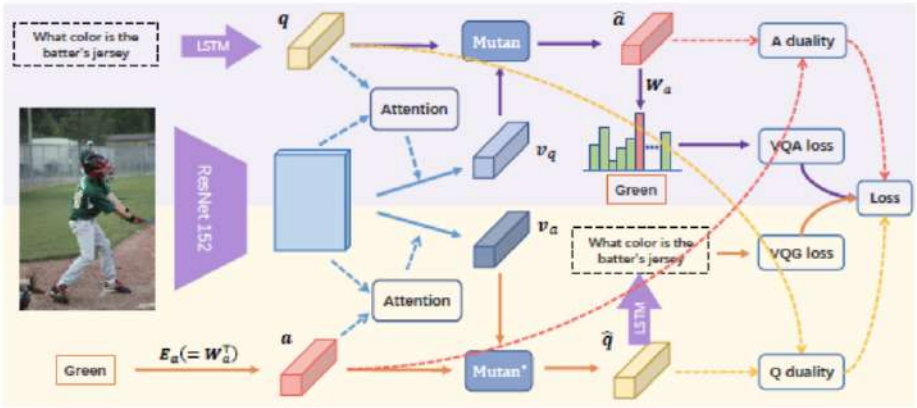
Shintaro Yamamoto

Visual Question Generation as Dual Task of Visual Question Answering

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang and Xiaogang Wang CVPR2018

概要

画像に関する質問に答えるVisual Question Answering(VQA)と与えられた答えになる質問を作るVisual Question Generation(VQG)を同時に扱うInvertible Question Answering Network(iQAN)を提案した。質問が与えられている場合は答えを、答えが与えられている場合は質問を推定することで学習をする。 その際、2つのタスクを独立した問題ではなく逆問題であると考え、質問と答え及びそれぞれを表現する特徴量間の変換に使用する重みを共有する。



新規性・結果・なぜ通ったか？

VQAに関しては、従来手法と比べて精度を向上することが可能となった。また、VQGによって生成した質問と答えのペアをVQAの学習に使用すると精度が向上することが分かり、VQGによってデータ数を増やすことが可能であると結論付けた。

コメント・リンク集

- 著者ホームページ

[#14]

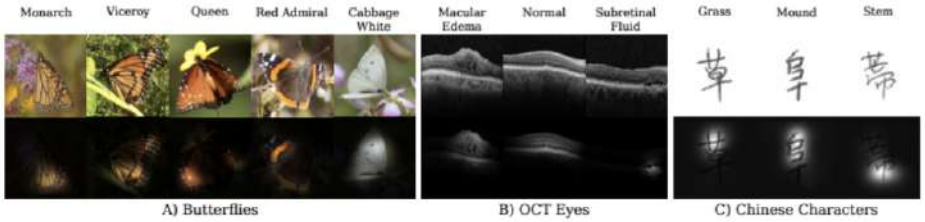
Shintaro Yamamoto

Teaching Categories to Human Learners with Visual Explanations

Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona and Yisong Yue CVPR2018

概要

画像に写っているもののカテゴリをコンピュータが人間に教えるためのシステムEXPLAINを提案。カテゴリを分類する上でどこに注目すればいいのか(例：蝶の種類を見分けるにはどこに注目すれば良いか）を提示することで人間がカテゴリを学習することを支援する。



新規性・結果・なぜ通ったか？

従来の手法ではカテゴリを表すラベルを提示するのみであったが、重要領域を提示することでより効率的に人間が学習することを可能とした。ユーザースタディにより人に学習してもらった内容に関するテストをしたところ、EXPLAINの方が短い時間で高い正答率を出すという結果を得られた。

コメント・リンク集

- link1

[#15]

Shintaro Yamamoto

Face Aging with Identity-Preserved Conditional Generative Adversarial Networks

Zongwei Wang, Xu Tang, Weixin Luo and Shenghua Gao CVPR2018

概要

人間の年齢変化顔を合成するIdentity-Preserved Conditional Generative Adversarial Networks (IPCGANs)を提案。合成画像が満たすべき特徴を、(1)目的の年齢に近づいている(2)変化前の人物と同一人物か(3)リアルな画像かの3つとした。 (1)(2)については、Generatorによって生成した画像を年齢推定及び同一人物性を評価するネットワークによって評価する。 (3)はDiscriminatorにリアルかどうかを判定させることで最適化を行う。

新規性・結果・なぜ通ったか？

ユーザースタディにより、Image Quality, Age Classification, Face Verificationの3つの観点を評価し、DNNベースの手法と比較してFace VerificationとImage Qualityの2つの観点で高い評価を得た。VGG-faceによりinception scoreを求め、比較対象の手法より高いスコアを得た。 また、計算時間についても劇的に良化した。

コメント・リンク集

- 論文

[#16]

Shintaro Yamamoto

Emotional Attention: A Study of Image Sentiment and Visual Attention

Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli and Qi Zhao CVPR2018

概要

画像に潜んでいる感情と注目を集める領域の関連を調査した。アイトラッキングのデータと、画像中に写っている感情に関連する物体(笑顔など)をアノテーションしたEMOtional attention dataset(EMOd)を構築した。 また、画像中の注目領域を抽出するDNNモデルであるCASNetを提案した。

新規性・結果・なぜ通ったか？

EMOdを用いて分析した結果、感情に関連する物体の方が人々の視線を集めることが判明した。その中でも、人間が関連する（笑顔など）場合がより視線を集めることが分かった。 従来のSaliencyを求める手法よりもCASNetの方が多くの指標で高いスコアを獲得した。また、感情に関連する物体の方がより注目を集めるという結果を出力したことからEMOdの分析結果を反映していることを確認した。

コメント・リンク集

- プロジェクトページ

[#17]

Shintaro Yamamoto

Categorizing Concepts with Basic Level for Vision-to-Language

Hanzhang Wang, Hanli Wang and Kaisheng Xu CVPR2018

概要

Vision and Languageのタスクに、Cognition分野で提唱されているbasic levelという概念を基にしたBasic Concept(BaC)を導入した。basic levelとは人間が幼少期に行う抽象化であり、本研究では物体のクラスを類似したもの同士を1つにまとめる。 始めに、MSCOCOのキャプションとImageNetのクラスをマッチングすることで、Salient Concept(SaC)というBaCに候補を決定する。 続いて、物体のクラス分類におけるConfusion Matrixを求め、混同されるクラス同士を1つにまとめることでBaCを決定する。

新規性・結果・なぜ通ったか？

Vision and Languageのタスクとして、Image CaptioningとVQAによって検証を行った。Image Captioningについては、ベースラインと比較してほとんどの指標において精度が向上し、向上しなかった指標についてもベースラインと大差ない数値を記録した。 VQAについては、ObjectとLocationについて精度の向上を確認した。

コメント・リンク集

- 論文

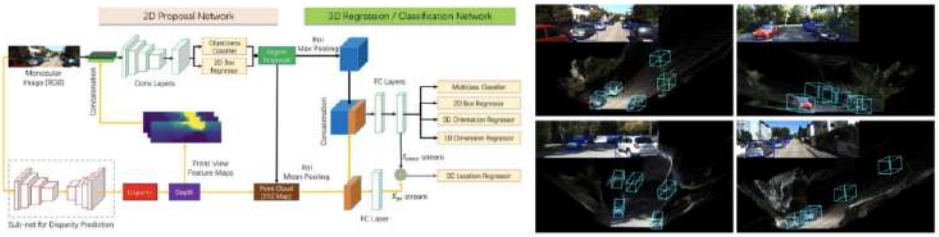
Yoshihiro Fukuhara

Multi-Level Fusion Based 3D Object Detection From Monocular Images

Bin Xu et al. CVPR 2018

概要

一枚のRGB画像から３次元物体認識を行う研究. region-based な２次元の物体検出器を３次元に拡張する一般的なフレームワークを提案し, end-to-end のネットワークで２次元と３次元の物体位置と物体のクラスを同時に推定することが可能. KITTI dataset を用いた評価実験では state-of-the-art の結果を達成した.



新規性・結果・なぜ通ったか？

- end-to-end のネットワークで単一のRGB画像から物体のクラスと２次元, ３次元の物体位置, ３次元の物体の方向などを同時に推定
- RGB画像に MonoDepth を用いて推定した Depth 画像を連結したものを CNN に入力し, Faster-RCNN と同様の方法で Region Proposal を生成
- また, Depth 画像から Point Cloud (XYZ Map)を推定
- 上記の２つを連結したものを全結合層に通して, 物体位置と物体のクラスの推定を行う
- KITTI dataset を用いた評価実験では　Mono3D, 3DOP, Deep3DBox などと比較して優位な結果を達成した

コメント・リンク集

- [論文] Multi-Level Fusion Based 3D Object Detection From Monocular Images

[#1]

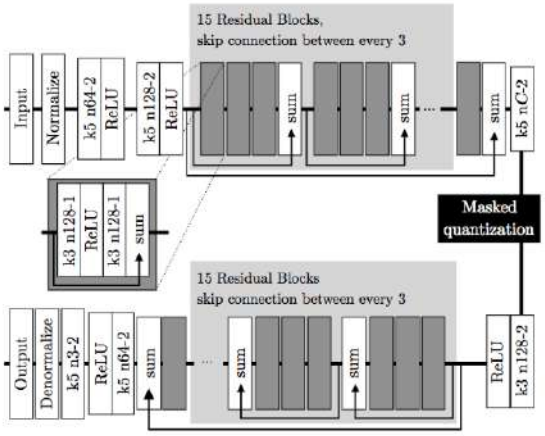
Hirokatsu Kataoka

Conditional Probability Models for Deep Image Compression

Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, Luc Van Gool CVPR 2018

概要

画像復元の問題は復元エラー（distortion）とエントロピー（rate）とのトレードオフであるが、本論文ではこのトレードオフをできる限り解消し、画像圧縮を行うAutoEncoderを提案する。著者らはコンテキストモデルから直接的に潜在表現のエントロピーを復元するモデルを考案して同問題に取り組んだ。AutoEncoderには条件付き確率モデルを学習した3D-CNNを適用。実験ではSSIMを用いて従来の畳み込みによるAutoEncoderモデルよりも良好な精度を実現した。



新規性・結果・なぜ通ったか？

3D-CNNにより条件付き学率モデルを学習したAutoEncoderモデルを考案したことが新規性であり、JPEG(2000)などよりも良い圧縮法であることを示し、Rippel&Bourdevらのモデルと同等レベルの精度を達成した。

コメント・リンク集

画像圧縮、超解像の違いがいまいちよくわからなくなってきた。評価方法の違い？

- 論文
- 著者

Hirokatsu Kataoka

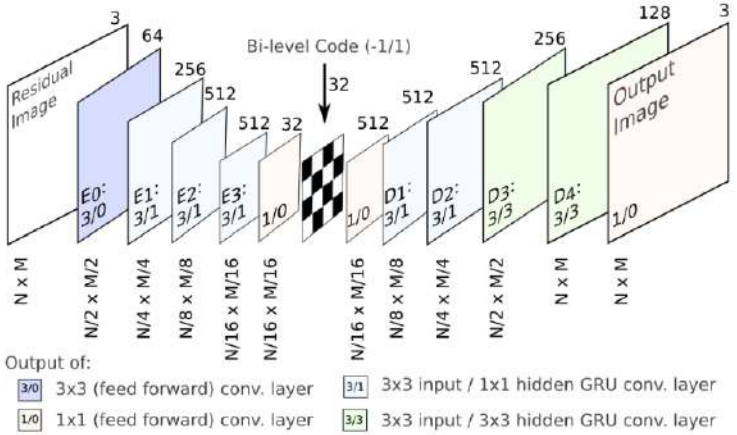
Improved Lossy Image Compression With Priming and Spatially Adaptive Bit Rates for Recurrent Networks

Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, George Toderici

CVPR 2018

概要

Recurrent/Convolutional Neural Networks（RNN/CNN）を用いた非可逆画像圧縮の手法を提案し、BPG(4:2:0), WebP, JPEG2000, JPEGよりも性能のよいものを提案した。3つの改善、(1)ニューラルネットにより空間的分散を効果的に捉えて情報量の劣化を防ぐ、(2)エントロピーコーディングの上に空間適応的ビット配置アルゴリズムを適用して効率的な画像圧縮とする、(3)SSIMによりピクセルごとの損失を計算して最適化することで圧縮数値を改善する、を加えて圧縮方法を提案。KodakやTecnickのカメラを用いてコーデックの評価を行った。



新規性・結果・なぜ通ったか？

従来の圧縮方法であるBPG(4:2:0), WebP, JPEG2000, JPEGなどよりも効率の良い圧縮方法を提案した。また、手法的にもCNN/RNNを応用し、さらに後処理として画質を改善するSpatially Adaptive Bit Rate (SABR)を提案したことが評価された。

コメント・リンク集

（数十年前からある問題という意味で）過去の問題と現在の手法が合わさって新規性を出している論文。

- 論文

[#3]

Kazuki Inoue

Deep Density Clustering of Unconstrained Faces

Wei-An Lin, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa CVPR 2018 Poster

概要

unconstrainedな顔に対してクラスタリングを行うDeep Density Clustering(DDC)を提案。顔画像をDNNによって単位超級面空間に射影する。続いて、各サンプル2点の類似度を測定する際に、 その2点の近傍に位置するサンプルを考慮することでクラスタの密度を推定することが可能となるため、これに基づいてクラスタリングを行う。

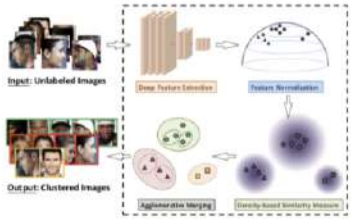


Figure 1: We introduce Deep Density Clustering (DDC) for unconstrained face images. DDC is a density-based clustering algorithm, which exploits the local structure of deep features for improved similarity measure.

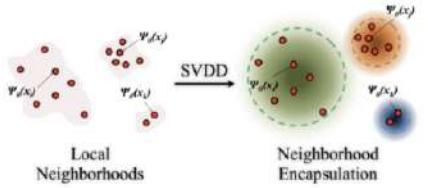


Figure 3: Neighborhood encapsulation. (left) Pink regions are the local neighborhoods of the points x_i, x_j, and x_k in feature space. (right) Encapsulations are learned by solving (3). The encapsulation is density-aware. In the figure, regions closer to the centers of the spheres have higher density.

新規性・結果・なぜ通ったか？

- YTF, LFW, IJB-Bデータセットを使用して評価。それぞれのデータセットには同一人物の画像が複数枚もつ。
- 評価指標はBCubed precision、Bcubed F-measure、NMIで評価。
- 提案手法と同等の精度を持つ既存手法のJULE、DEPICTはクラスタ数を指定する必要があるが、提案手法ではクラスタ数を指定する必要がない。
- クラスタリングの際の閾値の変更に対して、既存手法に比べてクラスタ数の変動が小さい。

コメント・リンク集

- 論文
- Supplementary material

[#4]

Kazuki Inoue

Pose-Guided Photorealistic Face Rotation

Yibo Hu, Xiang Wu, Bing Yu, Ran He, Zhenan Sun CVPR 2018 Poster

概要

入力顔画像に対して任意の画像を生成するネットワークを提案。顔向きのコンディションとしてランドマークのヒートマップを与え、U-Netによって画像を生成し、2つのdiscriminatorを用いることで画像を生成。 1つ目のdiscriminatorは入力画像をコンディションとして生成画像or正解画像を識別し、 2つ目のdiscriminatorはランドマークのヒートマップをコンディションとして生成画像or正解画像を識別する。 また人物IDを保存するためにLight CNNによる特徴量によるロスをとる。

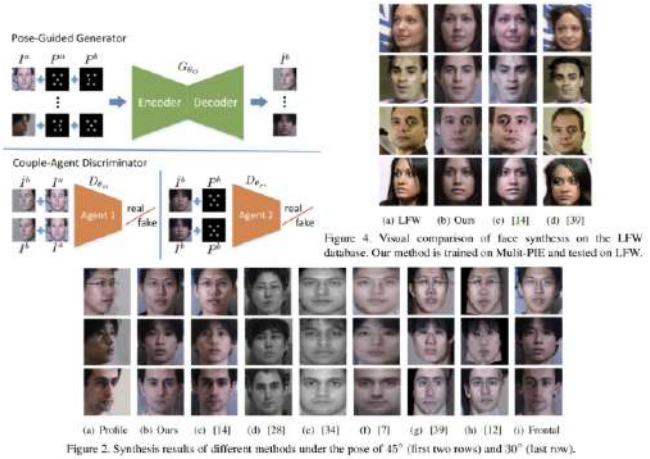


Figure 4. Visual comparison of face synthesis on the LFW database. Our method is trained on Multi-PIE and tested on LFW.

Figure 2. Synthesis results of different methods under the pose of 45° (first two rows) and 30° (last row).

新規性・結果・なぜ通ったか？

- ランドマークのヒートマップ、2つのdiscriminator、IDを保存するロスを用いて入力顔画像を任意の向きに回転させた画像を生成。
- 337IDそれぞれに対して20の照明環境と15種類の顔向きをもつMulti-PIEで検証。
- トレーニングには使用していないLFWで画像を生成したところ、既存手法による画像よりも見た目の良い画像が得られた。
- face verification、face recognitionにおいてSoTAを達成。
- ablation studyの結果、IDのロスがface recognitionに最も影響が高いことを確認。

コメント・リンク集

- 既存手法のように顔向きの角度を使うのではなくヒートマップを与えることでU-netの学習がしやすい、という上手い方法。
- IDのロスに使用する特徴量が最後のFC層に加えてプーリング層からも取得されておりIDについてはMS-Celeb-1Mでプリトレインした後Multi-PIEへとファインチューニングしているなど、かなり微調整を感じる論文。
- 論文
- Supplementary material

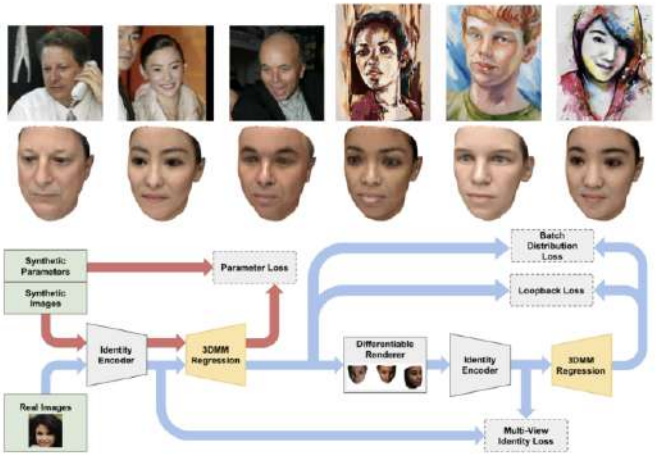
Kazuki Inoue

Unsupervised Training for 3D Morphable Model Regression

Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, William T. Freeman CVPR 2018 Poster

概要

それぞれ単独の実画像データセットと3D Morphable Model(3DMM)データセットを使用し、画像から3DMMを生成する手法を提案。トレーニングには実画像データセットVGG-Face、3DMMデータセットBasel Face 3DMMを使用。 IDが保たれることを念頭にネットワークを構築。Batch Distribution Lossでは、 Basel Face 3DMMのパラメタ分布が平均０、標準偏差1のガウス分布であるため、 実画像によって生成される3DMMのシェイプ、テクスチャパラメタがどちらも平均0、標準偏差1となるようにロスをとる。 Loopback Lossは画像/生成された3DMMのdecoderによる特徴量の差分を取り、よりリアルな3DMMかつ、 より現実的な3DMMパラメタを得ることを目的としている。



新規性・結果・なぜ通ったか？

- 画像、3DMMの対応がないデータセットを用いて、教師なしで画像から3DMMを生成する手法を提案。
- Batch Distribution Loss、Loopback Loss、Multi-view Identity Lossを学習することで教師なしであることを緩和している。
- MICC Florence 3D Faceデータセットで検証し、Mean error、Faceクラスタリング、Earth mover's distanceによる実画像と生成3DMMの顔類似度のそれぞれにおいてSoTA。

コメント・リンク集

- Basel Face 3DMMのパラメタ分布が平均０、標準偏差1のガウス分布という仮定はどこから来ている？
- 論文

Kazuki Inoue

Aligning Infinite-Dimensional Covariance Matrices in Reproducing Kernel Hilbert Spaces for Domain Adaptation

Zhen Zhang, Mianzhi Wang, Yan Huang, Arye Nehorai CVPR 2018 Poster

概要

ソースドメイン(SD)とターゲットドメイン(TD)のそれぞれのreproducing kernel Hilbert space(RKHS)における共分散を最適化することでdomain adaptation(DA)を行う手法。 既存のカーネルベースのDAはSDとTDのRKHS上の統計的分布の類似度に大きく依存することに着目。 共分散を最適化する方法としてkernel whitening-coloring map(KWC)とkernel optimal transport map(KOT)があり、これをRKHS上で計算で可能なように式変形を行うことでDAを行う。

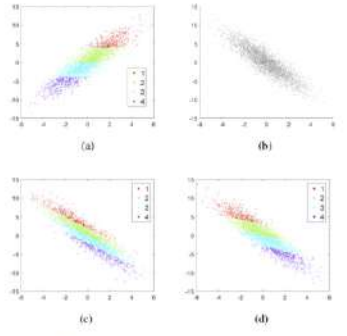


Figure 1: (a) The source samples $\boldsymbol{X}_s = [\vec{x}_1, \vec{x}_1, ..., \vec{x}_N]$. Different colors represent different classes. (b) The target domain. (c) The results of transforming the source samples by the whitening-coloring map (*i.e.*, $\vec{x}_i \to \boldsymbol{T}_{\mathrm{WC}}(\vec{x}_i)$). (d) The results of transforming the source samples by the optimal transport map (*i.e.*, $\vec{x}_i \to \boldsymbol{T}_{\mathrm{OT}}(\vec{x}_i)$).

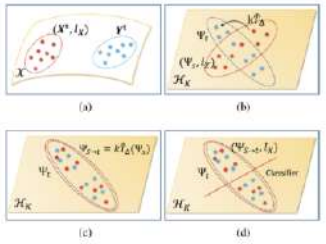


Figure 2: (a) The labeled dataset $(\boldsymbol{X}^s, \vec{l}_X)$ in the source domain and the unlabeled dataset $\boldsymbol{Y}^t$ in the target domain. Dots and stars represent different classes. (b) The centered source data, Ψ_s, and centered target data, Ψ_t, in RKHS. (c) Transform the source samples by the map $\mathrm{k}\tilde{T}_{\triangle}$ ($\triangle$ = WC or OT), and the resultant data are $\Psi_{s\to t} = \mathrm{k}\tilde{T}_{\triangle}(\Psi_s)$. (d) Train a classifier with $\Psi_{s\to t}$.

新規性・結果・なぜ通ったか？

- SDとTDのRKHS上の共分散を最適化することでDAを行う。
- 複数のDAのベンチマークデータセットにおいてKWC、KOTのいずれかがSoTAを達成。
- SoTAと比較して実行時間が短く、KWCは4分の１、KOTは10分の1程度。
- Out-of-Sampleによる推定においてもSoTAを達成。

コメント・リンク集

- 248パターンのDAを検証しており、本論文に載っていたのは34パターン
- 論文
- Supplementary material

[#7]

Kazuki Inoue

Cross-Dataset Adaptation for Visual Question Answering

Wei-Lun Chao, Hexiang Hu, Fei Sha CVPR 2018 Poster

概要

VQAのデータセットにおけるバイアスを調査した上で、VQAにおけるdomain adaptation(DA)を提案。提案手法では選択肢の中から解答を選択するVQAを扱う。VQAデータセットは画像、質問、解答選択肢＝正解＋誤答の要素からなる。 それぞれの要素を組み合わせた入力を用いて、その入力がどのデータセットに所属しているのかを調査した結果、 画像はほぼ無相関であることがわかり、質問と解答によってデータセット間にバイアスが生じていることを確認。 この結果に基づき、以下のようにDAを提案。ターゲットドメイン(TD)に質問/解答選択肢のみがある場合、 ソースドメイン(SD)の質問/正解(誤答は任意性があるため使用しない)の特徴量が持つ分布とTDの質問のDNNによる 特徴量が持つ分布のJensen-shannon Divergence(JSD)が小さくなるように学習。TDが質問と正解(＋誤答)を持つ場合、 SDが持つ質問・正解の特徴量分布とTDの質問・正解のDNNによる特徴料が持つJSDが小さくなるように学習。 さらにSDで事前学習を行った質問-正解識別をTDでfine-tuningを行う。

Figure 1. An illustration of the dataset bias in visual question answering. Given the same image, Visual QA datasets like VQA [4] (right) and Visual7W [50] (left) provide different styles of questions, correct answers (red), and candidate answer sets, each can contributes to the bias to prevent cross-dataset generalization.

新規性・結果・なぜ通ったか？

- 事前実験より与える情報によって、入力データがどちらのデータセットに所属しているかの識別率の変化を確認。画像、質問、正解解答、解答群(正解+不正解)を与え、与える要素を増やすほど識別率が高くなった。 この結果から、データセットによってバイアスがあることを確認。
- 質問に対する正答率を複数のデータセットにおいて既存手法であるADDA、CORALと比較した結果SoTAを達成。TDが解答選択肢のみ、質問と正解を持つ場合において高い精度を達成。

コメント・リンク集

- TDの正解、誤答のみを使用し質問を使用せずにDAを行った方が高い状況がいくつも確認できる。これはつまり質問と解答の相関がすでにSDで学習できており、SDの質問がノイズになってしまっているとを示唆している。
- VQAをDAしてみた、という実験的な論文であり比較している手法もDAのベンチマークの手法なので、まだまだ新規性を出すことができそう。
- 論文
- Supplementary material

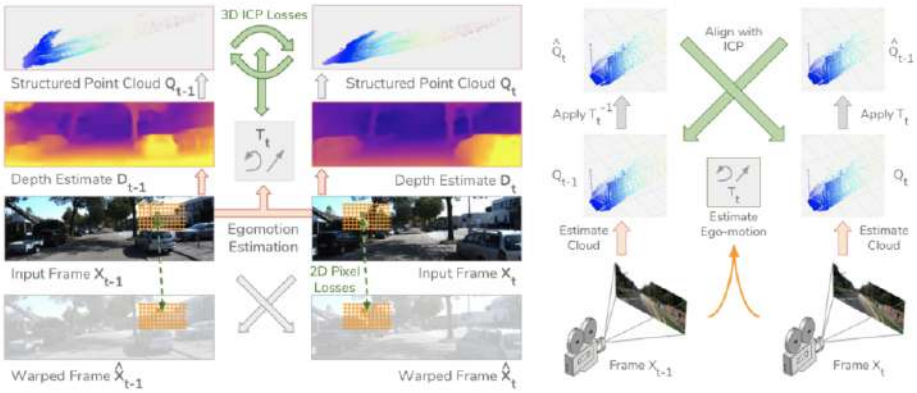
Yoshihiro Fukuhara

Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints

Reza Mahjourian et al. CVPR 2018

概要

教師なし学習で単眼の動画から Depth と Ego-Motion の推定を行う研究. 連続するフレーム間における 3D Geometry の一貫性を教師信号の代わりに利用して学習を行う.



新規性・結果・なぜ通ったか？

- 連続するフレーム間における 3D Geometry の一貫性を用いることで, 教師なし学習で単眼の動画から Depth と Ego-Motion の推定を行うことを可能とした
- 連続するフレームから推定された Point Cloud に対して Iterative Closest Point (ICP) を計算し, その Residual と Transform の大きさを 3D Loss として課す
- 3D Loss に加えて推定された Depth の滑らかさと, 推定結果を用いて復元した画像の誤差 (2種類) も Loss として課す
- KITTI dataset と mobile phone カメラで撮影した動画を用いて行った評価実験では　Trajectory と Depth の両方において先行研究よりも優位な結果を達成した

コメント・リンク集

- [論文] Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints

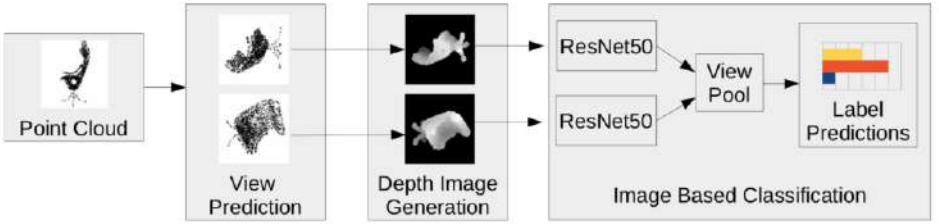
Yoshihiro Fukuhara

A Network Architecture for Point Cloud Classification via Automatic Depth Images Generation

RiccardoRoveri et al. CVPR 2018

概要

Point Cloud データのクラス分類についての研究. 順序不定の 3D Point Cloud データを 2D Depth 画像に変換し, ResNet でクラス分類を行う. 評価実験では PointNet より優位な結果となった.



新規性・結果・なぜ通ったか？

- Network は３つのモジュールで構成されており, joint training が可能
- １つ目のモジュールは PointNet を用いて PointCloud から有用な view direction を推定する
- ２つ目のモジュールは Gausiaan Interporation （Roveri＋18 の拡張版）によって推定された view direction からの Depth 画像を生成する
- ３つ目のモジュールは ResNet50 を用いて Depth 画像から Image Based Classification を行う
- ModelNet40 benchmark を用いて行った shape のクラス分類の評価実験では instance-based accuracy と class average accuracy の両方で PointNet よりも優位な結果となった

コメント・リンク集

- [論文] A Network Architecture for Point Cloud Classification via Automatic Depth Images Generation
- 3D の問題を既によく研究されている 2D 画像のクラス分類へと帰着させることで, 既存の強力な手法を用いる戦略

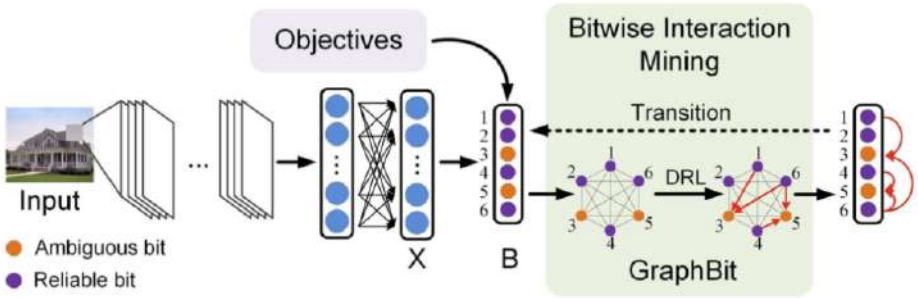
Yoshihiro Fukuhara

GraphBit: Bitwise Interaction Mining via Deep Reinforcement Learning

Yueqi Duan et al. CVPR 2018

概要

Deep binary descriptor においてバイナリを生成する際に0と1の境界に位置する曖昧なビット (ambiguous bit) の問題に取り組んだ研究. 強化学習によって学習したビット間の implicit な関係性を付加することで曖昧性を緩和する GraphBit を提案.



新規性・結果・なぜ通ったか？

- Binary descriptor における曖昧なビット (ambiguous bit) の問題を緩和するためにビット間の関係性を付加した GraphBit を提案
- CNNからの出力された正規化された特徴量（binary descriptor）に対して Grpah 構造を付加する
- ビット間の相互関係をマイニングする過程をマルコフ過程として定式化し, 強化学習（Policy Gradient）で学習
- State は現在の Graph の構造
- Atction は GraphBit に新しいエッジを 1 つ追加するか, 既存のエッジを 1 つ削除
- Reward は t ステップと t+1 ステップにおけるロス関数の減少度合いから計算
- CIFAR-10, Brown, HPatches dataset を用いた評価実験では mean average precision (mAP) の評価尺度でそれぞれ平均 9.64%, 8.84%, 3.22% の精度の向上を達成した

コメント・リンク集

- [論文] GraphBit: Bitwise Interaction Mining via Deep Reinforcement Learning

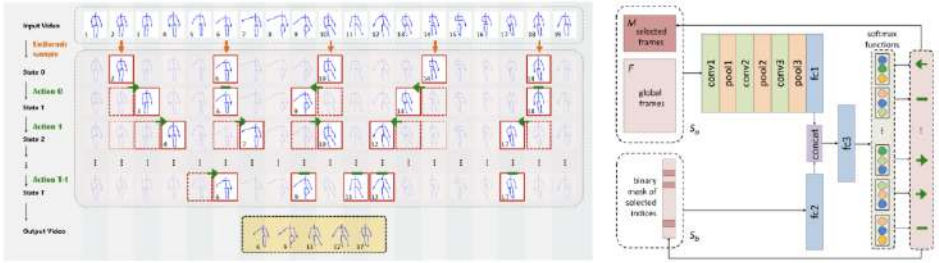
Yoshihiro Fukuhara

Deep Progressive Reinforcement Learning for Skeleton-based Action Recognition

Yansong Tan et al. CVPR 2018

概要

Skeleton-based action recognition の研究. 強化学習によって与えられた動画から最適な keyframe の組を選択する frame distillation network (FDNet) と graph-based convolution によって keyframe の skeleton 情報から行動認識を行う Graph-based CNN (GCNN) を提案.



新規性・結果・なぜ通ったか？

- 与えられた動画のシークエンスから最適な keyframe の組を選択する過程をマルコフ過程として定式化し, 強化学習 (policy gradient) を適用した
- State として Skeleton 動画全体と現在選択されてる keyframe の組の情報を使用
- Action は各 keyframe を1フレーム前後にずらすか, そのままかの３つ
- Reward は学習済みの GCNN を用いて計算
- また, keyframe から行動認識を行う際は gggraph-based convolution を用いることによって人間の関節の依存関係を考慮している
- NTU, SYSU, UT dataset を用いて評価実験では state-of-the-art とほぼ同等か, 優位な結果を示した

コメント・リンク集

- [論文] Deep Progressive Reinforcement Learning for Skeleton-based Action Recognition

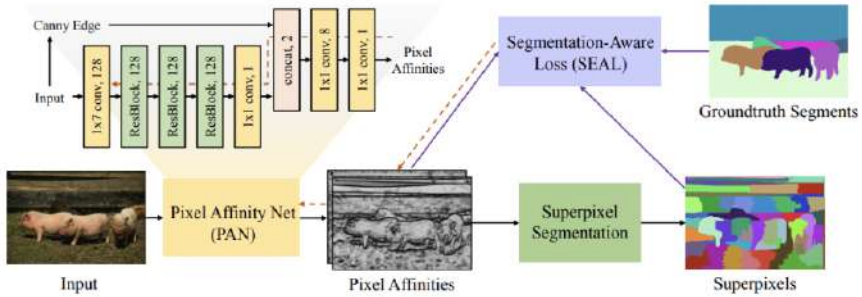
Hiroaki Aizawa

Learning Superpixels with Segmentation-Aware Affinity Loss

Wei-Chih Tu, Ming-Yu Liu, Varun Jampani, Deqing Sun, Shao-Yi Chien, Ming-Hsuan Yang, Jan Kautz CVPR2018

概要

superpixel segmentationのためにピクセルの類似性(pixel affinities)を学習するdeep learningベースの手法を提案。pixel affinitiesが同一物体に属する2つの隣接画素の尤度を測る。これまで、groundtruthがないこと、superpixelsのインデックスが交換可能であること、superpixelsの手法は微分不可であることからdeep learningベースのsuperpixelアルゴリズムは試みられていなかった。論文では、segmentation誤差から類似性を学習するsegmentation-aware loss(SEAL)と、pixel affinitiesを出力するPixel Affinity Net(PAN)を提案し、superpixelsとdeep learningを統合する。既存の手法より物体境界を保持したままsuperpixelsを計算することが可能になった。



新規性・結果・なぜ通ったか？

superpixels + deep learningが新しい。実験では単純なpretrained modelによる特徴量や、edge検出によるsuperpixelsとの統合はうまくいかないことを示している。手法に関しては、superpixelsを直接出力するのではなく、pixel affinitiesを計算、graph-basedのアルゴリズム(ERS)を経由し出力、そしてSEALを計算する。これにより、pixel affinitiesを出力するPANへ誤差を逆伝播することができる。

コメント・リンク集

より効果的に細部の情報をsuperpixelsとして保持することができるため、semantic segmentationの改善や計算量の削減につながるだろう。

- paper
- proj_page

Ryota Suzuki

Generating Synthetic X-ray Images of a Person from the Surface Geometry

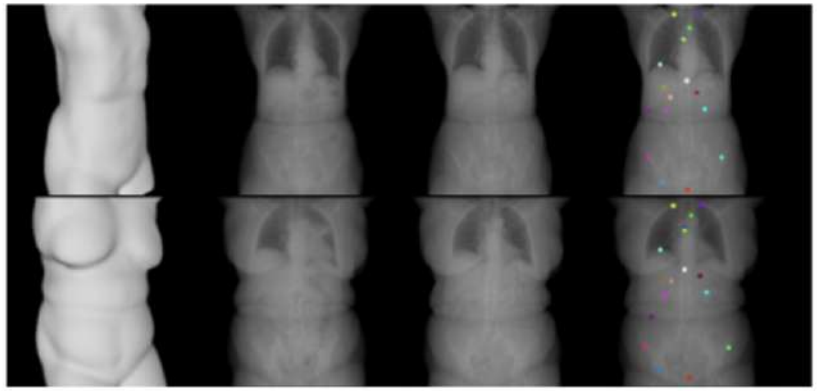
Brian Teixeira, Vivek Singh, Terrence Chen, Kai Ma, Birgi Tamersoy, Yifan Wu, Elena Balashova and Dorin Comaniciu CVPR2018

概要

人間の三次元輪郭形状から，見えない体の内側を解析してしまおうという話．本論文では，X線画像を生成する． さらに，X線画像はパラメタライズしておくことで，体のキーポイントの調節によるマニピュレーションも可能．

構造的には，2つのネットワークからなる．(1)部分画像といくつかのパラメータから，画像全体を生成するように学習， (2)全体画像が得られるような(1)のパラメータの推定． これら2つのネットワークを，一貫性が出てくるように反復的に学習させる．

生成した画像を使ってみて，画像補間に使ってみた．



元　　GT　　結果　　キーポイント

新規性・結果・なぜ通ったか？

体表面を計測しておくなどして，体表面形状のデータがあれば，X線画像をある程度任意に生成できる．逆に，体表面形状をいじることでそれに対応したX線画像も作れる． 学習データとして活用することができる可能性がある．

構造はGAN風だが，いい感じに変形している感じがウケているかもしれない．

コメント・リンク集

この時点での一番の貢献は，それっぽいX線画像が自動生成できる事だろう．SMPLと組み合わせていろいろやることを想定しているだろうか．

- 論文

[#14]

Kazuki Inoue

Fully Convolutional Adaptation Networks for Semantic Segmentation

Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, Tao Mei CVPR 2018 Poster

概要

スタイル特徴量を用いて画像の見た目を変換するネットワークとドメイン間で不変な特徴量を得るネットワークを用いて、domain adaptationを行うことで教師無しでセマンティックセグメンテーションを行うFully Convolutional Adaptation Networks (FCAN)を提案。画像の見た目を変換するAppearance Adaptation Networks (AAN)では ホワイトノイズから画像を生成し、ソースドメインの特徴量マップ、ターゲットドメインのもつスタイル特徴量が小さくなるように学習を行うことで、画像をもう一方のドメインの見た目になるように変換する。 ドメイン間で不変な特徴量を得るRepresentation Adaptation Networks (RAN)ではsemantic classificationと、 それぞれのドメインにから得られた特徴量マップに対するadversarial lossと、 ASPPによって得られた特徴量マップに対してピクセルごとにadversarial lossを適用。 ドメインとして実画像とゲーム画像で検証している。

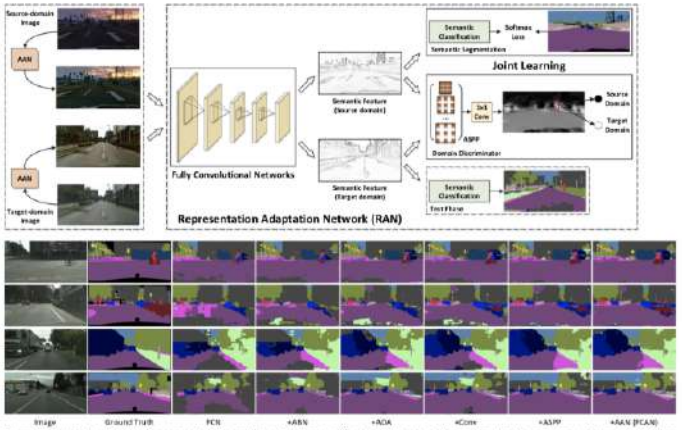


Figure 5. Examples of semantic segmentation results in Cityscapes. The original images, their ground truth and comparative segmentation results at different stages of FCAN are given.

新規性・結果・なぜ通ったか？

- style transferと同様の考え方でドメイン間の画像変換を行いsemantic classification、特徴量マップ、dilated convolutional layerから得られた特徴量マップに対する各ピクセルに対してadversarial lossをとることで教師無しでセマンティックセグメンテーションを行う。
- GTA5とcity spaceを用いて、セマンティックセグメンテーションの精度をstate-of-the-artと比較した結果、19クラスのうち17クラスで最も高い精度を達成。

コメント・リンク集

- 論文

[#15]

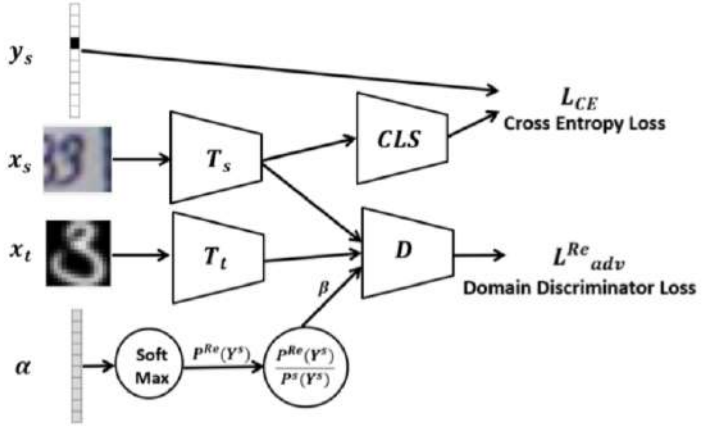
Kazuki Inoue

Re-weighted Adversarial Adaptation Network for Unsupervised Domain Adaptation

Qingchao Chen, Yang Liu, Zhaowen Wang, Ian Wassell, Kevin Chetty CVPR 2018 Poster

概要

Unsupervised Domain Adaptationを行うため、ドメイン間の特徴量分布を一致させるoptimal transportベースのEM distanceを導入し、ターゲットドメイン(T)のラベル分布をソースドメイン(S)のラベル分布に対してラベルごとに重み付けした分布で表現する手法を提案。 domain discriminatorをOTベースのEM distanceをロス関数とすることでドメイン間の特徴量分布を近づける。 一方でベイズの定理より、ドメイン間のラベルの事前分布と特徴量の事後分布は比例関係にありラベルは低次元かつ離散的であるので ドメイン間で類似度が高いと仮定し、Tにおけるラベルの事前分布をSのラベルの事前分布の重みを変更したもので表す。



新規性・結果・なぜ通ったか？

- ドメイン間で特徴量分布をOTベースのEM distanceの学習で、Tのラベル分布をSのラベル分布の重みを変更したもので表現することで、それぞれのdomain shiftを解消する手法を提案。
- 手書き文字データセットMNIST、USPS、SVHN、MINST-Mデータセット、19のラベルを持つ実画像、デプス画像のドメインを持つNYU-Dデータセットで検証。state-of-the-artと比較した結果、多くの状況で最も高い精度を達成。
- Sのラベル分布の重みの変更による有効性、ラベルごとの特徴量が分離できているかどうかも議論している。

コメント・リンク集

- 論文

Kazuki Inoue

Unsupervised Deep Generative Adversarial Hashing Network

Kamran Ghasedi Dizaji, Feng Zheng, Najmeh Sadoughi, Yanhua Yang, Cheng Deng, Heng Huang CVPR 2018 Poster

概要

教師無しで画像をバイナリに符号化するハッシュ関数であるHashGANを提案。ハッシュ関数が満たすべき条件は画像が変換されて同じハッシュ値を返すこと、異なる画像には異なるハッシュ値を与えることである。 既存の教師無しハッシュ関数は過学習のために精度がよくなかった。提案手法であるHashGANはgenerator、discriminator、 encoderからなる。学習はGAN loss、encoderによって生成されるハッシュ値のエントロピーが小さくなるように、 出現するハッシュ値が同じになるように、画像の変換によるハッシュ値が不変となるように、画像ごとのハッシュ値が固有となるように、 合成画像をエンコードした際のハッシュ値のL2ロス、実画像と合成画像を入力とした際のdiscriminatorの最後の層に対してfeature matchingを行う。またdiscriminatorはデータ固有の情報を識別し、encoderはデータ固有の情報を抽出しようとするため、 両者の目的が一致しているのでパラメタを共有して学習を行う。

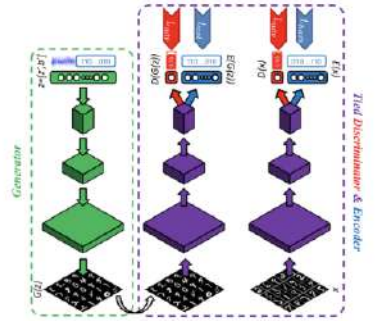


Figure 2: *HashGAN* architecture, including a generator (green), a discriminator (red) and an encoder (blue), where the last two share their parameters in several layers (red⊕blue=purple). The arrows on top represent the loss functions.

新規性・結果・なぜ通ったか？

- GAN、discriminatorとパラメタを共有しているencoder、ハッシュ関数が満たすべきロス関数を導入したHashGANを提案。
- image retrieval、image clusteringで手法の優位性を検討。image retrievalでは既存のunsupervised hash functionとの比較を行い、最も高い精度を達成。image clusteringではstate-of-the-artと同等の精度を達成。
- ablation testにより、特にadversarial loss, feture matching, L2ロス、画像変換によるハッシュの不変性の考慮の影響が大きいことがわかった。

コメント・リンク集

- 教師無し学習でもタスク特化の手法であり、ハッシュ関数の性質をよく考察した上でモデルを設計している。
- 論文

Kazuki Inoue

Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors

Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, Yaser Sheikh, CVPR 2018 Poster

概要

ランドマークのGT有り顔画像とラベルなし顔動画を用いて、現在フレームに対して直接推定されたランドマークと、トラッキングによって前フレームから推定されたランドマークの位置の誤差を学習することで顔画像に対してランドマークを推定する手法を提案。 人間によるランドマークのアノテーションは正確でないため、この誤差が学習や推定精度に影響を与えてしまう。 これに対して本論文ではランドマークの推定器に最適化によって計算されるオプティカルフローを教師情報として与える Supervision by Registration(SBR)を提案。ランドマーク位置を推定するCNNに対して、 Lukas-Kanade法によるトラッキング結果とランドマークの推定位置が同じになるように学習を行う。

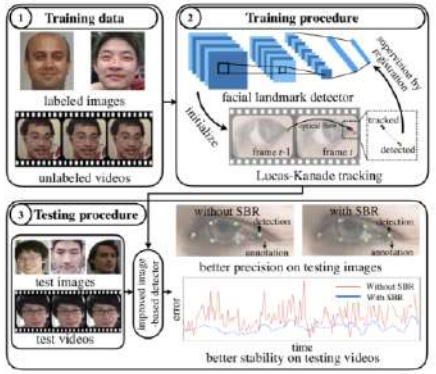


Figure 2. The **supervision-by-registration (SBR) framework** takes labeled images and unlabeled video as input to train an image-based facial landmark detector which is more precise on images/video and also more stable on video.

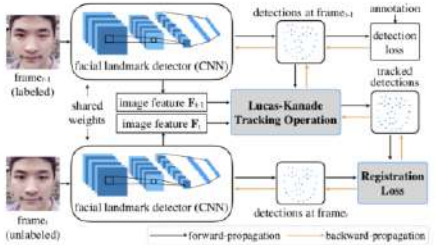


Figure 3. **The training procedure** of supervision-by-registration with two complementary losses. The detection loss utilizes appearance from a single image and label information to learn a better landmark detector. The registration loss uncovers temporal consistency by incorporating a Lucas-Kanade operation into the network. Gradients from the registration loss are back-propagated through the LK operation to the detector network, thus enforcing the predictions in neighboring frames to be consistent.

新規性・結果・なぜ通ったか？

- 人間のアノテーションよりも、より正確であるオプティカルフローを教師情報として使用することで顔画像に対するランドマークの推定手法を提案。
- 300-W、AFLWにおいてランドマーク推定手法であるCPMのアルゴリズムをSBRで学習させると、SBRを使用しない場合よりも精度が向上。
- 動画に対するランドマーク推定はstate-of-the-artに及ばなかった。ターゲットとなる人物をデータセットに含んでおくPersonalized Adaptation Modeling(PAM)を行うことで、state-of-the-artと同等の精度を達成。

コメント・リンク集

- 画像のランドマークを推定するために動画から得られるオプティカルフローを使用する、という発想の飛躍が面白い！最適化による正確な教師情報とCNNによる合わせ技。
- 論文

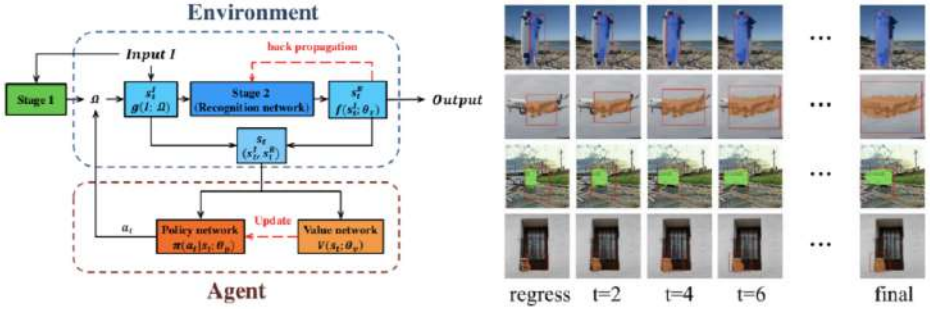
Yoshihiro Fukuhara

Environment Upgrade Reinforcement Learning for Non-differentiable Multi-stage Pipelines

Shuqin Xie et al. CVPR 2018

概要

微分不可能な multi-stage pipline において joint optimization を可能にする environment upgrade reinforcement learning (EU-RL) を提案. 2 段階の Instance segmentation と pose estimation のタスクで評価実験を行い, どちらも優位な結果を示した.



新規性・結果・なぜ通ったか？

- 微分不可能な multi-stage pipline の学習において問題であった上流への feedback が出来ないという点と end-to-end な最適化が出来ない点に取り組んだ研究
- 強化学習の agent が下流の出力を受けて上流の出力に変更を与える, environment upgrade reinforcement learning (EU-RL) を提案
- 強化学習の手法として actor-critic を Temporal Difference （TD) learning で学習
- State として 1 段階目（例えば物体認識）からの出力と 2 段階目からの出力（例えば semantic segmentation）を使用
- Action として 1 段階目からの出力結果を変更する操作の集合を使用（物体認識ならBounding Boxの位置の変更やスケールなど）
- Reward は 2 段目の出力の精度の向上度合いによって計算
- Instance segmentation と pose estimation のタスクで評価実験

コメント・リンク集

- [論文] Environment Upgrade Reinforcement Learning for Non-differentiable Multi-stage Pipelines
- 強化学習の応用先としても, アイデアとしても面白い. 今回の論文では 2 段階の pipeline についてのみ議論が行われていたが, 今後は 3 段以上の pipeline でも同様の議論が行われていく？

[#19]

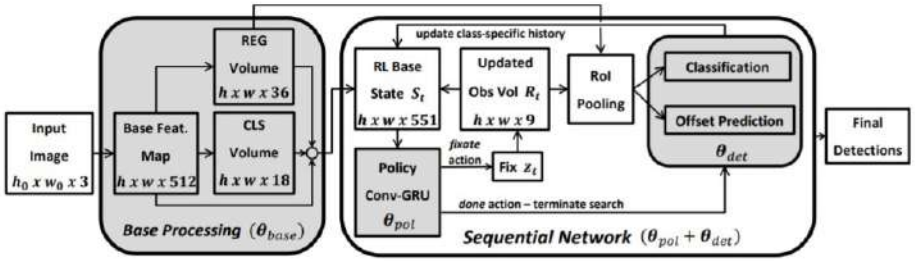
Ryosuke Araki

Deep Reinforcement Learning of Region Proposal Networks for Object Detection

Aleksis Pirinen and Cristian Sminchisescu CVPR2018 872

概要

Region proposal network（RPN）と深層強化学習（DRL）を組み合わせたdrl-RPNを提案する．通常のRPNがRoIを貪欲に選択するのに対し，DRLで学習されたsequential attention mechanismを用いて選択することで，最終検出タスクに最適化される．また，時間経過とともにクラス固有の特徴を蓄積し，分類スコアに良い影響を与えて検出精度が高めることを示す．また，学習をいつ停止するか自動的に判断する．



新規性・結果・なぜ通ったか？

RPNにDRLを導入して，attentionに即したRoIを選択できるようにした．VOC2007を用いた評価では，通常のRPNがmAP74.2%なのに対し，drl-RPNは76.4%を達成した．MSCOCOでも各指標・各セットで数%の精度向上が見られた．

コメント・リンク集

またまた高精度なRoIを検出するタイプの手法．ついにRLまで使うことになった．

- 論文

[#20]

Takumu Ikeya

A Closer Look at Spatiotemporal Convolutions for Action Recognition

Du Tran et al. CVPR2018 1711.11248

概要

- 動画解析のための時空間畳み込みの各手法が行動解析に及ぼす影響を調査した．
- Residual learningのフレームワークでは3D CNNsが2D CNNsよりも精度において優れていることを実験的に示した．
- 3D Convolution filterを空間と時間へ分割することで精度が向上することを示した．
- 新たな時空間畳み込みブロックの構造として”R(2+1)D”を提案した．

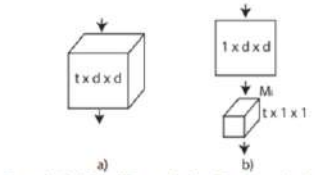


Figure 2. **(2+1)D vs 3D convolution**. The illustration is given for the simplified setting where the input consists of a spatiotemporal volume with a single feature channel. (a) Full 3D convolution is carried out using a filter of size $t \times d \times d$ where t denotes the temporal extent and d is the spatial width and height. (b) A (2+1)D convolutional block splits the computation into a spatial 2D convolution followed by a temporal 1D convolution. We choose the numbers of 2D filters (M_i) so that the number of parameters in our (2+1)D block matches that of the full 3D convolutional block.

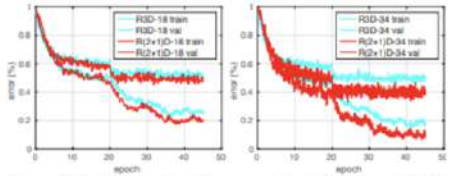


Figure 3. **Training and testing errors for R(2+1)D and R3D**. Results are reported for ResNets of 18 layers (left) and 34 layers (right). It can be observed that the training error (thin lines) is smaller for R(2+1)D compared to R3D, particularly for the network with larger depth (right). This suggests that the the spatial-temporal decomposition implemented by R(2+1)D eases the optimization, especially as depth is increased.

新規性・結果・なぜ通ったか？

- 新規の畳み込みブロックとして時空間の畳み込みブロックを時間と空間に分割する"R(2+1)D"を提案した．
- "R(2+1)D"はSports-1M，Kinetics,UCF101,HMDB51のデータセットでSOTAを達成した．

コメント・リンク集

- 論文

[#21]

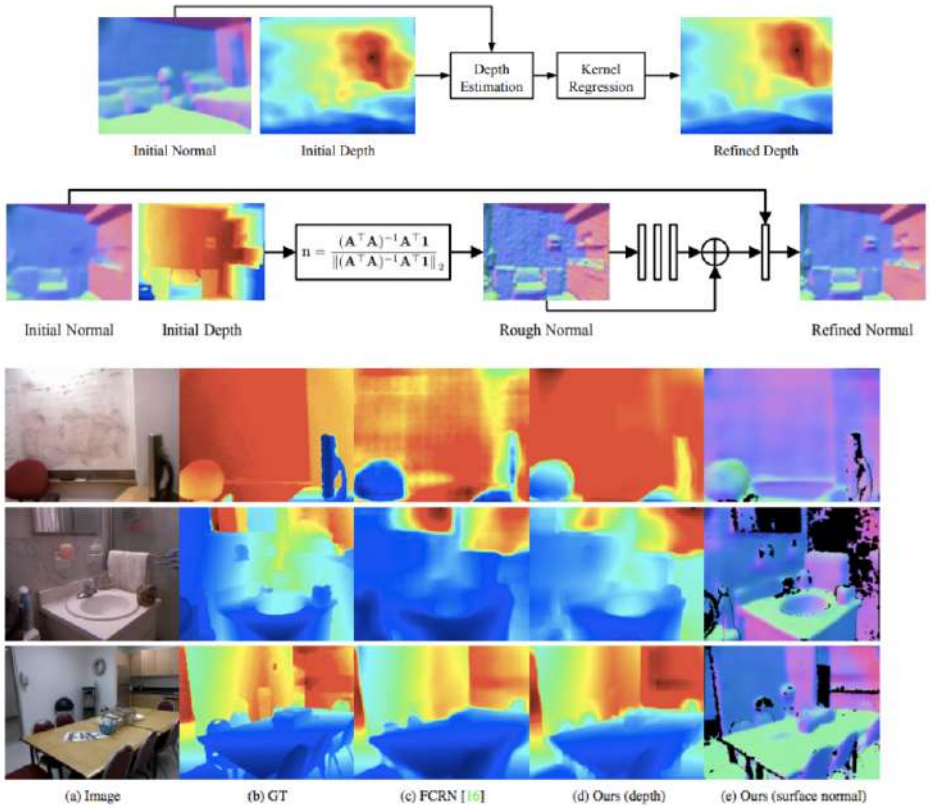
Yuta Matsuzaki

GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation

Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasu andJiaya Jia CVPR2018

概要

単眼の画像から深さ(depth)と表面の法線マップ(surface normal maps)を同時に予測する幾何ニューラルネットワーク(GeoNet)を提案．NYU v2 dataset、ではGeoNetが幾何学的に一貫した深度マップと法線マップを予測できることを確認．surface normal maps推定でSOTA、また既存のdepth推定方法と同等の精度を達成．



新規性・結果・なぜ通ったか？

- GeoNetは2つのストリームのCNNの上に構築されており、depthとsurface normal maps間の幾何学的な関係を構築．これによってdepthとsurface normal mapsを効率的に予測するための基礎となるモデルを構築し，高い一貫性と一致精度を達成することが可

コメント・リンク集

- 著者
- 論文

[#22]

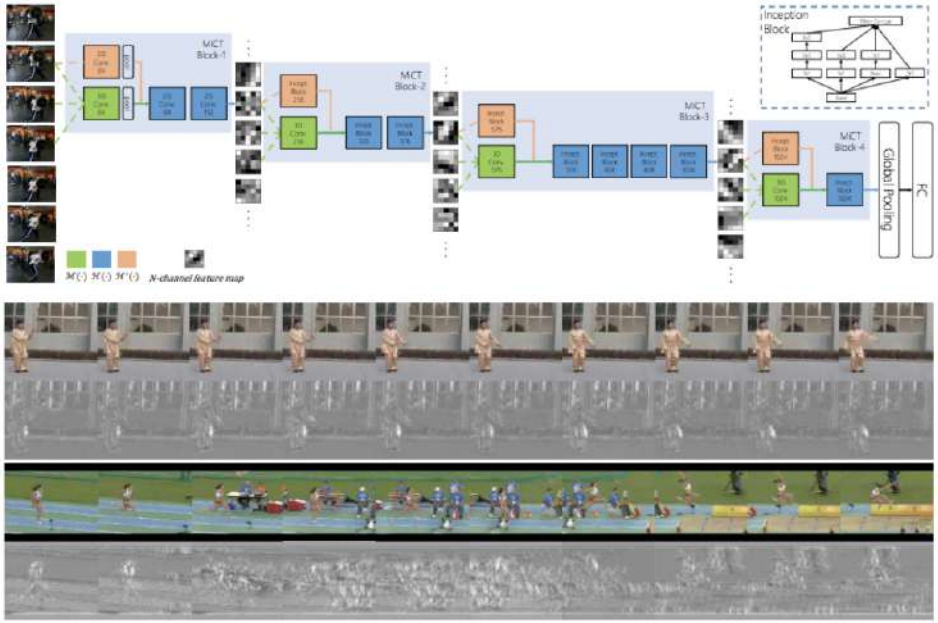
Yuta Matsuzaki

MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition

Yizhou Zhou, Xiaoyan Sun, Zheng-Jun Zha and Wenjun Zeng CVPR2018

概要

2D CNNと3D CNNの畳み込みモジュールを統合した行動認識のためのネットワークMixed Convolutional Tube(MiCT)を提案．3つの有名なベンチマークデータセット(UCF101，Sport1M，HMDB-51)においてMiCT-Netが元の3D CNNのみの手法より著しく優れていることを確認．UCF101とHMDB51での行動認識でSOTAの手法と比較し、MiCT-Netは最高の性能を発揮．



新規性・結果・なぜ通ったか？

- 2D CNNにおける手法を十分にリスペクトし，3D Convと融合した新規のネットワークを構築
- MiCT-Netによって時空間融合の各ラウンドにおける学習の複雑さを軽減しつつ、より深くより有益な特徴マップを生成可能
- UCF101とHMDB51においてSOTA

Method	UCF101	HMDB51
Slow fusion [15]	65.4%	-
C3D [30]	44.0%[1]	43.9[2]%
LTC [31]	59.9%	-
Two-stream [25]	73.0%	40.5%
Two-stream fusion [11]	82.6%	47.1%
Two-stream+LSTM [40]	82.6%	47.1%

コメント・リンク集

- 論文

[#23]

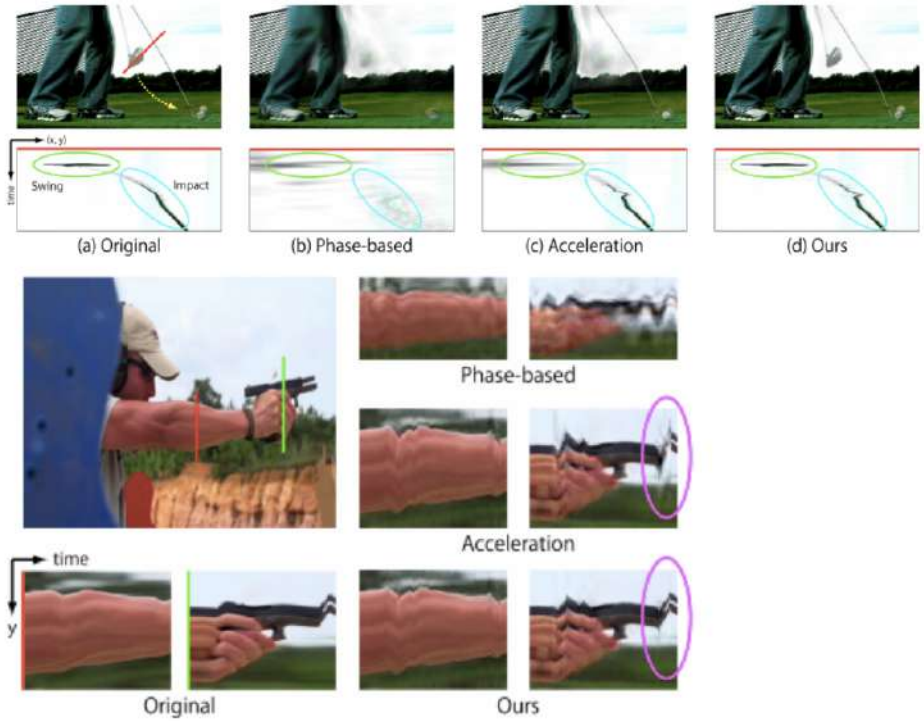
Yuta Matsuzaki

Jerk-Aware Video Acceleration Magnification

Shoichiro Takeda, Kazuki Okami, Dan Mikami, Megumi Isogai and Hideaki Kimata CVPR2018

概要

高速で大きな動きに対して加速度法の出力を頑健にするための、ジャーク(振動，ぶれ)の新規利用方法について言及．微小な変化は時間的スケールでの高速な大きな動きよりも滑らかであるという観点・観測に基づき、高速で大きな動きの下でのみ微妙な変化を通過させるジャークフィルタを設計．



新規性・結果・なぜ通ったか？

ジャークフィルタを加速度法に適用することで、最先端のものより優れた結果を確認．

コメント・リンク集

- link1

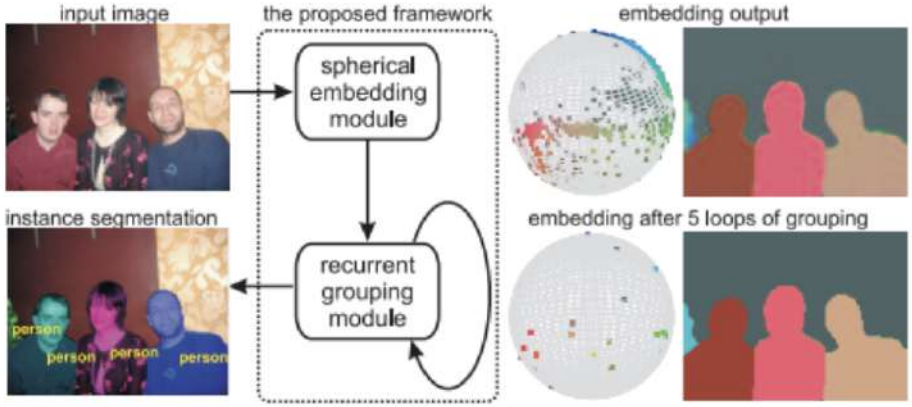
Hiroaki Aizawa

Recurrent Pixel Embedding for Instance Grouping

Shu Kong, Charless Fowlkes CVPR2018

概要

Instance segmentationのような画素単位のグループ分け問題を行うEnd-to-Endで学習可能な枠組みを提案。同じグループの画素はcosine similarityが高くなるように、異なるグループはmargin以下の値になるように超球面上に回帰(Spherical Embedding Module)し、そこでRNNによるMean-shift clusteringを実行すること(Recurrent Grouping Module)で実現。



新規性・結果・なぜ通ったか？

既存のregion proposalやbboxによる組み合わせたinstance segmentationの手法とは大きく異なり新しい。またこれをRNNでMean-shift clusteringを表現することで実現し、End-to-Endな学習を可能としている。加えてhyperparameterの設定に関する理論的分析も提供。instance segmentationやsemantic segmentationだけでなく、様々なpixel-levelのドメインタスクへ応用可能。

コメント・リンク集

手法もシンプルでかつ効果的で応用先も広い。Fig.11の結果からsemantic segmentationにおいてもinstanceの情報が効果的に利用できそうで試してみたい。

- arxiv
- project_page
- GitHub

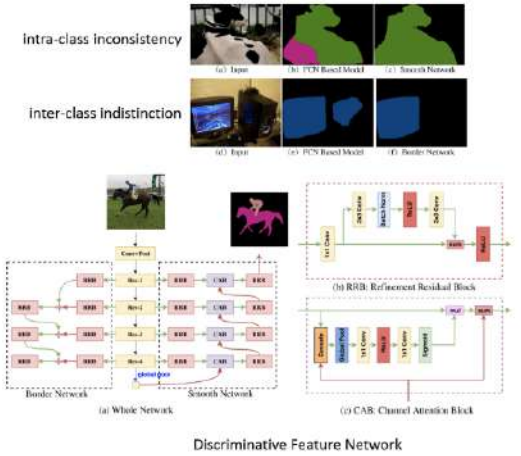
Hiroaki Aizawa

Learning a Discriminative Feature Network for Semantic Segmentation

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, Nong Sang CVPR2018

概要

Semantic Segmentationにおけるintra-class inconsistencyとinter-class indistinctionの問題を、Discriminative Feature Network(DFN)によって対処。intra-class inconsistencyは図の牛の一部を馬と誤認識するような現象。inter-class indistinctionは、図のコンピュータのように外見が似ている対象の区別することが難しい現象。前者の問題をmulti-scaleかつglobal contextな情報を抽出するChannel Attention Block(CAB)を持つSmooth Networkにより、後者の問題をbottom-upなBorder Networkにより緩和する。



新規性・結果・なぜ通ったか？

Semantic Segmentationをpixel単位のラベル付けだけではなく、物体の1つのカテゴリに対して一貫したセマンティックラベル付けをするタスクとして考えた。それゆえのBorder Networkと考える。上記の2つの問題は、必要な情報が異なるゆえ、対処の仕方をCABとU-Net構造に似たSmooth NetworkとBottom-upなBorder Networkとうまく分解している。PASCAL VOC 2012でmean IoU 86.2%、Cityscapesで80.3%を達成。

コメント・リンク集

実験で各モジュールの効果を検証していたが何が効いているのかよくわからない。直感的にはBorder NetworkとSmooth Networkの分離は良いアイデアと感じたが、この分離による効果は1%未満。

- arxiv

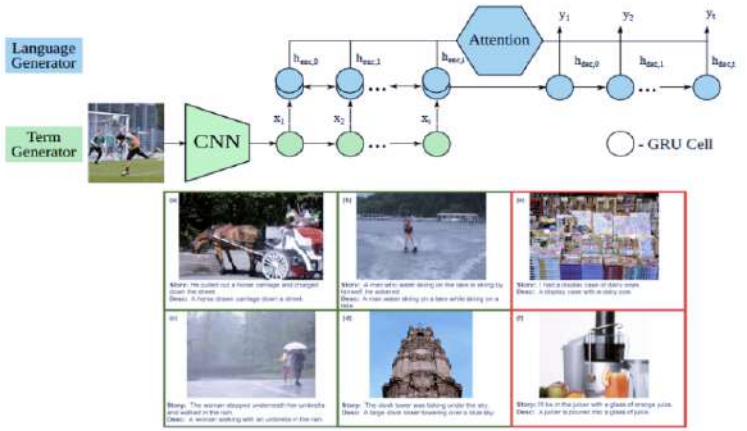
Kota Yoshida

SemStyle: Learning to Generate Stylised Image Captions using Unaligned Text

A.Mathew, L.Xie and X.He CVPR2018 arXiv:1805.07030

概要

書面上のコミニュケーションをする上で文書のスタイルは魅力と明快さに影響する．同一の画像からスタイルの異なるキャプションを生成するという研究．様々なスタイルの単語の選択肢とは異なる構文をもつ文章をデコードするための統一された言語モデルを開発した．



新規性・結果・なぜ通ったか？

- Semanticな用語を用いて文章の柔軟性を備えたキャプションの生成
- スタイルと記述両方のコーパスを用いて文章レベルのスタイルを模倣するための学習
- SemStyleのキャプションが画像の意味を保持し、記述的で、スタイルもシフトできていることを示した

コメント・リンク集

- 連続する写真からより豊富なキャプションを生成できる可能性を秘める
- Paper

[#27]

Kota Yoshida

Smart, Sparse Contours to Represent and Edit Images

T.Dekel, C.Gan, D.Krishnan, C.Liu and W.T.Freeman CVPR2018 arXiv:1712.08232

概要

元画像の輪郭情報から画像を再構成する手法を提案.GANをベースとして，入力情報が与えられない領域のテクスチャと細部を合成する.実験では，顔認証システムや人間を対象にして元画像と再構成された画像と区別されないという結果となった.

Figure 6. The source image (a) is reconstructed from different representations kept at the same pixels marked in red (b), using the following methods: (c) Diffusion [13] based solution that propagates RGB values sampled at both sides of each contour pixel. (d) Pix2pix [18] which uses only binary contours as input. (e) Our LFN output using gradient features stored at each contour pixel and (f) our final HFN output.

新規性・結果・なぜ通ったか？

- Pix2pixなどの既存の手法よりも大幅に向上している.
- 2つのネットワークで構成されており，1つ目のネットワークでは，画像全体の構造，色を再構成，2つ目のネットワークでは画像のテクスチャと細部の表現をしている.
- 直感的な操作が可能で，顔のパーツを移動させたり，追加させることもできる.

コメント・リンク集

- 入力情報がない輪郭と輪郭の間の画像部分の再構成にも力を入れてる

- Paper

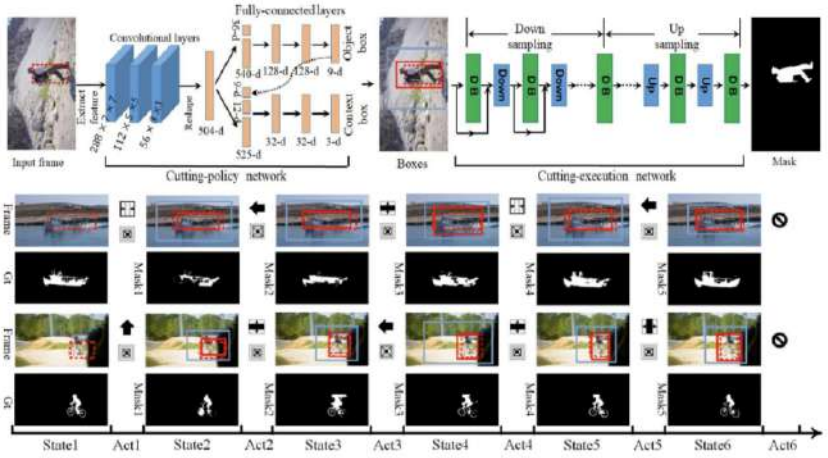
Yoshihiro Fukuhara

Reinforcement Cutting-Agent Learning for Video Object Segmentation

Junwei Han et al. CVPR 2018

概要

Video Object Segmentation (VOS) を強化学習によって行う研究. Object Segmentation では主に物体の領域とそれらの(周辺との)関係性が重要であるという推量に基づいて, VOS をマルコフ過程として定式化し, Deep Q-Learning を適用した. 評価実験では, state-of-the-art とほぼ同等の結果を達成した.



新規性・結果・なぜ通ったか？

- Video Object Segmentation (VOS) をマルコフ過程 (MDP) として定式化した
- State は動画の現在のフレームの特徴量と過去 k （論文では k=4） フレーム分の action のヒストリーを使用
- Action は object searching (9次元) と context embedding (3次元) を使用
- Reward は ground truth のマスクと推定されたマスクの IoU の差で評価
- 強化学習は Deep Q-Learning (DQN) を使用
- DAVIS dataset と YouTube-Objects dataset を用いた評価実験では, state-of-the-art とほぼ同等の結果を達成した

コメント・リンク集

- [論文] Reinforcement Cutting-Agent Learning for Video Object Segmentation
- [Dataset] DAVIS dataset
- [Dataset] YouTube-Objects dataset
- Future work として同様の手法が　Semantic Segmentation, Object Localization, Saliency Estimation, 3D Shape Learning などに適用できる可能性を示唆

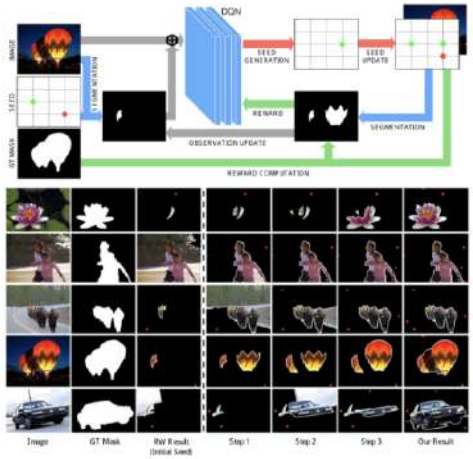
Yoshihiro Fukuhara

SeedNet: Automatic Seed Generation with Deep Reinforcement Learning for Robust Interactive Segmentation

Gwangmo Song et al. CVPR 2018

概要

インタラクティブセグメンテーションに強化学習を適用した研究. 入力画像と初期 seed から自動で新しい seed を順次生成する SeedNet を提案. 評価実験では state-of-the-art の結果を達成すると共に, 教師あり手法と比較しても優位な結果を達成した.



新規性・結果・なぜ通ったか？

- Interactive Segmentation のタスクをマルコフ過程として定式化し, 強化学習（Deep Q-Learning）を用いて学習を行った
- State には入力画像の画素情報と seed の位置とラベル， mask 画像を用いる (seed の位置を state に陽に加えることによって, 生成される mask が seed 位置の変化についてロバストになるらしい)
- Action は state の情報から新しい seed の位置とラベルの決定（自由度を削減するために 20x20 のグリッド上から位置を選択, seed の数が10点になった段階で終了）
- Reward は生成された Mask と Ground Truth の Mask の IoU（exp 型を提案）に加えて, SeedNet によって追加された新 seed のラベルと位置が適切かの２点を考慮して決定

コメント・リンク集

- [論文] SeedNet: Automatic Seed Generation with Deep Reinforcement Learning for Robust Interactive Segmentation
- 強化学習を新タスクに適用してみました系列の論文
- 他の同系列の論文に見られる傾向と同じく, MDPによる定式化と Reward の計算方法を主な貢献としている
- 特に本論文は, 教師ありでは学習するのが難しい問題を上手く見つけている（seed の打ち方は user によって千差万別なのでトレーニングデータを作るのが難しい）

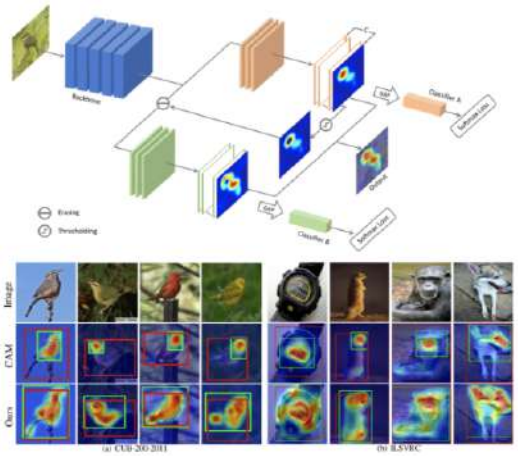
Yoshihiro Fukuhara

Adversarial Complementary Learning for Weakly Supervised Object Localization

Xiaolin Zhang et al. CVPR 2018

概要

弱教師ありの Object Localization の研究. 2つの Classifier を並列に配置し, 片方の classifier で注目された領域を他方の入力から取り除いておくことで, それぞれが異なる領域に反応するような構造となっている. 評価実験では ILSVRC dataset の localization　のタスクで 45.15% (new state-of-the-art) の誤差率を達成した.



新規性・結果・なぜ通ったか？

- 全結合層の最後に畳み込み層を1つ追加することで, CAM [Zhou＋16] と同等の object localization maps を事後処理無しで得られることを数式で示した
- 画像から畳み込み層によって抽出した特徴量を, 並列に配置した classifier に入力する
- 片方の classifier から出力された object localization map で注目されていた領域を消去したものを, 他方の入力とすることで両方の classifier を異なる領域に反応させる
- ILSVRC dataset 等を用いて行った評価実験では Localization と Classification の両タスクにおいて, state-of-the-art [Zhou+16, Singh+17] と同等か優位な結果を達成した

コメント・リンク集

- [論文] Adversarial Complementary Learning for Weakly Supervised Object Localization

[#31]

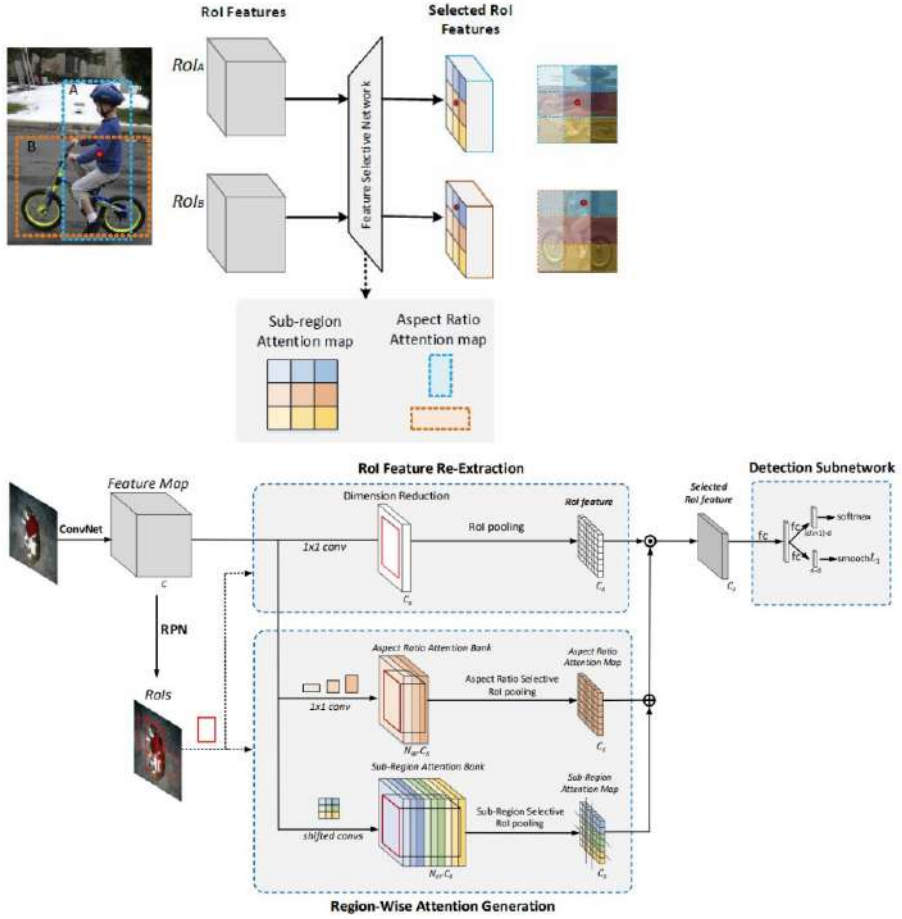
Ryosuke Araki

Feature Selective Networks for Object Detection

Yao Zhai, Jingjing Fu, Yan Lu, Houqiang Li CVPR2018 538

概要

物体検出時に用いるRegion-of-Interest（RoI）を，sub-regionとアスペクト比の差を用いて再構成するFeature selective netsを提案．画像全体に対してsub-regionのattention bank（すべてのattention mapを記憶するbank）とアスペクト比のattention bankを生成する．Attention mapはbankから選択的にpoolされ，RoIの改善に使用される．処理の手順は(1)CNNから得られた特徴マップをRPNに入力しRoIを得て，(2)特徴マップのチャンネル数を削減してRoIプーリングを行い，圧縮されたRoI特徴を得る．(3)削減される前のRoIをregion-wise attention生成モジュールに入力する．特徴マップを用いてアスペクト比attention bankとsub-region attention bankを得る．(4)各bankにselective RoIプーリングを行う．そして，(2)と(4)で得られたRoI特徴と各attention mapを結合して検出サブネットワークに入力する．



新規性・結果・なぜ通ったか？

RoIをattentinを用いて補正する．VGGだけではなくGoogLeNetやResNetにも適用可能である．VOC2007を用いた評価では，mAP: 82.9%, 76.8%, 74.3% （Res101, GoogLe, VGG-16）を達成し，

コメント・リンク集

Attentionを用いた物体検出が増えてきている．Mask R-CNNみたいにRoIに注目する手法も多い？

[#32]

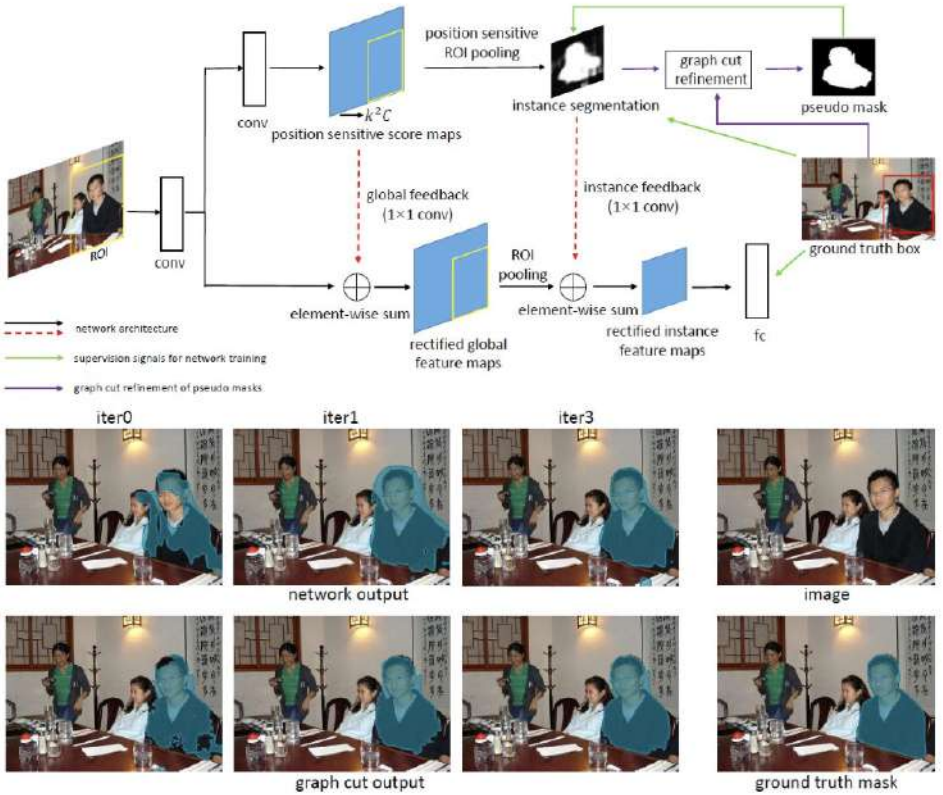
Ryosuke Araki

Pseudo Mask Augmented Object Detection

Xiangyun Zhao, Shuang Liang, Yichen Wei CVPR2018 530

概要

Bounding boxでの物体検出でグラフカットを用いて擬似的なマスク（セグメンテーション）のrefinementを行う．インスタンスセグメンテーションの学習を行うことで擬似的な物体マスクを推定できるようにネットワークパラメータを最適化する．フレームワークは検出ネットワークと擬似的なマスクのrefinementを行うグラフカットベースのモジュールからなる．RoIを入力として，ベースネットワークの特徴マップからインスタンスセグメンテーションを行い，それをグラフカットモジュールに入力して擬似的なマスクを得る．インスタンスセグメンテーションの結果はbounding boxの修正にも用いられる．



新規性・結果・なぜ通ったか？

流行りの物体検出＋セグメンテーションの手法．マスクを単に特徴マップから得て終わりではなく，グラフカットでrefineする部分は新しいところ．グラフカットを数iter行うことで，よりきれいなマスクを得ることができる．VOC2007/2012を用いた物体検出の精度はmAP74.4%（VGG-16）で，Faster R-CNN（70.4%）やHyperNet（71.4）よりも良い．VOC2012SDSを用いたセグメンテー

コメント・リンク集

セグメンテーションタスクの精度向上のためグラフカットでマスクのrefineを繰り返し行うのは面白いと思った．Iter0とiter3でマスクの結果を比較するとかなりきれいになっている．

- 論文
- arXiv

[#33]

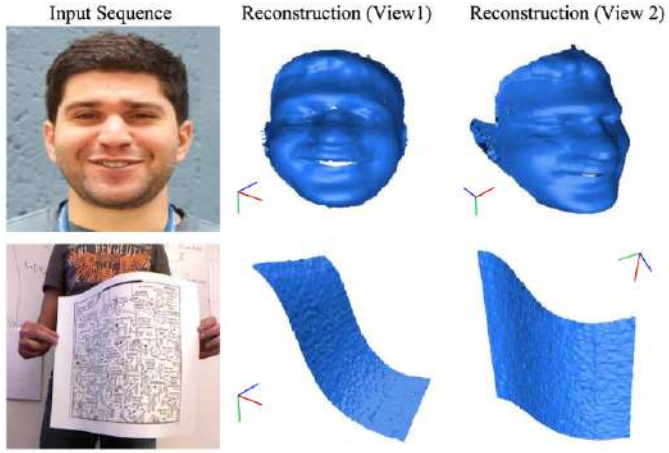
Hirokatsu Kataoka

Scalable Dense Non-Rigid Structure-From-Motion: A Grassmannian Perspective

Suryansh Kumar, Anoop Cherian, Yuchao Dai, Hongdong Li CVPR 2018

概要

複数画像を使用した非剛体のSfM (Non-Rigid Structure-from-Motion)に関する研究である。右図は非剛体の表面形状復元結果の一例であり、顔のように時系列的に変化する形状を、多様体の概念をSfMに導入することにより問題解決を図っている。非剛体の形状変化を、空間的・時間的な部分空間としてすいていすることでSfMを実行する。



新規性・結果・なぜ通ったか？

非剛体物体の表面形状復元に関するSfM問題を、グラスマン多様体（Grassman Manifold）の問題と捉えて解決している点が新規性として挙げられる。柔軟に表面形状復元ができている様子は動画にて確認可能である。

コメント・リンク集

DynamicFusionからこの手の問題は出て来たのだが、どのような違いがある/どのように展開されているのか？

- 論文
- 著者
- YouTube

[#34]

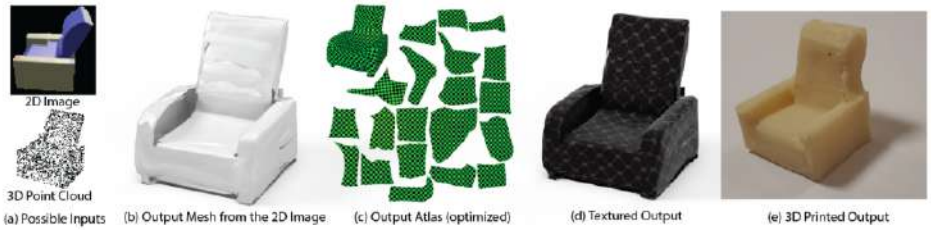
Hirokatsu Kataoka

A Papier-Mâché Approach to Learning 3D Surface Generation

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, Mathieu Aubry CVPR 2018

概要

2次元画像、もしくは3次元点群からメッシュや分解構造を生成し、テクスチャありのメッシュや3次元プリント物体を出力する。この枠組みはAtlasNetと呼ばれ、同タスクのPrecision向上と一般化の面で性能改善を行い、3次元形状を集めたデータベースであるShapeNet上で形状をAuto-Encoding、単眼画像からの形状復元を行った。その他、AtlasNetを用いてモーフィング、パラメトライゼーション、超解像、形状マッチング、共セグメンテーションを実施した。



新規性・結果・なぜ通ったか？

3D表面形状生成器であるAtlasNetを構築したことが最も大きな新規性である。形状に関するパラメータを学習可能にした。さらに、AtlasNetをGitHub上で公開して使用できる形式にしている。復元したメッシュ形状も、提案手法がもっともノイズが少なく、良好な復元結果となった。

コメント・リンク集

数年前は型崩れの多い3次元形状を出力するGeneratorであったが、徐々によくなりつつある。この研究もまだ過程にしか過ぎない？

- 論文
- Project
- GitHub

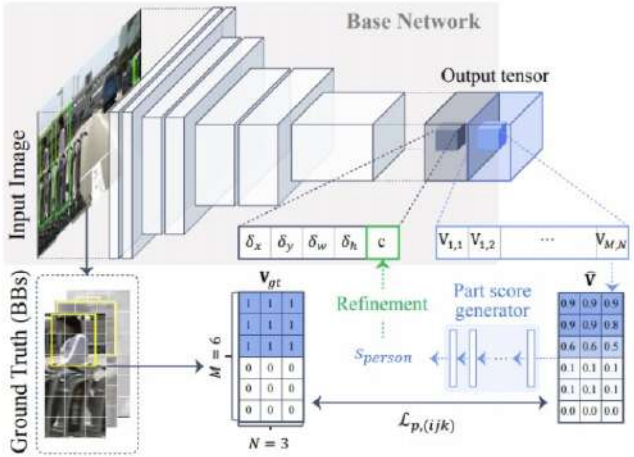
Munetaka Minoguchi

Improving Occlusion and Hard Negative Handling for Single-Stage Pedestrian Detectors

Junhyug Noh, et al.

概要

歩行者検出におけるオクルージョンやハードネガティブを改善するための提案。本提案手法は、シングルステージ物体検出手法に適応可能。オクルージョン処理のために、ベースモデルの出力テンソルを更新してパートスコアを推定し、オクルージョン認識スコアを算出する。ハードネガティブの混同を軽減するために、 average grid classifiersをpost-refinement classifiersとして導入。



新規性

SqueezeDetやYOLOv2、SSD、DSSDを含むシングルステージ物体検出手法に適応でき、オクルージョンやハードネガティブを改善する。本論文では歩行者検出におけるオクルージョンにフォーカスを当てているが、一般物体検出にも適応できる可能性がある。

結果・リンク集

CaltechPedestrianとCityPersonsデータセットで評価。4つのモデルのパフォーマンス向上を確認。重度のオクルージョン設定において、最良のパフォーマンス。

- 論文

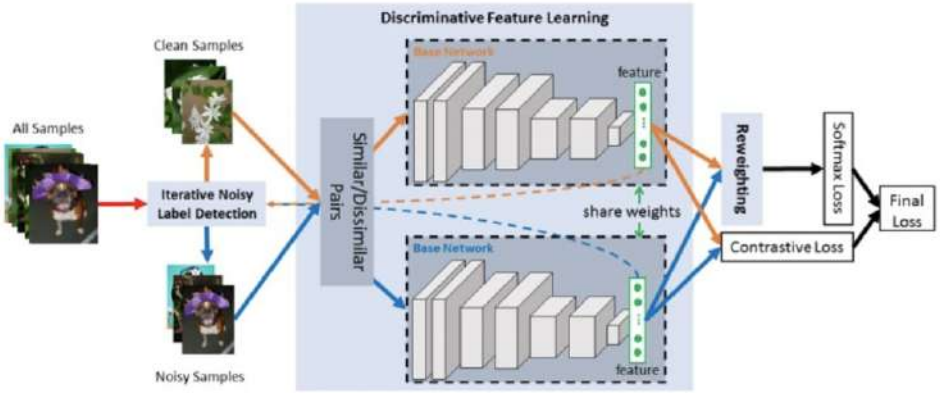
Munetaka Minoguchi

Iterative Learning with Open-set Noisy Labels

Yisen Wang, et al. 1804.00092

概要

ノイズのあるラベルを含んだデータセットを使い、CNN学習を高精度に行うための新しい反復学習フレームワークの提案。反復的なノイズラベル検出、特徴学習、および再重み付けの3段階のフレームワークでノイズの多いラベルを検出しつつ、識別器を反復的に学習。再重みづけでは、クリーンなラベルの学習を重視し、ノイズの場合には低減させる。



新規性

綺麗なラベルアノテーション付き大規模データセットによる学習は非常に重要だが、人の手間がかなりかかる他、ヒューマンエラーを含む可能性が否めない。本研究では、あえてノイジーなデータセットに挑戦することで、これらの問題を解決する。

コメント・リンク集

データセットの収集コストや信頼性の問題に伴って、自ら良いデータを選択して学習する需要が高まっている印象。

- 論文

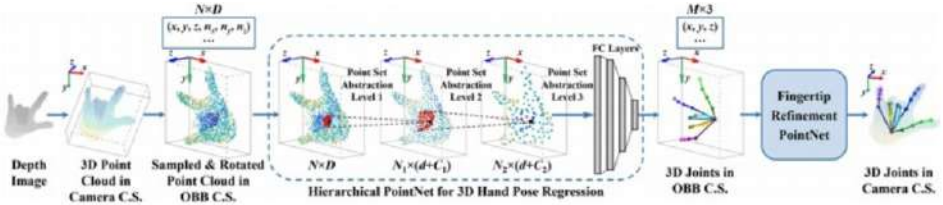
Munetaka Minoguchi

Hand PointNet: 3D Hand Pose Estimation using Point Sets

Liuhao Ge, et al.

概要

正規化されたポイントクラウドを入力として、複雑な手構造を捕捉し、手の姿勢の低次元表現を正確に回帰させることができるHand PointNetの提案。Oriented Bboxでポイントクラウドを正規化し、ネットワーク入力をよりロバストにする。その後、階層的なPointNetに入力し特徴抽出。PointNetを細分化することにより、指先に対する推定精度を向上させる。



新規性

CNNを用いた従来の奥行き画像における3次元手姿勢推定手法とは異なり、本研究では三次元点群に着目している。データは、奥行き画像をポイントクラウドデータに変換してから使用している。

結果・リンク集

3つのハンドポーズデータセットにて実験し、リアルタイム性に優れていることを示唆。

- 論文

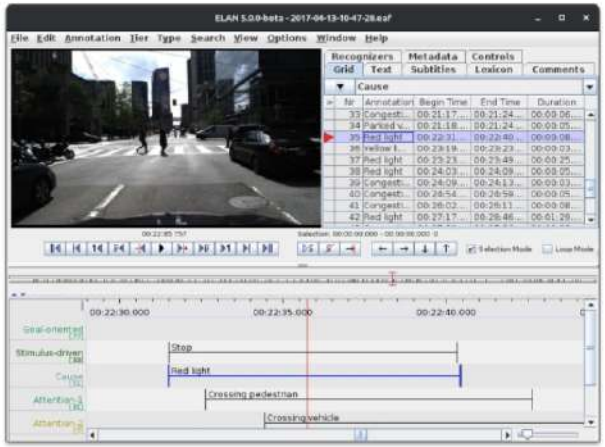
Munetaka Minoguchi

Toward Driving Scene Understanding:A Dataset for Learning Driver Behavior and Causal Reasoning

Vasili Ramanishka, et al.

概要

自動車の運転シーン理解のためのデータセットであるHonda Research Institute Driving Dataset(HDD)の提案。本データセットはサンフランシスコ・ベイエリアにて、様々なセンサーを備えた自動車を人間が運転したデータが104時間分含まれる。センサはグラスホッパーカメラ、LiDAR、ダイナミックモーションアナライザ、Vehicle Controller Area Network (CAN)の4つ。これらのデータから運転者の行動を基にアノテーションを付加している。



新規性

様々なセンサを用いて、大規模データを収集しただけでなく、ヒューマンファクタや認知科学に基づいてアノテーションを行っている。アノテーションは、Goal-oriented action, Stimulus-driven action, Cause, Attentionの4つ。

コメント・リンク集

LSTMを用いたベースラインにおいて、センサを増やすことによって表現力の向上が見られた。評価が難しいアノテーションデータが含まれ、チャレンジングなデータセット。

- 論文

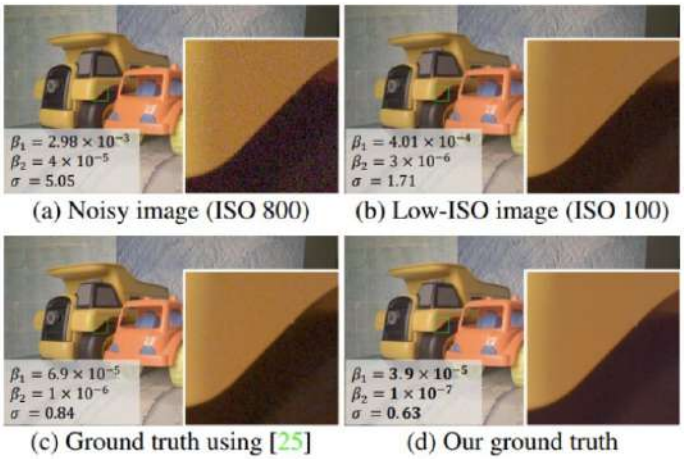
Munetaka Minoguchi

A High-Quality Denoising Dataset for Smartphone Cameras

Abdelrahman Abdelhamed, Stephen Lin, Michael S. Brown

概要

スマートフォンで撮影したノイズの多い画像で構成したデータセットSmartphone Image Denoising Dataset (SIDD)の提案。 5つの代表的なスマホカメラを使用し、様々な照明条件下で約30,000枚のノイズの多い画像を収集。ノイズの多い画像だけでなく、ノイズを除去した画像をground truthとして提案。



(a) Noisy image (ISO 800) (b) Low-ISO image (ISO 100)
(c) Ground truth using [25] (d) Our ground truth

新規性

過去10年間で、撮影される画像は一眼レフやコンデジから、スマートフォンに切り替わったことに着目。しかし、口径やセンサーサイズが小さいため、スマホの写真はノイズを多く含んでいる。このような、ノイズを多く含んだスマホ画像を集めることで新たなデータセットを提案する。

コメント・リンク集

やはりノイズを含むスマホ画像でのトレーニングよりも、高品質な画像でトレーニングした方が、CNNで高い精度を得た。現在のタスクにおいて「スマホの画像だから精度が出ない」というのはあまり考えにくいが、日常的なアプリケーションには有用なデータセットではないか。

- 論文

[#40]

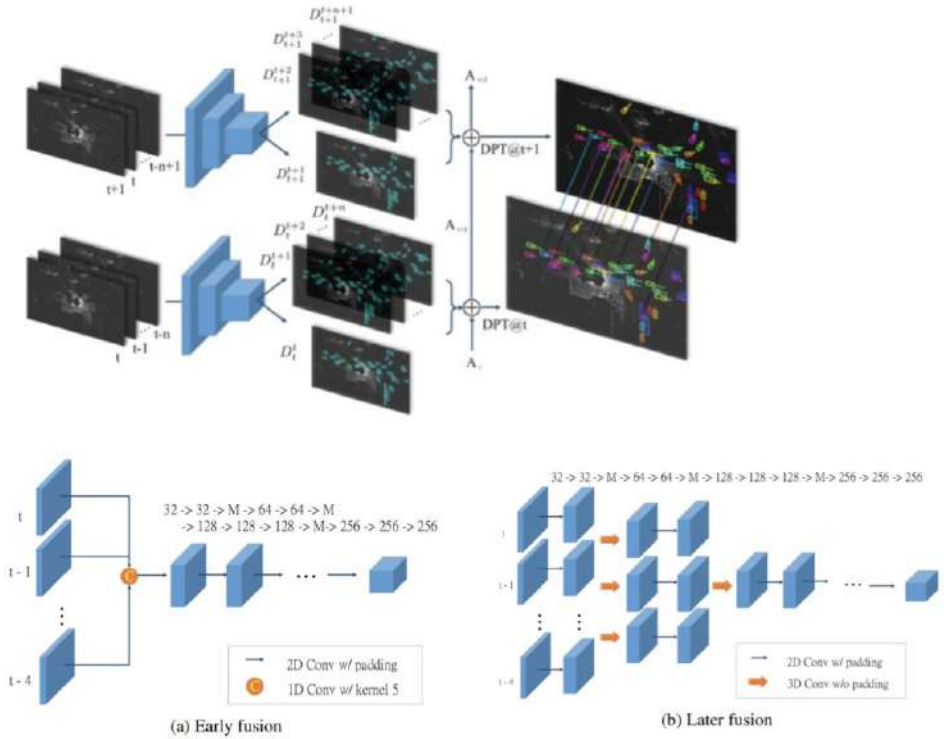
Ryosuke Araki

Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net

Wenjie Luo, Bin Yang, Raquel Urtasun CVPR2018 437

概要

3Dセンサで得られた点群から3D物体検出や追跡を行う新しいDNN「Fast and Furious（FaF）」を提案．検出と追跡，さらに短期の経路予測を同時に推論でき，Sparse dataやオクルージョンに頑健な検出ができる．3D点群と時間の4Dテンソルを入力として，空間と時間に対して3D畳み込みを行う．4DテンソルはEarly FusionまたはLate Fusion（図中ではLater）で時間情報を結合している．これらは精度と効率のトレードオフ関係にある．



(a) Early fusion (b) Later fusion

新規性・結果・なぜ通ったか？

物体検出から追跡，さらに経路予測までend-to-endで行えるモデル．全体の検出時間はわずか30ms以下である．約55万フレームからなるLiDARのデータセットを作成し，車両に3D bboxとトラッキング用IDをラベリングして学習および評価に用いる．物体検出の結果はSSDのIoU 77.92mAPを上回る83.10mAPである（Late Fusionを用

コメント・リンク集

タイトルが某カーアクション映画みたいでカッコいい．内容も名前負けしておらずよく作り込まれておりOralで採択されている．インパクトのあるタイトルは大切．

- 論文

Munetaka Minoguchi

Low-Shot Learning from Imaginary Data

Yu-Xiong Wang, et al. 1801.05401

概要

人間の想像力に着目することで、メタ学習におけるLow-Shot Learningを可能にするアーキテクチャの提案。コンピュータビジョンに幻覚(想像)を抱かせることで、少ないデータから新しい視覚的概念を学習させる。アプローチとしては、メタ学習を取り入れており、 meta-learnertとhallucinator(幻覚者)を組み合わせて共同で最適化。hallucinatorは、通常のトレインセットとノイズベクトルから幻覚トレーニングセットを出力する。通常のトレーニングセットに加えて、幻覚トレーニングセットを学習することで精度向上を図る。

新規性

人間は新しい視覚的情報を素早く学習できる。これは、「物体がさまざまな視点から見たときにどのように見えるかを想像できるから」と仮定。そのうえで、人間の想像力をモデルとし、システムに組み込むことでLow-Shot Learningを可能にしている。

コメント・リンク集

AIに幻覚を見せられる時が来た模様。さまざまなメタ学習手法に組み込むことができ、精度を向上させられるらしい。

- 論文

[#42]

Hirokatsu Kataoka

Multi-View Harmonized Bilinear Network for 3D Object Recognition

Tan Yu, Jingjing Meng, Junsong Yuan CVPR 2018

概要

3次元物体認識を実行するMulti-view Harmonized Bilinear Network (MHBN)を提案する。異なるビューの特徴量を学習するために基本的にはパッチベースでマッチングを行う。Polynomial Kernel/Bilinear Poolingの関係性を記述するために、畳み込みによる3次元物体表現とBilinear Poolingを実行する。MHBNの枠組みはEnd-to-Endでの学習が可能である。構造は右図のように示され、畳み込みにより特徴マップ（3次元物体表現）を生成、最後にBilinear Poolingを通り抜けて識別を実行。

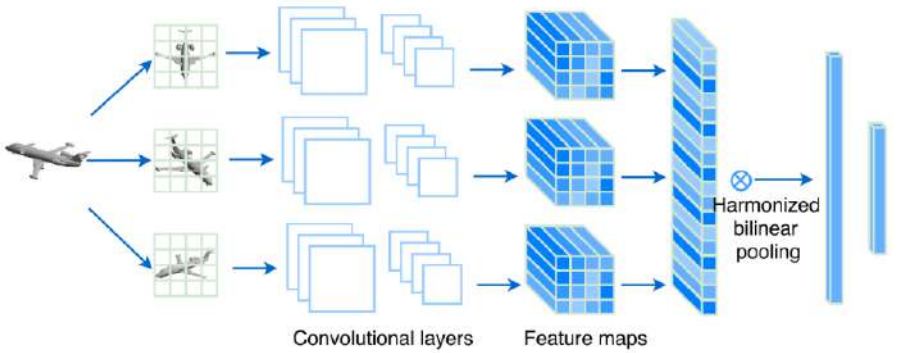


Figure 1. The architecture of the proposed Multi-view Harmonized Bilinear Network (MHBN).

新規性・結果・なぜ通ったか？

3次元物体認識の場面においてSoTA。ModelNet40, ModelNet10ではそれぞれ94.7 (Instance)/93.1 (Class), 95.0 (Instance)/95.0 (Class)である。

コメント・リンク集

3次元物体認識ではホントの意味での大規模DBはないのだろうか？ModelNetにしてもShapeNetにしてもCADをベースにしている？

- 論文
- 著者
- MVCNN

[#43]

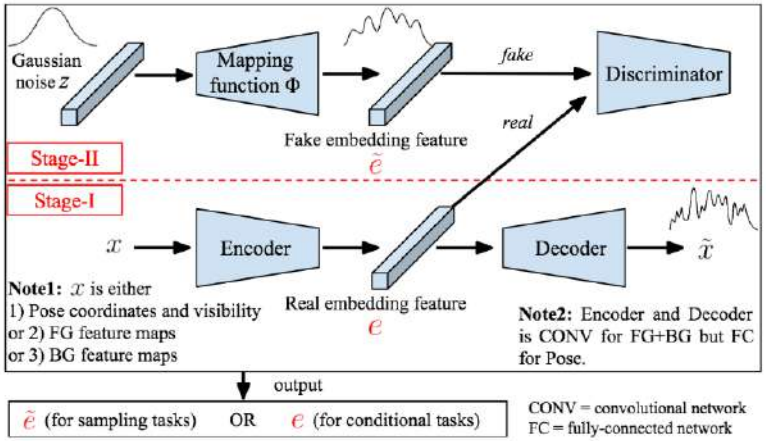
Hirokatsu Kataoka

Disentangled Person Image Generation

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, Mario Fritz CVPR 2018

概要

アピアランス/ビューポイント/背景など、分解された（Disentangled）人物画像の生成を行うための研究である。この目的のため、2ステージの生成手法を考案した（右図を参照）。1ステージ目はリアルの埋め込み特徴（Embedding Features）を獲得する学習を行い、前景/背景や姿勢などを表現。次に2ステージ目は敵対的学習により生成的特徴学習を行いガウシアンノイズから中間表現にマッピング、特徴変換を行う。



新規性・結果・なぜ通ったか？

姿勢ベースの人物画像を生成し、人物再同定（Person Re-Identification; ReID）の学習に適用。人物画像生成自体も誤差が少なく、ReIDのためのにおいても良好な精度を実現した。

コメント・リンク集

学習画像がコントロールできるということで注目される技術。ある程度の知見を学習しておけば、そのうちリアル画像のデータがいらない時代になる？

- 論文
- GitHub

[#44]

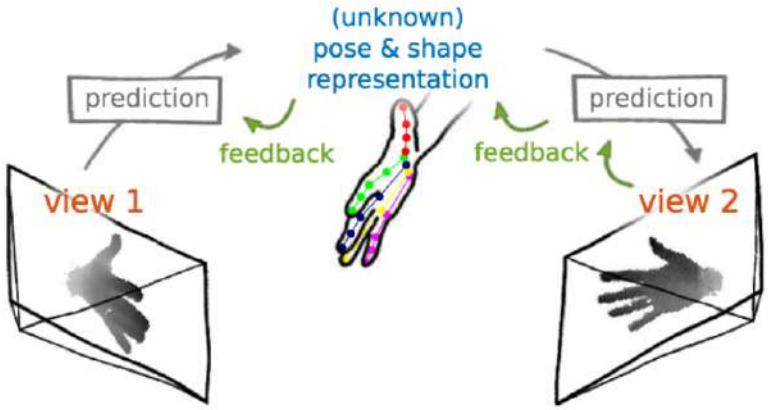
Hirokatsu Kataoka

Learning Pose Specific Representations by Predicting Different Views

Georg Poier, David Schinagl, Horst Bischof CVPR 2018

概要

異なるビューポイントの距離画像入力から、低次元の潜在表現を利用して手部領域追跡の学習を実行する研究である。ビューポイント推定の誤差をフィードバックして、教師なしでも手部の姿勢推定に必要な潜在表現を獲得する。これにより、必要なのは対象となるビューポイントではなく、第二のビューポイントのみであり、ラベルあり/ラベルなしの場合においても効果的に学習することができる（Semi-supervised Learningの枠組みで学習可能）。



新規性・結果・なぜ通ったか？

あるビューポイントの距離画像が手に入れば、異なるビューポイントに関する手部領域の姿勢推定が可能になるSemi-supervised Learningを提案。異なるビューポイントの低次元潜在表現を学習し、3Dの関節位置を推定することができる。NYU-CS dataset/MV-hands datasetにてState-of-the-artな精度を達成。

コメント・リンク集

中間表現（本論文の場合には低次元潜在空間）を学習して、異なるドメイン間の学習に応用したい。このような問題は意外と簡単にできるのだろうか？

- 論文

[#45] okayasu

Fine-grained Video Captioning for Sports Narrative

Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, Xiaokang Yang CVPR 2018

概要

Fine-grainedなスポーツ動画キャプショニング

img(src=`${figpath}Fine-grained_Video_Captioning_for_Sports_Narrative.png`,alt="Fine-grained_Video_Captioning_for_Sports_Narrative")

新規性・結果

- youtubeから2Kのスポーツ動画とキャプションからなるFine-grained Sports Narrative dataset(FSN)の提案
- スポーツビデオのキャプショニングの新しい評価指標Fine-grained Captioning Evaluation(FCE)の提案
- スポーツビデオのキャプショニングの新しいフレームワークの提案(骨格情報とオプティカルフローで詳細な動作のエンコード，オプティカルフローと選手のローカライズ結果で人物間のインタラクションをエンコードそれらのエンコードされたベクトルを階層的RNNで言語化)

コメント・リンク集

- 論文

[#46]

Hirokatsu Kataoka

GANerated Hands for Real-Time 3D Hand Tracking From Monocular RGB

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, Christian Theobalt

CVPR 2018

概要

RGBのみの動画入力からリアルタイムに3次元手部関節位置推定を実行する手法を提案。YouTubeのようなコントロールされていない場面においても3次元手部関節位置推定を行うことができる。本論文では3次元のハンドモデルとCNNを組み合わせることによりトラッキングを実行しており、GANによる生成ベース（手の3次元合成データをリアルに変換していることに相当）の手法によりオクルージョンやビューポイントの違いに頑健である。GANはAdversarial LossとCycle-consistency Loss、さらには幾何学的な整合性を保つためにGeometric Consistency Lossを最適化するよう学習。

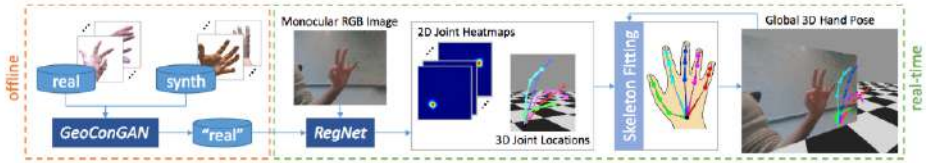


Figure 2: Pipeline of our real-time system for monocular RGB hand tracking in 3D.

新規性・結果・なぜ通ったか？

GANをベースとして合成データからリアル画像を生成、同データで学習したモデルは、RGB-onlyな3次元ハンドトラッキングにおいてState-of-the-artである。敵対的学習を用いたデータ生成手法、YouTube等のあまり校正されていないデータにおいても良好な精度を実現していることが採択された理由であると考える。

コメント・リンク集

3Dデータを自由に生成できることは、次世代のアイディアを実現するための大きなポイントである。3次元トラッキングのみならず面白いこと考えたい。

- 論文

[#47]

Hirokatsu Kataoka

A Certifiably Globally Optimal Solution to the Non-Minimal Relative Pose Problem

Jesus Briales, Laurent Kneip, Javier Gonzalez-Jimenez CVPR 2018

概要

キャリブレーション済みの２カメラにおける相対姿勢の推定問題を解くための全体最適化法（Globally Optimal Solution）を提案する。局所最適解ではなく、グローバルな最適化が計算できることが新規性である。本論文では、凸最適化の問題においてあらかじめ定義された問題（Shor's Convex Relaxation）としてQuadratically Constrained Quadratic Program (QCQP)を扱うことを実施する。ここに対して、理論的かつ実験的な解答法を提示したことが本論文の貢献である。

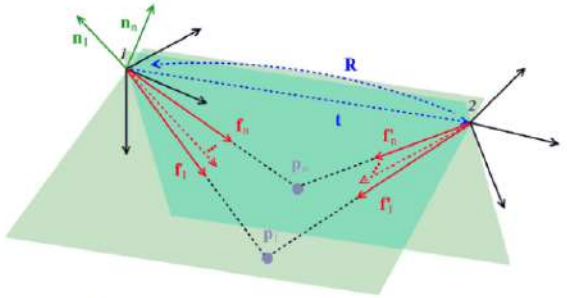


Figure 1. The relative pose problem. The measurements are given by correspondence pairs of unit bearing vectors $\{\mathbf{f}_i, \mathbf{f}'_i\}$, and the unknown variables are given by the relative orientation $\mathbf{R}$ and the direction of the relative translation $\mathbf{t}$.

新規性・結果・なぜ通ったか？

２カメラの相対姿勢問題の解決のために従来の凸最適化手法を適用して、理論的かつ実験的に解決できることを示したことが新規性であり、CVPRに採択された理由である。

コメント・リンク集

（あまり深く読めていないのと、知識が足りなくて自信がないです。。）

- 論文

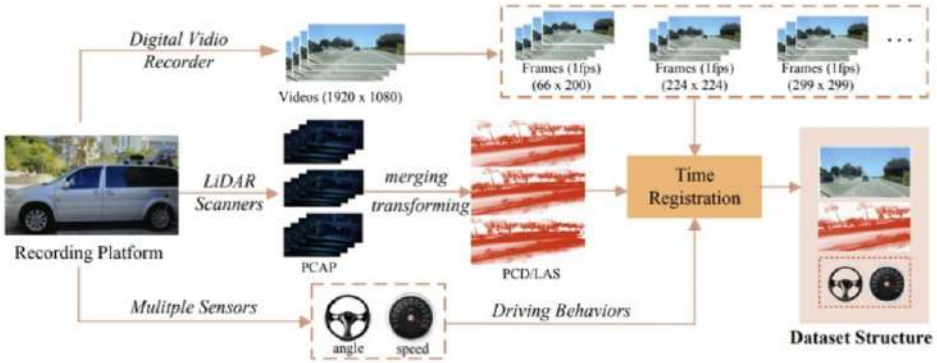
Munetaka Minoguchi

LiDAR-Video Driving Dataset: Learning Driving Policies Effectively

Yiping Chen, et al.

概要

LiDERで取得したポイントクラウド、車載カメラ映像、および一般ドライバーの運転動作からなるLiDAR-Videoデータセットの提案。運転動作は、ハンドルの傾きと自動車の走行速度情報によるもの。また、これらのデータを使い、自律走行における運転手段を決定するためのPolicy Learningを提案。 これは、DNN+LSTMで構成されるアーキテクチャである。3種類のデータの対応時間を登録することでどのように運転するかをベンチマークする。



新規性

自律走行において、これまではカメラとレーザースキャナー、運転動作を組み合わせたデータやアプローチがなかった。本論文ではデータベースを構築したうえで、自律走行に対するアプローチを提案している。

結果・リンク集

単一のデータよりも3つのデータを組み合わせることで精度が向上していることを示唆。また、DNN単体よりも長いtermで処理できるDNN+LSTMの方が精度向上につながることも示唆。

- 論文

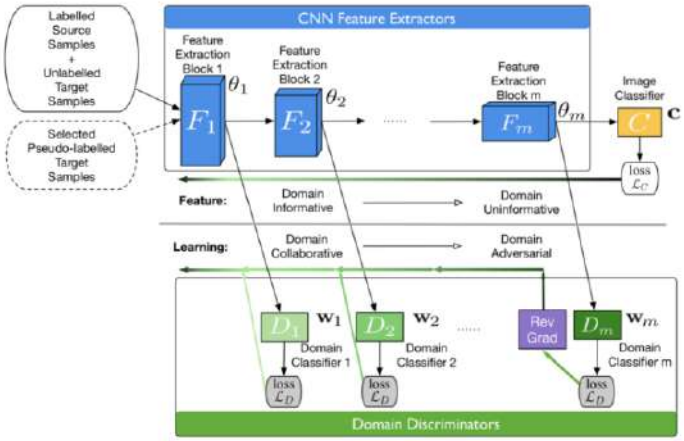
[#49] Kazuki Inoue

Collaborative and Adversarial Network for Unsupervised domain adaptation

Weichen Zhang, Wanli Ouyang, Wen Li, Dong Xu CVPR 2018 Poster

概要

CNNの浅い層ではドメイン固有の特徴量を、深い層ではドメインに不変な特徴量を取得することでdomain adaptationを行うCollaborative and Adversarial Network(CAN)を提案。 従来のDomain Adversarial Training of Neural Network(DANN)ではドメインに不変な特徴量を学習することができるものの、ターゲットドメイン固有の特徴量を得ることが難しいという問題があった。 提案手法では、CNNの浅い層では低次の特徴量を、深い層では高次の特徴量を取得することができることに着目し、 CNNのそれぞれのブロックに対するdomain discriminatorに対して、浅いブロックではソースドメインとターゲットドメインを識別可能となるように、 深いそうでは識別が不可能となるように学習を行う。ソースドメインに対してはクラスの識別も行う。 またテストデータに対してpseudo labelingを行うIncremental CAN(iCAN)も提案。 ターゲットドメインのサンプルのうち、高いconfidenceでソースドメインであると判定され、 かついずれかのラベルに対するconfidenceが高いものに対してpseudo labelingを行うことで、データセットを拡張しdomain shiftを解消する。



新規性・結果・なぜ通ったか？

- CNNの浅いブロックで得られる特徴量に対してはドメイン識別が可能なように、深いブロックで得られる特徴量に対してはドメイン識別が不可能なように学習を行うCANを提案。 またターゲットドメインに対してpseudo labeingを行うiCANも提案。
- 実験で使用したのはpretrained RenNet50であり、10層目、22層目、40層目、49層目のそれぞれに対してdomain discriminatorを適用。41~49層からなるブロックからドメインに不変な特徴量を

コメント・リンク集

- シンプルな発想だが面白い手法！似たアイディアで画像の生成もできないだろうか？
- 論文

[#50]

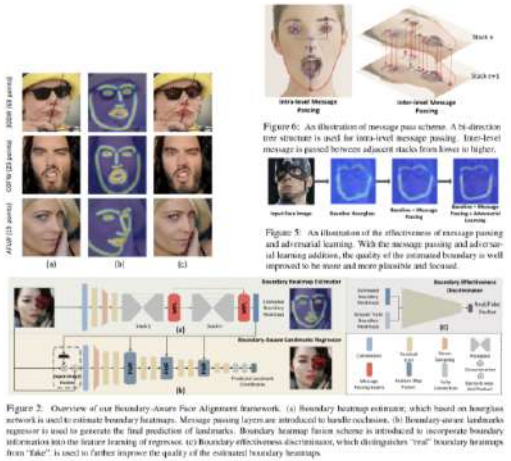
Kazuki Inoue

Look at Boundary: A Boundary-Aware Face Alignment Algorithm

Author CVPR 2018 Poster

概要

顔の境界線を事前分布として使用することで、顔のランドマークを推定する手法を提案。既存手法でジゼ情報として使用されている顔のパーツは情報が離散的であり、 顔に対するセマンティックセグメンテーションであるface parsingは鼻に対する精度が良くない。 一方で顔の境界線は定義がはっきりしており、かつ顔の形状から推定することが可能。 提案手法では顔の境界線をstacked hourglassをベースとして、オクルージョンに対して頑健になるようにmessage passing layer、 推定精度の向上のためにadversarial netを導入している。推定された顔の境界線を元に、顔のランドマークを推定する。



新規性・結果・なぜ通ったか？

- 事前実験によって顔の境界線を用いたランドマーク推定がstate-of-the-artよりも優っていることを確認した上で手法を提案。
- 300W, COFW, AFLWなどのデータセットにおいてstate-of-the-arttと比較した結果、全ての場合において提案手法が優位となった。また境界線のGTを使用したランドマーク推定をOracleとして示しており、 Oracleによる推定精度が最も高くなった。
- WIDER Faceデータセットをベースにしたwider Facial Landmarks in-the-wild(WFLW)データセットを構築しており、10000枚の画像に対して98点のランドマーク、オクルージョン、メイク、照明環境、ブラー、表情のアノテーションを持つ。

コメント・リンク集

- 事前実験やOracleによって精度向上の理由が明確になっていルため、手法の優位性がはっきりと伝わってくる。
- 論文
- Project page(Supplementary material, Demo, Code)

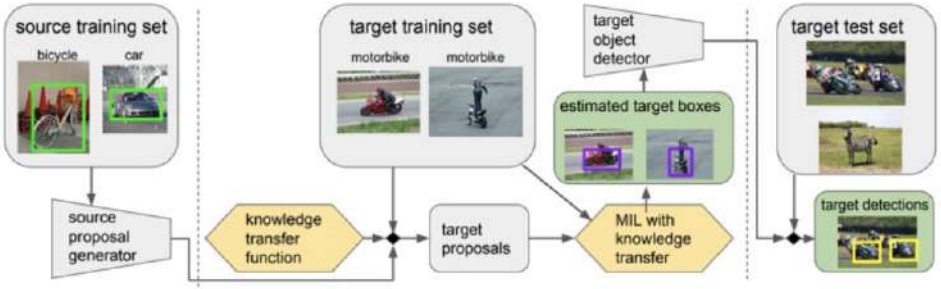
Munetaka Minoguchi

Revisiting knowledge transfer for training object class detectors

Jasper Uijlings, Stefan Popov, Vittorio Ferrari 1708.06128

概要

ソースクラスのBBoxアノテーションを使って、弱教師付きのトレーニング画像からターゲットの物体検出器を学習する知識転移手法の提案。まず、ソーストレインセットでproposal generatorをトレーニングし、それをターゲットトレインセットに適用。次に、画像のクラスラベル(Bboxなし)を使用し、知識転移でMultiple Instance Learning(MIL)を実行。 MILによって、物体検出器をトレーニングするために使用する、ターゲットクラス用のBBoxを生成。最後に、ターゲットの物体検出器をターゲットテストセットに適用。



新規性

物体候補とクラスを段階的に知識伝達していくフレームワーク。これにより、固有のクラスやジェネリックなクラスに渡る、広い知識伝達を可能にすることができる。

結果・リンク集

段階的な知識伝達によって、良質な物体候補を出力できる。

- 論文

[#52]

Hirokatsu Kataoka

Fight Ill-Posedness With Ill-Posedness: Single-Shot Variational Depth Super-Resolution From Shading

Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, Daniel Cremers CVPR 2018

概要

距離空間/距離画像の超解像を行う（Super-Resolution）を行う技術を提案。従来はShape-from-shadingにより行って来たが、形状の複雑性（誤りを含む）が存在していたため、これを改善する手法を提案した。

新規性・結果・なぜ通ったか？

距離画像における超解像を行うための最適化手法を提案した。結果は図に示すとおりである。

リンク集

- 論文
- Single-Image-Super-Resolution

[#53]

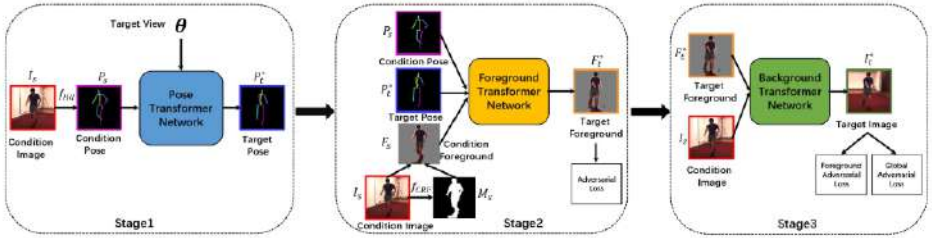
Hirokatsu Kataoka

Multistage Adversarial Losses for Pose-Based Human Image Synthesis

Chenyang Si, Wei Wang, Liang Wang, Tieniu Tan CVPR 2018

概要

人物の姿勢を事前情報として、ある視点の人物画像の入力からビューポイントを変更した人物画像を合成する手法を提案する。右図では3ステージのフレームワークについて示しており、最初のステージでは角度情報を挿入した姿勢変換、次のステージでは角度変化した人物にアピアランスを挿入、最後に背景を自然に挿入するステージ、という感じで変換が進んで行く。どう枠組みを実行するため、特にステージ２ではAdversarial Lossが、ステージ３では Foreground/Global Adversarial Lossを適用して誤差を計算する。



新規性・結果・なぜ通ったか？

評価は生成した画像のPSNR（シグナル・ノイズ比）、正解値との誤差SSIMを計算して、提案手法がもっとも優れた数値を出していることを明らかにした（SSIM: 0.72, PSNR: 20.62）。

コメント・リンク集

データセットの環境が固定だからできる？背景モデルの空間が非常に小さいので変換した際にもテクスチャが崩れずに生成できる？

- 論文

[#54]

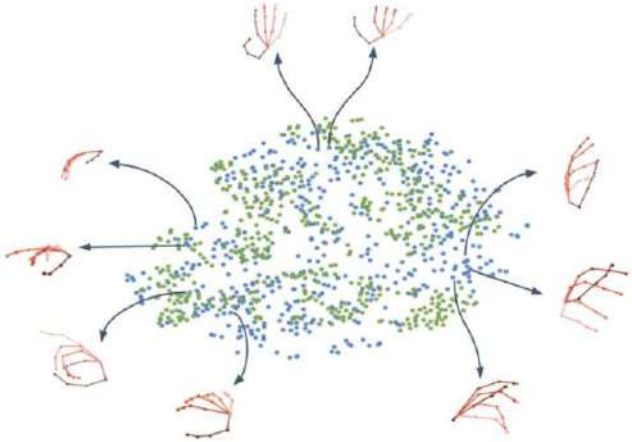
Hirokatsu Kataoka

Cross-Modal Deep Variational Hand Pose Estimation

Adrian Spurr, Jie Song, Seonwook Park, Otmar Hilliges CVPR 2018

概要

2次元画像と3次元手部モデルを同様の空間で扱うことができるCross-modal latent spaceを提案して、手部姿勢推定を実行する。別々にクラスタリングするのではなく、同一の空間で扱う（2DRGB-3D空間関係なく、同じ姿勢は同じような空間位置に投影される）方がマッチングの際にも便利。この特徴空間を学習するためにVariational Auto-Encoder（VAE）の枠組みで、Cross-modalのKL-divergenceを学習する。



新規性・結果・なぜ通ったか？

2D-3Dの共通空間を学習することで、2D画像からダイレクトに手部の3D関節点推定に成功した。距離画像との単一空間も学習可能とした。同一空間上で扱えるようにして、かつ従来法よりも精度向上が見られたため、CVPRに採択された。

コメント・リンク集

異なるモダリティを同一の枠組みで行ってしまう（2d-3dを同じ空間で）学習は他にもありそう？

- 論文

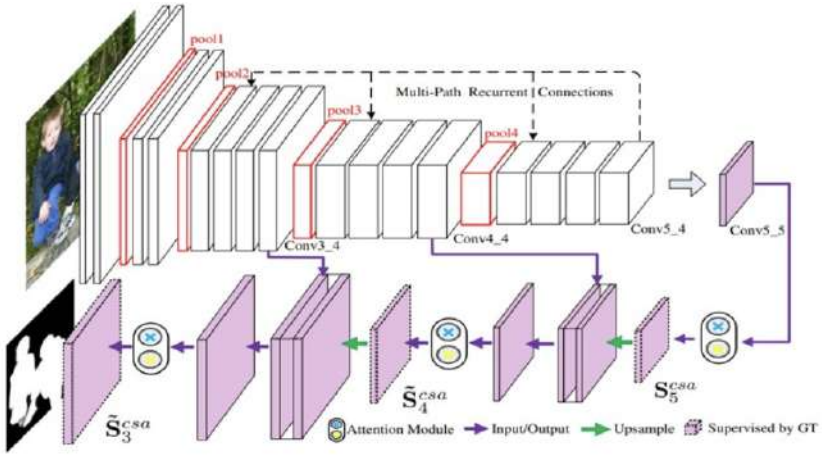
Munetaka Minoguchi

Progressive Attention Guided Recurrent Network for Salient Object Detection

Xiaoning Zhang, et al.

概要

マルチレベルのコンテクスト情報を選択的に統合する、顕著性のためのProgressive Attention Guided Recurrent Networkの提案。Attention Moduleを複数組み込み、その出力をステップ形式で統合していく。高レベルのfeatureを使って、低レベルのfeatureをガイドするイメージ。また、ネットワーク全体を最適化するための multi-path recurrent feedbackを提案。これにより、上部の畳み込み層からのセマンティック情報を、浅い層に転送することができる。



新規性

顕著性推定のための学習方法の提案。 従来のFCNベースの方法では、情報を区別せずに多レベルの畳み込み特徴を直接適用してしまうため、精度が上がらないと指摘。複数の層、複数のAttention Module出力を使い、コンテキスト情報を統合するので強力な特徴を抽出できる。

結果・リンク集

6種類のデータベースで精度評価。従来手法と比較して、ほぼ全てで最良の結果。

- 論文

[#56]

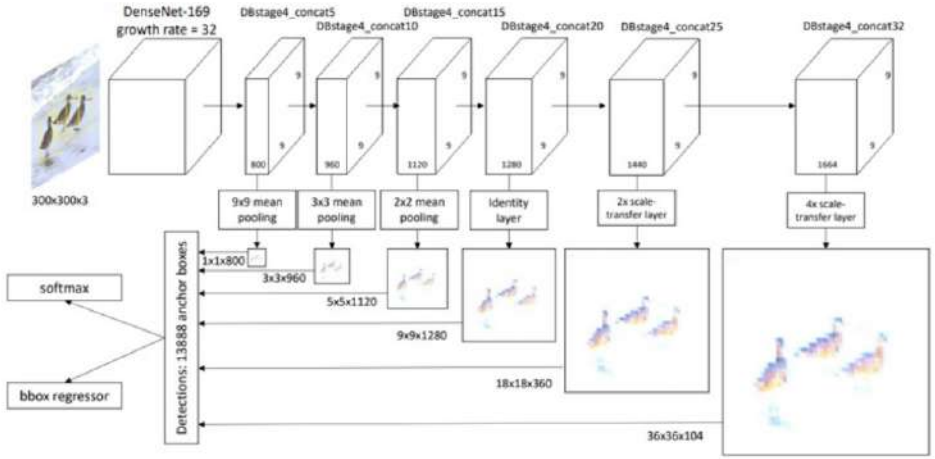
Munetaka Minoguchi

Scale-Transferrable Object Detection

Peng Zhou, et al.

概要

マルチスケールに対応した物体検出器であるScale-Transferrable Object Detection(STDN)の提案。STDNは DenseNet-169をベースとし、複数の物体スケールに対応するためのsuper-resolution layersを搭載。このsuper-resolution layersによってアップサンプリングすることで高解像度のfeature mapを得られるので小さな物体に対応し、大きな物体にはpooling層で対応する。



新規性

従来の物体検出手法では、様々なサイズのfeature mapを組み合わせるなどして、スケールに対応していたが、やはり小さな物体は苦手。本手法では、super-resolution layersという新たな手法によって改善を図る。

結果・リンク集

PASCAL VOCやMS COCOなどで精度向上を示している。個人的には、物体検出が苦手とする小さな物体に着目したデータセットなどを用意したうえで精度を比較してみたい。

- 論文

[#57]

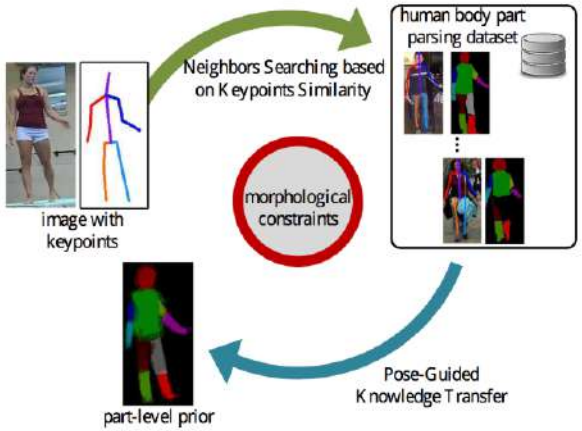
Hirokatsu Kataoka

Weakly and Semi Supervised Human Body Part Parsing via Pose-Guided Knowledge Transfer

Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, Cewu Lu CVPR 2018

概要

人物姿勢推定において「似たような姿勢はほぼ同じセグメント結果を保有する」という前提で弱教師付き/半教師あり学習を実行する。ある対象画像が入力された際にはほぼ同じ姿勢のデータをDBから検索して知識を転用（Pose-guided Knowledge Transfer）学習を実行する。その際に姿勢による拘束条件（Morphological Constraints）を入れ込むことでピクセルベースの姿勢のセグメンテーションを実行。モデルは全層畳み込みネット（Fully Convolutional Networks; FCN)を適用。



新規性・結果・なぜ通ったか？

弱教師付き学習（類似の姿勢を検索して対応づける）/半教師付き学習（少量のデータがあれば学習を実行）、いずれの手法でも姿勢学習を実行することができる。その上でデータ量を確保することに成功し、PASCAL-Part datasetにてmAPが3ポイント向上した。

コメント・リンク集

より少量のアノテーションで、かつ複数の枠組みで（本論文の場合は弱教師付き学習/半教師あり学習）学習が実行できる枠組みが増えてきた。そればかりか、教師あり学習のみよりも精度の高いものができあがりつつある。

- 論文
- 著者
- GitHub

Munetaka Minoguchi

Occluded Pedestrian Detection Through Guided Attention in CNNs

Shanshan Zhang, et al.

概要

オクルージョンに頑健な、Faster R-CNNベースの歩行者検出手法の提案。歩行者検出について解析することで、CNN特徴の各チャンネルがそれぞれ異なる身体部分を活性化していることに着目。(実際にチャンネルごとにアテンションを取ることで確認)各チャンネルが異なる身体部位を表現しているならば、オクルージョン発生時に身体部位の特定の組み合わせを定式化することができる。

新規性

歩行者検出器におけるCNN特徴について解析することで、歩行者に特化した物体検出を可能にしている。Faster R-CNNにAttention Networkを追加したアーキテクチャを提案。これにより、上位featureの重みパラメータを調節。

結果・リンク集

アーキテクチャをあまり複雑化せずに精度を向上させている。動物や虫などでも、CNNチャンネルごとに異なる身体部位を表現しているのだろうか。

- 論文

Kazuki Inoue

FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis

Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, Xiaoou Tang CVPR 2018 Poster

概要

IDを保った任意の顔向き画像をGANで生成するために、実画像ドメインと合成画像ドメインのそれぞれのIDを識別するclassifierを導入したFaceID-GANを提案。従来のGANではgeneratorとdiscriminatorが競い合うだけでclassifierは補助的な機能を果たしていたが、 提案手法におけるclassifierは実画像に対しては実画像ドメインのID番号を、 合成画像に対しては合成画像ドメインのID番号を識別させる、というようにデータセットに含まれるN個のラベルに対して、 2Nのラベル識別を行う。 他にも実画像のIDを表す特徴量と合成画像のIDを表す特徴量のコサイン類似度をロス関数として使用することで、異なるドメインに属する特徴量の類似度を高める。generatorには顔の形状特徴量、顔向き特徴量、ランダムノイズを入力とする。

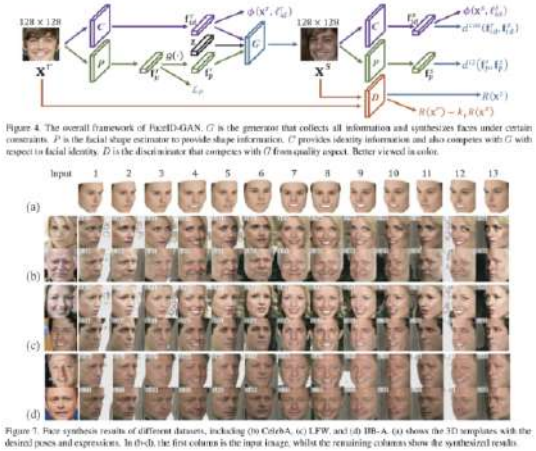


Figure 4. The overall framework of FaceID-GAN. G is the generator that collects all information and synthesizes faces under certain constraints. P is the facial shape estimator to provide shape information. C provides identity information and also competes with G with respect to facial identity. D is the discriminator that competes with G from quality aspect. Better viewed in color.

Figure 7. Face synthesis results of different datasets, including (b) CelebA, (c) LFW, and (d) IJB-A. (a) shows the 3D templates with the desired poses and expressions. In (b)-(d), the first column is the input image, whilst the remaining columns show the synthesized results.

新規性・結果・なぜ通ったか？

- 実画像、合成画像のそれぞれのドメインにおいてID識別を行うclassifierをGANに導入することで、generator VS. discriminator & classifier の構図を持つFaceID-GANを提案。
- CASIA-WebFace494414枚(10575人のID)の画像でトレーニングを行い、LFW, IJB-A, CelebA, CFPで検証した。
- state-of-the-artと横顔を入力とした正面顔画像生成、水平方向の視点移動、face verificationの精度を比較した結果、最も高い精度を達成した。

コメント・リンク集

- 論文
- Demo

Kazuki Inoue

Unsupervised Sparse Dirichlet-Net for Hyperspectral Image Super-Resolution

Ying Qu, Hairong Qi, Chiman Kwan CVPR 2018 Poster

概要

高解像度かつ短いスペクトルバンド幅で撮影された画像である hyper resolution hyperspectral image(HR HSI)を、HR HSIの正解データなしで、広いスペクトルバンド幅で撮影された高解像度画像(HR MSI)と、短いスペクトルバンド幅で撮影された低解像度画像(LR HSI)を用いて生成する手法を提案。 高解像度かつ短いスペクトルバンド幅で写真を撮影することはハードウェア的に困難であり、データセットの構築も難しい。 提案手法ではHR MSIとLR HSIをトレーニングデータとして2つのencoder-decoderを用いる。 HR MSIとLR HSIにはそれぞれ独立のエンコーダーが適用されるが、LR HSIから得られるスペクトル情報を共有するため、 デコーダーは共有する。 またスペクトル係数の総和は1という物理的な制約を実現するために潜在変数がディリクレ分布に従うようにする。 また推定されたスペクトルに対し得てスペクトル空間上の角度の差が小さくなるように学習を行う。

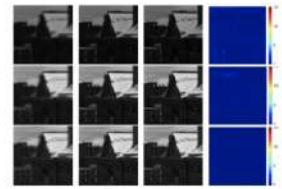


Figure 5. Reconstructed images from the CAVE (top) and Harvard dataset (bottom) at wavelength 460, 540 and 620 nm. First column: LR images (16 × 16). Second: estimated images (512 × 512). Third: ground truth images. Fourth: absolute difference.

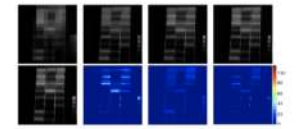


Figure 6. Reconstructed images of two examples (top two rows and bottom two rows) from the CAVE dataset at wavelength 670 nm. The first column shows the LR image (top) and the ground truth image (bottom). The second, third and fourth columns are the reconstructed results (top) and the absolute difference (bottom) from CSU, BSR and uSDN, respectively.

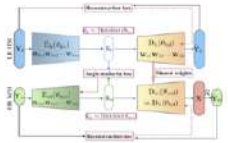


Figure 2. Simplified architecture of uSDN.

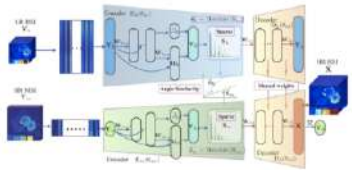


Figure 3. Details of the encoder nets.

新規性・結果・なぜ通ったか？

- CAVE、Harvardデータセットにて検証を行い、state-of-the-artとRMSE、SAM(スペクトル空間のベクトル類似性)比較して最も高い精度を達成。
- 教師無し学習が行えた理由として、古くから取り扱われている問題設定であったため、問題の性質をよく知っていたことがあげられる。

コメント・リンク集

- 論文

[#61] Yue Qiu

3D Semantic Segmentation with Submanifold Sparse Convolutional Networks

Benjamin Graham, Laurens van der Maaten, Martin Engelcke CVPR 2018 1248

概要

- スパース性が持ったデータ(ポイントクラウドなど)をより効率的で畳み込むsparse convolutional operationsを提案した．また，提案operationsを用いて新たな高次元スパースデータを有効的に処理できるsubmanifold sparse convolutional networks(SSCNs)を提案した．
- 従来の問題点：従来のCNNをsparse dataに用いたら計算及びメモリーの効率が良くない問題点がある．また，従来のスパースデータのためのネットワークは主に”full convolution”を行うためスパースデータをdilateしてしまう問題点がある．また，従来のCNNは層が深まることにより，active sitesが大幅に増加してしまうような“submanifold dilation problem”がある．
- 以上の様々な問題から，“ネットワークの異なる層で同じレベルのactive sitesのスパース性を保つ”をベースな考えとした新たなconvolution operations:SSCを提案した．こういうような性質から，SSCを用いたらより深い層構造持ったネットワークの学習を可能にした
- 具体的なssc：①プーリーングとstrided畳み込み操作と合併②入力のactive sitesだけに対して畳み込みし，active sitesを出力．Ground stateの入力を0と取り扱い畳み込みを廃棄のような設定がある

Figure 2: Example of “submanifold” dilation. **Left:** Original curve. **Middle:** Result of applying a regular 3×3 convolution with weights $1/9$. **Right:** Result of applying the same convolution again. Regular convolutions substantially reduce the feature sparsity with each convolutional layer.

Figure 3: SSC$(\cdot,\cdot,3)$ receptive field centered at different active spatial locations. Active locations in the field are shown in green. Red locations are ignored by SSC so the pattern of active locations remains unchanged.

Active	Type	C	SC	SSC
Yes	FLOPs	$3^d mn$	amn	amn
	Memory	n	n	n
No, $a > 0$	FLOPs	$3^d mn$	amn	0
	Memory	n	n	0
No, $a = 0$	FLOPs	$3^d mn$	0	0
	Memory	n	0	0

Method	Average IoU
NN matching with Chamfer distance	77.57%
Synchronized Spectral CNN [18]	84.74%
Pd-Network (extension of Kd-Network [10])	85.49%
Densely Connected PointNet (extension of [17])	84.32%
PointCNN	82.29%
Submanifold SparseConvNet (Section 6.5)	**85.98 %**

新規性・結果・なぜ通ったか？

- 提案のSSCがスパース性持ったデータの高効率CNNを可能にした．また，計算量とメモリー消耗の大幅削減及び深い層ネットワークの構築などに用いられる．

コメント・リンク集

- 論文がとても読みやすかった．しかし想像力が貧乏なので，うまくまとめられない．発表ビデオやコードで具体的なsparse convolutional operations操作を勉強したい

[#62]

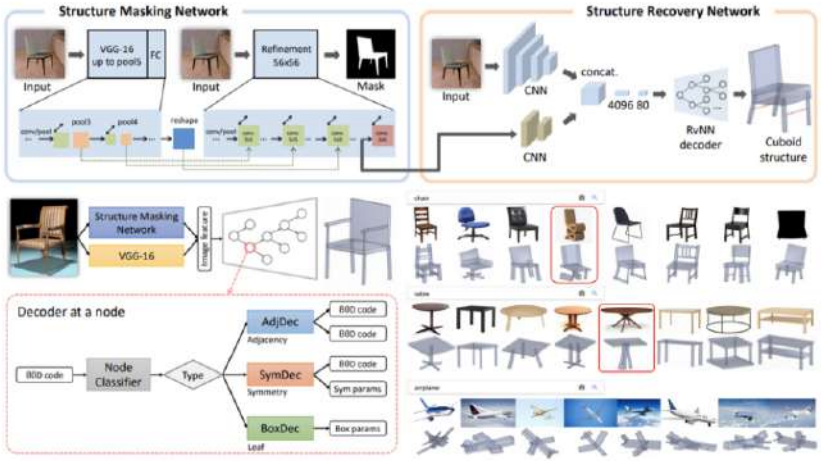
Yue Qiu

Im2Struct: Recovering 3D Shape Structure from a Single RGB Image

Chengjie Niu, Jun Li, Kai Xu CVPR 2018 578

概要

- 1枚のRGB画像から3次元形状構造(直方体で物体パーツを表示し，構造をパーツ間の接続性や対称性などの関係で表す)を復元するネットワーク構造を提案した．
- 従来1枚のRGB画像からボリューメトリックの復元が広く研究されている．しかし従来の様々な手法より復元された物体はトポロジーや構造が崩れる問題点が多く存在する（特に入力モデルの構造欠損がある場合）．提案手法は画像から形状構造復元を行うため，従来の体積復元の更なる精度向上や3次元形状構造の編集や高レベル画像編集など様々なところに応用できる．
- 提案手法のネットワークは①構造マスクを推定するネットワーク②再帰的オートエンコーダーを用いた直方形階層の構造復元ネットワークで構成される．具体的①はskip連結付きなマルチスケールCNNを用いた．②は①の抽出特徴及び元画像の特徴から再帰的なデコーダーを用いた．学習データは3D CADモデルからレンダリング及び構造抽出により作成した．



新規性・結果・なぜ通ったか？

- 提案手法が初めての1枚RGB画像から詳細3次元形状構造を復元する手法と指摘した．
- 提案の形状構造復元手法がパーツ間の連結や対称性など関係の復元を学習するので，復元された形状の構造の妥当性と汎用性が保証できる．
- 構造駆動型3次元体積補間及び構造awareなインタラクティブ画像編集の2つのアプリを開発し，提案手法により復元された形状構造の有効性および妥当性を示した．

コメント・リンク集

- 画像からの3次元形状構造復元がvolume復元と比べパラメータ数が圧倒的少ないので，問題自体の難しさも低い．しかし実応用を考えると，構造復元がかなり応用場面が多いと思う．問題設定がとても良いと思う
- 逆に今までどうしてやる人がなかったのが分からない

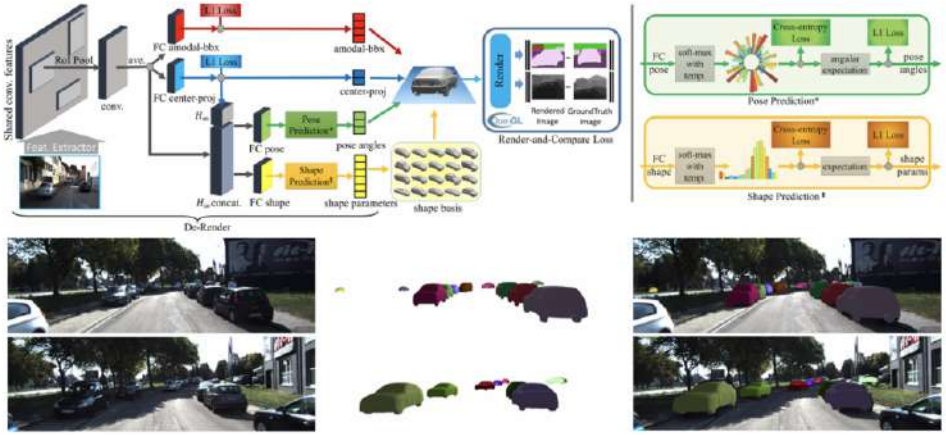
[#63] Yue Qiu

3D-RCNN: Instance-level 3D Object Reconstruction via Render-and-Compare

Abhijit Kundu, Yin Li, James Rehg CVPR 2018 436

概要

- RGB画像からインスタンスレベルの物体full3次元形状及び姿勢を行う”inverse graphics”なend-to-endなネットワーク構造の提案．物体のカテゴリ検出の結果が与えられたことを仮定し，画像中の物体2次元観測から物体の3次元パラメータの推定を行う．
- 提案手法の主な貢献としては①3次元表示：物体の3次元形状がクラス内で共通性が高いことから，大量なCADモデルから低次元なclass-specificな形状priorsを学習する．②2D-3Dマッピングを効率的行える新たなshape,poseの表示を提案した．(例:egocentricではなくallocentric視点を用いるなど)③提案手法を2D監督信号で学習可能にする予測した3次元形状を2次元にレンダリングし2次元のgtと比較することをベースとしたRender-Compareロス関数を提案した．



新規性・結果・なぜ通ったか？

- 従来のシーン理解は主にシーンに対しセマンティックセグメンテーションや物体検出などを行う．3次元空間のreasoningなどのタスクにおいては3次元のrepresentationが必要となる．また，従来の画像から3次元情報復元に関する研究は主に簡単なシーンから一つの物体に対し推定を行う．提案手法はより複雑なシーンの2次元画像から全部の物体インスタンスに対し3次元情報を推定できるため，自動運転の車・人の3次元情報推定などの様々な複雑なタスクに用いられる．
- ジョイント物体検出と姿勢推定、バウンディングボクス領域内の物体三次元姿勢推定の2つのタスクにおいて，Pascal 3D+,KITTIデータセットでstate-of-the-artな精度を達成した．

コメント・リンク集

- 今後”analysis by synthesis”,”inverse graphics”などの概念の引用が増やしそう
- かなり様々なところで工夫をしている．
- 論文

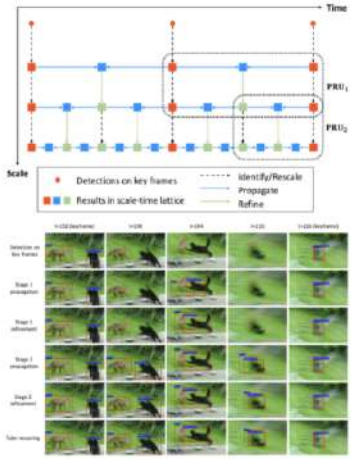
Yoshihiro Fukuhara

Optimizing Video Object Detection via a Scale-Time Lattice

Kai Chen et al. CVPR 2018

概要

動画中の物体検出において精度とコストの柔軟な trade-off が可能となる Scale-Time Lattice を提案. Propagation and Refinement Unit を用いて時間とスケールについての upsampling を階層的に行う. ImageNet VID dataset を用いた評価実験では先行研究と同等の精度の結果を Realtime で得られた.



新規性・結果・なぜ通ったか？

- Propagation and Refinement Unit は入力された 2つのフレームの中間の時間のフレームでの推定結果を Motion History Image [Bobick+ 2001] を用いて推定し, その結果をもとにより大きなスケールでの推定を行う.
- Propagation と Refinement を 2 段階行ったあとは, 残りの全フレームに対して線形補間を行う.
- 1段階目の入力となる Keyframe は, まず粗く一様にサンプリングした後, Keyframe 間の Propagation　の容易さ（物体の大きさが小さく, 動きが早いほど難しい）を評価し閾値を超えたら新しい中割りの Keyframe を動的に追加する.
- ImageNet VID dataset を用いた評価実験の結果は 20fps のとき 79.6mAP, 62fps のとき 79.0 fps と先行研究([Feichtenhofer+ 17] が 5fps で 79.8mAP)と同等の高い推定精度を維持したまま

コメント・リンク集

- [論文] Optimizing Video Object Detection via a Scale-Time Lattice
- [Project page] Optimizing Video Object Detection via a Scale-Time Lattice

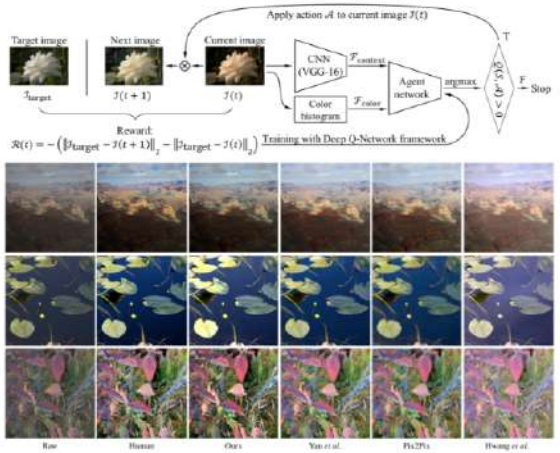
Yoshihiro Fukuhara

Distort-and-Recover: Color Enhancement using Deep Reinforcement Learning

Jongchan Park et al. CVPR 2018

概要

強化学習(DQN)を用いて automatic color enhancement を行う研究. 編集後の画像のみを利用して学習を行う方法（distort-and-recover scheme）を提案し, この学習方法の場合は従来の教師あり学習の手法よりも, 強化学習を用いる方が適していることを検証した. また, 評価実験では先行研究と同等か優位な結果を達成した.



新規性・結果・なぜ通ったか？

- color enhancement の工程をマルコフ過程としてモデル化し, 強化学習(DQN)を用いて解いた.
- 従来手法のように編集前後の画像の組では無く, 編集後の画像のみを利用して学習を行う方法（distort-and-recover scheme）を提案.
- action は様々な色調整の操作, reward は教師画像に特徴量がどれだけ近づいたかによって計算.
- MIT-Adobe FiveK dataset を用いた評価実験やユーザースタディーでは先行研究と同等か優位な結果を達成した.

コメント・リンク集

- [論文] Distort-and-Recover: Color Enhancement using Deep Reinforcement Learning
- [Project Page] Distort-and-Recover: Color Enhancement using Deep Reinforcement Learning

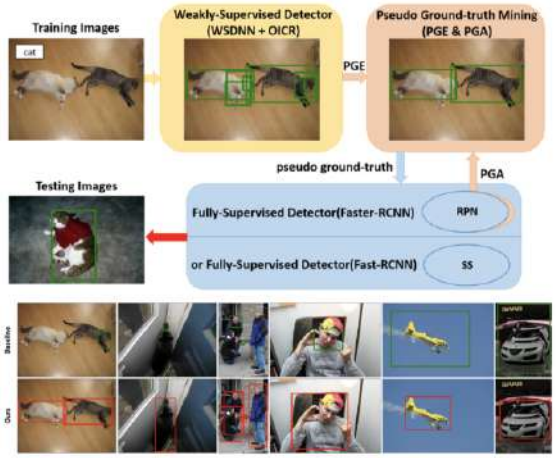
Yoshihiro Fukuhara

W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection

Yongqiang Zhang et al. CVPR 2018

概要

弱教師ありの物体認識の学習を使用して, 教師あり物体認識を学習を行う研究. 弱教師ありの物体認識は物体中の最も特徴的な領域や, 複数の領域を抽出してしまう傾向があるが, それらの結果から教師データとして最もらしい Pseudo ground-truth を生成する方法を提案. PASCAL VOC 2007 と 2012 を用いた評価実験では先行研究よりも優位な結果となった.



新規性・結果・なぜ通ったか？

- WSDNN [Bilen+ 16] の結果を OICR [Tang+ 17] を用いて改善したものを弱教師ありの物体認識の結果として使用.
- 上の結果に対して Pseudo ground-truth excavation (PGE) というアルゴリズムを適用することで, 物体全体を囲う Bounding Box を生成する.
- 更に, region proposal network [Ren+ 15] を用いて上の結果を改善したものを　Pseudo ground-truth とする.
- Pseudo ground-truth を用いて, Fast RCNN [Girshick 15] や faster RCNN [Ren+ 15] などの教師あり物体認識の手法の学習を行う.
- PASCAL VOC 2007, 2012 を用いて行った評価実験では先行研究 [Tang+ 17] [Krishna+ 16] と比較して mAP に置いて 5% 程度優位な結果となった.

コメント・リンク集

- [論文] W2F: A Weakly-Supervised to Fully-Supervised Framework for Object Detection

[#67] Yue Qiu

Learning Descriptor Networks for 3D Shape Synthesis and Analysis

Jianwen Xie, Zilong Zheng CVPR 2018 1093

概要

- 3次元ボリュームデータの形状特徴をモデリングできる深層畳み込みエネルギーベースなdescriptorネットワークを提案した．
- 提案の3D DescriptorNetがvoxelized形状の3D形状特徴を抽出できる．具体的には，voxelized形状のprobability density functionを定義した．また，3次元形状を特徴にマッピングできるボトムアップなボリューメトリックConvNetで特徴の統計またはエネルギー関数を定義した．
- 提案手法の貢献としては①ボリュームベースな3次元形状特徴をモデリングできる3D DescriptorNetを提案．②提案手法の学習プロセスをモードseeking,shiftingと解釈した．③形状検索に用いられるconditional 3D DescriptorNetを提案した．④3D形状生成モデルの新たな評価メトリクスを提案した．⑤3D GANを代替できる3D cooperative training schemeを提案した．

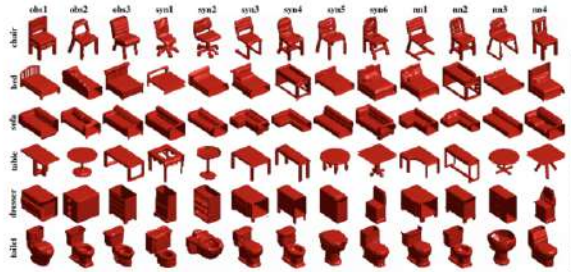


Figure 1: Generating 3D objects. Each row displays one experiment, where the first three 3D objects are some observed examples, columns 4, 5, 6, 7, 8, and 9 are 6 of the synthesized 3D objects sampled from the learned model by Langevin dynamics. For the last four synthesized objects (shown in columns 6, 7, 8, and 9), their nearest neighbors retrieved from the training set are shown in columns 10, 11, 12, and 13.

Table1. 3D object classification on ModelNet10 dataset

Method	Classification
Geometry Image	88.4%
PANORAMA-NN	91.1%
ECC	90.0%
3D ShapeNets	83.5%
DeepPano	85.5%
SPH	79.8%
VConv-DAE	80.5%
3D-GAN	91.0%
3D DescriptorNet (ours)	**92.4%**

新規性・結果・なぜ通ったか？

- 従来あまり提案されていないエネルギーベースな3次元形状descriptorを提案した．
- 提案の3D DescriptorNetを3次元形状生成，3次元形状検索，3次元形状スーパー解像度，3次元物体認識などタスクにおいて実験を行った．それぞれstate-of-the-artな性能を得られた．

コメント・リンク集

- コードで実際のネットワーク構造を確認したい．

- 論文

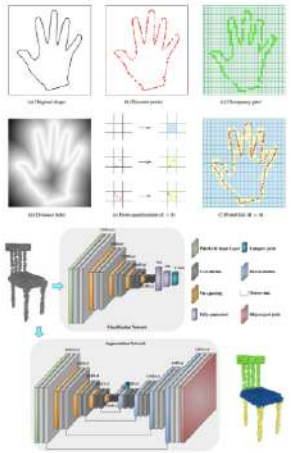
[#68]

Yue Qiu

PointGrid: A Deep Network for 3D Shape Understanding

Truc Le, Ye Duan CVPR 2018 1246

概要



- 3D CNNに用いられる新たな3次元データの表示方法(volumetric grid及びpoints表示をコンバインした表示方法)及び3DCNNネットワークPointGridを提案した．提案の3次元データ表示方法は畳み込みができるregular構造でありながら，ポイントクラウドのローカル幾何情報を抽出できる．
- 提案PointGridの処理ポロセスは：①ポイントクラウドを-1,1の区間のユニットボクスに正規化する②cellでユニットボックスを分割し，cellごとのポイント数をKまたは0にダウンサンプリング（増強の場合もある），cell内のKポイントのx,y,zを3チャンネルの特徴として取り扱う．③前述した処理後の表示を3D encoderまたは3D U-Netにより物体識別、パーツセマンティックセグメンテーションに適用する．

新規性・結果・なぜ通ったか？

- 従来の3次元表示方法の①occupacy gridやdistance fieldなどはレギュラー構造であるが，3次元形状の近似方法の特徴により低レベルの3次元局所情報しか表示できない，高レベルの特徴を表示するには高解像度が必要だが，CNNに用いたら処理・メモリーコストが極めて高くなる．②PointNetがポイントクラウドを直接CNN処理を行えるが，max poolingだけでグローバル特徴の抽出を行っているので，局所的な情報抽出が弱い．以上の問題点から， CNN処理を行えるグリッドとポイント表示をコンバインした構造を提案し，occupacy gridより低解像度で豊かな情報を表示でき， PointNetより局所的情報の抽出が強いPointGridを提案した．

コメント・リンク集

- PointNetの考え方を従来のボリューメトリック方法の解像度削減に利用し，16，16，16解像度でも良い性能を得られるのが魅力的
- 提案のPointGridが構造的簡潔でほかのネットワークにも前処理の一部として用いられそう
- 論文

[#69]

Hirokatsu Kataoka

Hybrid Camera Pose Estimation

Federico Camposeco, Andrea Cohen, Marc Pollefeys, Torsten Sattler CVPR 2018

概要

キャリブレーション済みのピンホールカメラにおいてカメラ姿勢推定問題を解く。例としてStructure-from-Motion (SfM)の2D-3Dマッチングを2D-2Dマッチングのように行う問題である。従来は構造ありの2D-3Dマッチングを解く絶対的なカメラ姿勢推定（absolute pose approaches）か、構造なしのテスクチャベースで2D-2Dマッチング（relative pose approaches）を行なっていたが、両者のいいとこ取りをする。本稿では新規にRANSACベースの手法を提案することで繰り返し最適化を行い、同問題の解決に取り組んだ。提案手法は、2D-3D/2D-2Dマッチングを同時にRANSACの要領で繰り返し最適化することができる（図を参照）。

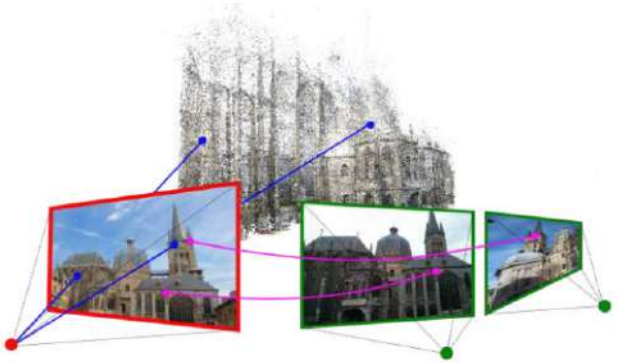


Figure 1. Visualization of 2D-2D matches (pink) and 2D-3D matches (blue) used by one of our hybrid pose solvers. The query camera is represented in red and SfM cameras in green.

新規性・結果・なぜ通ったか？

Structure-based/Structure-lessなマッチング（それぞれ2D-3D/2D-2Dに対応）を同時に解決する手法であるHybrid-RANSACを提案して、SfMの問題に対して適用した。両者のマッチングを単一の枠組みで実装しただけでなく、両者のいいとこ取りができる手法として完成させた。CVPRオーラルとして採択された。

コメント・リンク集

SfMのことはそこまで詳しくないのだが文章から「凄さ」が伝わってくる論文だった。

- 論文

[#70]

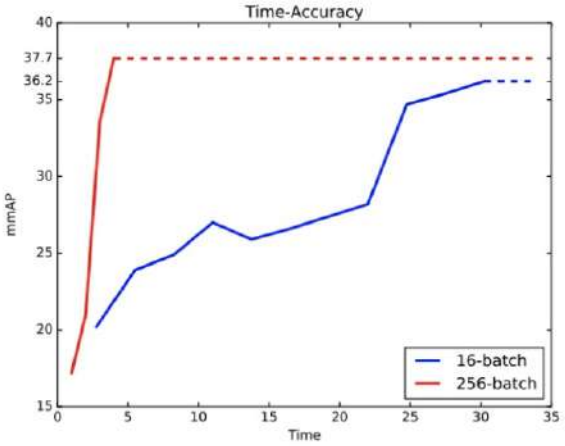
Munetaka Minoguchi

MegDet:A Large Mini-Batch Object Detector

Chao Peng, et al. 1711.07240

概要

16~256のような大きなバッチサイズでも学習することができる、物体検出手法MegDetの提案。ミニバッチ数を上げられることから、GPUを効率的に使用することができ、学習速度を向上。複数のGPUからうまくバッチ正規化を行う、Cross-GPU Batch Normalizationを提案。これにより、33時間の学習を4時間に短縮、かつ高精度にうまいこと学習できる。



新規性

2018年現在の著名な物体検出アルゴリズム(Faster R-CNNやMask R-CNNなど)は、全体のフレームワークやロスの設計に力を入れている。本研究では、手薄と思われるバッチサイズに着目し，新しいアプローチで精度向上を図っている。

リンク集

GPUの性能(メモリ数)の向上に伴って、この研究は生きてくる可能性がある。学習速度を上げながらCOCO2017一位はすごい。

- 論文

[#71]

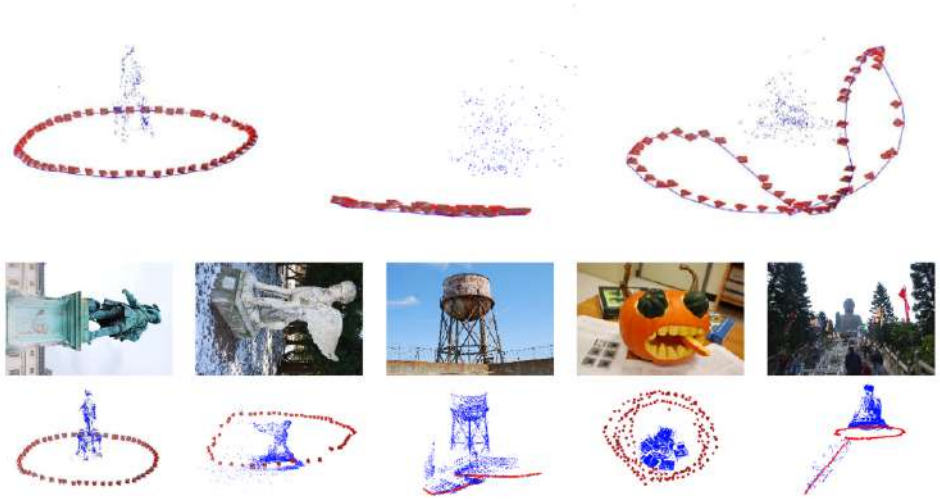
Hirokatsu Kataoka

Rotation Averaging and Strong Duality

Anders Eriksson, Carl Olsson, Fredrik Kahl, Tat-Jun Chin CVPR 2018

概要

本稿では非凸問題の一種であるRotation Averagingに対して Lagrangian Dualityを用いる。3次元再構成問題において、その画像群が「どこで、どのカメラ角度で、いつ撮影されたか？」に依存して再構成されるモデルが局所最適解に陥るという問題がRotation Averagingである（Rotation averaging）。図のようにカメラの移動軌跡やそのカメラアングルが変化した状態だと3次元再構成の局所解は大きく異なる（3次元再構成が表面のみ捉えていることに依存する）。



新規性・結果・なぜ通ったか？

Structure-from-Motion (SfM)の重要タスクであるRotation Averagingの問題解決についてLagrangian Dualityを用いた全体最適化（局所最適解をできる限りの場面で脱することができた）を行ったことがもっとも大きな新規性である。シンプル/スケーラブルなアルゴリズムであり、大規模空間に対するSfMにも応用可能である。結果は下の図の通りであり、局所最適解を脱してより詳細な形状復元を行うことに成功した。

コメント・リンク集

ディープラーニングを使っていない側の問題！SfMの未解決問題？であるRotation Averagingを高いレベルで改善している。

- 論文
- Rotation Averaging

[#72]

Kazuki Inoue

An Unsupervised Learning Model for Deformable Medical Image Registration

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, Adrian V. Dalca CVPR 2018 Poster

概要

脳の平均3D形状である脳アトラスの各ボクセルが患者の脳3次元データのどの位置に対応するか、という画像位置合わせ(image registration)をUnetを用いて正解データ無しの教師無し学習で行う手法を提案。 既存手法は最適化ベースだったが、学習ベースの画像位置合わせを初めて提案。トレーニング、検証で使用されているのは脳のMRIデータだが、 他のデータに対する画像位置合わせにも適用することが可能。

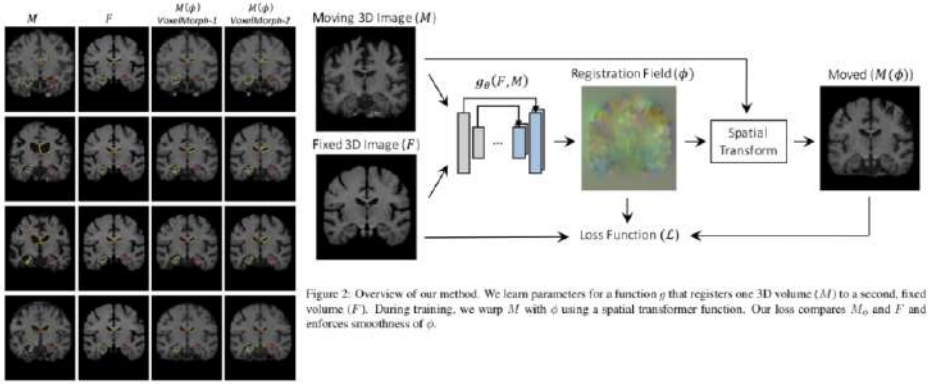


Figure 2: Overview of our method. We learn parameters for a function g that registers one 3D volume (M) to a second, fixed volume (F). During training, we warp M with ϕ using a spatial transformer function. Our loss compares M_ϕ and F and enforces smoothness of ϕ.

新規性・結果・なぜ通ったか？

- U-netを用いた学習ベースの3次元画像における画像位置合わせ手法を提案。
- 比較は最適化ベースの手法であるSyNと行った。 SyNと同等の精度を達成し、一方で実行時間はCPU上では約160倍、GPU上では更にその156倍の速度で実行可能。
- 教師無し学習のため出力された脳アトラスの全体的な形状は異なっているが、各器官の位置はかなり高い精度で推定できていることが驚き。

コメント・リンク集

- 選択分野の勝利？手法に新規性は無く、検証で比較した手法も2008年のものとかなり古いが、それでも同等の精度で実行時間が速くなれば、それはCV分野としてはOKと判断されたのか？
- 論文
- GitHub

[#73]

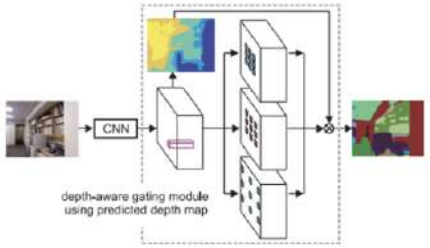
Hiroaki Aizawa

Recurrent Scene Parsing with Perspective Understanding in the Loop

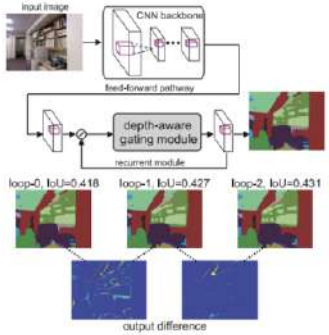
Shu Kong, Charless Fowlkes CVPR2018

概要

固定解像度で処理する画像認識システムでは、遠近感を持つシーンの画像において物体が任意のスケールを持つことが問題となる。(距離によって物体のスケールが変わる。カメラから遠いほど物体は小さく、近いほど大きい。)これ解決するために、物体のスケール(Depthに反比例)によってPoolingサイズを可変にするdepth-aware pooling moduleを提案。遠くの物体の細部は保持され、近くの物体は大きな受容野を持つことができる。 Depth画像は与えられるか直接RGB画像から推定され、Depth情報と意味的予測を利用するRecurrent Refinement Moduleにより、Semantic Segmentationを反復的に精錬する。



depth-aware gating module



recurrent refinement module

新規性・結果・なぜ通ったか？

受容野のサイズを変化させるためにDepth情報を利用しこれを自然にCNNに組み込んだこと(geometricな情報を利用する先行研究はあり)。またこのDepth予測をSemantic Segmentationと互いに補い合う用にRecurrent Refinement Moduleを組み込んだこと。NYU-depth-v2の単眼深度推定においてstate-of-the-artな性能とSemantic Segmentationの性能改善を確認。

コメント・リンク集

Recurrent refinement moduleのLoopにより物体の事前情報を捉えることができるが、Loopによる精度変化が小さい。Curriculum Learningと組み合わせるとおもしろそう。ResNetから得られる特徴はすでにスケールを考慮した特徴が抽出できているようにも思え、depth-aware pooling moduleが活かされているかというと疑問。

- 論文
- Project Page
- GitHub

[#74]

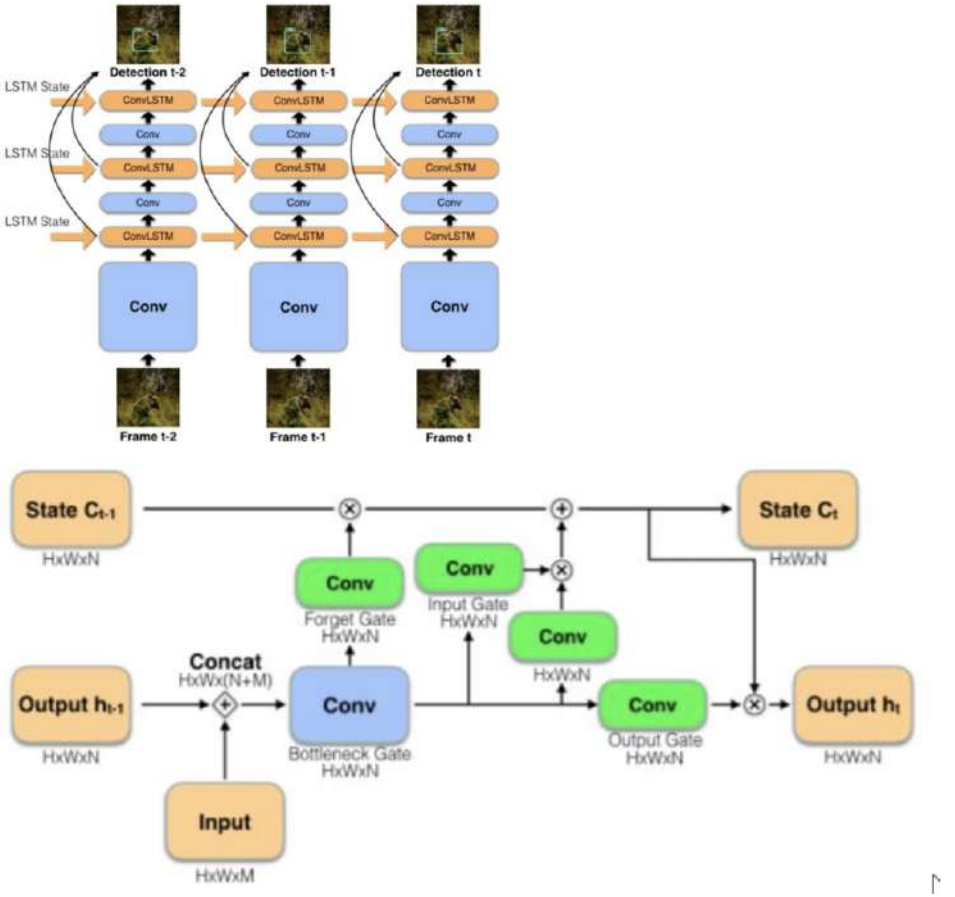
Ryosuke Araki

Mobile Video Object Detection with Temporally-Aware Feature Maps

Mason Liu and Menglong Zhu CVPR2018 698

概要

モバイルや組み込み機器上で低消費電力かつリアルタイムに動作する物体検出のオンラインモデル．Single-Shotベースの物体検出モデルとLSTMを組み合わせたモデルである．また，通常のLSTMよりも計算コストを大幅に削減できるBottleneck-LSTMを提案する．Bottleneck-LSTMは，NチャンネルのBottleneck特徴マップ（Bt）を計算してすべてのゲートの入力をBtに置き換える．これによるゲート内の計算が減る．LSTM自体をDeepな構成にしても標準LSTMより効率的な計算が可能である．



新規性・結果・なぜ通ったか？

従来のVideo object detectionはフレームごとの検出に依存しているため，時間的情報を利用することができなかったが，本研究では検出器の速度を犠牲にせず時間的な情報を組み込んだ．ImageNet VIDデータセットでmobilenet-SSDよりも高精度（54.4mAP）に検出可

コメント・リンク集

Googleでのインターン成果とのこと．リアルタイム検出は時系列情報があれば精度がよくなるが，それを入れることで速度の低下が起きてしまうのでこの2点のトレードオフになっている？

[#75]

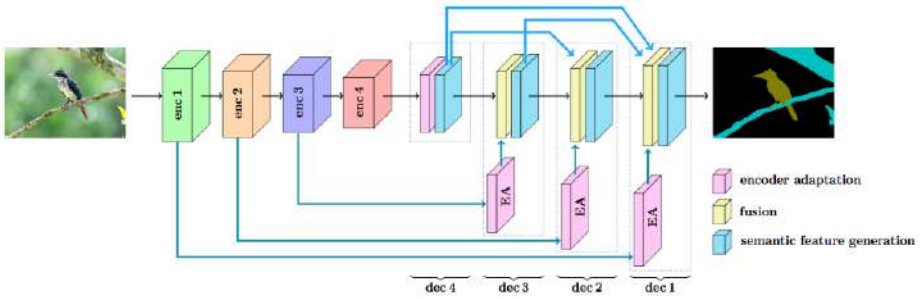
Hirokatsu Kataoka

Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation

Piotr Bilinski, Victor Prisacariu CVPR 2018

概要

ResNeXtを用いたEncoder-Decoder（エンコーダ-デコーダ）構造、かつシングルパスのセマンティックセグメンテーション手法を提案する。エンコーダとデコーダは折り返したような構造になっており、エンコーダの特徴は図のように対称となる/同じサイズのデコーダ位置に統合される（enc1-dec1が対応）。今回は特にデコーダ側に改善があり、(1)コンテキスト情報を抽出、(2)セマンティック情報を生成、(3)異なる解像度の出力を適宜統合という新規性がある。これを実現するため、DenseNetを参考にしたDense Decoder Shortcut Connectionsを提案し、デコーダにおいてコンテキスト特徴を全て後段に渡すようにした。



新規性・結果・なぜ通ったか？

デコーダにおいてDenseNetを参考にしたDense Decoder Shortcut Connectionsを提案、コンテキスト情報を後段に渡して精度を向上させた。ResNeXtの構造適用と合わせて各データセットにてState-of-the-artな精度を達成。NYUD datasetにて48.1（mean IoU）、CamVid datasetにて70.9（mean IoU）となった。PascalVOC2012においても81.2であった（SoTAはPSPNetの82.6）。

コメント・リンク集

セマンティックセグメンテーションの覇権争いが激化。ここら辺まで精度が向上すると確率的にSoTAになったりならなかったりする（回す回数が多いと一回くらい精度が高いモデルが学習される）？逆に、学習しやすい（誰が、どんなパラメータで回しても同じくらいの精度が出る）アーキテクチャというのが提案されてもよいかも。

- 論文
- ResNeXt(facebookresearch)
- DenseNet
- PSPNet

[#76]

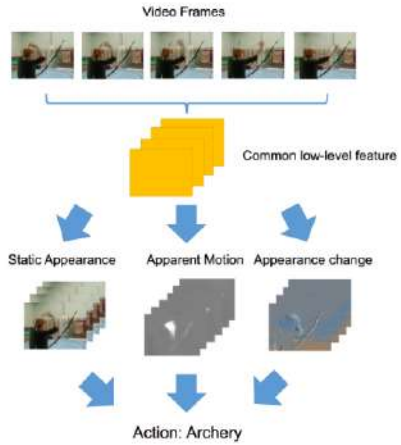
Hirokatsu Kataoka

Recognize Actions by Disentangling Components of Dynamics

Yue Zhao, Yuanjun Xiong, Dahua Lin CVPR 2018

概要

人物行動認識のための表現に対して、モーションとアピアランスの共起表現（Disentangling Components of Dynamics）を提案する。従来の人物行動認識に限らず動画認識ではRGBを入力とするアピアランス、オプティカルフローを画像に投影したフロー画像が用いられていたが、本論文ではそれらの共起表現を新たに提案した。フロー画像とは異なり、特に「アピアランスの変化」をカラー付きで表現できる。さらに、3Dプーリングを提案し、上記３つのチャンネルからの特徴を蓄積する手法についても考案した。



新規性・結果・なぜ通ったか？

人物行動認識の文脈において、新規の特徴表現方法であるDisentangling Components of Dynamicsを提案した。同手法はフローとは異なり、RGB値の変化を効果的に捉える方法である。さらに、3Dプーリングも提案し、RGB/Flowも合わせた3チャンネルの特徴を適切にプーリングすることができる。フルモデルを用い、さらにKineticsにて事前学習を行った実験では、95.9%@UCF101を達成、従来の行動認識の大部分よりも高い精度を実現。

コメント・リンク集

Kinetics Datasetの事前学習特徴が（やはり）強い。ImageNetでは91.8%だったものがImageNet+Kineticsで95.9%。転じて、やはりアルゴリズムなどよりもデータを用意するのがもっとも効果的。

- 論文
- 著者

[#77]

Ryosuke Araki

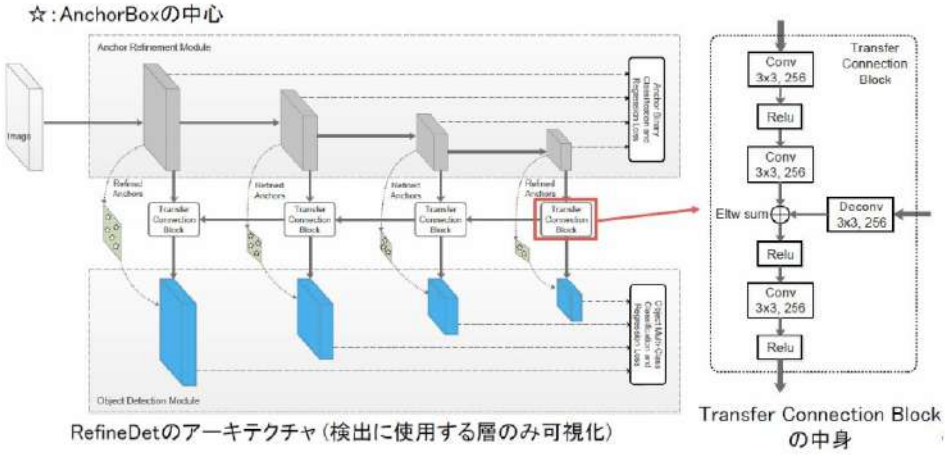
Single-Shot Refinement Neural Network for Object Detection

Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li CVPR2018 545

概要

SSDをベースにした2つのモジュールから構成されるSingle-shotベースの物体検出アルゴリズム「RefineDet」を提案．Anchor Refine Module (ARM) とObject Detection Module (ODM) と呼ばれるモジュールと，2つを繋いで特徴マップを転送するTransfer Connection Block (TCB) からなる．ARMは物体が存在しない領域を示すNegative Anchor(※)の削減や，Anchorの粗い調整を行う．ODMはTCBを通じて特徴マップを受け取って座標の回帰およびクラス推定を行う．

※物体候補領域を示すBounding-boxをAnchorと呼ぶ．SSDでDefault boxと呼ばれているものと同じ．



RefineDetのアーキテクチャ（検出に使用する層のみ可視化）

Transfer Connection Block の中身

新規性・結果・なぜ通ったか？

SSDで細かい物体をより精度よく検出するために，一度畳み込んだ特徴マップをDeconvしたりUp samplignしたりする手法がいくつかあるが，この手法はTCBで特徴マップを転送するときに1つ前 (=出力側) の特徴マップをDeconvして足している．Single-shotでありながら2つの役割分割されたモジュールがうまく連携している．推論速度は入力320x320で24.8ms (40.3FPS)，512x512で41.5ms (24.1FPS) @TITAN Xと非常に高速である．精度もDSSDより高性能 (VOC2007: 83.8mAP, MSCOCO: 41.8AP)である．

コメント・リンク集

Single-Shotベースの物体検出は前層の特徴マップを持ってくる系が流行り？精度も良い．

- arXiv
- GitHub

[#78]

Kazuki Inoue

Neural Kinematic Networks for Unsupervised Motion Retargetting

Ruben Villegas, Jimei Yang, Duygu Ceylan, Honglak Lee CVPR 2018 Oral

概要

異なるキャラクタに対するモーションのリターゲティングをRNN、Cycle consisteny lossを用いることで教師なしで学習する手法を提案。RNNのencoder-decoderを用いて入力された関節位置、局所座標の原点の4次元モーションから、 各関節のクォータニオンと局所座標の4次元モーションを出力しそれをForwad Kinematicsによってターゲットキャラクターに転写する。 これを教師なしで行うためにCycle consistency loss、GAN lossを導入する。 これによって同じモーションを持った異なるキャラクタのデータが無い場合にも、モーションのリターゲティングを行うことが可能となる。

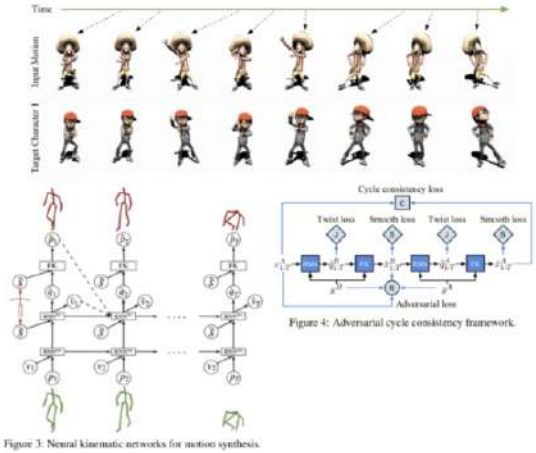


Figure 3: Neural kinematic networks for motion synthesis.

Figure 4: Adversarial cycle consistency framework.

新規性・結果・なぜ通ったか？

- RNNのencoder-decoder、Cycle consistency lossを用いることで同じモーションを持った異なるキャラクタのデータが無い場合にも、モーションのリターゲティングが可能な手法を提案。
- モーションのリターゲティングはオンラインで実行可能。
- Mixamo animation dataを用いて、トレーニングは同じモーションを持たない７体のキャラクタの計1646のモーションを使用し、テストには６体のキャラクタを使用した。
- RNN、RNNからrecurrent connectionを削除したMLP、入力モーションを単純にコピーした結果、ablation testを行い推定された関節位置のMSEを比較した結果、提案手法が最も高い精度を達成した。
- 特に入力モーションを単純にコピーした場合にはターゲットキャラクタの足が床中に埋まってしまったが、提案手法ではこれを防ぐ

コメント・リンク集

- クォータニオンの出力で止めているのは、クォータニオンがスケルトンに不変であることと、ボーンの回転角を制限するロス関数twist lossを取るためだと考えられる。
- 異なるキャラクタで同じモーションのGTがあるようなので、教師あり学習との比較を見てみたかった。一方でことモーションに関しては数値的には悪くても見た目では良し悪しがつかないということもあるので、これを考慮したのかもしれない。
- Most of this work was done during Ruben' internship at Adobe.
- 論文

Kazuki Inoue

Cross-Domain Weakly-Supervised Object Detection through Progressive Domain Adaptation

Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, Kiyoharu Aizawa CVPR 2018 Poster

概要

インスタンスレベルのアノテーションを持つソースドメイン(S)とイメージレベルのアノテーションを持つターゲットドメイン(T)を用いてdomain adaptationを行い、Tに対する物体検出を行う手法を提案。Sを用いて物体検出器のプリトレーニングを行い、 Cycle GANによってSをTに変換した画像を用いて物体検出器のfine-tuningを行う。 続いてSとそのイメージレベルのアノテーションを用いて半教師学習を行いSに対する物体検出を行う。 半教師学習を行う際にインスタンスレベルのアノテーションが施されたデータセットが必要なため、 クリップアート、水彩画、漫画のデータセットの構築も行っている。

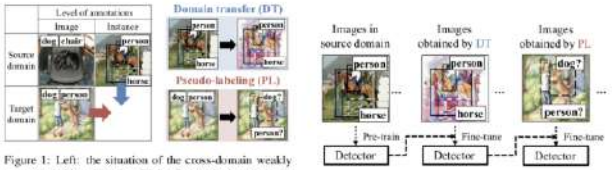
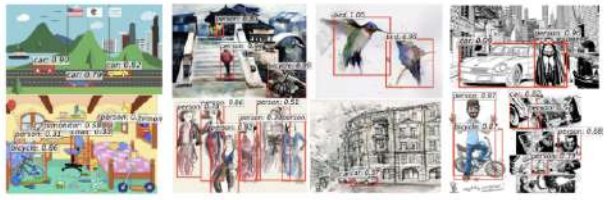


Figure 1: Left: the situation of the cross-domain weakly supervised object detection; Right: Our methods to generate instance-level annotated samples in the target domain.

Figure 3: The workflow of our framework.



(a) Clipart1k (b) Watercolor2k (c) Comic2k

Figure 5: Example outputs for our DT+PA in the test set of each dataset. We only show windows whose scores are over 0.25 to maintain visibility.

新規性・結果・なぜ通ったか？

- Cycle GANによる検出器のfine-tuning、半教師学習による物体検出というステップをへてイメージレベルのアノテーションを持つ実画像ではないドメイン(クリップアートなど)に対する物体検出手法を提案。
- Clipart1k, Watercolor2k, Comic2kという、それぞれクリップアート1000枚、水彩画2000枚、漫画2000枚の画像に対してインスタンスレベルのアノテーションを施したデータセットを構築。
- 自ら構築した三種のデータセットにおいて教師なし学習、半教師学習、SSD300、YOLOv2と比較した結果、最も高い精度を達成。

コメント・リンク集

- 検証しているラベル数が最大でも20と少ないことが気になった。これはターゲットドメインの構築が難しかったからであり、データさえあればラベルを増やすことができるのだろうか？
- 論文
- Project page
- GitHub

[#80]

Kazuki Inoue

Real-Time Monocular Depth Estimation Using Synthetic Data With Domain Adaptation via Image Style Transfer

Amir Atapour-Abarghouei, Toby P. Breckon CVPR 2018 Poster

概要

合成画像とそのデプス画像、そして実世界画像を用いてunsupervised domain adaptaionを行うことで、実世界画像に対するデプス画像を生成する手法を提案。 実世界画像に対するデプスのアノテーションは困難であり、かつ枚数も多くない。 一方合成画像に対するデプスのアノテーションは完璧だが、 実世界画像に対する推定を行うときにドメインシフトが起きてしまう。 提案手法ではUnetによって合成画像からデプスを推定し、Cycle GANによって実世界画像を合成画像に変換することでデプスを推定する手法を提案。 GPUを用いることで44FPSで実行することが可能。

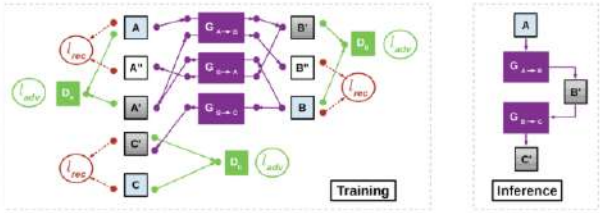


Figure 2: Our approach using [38]. Domain A (real-world RGB) is transformed into B (synthetic RGB) and then to C (pixel-perfect depth). A, B, C denote ground truth, A', B', C' generated images, and A'', B'' cyclically regenerated images.

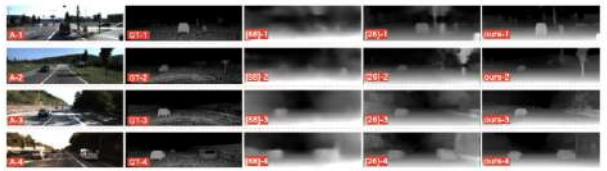


Figure 3: Qualitative comparison of our results against the state-of-the-art methods in [38, 23] over the KITTI split. GT denotes ground truth. Our approach produces sharp and crisp results with no blurring or additional artefacts.

新規性・結果・なぜ通ったか？

- ラベルなし実世界画像とラベルあり合成画像に対してCycle GANによるスタイルトランスファーによりdomain adaptaionを行うことで、実世界画像のデプスを推定する手法を提案。
- 合成画像、KITTIデータセットでトレーニングを行い、KITTIデータセットの推定精度をstate-of-the-artと比較した結果、最も高い精度を達成。
- Cycle GANによるスタイルトランスファーでは急激な照明変化や影を物体として認識してしまうといったリミテーションが存在する。

コメント・リンク集

- Cycle GANによってdomain adaptationを行う割合ベーシックな手法だが、その推定精度がstate-of-the-artに優っている。
- 論文
- Project page
- Vimeo

[#81]

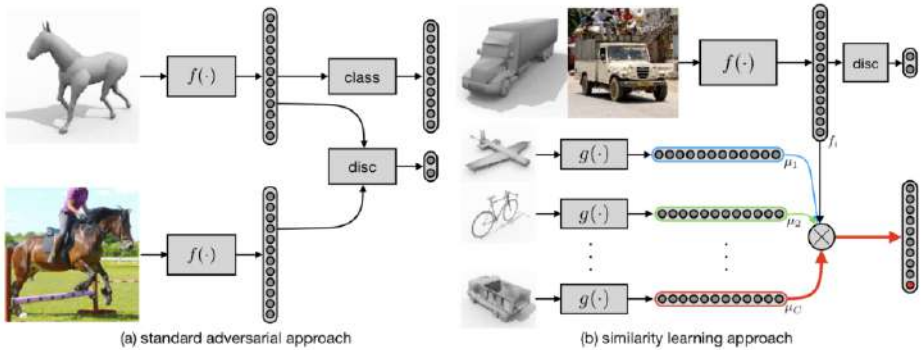
Kazuki Inoue

Unsupervised Domain Adaptation with Similarity Learning

Pedro Pinheiro CVPR 2018 Poster

概要

ソースドメイン(S)の各カテゴリの重心ベクトルと、S・ターゲットドメイン(T)から得られたadversarial featuresの行列積を用いることでdomain adaptation(DA)を行う手法を提案。 従来のDAではSとTのそれぞれから得られる特徴量をGANによってdomai-confusionを行い、 Sで学習したラベル識別器をTに適用するという手法だった。提案手法ではadversarial-confusionに加えて、 Sの各カテゴリにおける重心ベクトルとgeneratorから得られる特徴量の類似度を高くするように学習しDAを行う手法を提案。



新規性・結果・なぜ通ったか？

- domain-confusionに加えてラベルごとの重心ベクトルとgeneratorから得られる特徴量の類似度を高くするように学習しDAを行う手法を提案。
- MNIST・USPS・MISNT-M、Officde-31, VisDAデータセットで検証。11のdomain adaptationにおいて、9つの設定においてstate-of-the-artよりも高い精度を達成。

コメント・リンク集

- この論文に限らずDAを提案する論文ではdomain-confusionを可視化しており、数値評価だけではなく、ドメインの分布の可視化画像を載せることも重要だと思われる。
- 論文

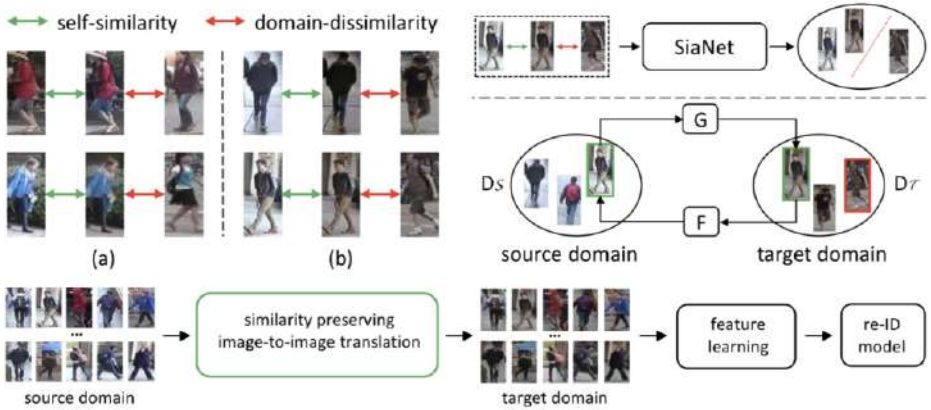
Kazuki Inoue

Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification

Weijian Deng, Univ. of Chinese Academy; Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, Jianbin Jiao CVPR 2018 Poster

概要

人物認証(person re-ID)の精度が落ちないようにソースドメインの人物画像をターゲットドメインの画像に変換するSimilarity Preserving GAN(SPGAN)を提案。ドメイン間の変換をCycleGANで行う。 またそれぞれのperson re-IDのデータセットには基本的に同じ人物は写っていないということを利用して、 ソースドメインとターゲットドメインで異なるデータセットを使用し、 ターゲットドメインへと変換された画像はIDが保たれ、かつターゲットドメインのどの人物のIDとも一致しないように学習を行った。



新規性・結果・なぜ通ったか？

- person re-IDデータセットの特徴を生かしドメイン変換された画像はターゲットドメインの人物画像とは一致せず、かつ元々のIDを生かすように学習を行い、ドメイン間で人物画像の変換を行うSPGANを提案。
- Market-1501、Duke-MTMC-reIDデータセットで検証を行い、一方のデータセットの人物画像をもう一方のドメイン画像に変換した際に正しくre-IDができるのかを検証した。
- ベースラインであるCycleGANや教師なし学習のstate-of-the-artと比較して最も高い精度を達成。

コメント・リンク集

- person re-IDのタスクの中でもソースドメインの人物がターゲットドメインに存在する場合にも発見する、というタスクを解いている。
- ソースドメインとターゲットドメインに含まれるIDが全く違う、ということを逆手にとった手法。
- 論文

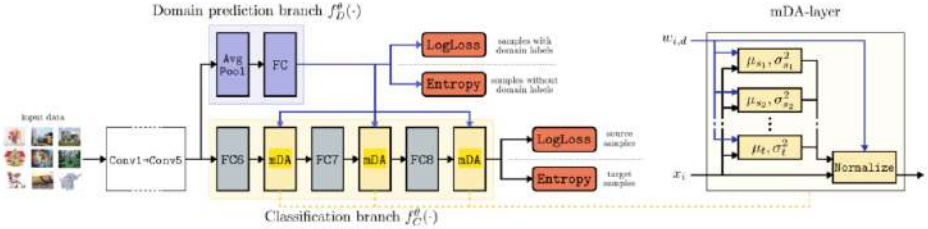
[#83]

Kazuki Inoue

Boosting Domain Adaptation by Discovering Latent Domains

Massimiliano Mancini , Lorenzo Porzi, Samuel Rota Bulò, Barbara Caputo, Elisa Ricci CVPR 2018 Poster

概要



domain adaptaion(DA)に対して、ソースデータは潜在的に複数のドメインで構成されていると仮定し、ソースサンプルがどのドメインに所属しているかを精度よく識別するためにMulti-domain DA layer(mDA-layer)を導入することで、 ターゲットのラベルの識別精度を向上させる手法を提案。 実験ではmulti-soure domain adaptationを行うことでその有効性を検証している。 ソースデータないのドメインを識別するCNNの特徴量を用いることで、ターゲットドメインのラベル識別の精度が向上している。

新規性・結果・なぜ通ったか？

- mDA layerによってマルチソースドメイン内のドメインを識別する学習を行うことで、ターゲットドメインのラベル識別に有効な特徴量を獲得。
- MNIST・MISNT-m・USPS、Office-31、Office-Caltech、PACSデータセットで提案手法の有効性を検証。state-of-the-artのmulti-source domain adaptation(DA)よりも高い精度を達成。
- ソースサンプルにドメインのラベルが全くない場合とラベルがない場合でも、精度は1%ほどしか変わらない。

コメント・リンク集

- 論文

[#84]

Kazuki Inoue

Large Scale Fine-Grained Categorization and the Effectiveness of Domain-Specific Transfer Learning

Yin Cui Yang Song, Chen Sun, Andrew Howard, Serge Belongie CVPR 2018 Poster

概要

鳥の種族などより細かいラベルを推定するdomain-specific fine-grained visual categorization(FGVC) taskにおいて、効果的なトレーニングデータセットの構築方法を提案。 事前実験からターゲットドメインの画像の見た目に近い画像を含むソースドメインでトレーニングするほど、 識別精度が高くなるということを発見している。ターゲットドメインに含まれる画像の見た目に近い画像を多く持つソースドメインのクラスをいくつか選択することで トレーニングデータセットを構築する。画像の見た目はEarth Mover's Distanceで測定され、 7つのfine-grainedデータセットにおいて提案手法が効果的であることを示した。

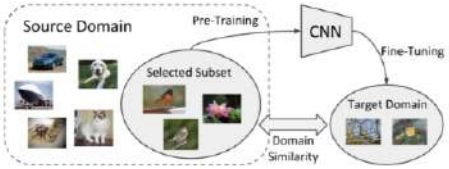
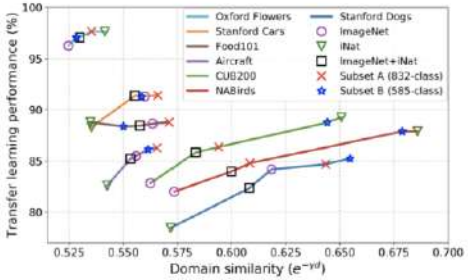


Figure 1. Overview of the proposed transfer learning scheme. Given the target domain of interest, we pre-train a CNN on the selected subset from the source domain based on the proposed domain similarity measure, and then fine-tune on the target domain.



新規性・結果・なぜ通ったか？

- FGVCを行う際のトレーニングスキームとして、ImageNetのような大規模データセットやクラスごとのデータ数が偏っているiNatを学習するのではなく、 より効果的なトレーニングデータセットを構築する手法を提案。
- fine-grainedデータセットCUB200、Stanford Dogs、Flower-102、Stanford Cars、Aircraft、Food101、NABirdsで検証した結果、5つのデータセットにおいて提案手法によって構築されたトレーニングデータセットで 学習した場合に最も高い精度を達成。
- classificationで使用したネットワークはResNet、Inception、Squeeze-and-Excitationであり識別ネットワーク自体には依存しないことも検証している。

コメント・リンク集

- 手法自体は単純ながら、事前実験に基づく論文展開や既存手法に対して投げかけた疑問を回収できたところが評価されたと思われる。
- 論文

[#85]

Kazuki Inoue

Residual Parameter Transfer for Deep Domain Adaptation

Artem Rozantsev, Mathieu Salzmann, Pascal Fua CVPR 2018 Poster

概要

ソースドメインを学習したネットワークのパラメタを残差ブロックで変換することでターゲットドメインへdomain adaptaionを行う手法を提案。 既存手法ではドメインに普遍な特徴量を学習していたためにネットワークのパラメタが多すぎてしまう。 提案手法は学習時には残差ブロックとソースドメインを学習するネットワークのファインチューニングを行い、 ソースドメインに対するラベルの識別と2つのドメインに対してadversarial domain adaptationを行う。

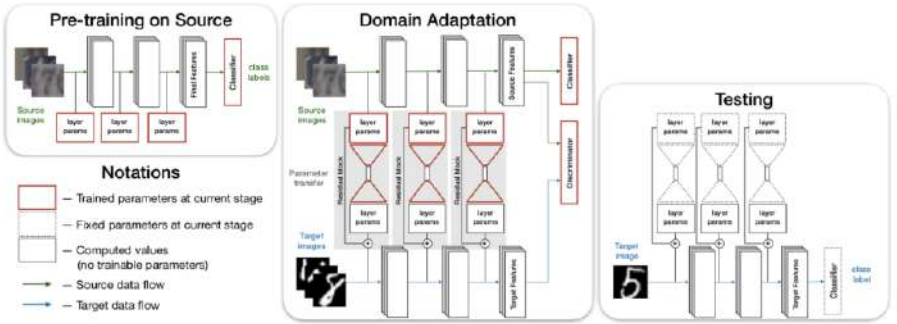


Figure 2: **Approach overview.** We first pre-train the network on the source data. We then jointly learn the source stream parameters and their transformations using adversarial domain adaptation. Finally, at test time, we use the network with transformed parameters to predict the labels of images from the target domain. (Best seen in color)

新規性・結果・なぜ通ったか？

- ドメインに普遍な特徴量を学習するのではなく、ソースドメインを学習したネットワークの重みをソースドメイン用に変換することでパラメタ数を抑えかつ精度の高い domain adaptationを実現。
- state-of-the-artと比べて、SVHN・MNIST、UAV-200データセット、Officeデータセットにおいてもっとも高い精度を達成。
- ソースドメインを学習するネットワークがResNetのような深いネットワークの場合にも有効であることを主張。

コメント・リンク集

- 論文

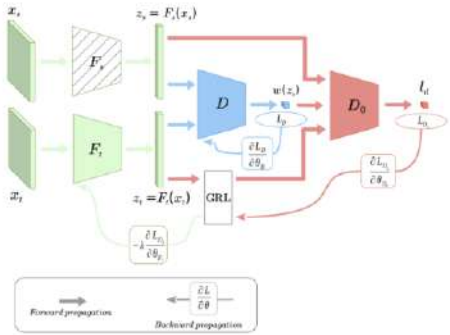
[#86]

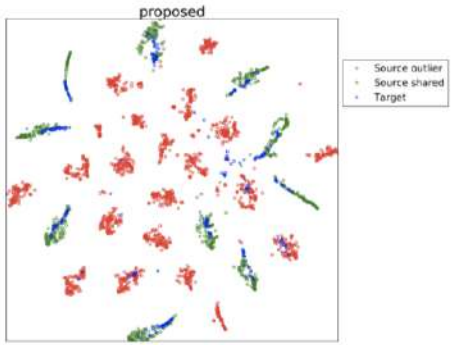
Kazuki Inoue

Importance Weighted Adversarial Nets for Partial Domain Adaptation

Jing Zhang, Zewei Ding, Wang Ding, Wanqing Li, Philip Ogunbona CVPR 2018 Poster

概要





ターゲットドメインがソースドメインが所持するクラスの一部しか持たずかつラベルがない場合であるpartial domain adaptationをadversarial netベースで行う手法を提案。 adversarila netの手前いにドメインを識別するclassifierを用意し、 このclassifierが精度良く判別可能なソースサンプルはターゲットドメインには含まれていないクラスに所属している可能性が高いので重みを小さくし、 逆にconfidenceが低いソースサンプルはターゲットにも存在するクラスに所属している可能性が高いので重みを大きくする。 この重みとソースサンプルを掛け合わせたものとターゲットサンプルをadversarial netで学習させる。

新規性・結果・なぜ通ったか？

- 4つのドメインを持つOffice+Caltech-10において、ソースは各ドメインで10のラベル、ターゲットは各ドメインで5つのラベルを使用。同様の設定でOffice-31データセット、Caltech256→Office10データセットで実験を行った。
- partial domain adaptationのstate-of-the-artであるSANと比較して8つの実験のうち4つの設定でより高い精度を達成。
- SANではソースのクラスの数だけclassifierを必要とするが、提案手法で必要なclassifierは2つのみ。

コメント・リンク集

- 論文

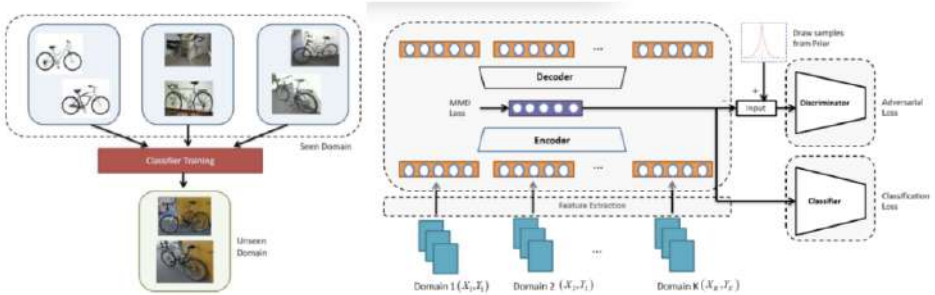
[#87]

Kazuki Inoue

Domain Generalization with Adversarial Feature Learning

Haoliang Li, Sinno Jilain Pan, Shiqi Wang, Alex Kot CVPR 2018 Poster

概要



Adversarial Autoencoder(AAE)に対してMaximum Mean Discrepancy(MMD)を導入することでトレーニングデータを過学習することなくdomain generalizationを行う手法を提案。 domain generalizationとは、複数ドメインのラベル付きデータセットを学習し、 テスト時にはデータセットに含まれていないドメインのデータセットにおける識別や生成タスクを行うことを指す。 複数のソースドメインで不変な特徴量を取得するmulti-task learningに対して、提案手法ではMMDベースでドメイン間の差分をとることと、AAEによって特徴量空間に対して事前分布が押し込むことでソースドメインに対する過学習が防ぐ。

新規性・結果・なぜ通ったか？

- AAEに対してMMDを組み込むことで、ソースドメインを過学習することなくdomain generalizationを行う。
- domain generalizationのstate-of-the-artと識別タスクにおいて比較。
- MNISTを15度刻みで回転させた場合の認識精度、VLCSデータセットにおける物体認識、IXMASにおける行動認識においてstate-of-the-artよりも高い精度を達成。
- AAEにおける事前分布の違いによる精度も議論しており、ラプラシアン分布が最も精度が良かったと主張。

コメント・リンク集

- 論文

[#88]

Kazuki Inoue

Adversarial Feature Augmentation for Unsupervised Domain Adaptation

Riccardo Volpi, Pietro Morerio, Silvio Savarese, Vittorio Murino CVPR 2018 Poster

概要

特徴量空間におけるデータオーギュメンテーションとソースドメインとターゲットドメインに不変な特徴量を取得することでunsupervised data adaptationを行う手法を提案。 右図にあるようにstep1で、ソースドメインとノイズをデコードして生成されたベクトルをGANにかけ、 特徴量空間においてソースドメインに対するオーギュメンテーションを行う。 続いてstep2において、ソースドメインとターゲットドメインを同一のエンコーダーに入力することでドメインに不変な特徴量を取得する。 ベースラインである Adversarial discriminative domain adaptationではドメインごとにエンコーダーを使用していたが、提案手法ではエンコーダーは一つ。

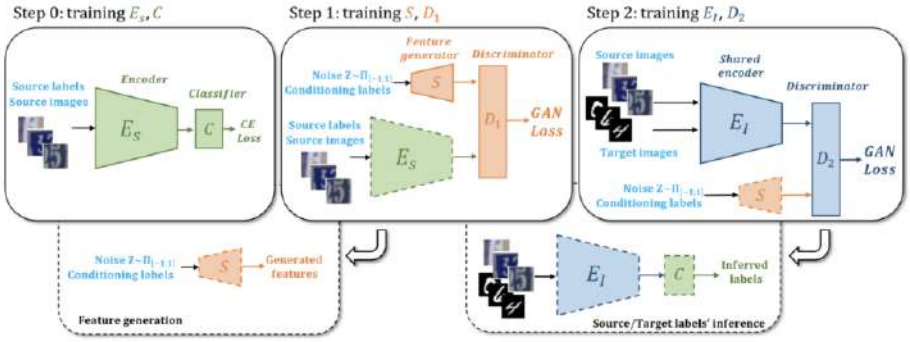


Figure 1. Training procedure, representing the steps described in Section 3.1. Solid lines indicate that the module is being trained, dashed lines indicate that the module is already trained (from previous steps). All modules are neural networks, whose architectures are detailed in Section 5.1. Smaller, dashed panels in the bottom indicate how to generate features (*left*) and how to infer source or target labels (*right*).

新規性・結果・なぜ通ったか？

- GANを用いてソースドメインの特徴量空間でデータオーギュメンテーションを行い、かつソースドメインとターゲットドメインに不変な特徴量を推定することで、unsupervised data adaptationを行った。
- ベースラインであるAdversarial discriminative domain adaptationに対して上記の2つの拡張の有効性を議論している。
- state-of-the-artと比較して、数字の識別、物体の識別において既存手法と同等かそれ以上の精度を達成。

コメント・リンク集

- Limitationにも書かれているようにsourceとtargetのラベルが同じになる保証はなく、最終的な精度はsourceのエンコーダーがどれほどうまく学習できているかに強く依存する。
- 論文
- GitHub

[#89]

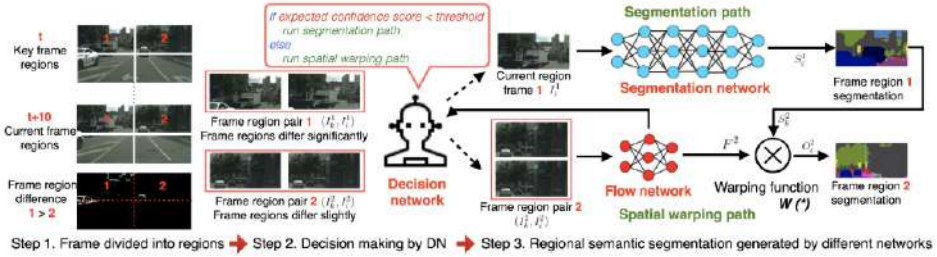
Hirokatsu Kataoka

Dynamic Video Segmentation Network

Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, Chun-Yi Lee CVPR 2018

概要

動画像セグメンテーションの問題に対してネットワーク選択（Decision Network）を行い適応的にCNNモデルを処理するDynamic Video Segmentation Network (DVSNet)を提案する。同手法では性質の異なるふたつのネットワーク（深くて精度が高いが低速/浅くて精度は低いが高速）を組み合わせて交通シーンにおけるシーン解析にて高速な処理を実現する。



新規性・結果・なぜ通ったか？

DVSNetは低速なもので70.1%/20fps、高速なものだと65.2%/34.4fps（いずれもCityScapes datasetにて処理した結果）を達成する。両者を、トレードオフを考慮してあらゆる場面に適応することができるという意味で新規性がある。

コメント・リンク集

こういう通し方があったのか、と勉強になる。実利用を想定し、トレードオフを考慮、それを解決することも重要な問題である。

- 論文
- 論文（査読ver.）

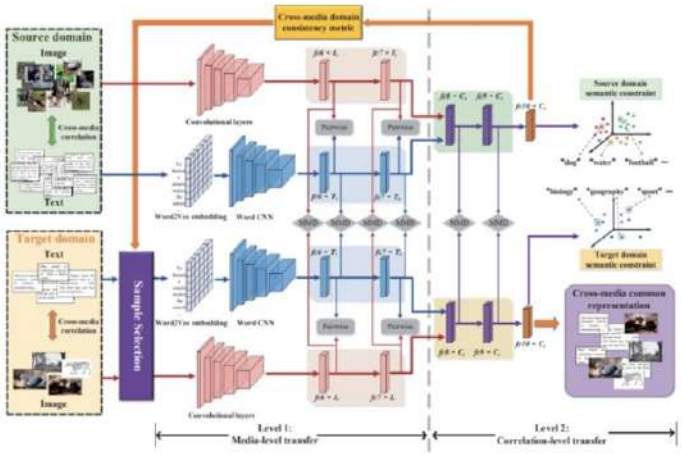
Munetaka Minoguchi

Deep Cross-media Knowledge Transfer

Xin Huang, et al. 1803.03777

概要

画像とテキストなどの異なるメディアタイプ間で検索する、クロスメディア検索手法のcross-media knowledge transfer(DCKT)の提案。大規模なクロスメディアデータセットの知識を、小規模なデータセットのモデルに転移学習する。メディアレベルと相関性レベルでのドメインの違いを最小化するために、2レベルでドメイン変換することで精度向上。また、ドメインの違いを徐々に減らすようにトレーニングサンプルを選択することで、モデルがより頑健になる。



新規性

マルチメディア分野における検索。既存の手法では、ラベル付きデータを学習する方法が多いが、大規模なデータの収集とラベル付けは手間取るため問題とされる。そこで、既存のデータを転移して解決する。

リンク集

- 論文

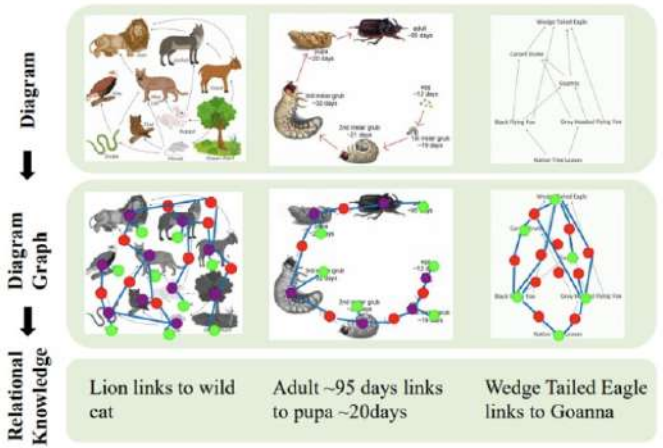
Munetaka Minoguchi

Dynamic Graph Generation Network: Generating Relational Knowledge from Diagrams

Daesik Kim, et al. 1711.09528

概要

視覚情報とテキストの情報が抽象的に統合された図であるダイアグラムを解析するためのunified diagram parsing network(UDPnet)の提案。入力は様々なイラストやテキスト、レイアウトを持つ図のみ。物体検出器によって、図内のグラフ構造を推論し、新手法であるdynamic graph generation network(DGGN)によってグラフを生成。生成されたグラフからテキストで関係性を出力する。



新規性

ダイアグラムのような図には、豊富な知識が含まれているが、固有の特性やレイアウトの問題から、コンピュータに自動的に理解させる方法はあまり提案されていない。本手法では、物体検出器やRNNを統合し、ダイアグラムから知識をテキストとして生成する。

結果・リンク集

自然画像でなく，人間による作為的なグラフ理解において優れている。人間の意図や、人間にとって自然な解釈を学習できているのではないか。

- 論文

[#92]

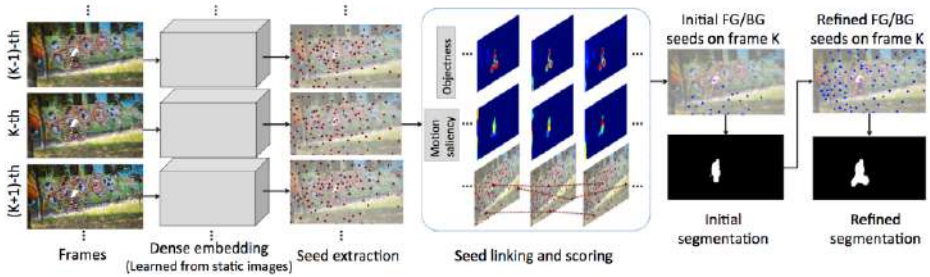
Hirokatsu Kataoka

Instance Embedding Transfer to Unsupervised Video Object Segmentation

Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, C.-C. Jay Kuo CVPR 2018

概要

物体インスタンス特有の特徴（同じ物体領域に属しているか？）を捉えることでビデオに対する教師なしの物体セグメンテーションを実施する。ここでは静止画で捉えた特徴を、ビデオに表れる物体候補/オプティカルフローと組み合わせて物体のインスタンスセグメンテーションを実施。本論文ではさらに、ビデオに対するfine-tuningなしに高精度なセグメンテーション手法を構築したと主張している。



新規性・結果・なぜ通ったか？

静止画の学習パラメータを動画に適用していく、その際に物体候補/オプティカルフローと統合していくことで動画的な表現を教師なしで獲得していく。DAVIS datasetを用いた評価で78.5%、FBMS datasetにて71.9%（いずれもmean Intersection-over-Union (mIoU)の評価にて）を達成し、それぞれのデータセットでState-of-the-art。

コメント・リンク集

"Without finetuning"というのもアピールになるということを勉強した（ただしそれでstate-of-the-artである必要がある？）。

- 論文
- GitHub(Semi-Supervised Object Segmentation)

[#93]

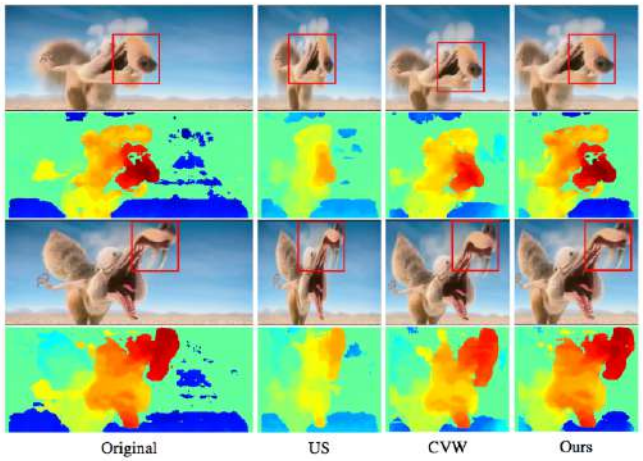
Hirokatsu Kataoka

Depth-Aware Stereo Video Retargeting

Bing Li, Chia-Wen Lin, Boxin Shi, Tiejun Huang, Wen Gao, C.-C. Jay Kuo CVPR 2018

概要

ステレオビデオ（Stereo Video）に対するリターゲティング（Retargeting）を扱う。ステレオ（かつビデオ）に対するリターゲティングは従来のリターゲティングと比較すると、動画中の顕著性が高い物体の把握やダイナミクスを含むためまだ新しくチャレンジングな課題である。ここに対して、Depth-aware Fidelity Constraint（距離画像から推定される信頼性のようなもの）を適用することで物体の顕著性を把握しつつ3次元空間を再構成することができる（リターゲティングと3次元再構成の同時推定問題）。最適化にはTotalCost関数を適用して物体の顕著性を把握しつつ形状、時間情報、距離画像のディストーションを推定。



新規性・結果・なぜ通ったか？

ステレオビデオの入力から、顕著性の把握、形状推定、時間情報、距離画像のディストーションを同時推定し、従来法であるCVWよりも綺麗なリターゲティング画像を生成することに成功した。

コメント・リンク集

VR/AR、3D映画などに使える！より自然に見せることで映像酔いを軽減することができる？

- 論文
- 著者
- リターゲティング（マイナビ記事）
- CVW（従来手法）

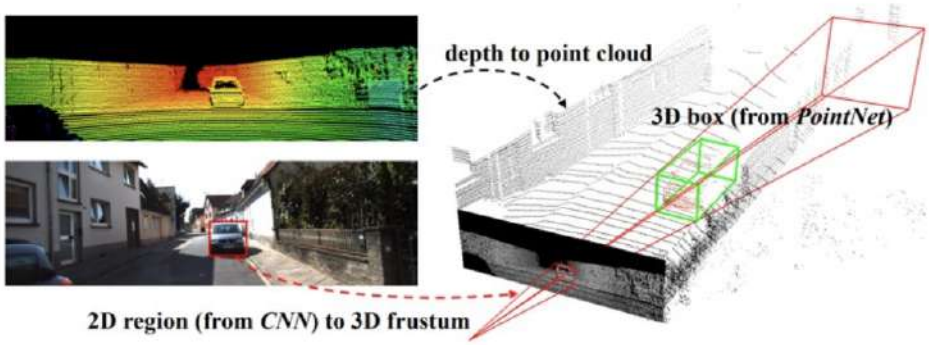
Munetaka Minoguchi

Frustum PointNets for 3D Object Detection from RGB-D Data

Charles R. Qi, et al. 1711.08488

概要

屋内および屋外シーンにおける3D物体検出手法のfrustum PointNetsの提案。まず、RGBデータからCNNで2Dの物体候補領域を推定する。次に、点群の深度情報を用いて、各物体領域の視錐台(viewing frustum)を推定する。最後に、frustum PointNetsによって3Dバウンディングボックスを推定。



新規性

従来の手法では、画像や3Dボクセルに処理を加えて、3Dデータの自然なパターンや不変性を曖昧にしている。本手法では、RGB-Dスキャンによって生の点群データを直接操作する。

結果・リンク集

2Dと3Dで別々のネットワークを使うことで、小さな物体やオクルージョン、まばらな点群についても正確に推定することができる。リアルタイムも実現。

- 論文

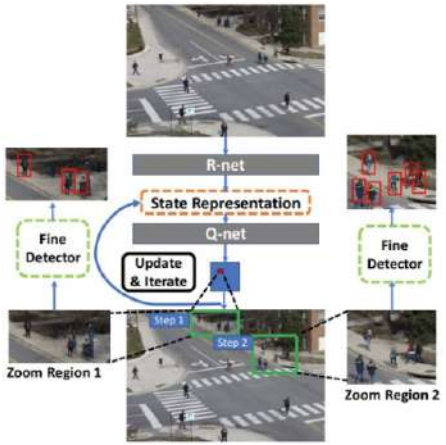
Munetaka Minoguchi

PhaseNet for Video Frame Interpolation

Mingfei Gao, et al. 1711.05187

概要

高解像度画像に出現する様々なサイズの物体を、精度の維持と処理コストの低減を実現しながら検出するフレームワークの提案。最初はダウンサンプリングされた粗い画像から、次に高解像度の細かい画像から検出する。強化学習を用いた2つのネットワークで構成。R-net：低解像度の画像を入力し、その検出結果を用いて高解像度領域を解析する。これにより、どの順番にズームインすべき判断できる。Q-net：ズームの履歴を使用し、拡大領域を順次選択。



新規性

しっかり検出する範囲を絞ることで処理量を低減、効率化を図ることができる。基本的な検出の構造はいじっていない。処理する画素数を約70%、処理時間を50%以上短縮し、なおかつ高い検出性能を維持できる。

結果・リンク集

YOLOやSSDなどの物体検出手法の精度向上にも使える。

- 論文

[#96]

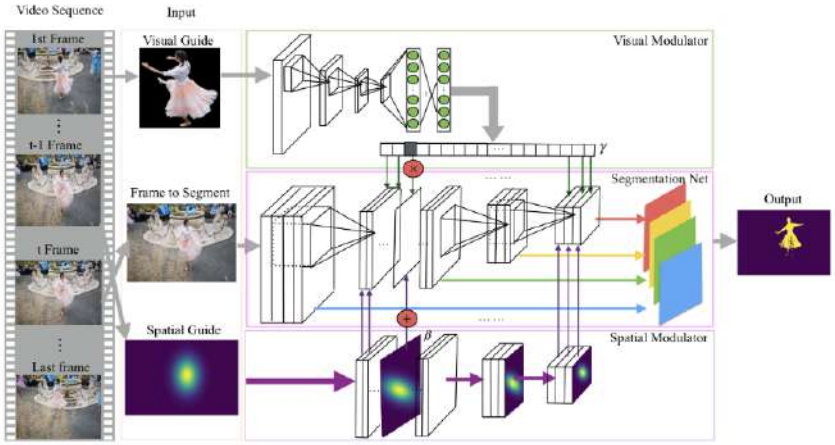
Hirokatsu Kataoka

Efficient Video Object Segmentation via Network Modulation

Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, Aggelos K. Katsaggelos CVPR 2018

概要

セグメンテーションを実行する際に任意のアノテーション済み物体を事前情報（Spatial Prior）として高精度化を図るための技術を提供する。本論文では、最初の一フレームに対してセグメンテーションを行うだけで、動画中の物体に対してセグメンテーションを行うモデルを提案する。アノテーションから抽出した事前情報はニューラルネットの中間層にて情報を挿入して抽象化を行う。図は提案のフレームワークを示しており、VisualModulator（初期フレームのアノテーションから視覚的なガイドを行う）、SegmentationNet（VisualModulator/SpatialModulatorの補助を受けつつ、RGB画像の入力からセグメンテーションを実行）、SpatialModulator（空間的にどこらへんに対象物体があるかをサポート）の３つのコンポーネントから構成される。



新規性・結果・なぜ通ったか？

最初のフレームのアノテーションのみから動画セグメンテーションを実行するという問題を提供した、さらに視覚的な特徴量/位置的な事前知識をセグメンテーションのネットワークに導入し、動画セグメンテーションを高精度化した点が評価された。動画セグメンテーションタスクであるDAVIS2016にて74.0、YoutubeOjbsにて69.0（処理速度は0.14second/image）であった。State-of-the-artには劣る（それぞれ79.8, 74.1）が、処理速度では優っている（提案0.14 vs. 従来 10.0）。

コメント・リンク集

メタ学習の枠組みを使用している。

- 論文
- GitHub

[#97]

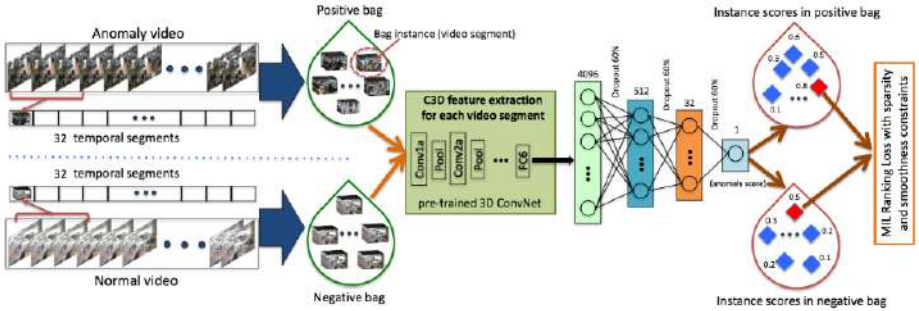
Hirokatsu Kataoka

Real-world Anomaly Detection in Surveillance Videos

Waqas Sultani, Chen Chen, Mubarak Shah CVPR 2018

概要

監視カメラの文脈において異常検出を実行する研究である。ここで、異常検出においてビデオに対して時間のアノテーションを付与するのは非常にコストのかかる作業であるが、ここに対して弱教師付き学習の一種であるMultiple Instance Learning (MIL)を適用して正常/異常ラベルが付いたビデオから異常検出を行うモデルDeep Anomaly Ranking Modelを提案する。さらに、13種類の異常シーン（e.g. road accident, robbery）を収集したデータセットを提供することで同問題の解決を実践した。



新規性・結果・なぜ通ったか？

弱教師付き学習であるMILをベースとして異常検出を行なった、おそらく初めての例であり、その精度は従来法による精度を上回りState-of-the-artとなった（AUCにて75.41を達成）。また、1900の動画に対して13種類の異常を収集したデータセットを構築し、公開した。同データセットは合計で128時間にも及ぶ。

コメント・リンク集

異常の動画データセットを公開したことが評価できるポイント。現在ではYouTube検索とダウンロードである程度のデータセットは構築できそう？（ここらへんを効率化する研究自体があってもよい）

- 論文
- Project
- DB

[#98]

Ryota Suzuki

Normalized Cut Loss for Weakly-supervised CNN Segmentation

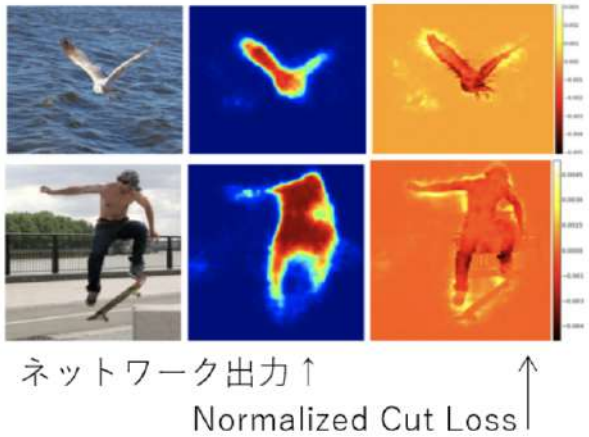
M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, C. Schroers CVPR2018

概要

Weakly-supervisedなセマンティックセグメンテーション手法があって，その方針はインタラクティブに部分的に正解（シードとか）を与えるというものである．そこで，よく用いられるロス関数（クロスエントロピー等）で評価しようとすると，教示の塗りミスが致命的になったりする．そもそも設計的にエラーが考慮されていないからである．

本論文では，非Deepな手法で行われていた評価指標に基づく新たなロス関数Normalized Cut Lossを提案．

従来法と違うところは，提案するロス関数におけるクロスエントロピーの部分は，ラベルが既知のシードの部分での評価だけやっているという点．Normalized Cutはゆるく全ピクセルに対する一貫性の評価を行う．



新規性・結果・なぜ通ったか？

Fully-supervisedな手法と同レベルの性能を実現できた．

従来法の知見を活かした橋渡し的手法．

コメント・リンク集

Disney Researchのインターンでやった模様．

- arXiv

[#99]

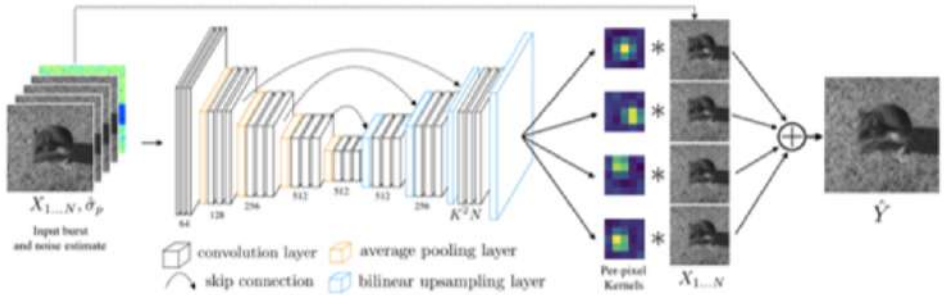
Ryota Suzuki

Burst Denoising with Kernel Prediction Networks

B. Mildenhall, J.T. Barron, J. Chen, D. Sharlet, R. Ng, R. Carroll CVPR2018

概要

携帯含む最近のカメラは連写機能が付いているので，手ブレのあるようなハンドヘルドカメラの連写で撮ったノイズ入り画像をデノイズしようという話．**連続撮影における手ブレに頑健なデノイズCNN**を提案する．



写実的ノイズ定式化に基づく，インターネットから拾ってきた加工済み画像からカメラで撮ったような写実的画像を生成する合成データ生成手法で学習データを作成．学習中に空間的に変化するカーネルを使い，位置調整とデノイズを実現． 不慮の局所解落ち回避のための，焼きなましロス関数をガイドとした最適化．

新規性・結果・なぜ通ったか？

流行に乗った手法（合成データによる学習，適応的パラメータ調整）を使って実現．問題設定も地に足がついている感じがする．

コメント・リンク集

Google Researchのインターンでやった模様．

- arXiv
- プロジェクトページ

[#100]

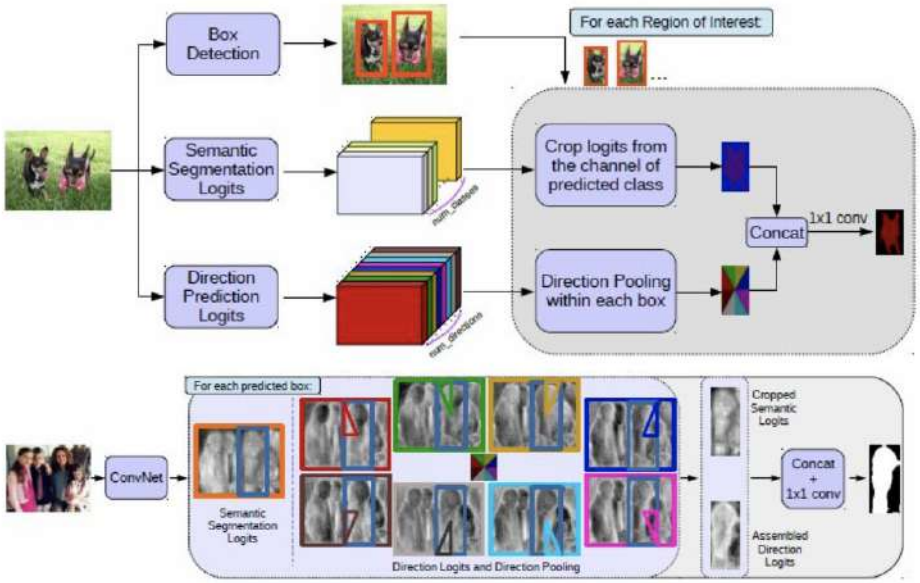
Ryosuke Araki

MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features

Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang and Hartwig Adam CVPR2018 525

概要

物体のBounding-box detection, Semantic segmentationとDirection predictionを同時に行うモデル「MaskLab」を提案する．Faster R-CNN・ResNet-101をベースに，Bounding-box内の前景と背景をわけることでSegmentationを行う．Mask R-CNNと違い，Segmentationを行うときは単純に前景背景分割をするだけでなくクラス分類も行い，また，各ピクセルのDirectionを予測して同じクラスの重なっている物体のInstance segmentationも可能である．また，検出されたBox内でさらに切り出しを行い，小さな物体の検出をしやすくする仕組みも入れている．



新規性・結果・なぜ通ったか？

Object detectionとSemantic segmentationを同時にEnd-to-endで解くモデルの提案．それだけでなく，Semantic segmentationではDirectionを考慮して高精度な認識が可能である．MSCOCOで性能評価を行い，FCIS+++（mAP，Seg：33.6），Mask R-CNN（Seg：35.7，Det：38.2）よりも高い性能（学習時にScale augmentationを行いSeg：38.1，Det：43.0）を達成した．Res-NeXtを用いたMask R-CNN（Seg：37.1，Det：39.8）よりも高性能である．

コメント・リンク集

最近，Detection + Segmentationがいくつか出てきているので今後に注目．検出速度に関する記述は見当たらなかったが，Faster R-CNNベースなのでそれ相応の速度だと思われる．ワンショット系の検出器に適応してこの精度を保ちつつ高速な検出ができればウケそう？

- arXiv

[#101]

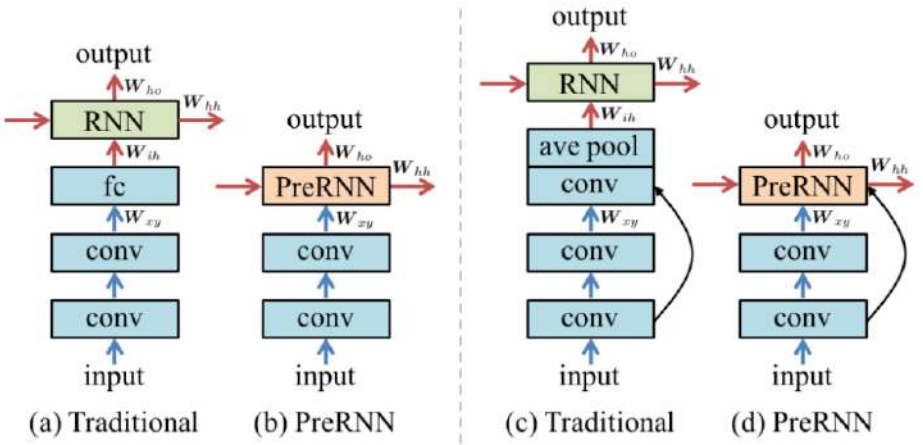
Hirokatsu Kataoka

Making Convolutional Networks Recurrent for Visual Sequence Learning

Xiaodong Yang, Pavlo Molchanov, Jan Kautz CVPR 2018

概要

RNNの改良であり、畳み込み層や全結合層の役割を前処理として構造に入れ込むPreRNNを提案した。従来のRNNとPreRNNの違いは図に示すとおりである（従来型TraditionalなRNNは構造内にfc/conv+avepoolを要するが、PreRNNではそれらを内包している）。このPreRNNを用いて、より有効だと思われるタスクーSequential Face Alighnment, Dynamic Hand Gesture Recognition, Action Recognitionにて適用した。



新規性・結果・なぜ通ったか？

従来型のRNNを改善して、fc-layer/conv+avepool-layerをその構造の中に取り込んだPreRNNを提案し、複数タスク（顔アライメント推定、ジェスチャ認識、人物行動認識）にて従来法よりも高い精度を達成した。

コメント・リンク集

画像キャプションなどにも効果あり？どのように説明文が改善されるのか試してみたい。

- 論文
- SupplementaryMaterial

[#119]

Hirokatsu Kataoka

Inferring Shared Attention in Social Scene Videos

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, Song-Chun Zhu CVPR 2018

概要

複数人いる人物が同時に同領域に注意を向けることをCo-attention/Shared-attentionといい、本論文では三人称視点の入力からこの推定に取り組む。ここに対してConvLSTM（Convolutional Long-Short Term Memory）を用いたモデルを適用、さらにはVideoCoAttと呼ばれるTV番組をメインとしたビデオからデータ収集を行なった。モデルは視線推定（YOLOv2による顔検出も含む）、領域推定（Region Proposal Map）、空間推定（Convolution）と時系列最適化（LSTM）から構成される。データは380ビデオ/492,000フレームから構成される。

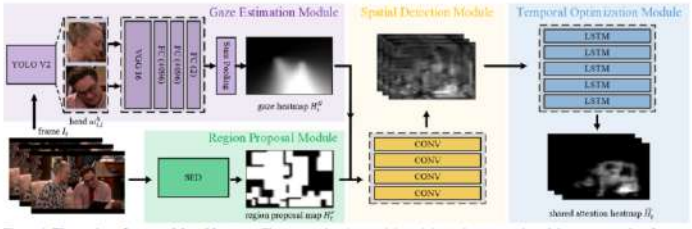


Figure 4. **Illustration of our model architecture.** The gaze estimation module and the region proposal module extract two key features of individuals and the scene context from raw input videos. The subsequent spatial detection module integrates the outputs from the two base modules to perform shared attention detection on a single frame. The temporal optimization module utilizes temporal constraints to optimize the predicted shared attention heatmap.

Figure 5. **Illustration of gaze heatmap H_t^g generation procedure.** With detected head position $q_{t,j}$ (red rectangles in (a)(c)) and corresponding predicted gaze direction $d_{t,j}$ (yellow arrows in (a)(c)), we first generate individual gaze heatmap $H_{t,j}^g$ in (b) and (d), and then get the final gaze heatmap H_t^g in (e) via sum-pooling all the gaze heatmaps in (d).

新規性・結果・なぜ通ったか？

新しい問題である、三人称視点からの共注視を設定し、データとモデルを公開したことが採択された理由である。また、実験により従来法を抑えて、提案法が71.4%の精度かつ誤差がもっとも小さい手法であることを明らかにした。

コメント・リンク集

共注視、面白い！（が、ビデオを見てみると曖昧な部分もありもうすこしアノテーションなどに改善の余地がある？）

- 論文
- YouTube
- Project

[#120]

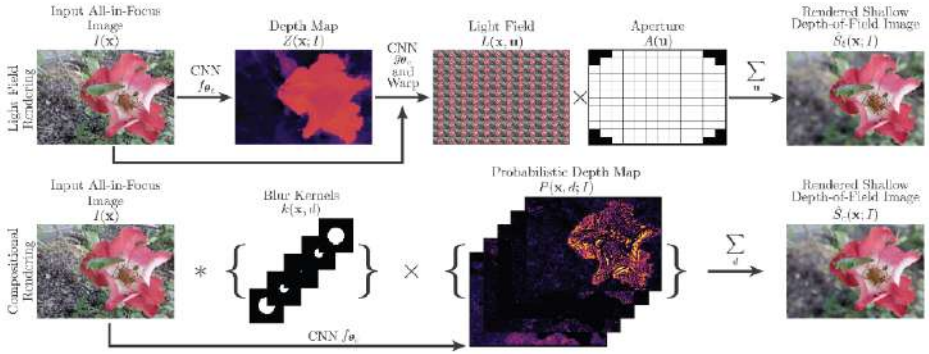
Hirokatsu Kataoka

Aperture Supervision for Monocular Depth Estimation

Pratul P. Srinivasan, Rahul Garg, Neal Wadhwa, Ren Ng, Jonathan T. Barron CVPR 2018

概要

Aperture Supervision（カメラのフォーカスによる教示）により単眼画像からデプスマップを推定する研究である。これを推定するために、Focus/Defocusを処理して、領域ごとの反応を確認することでデプスの教示に相当する。CNNベースの距離画像推定では、確率的距離マップ、Shallow Depth-of-field（各距離における重み付けされたマップ）を適用する。図は本論文における単眼カメラによる距離画像推定のパイプラインである。



新規性・結果・なぜ通ったか？

RGB-Depthを変換する、いわゆるダイレクトな距離画像推定では計算コストも高く、かつ解像度も低かったが、本論文ではフォーカスに関係する教示によりこの問題を解決し、単眼による距離画像推定を実現した。

コメント・リンク集

距離画像を直接的には使わなくても、LightFieldなどの情報から距離画像を推定することができるので、他の関連手法とは異なるアプローチを与えている。

- 論文
- GitHub

[#121]

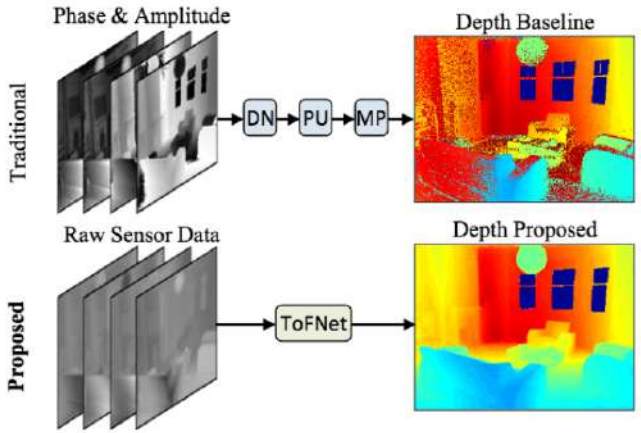
Hirokatsu Kataoka

Deep End-to-End Time-of-Flight Imaging

Shuochen Su, Felix Heide, Gordon Wetzstein, Wolfgang Heidrich CVPR 2018

概要

End-to-EndでセンサデータからToFセンサの出力を行うToFNet (Time-of-Flight Network)を提案する。従来のシステムであh、センサーデータの入力からデノイジング、Phase Unwrapping (PU)やMultipath Correction (MP)を行っていたが、ToFNetでは一括処理が可能となるだけでなく、ノイズがない鮮明な画像を出力可能、リアルタイムで動作可能である。ToFNetはPatchGANという枠組みにより最適化が行われる。PatchGANはEncoder-Decoderの構造をした生成器と非常にシンプルな構造の識別器により構成される。誤差はL1+DepthGradient+Adversarialと、その重み付き和により計算される。



新規性・結果・なぜ通ったか？

従来のカスケード型処理（デノイジング、PU、MP）ではノイズが蓄積してしまいがちだが、提案のToFNetは一括での処理を行い、(1)ノイズを鮮明に除去できるのみならず(2)リアルタイムでの処理が可能である。主にこの２点が採択された理由であると考える。

コメント・リンク集

Depth推定、すでに数値や見た目による判断が曖昧になりつつある？屋内だけでなく、多様なドメインでの適応が待たれる。

- 論文
- Project

[#122]

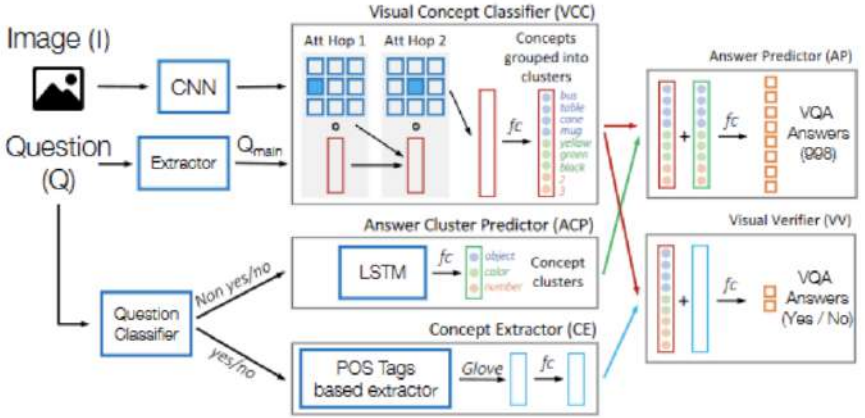
Shintaro Yamamoto

Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrel and Dawn Song CVPR 2018

概要

VQAの学習は学習データの答えの分布に依存してしまう。そこで、答えの分布が異なる学習データを用いて学習した場合でもGrounded Visual Question Answering(GVQA)を提案した。 GVQAでは質問に答える上で、(1)必要な情報を認識する（例：物体の色を聞かれている場合対象となる物体を認識する)(2)必要な答えを推測する(例：物体の色を聞かれている場合色を答える)の2つが重要であると仮定する。 そこで、画像から質問に答えるために必要な情報を抽出する部分と答えを推定する部分の2つに分けたモデルを構築した。その際、質問から質問のタイプ(yes/noで答えられるか)を推定することで、質問の答えを異なるネットワークによって出力させる。



新規性・結果・なぜ通ったか？

質問の答えの分布を学習データとテストデータで異なる分布にしたVQA-CPデータセットを提案した。同データセットを用いて従来手法及びGVQAの精度を調べたところ、従来のデータセットと比べた際の従来手法の精度低下及びGVQAの方が高い精度を記録したことを示した。 また、GVQAによって答えの根拠を説明することが可能となった。

コメント・リンク集

- 論文

[#123]

Shintaro Yamamoto

Fooling Vision and Language Models Despite Localization and Attention Mechanism

Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrel and Dawn Song CVPR 2018

概要

Adversarial attackが、VisionとLanguageの融合問題のようにより複雑な問題に対しても有効であるかを調査した。対象とするタスクは、画像キャプショニング及びVQAとして画像のAdversarial exampleによる出力の変化を調べた。 また、これらの手法におけるlocalizationがAdversarial Attackに影響されるかを確認した。

Figure 2: Adversarial examples generated from different images with the target caption to be "a window on a building".

Figure 3: Adversarial examples generated from Image 4 with different target captions (shown as sub-figure captions).

新規性・結果・なぜ通ったか？

Dense Captionについては、97%の確率で騙すことに成功した。同じ画像の同じ領域に対しても目標とするキャプションが異なると異なるキャプションを出力させることが可能なことを確認した。 VQAについてもごく一部を除いて騙すことができることを確認した。Attention Mapを確認すると、Adversarial exampleを入力した場合異なる領域に注目していることが明らかになった。

コメント・リンク集

- 論文

[#124]

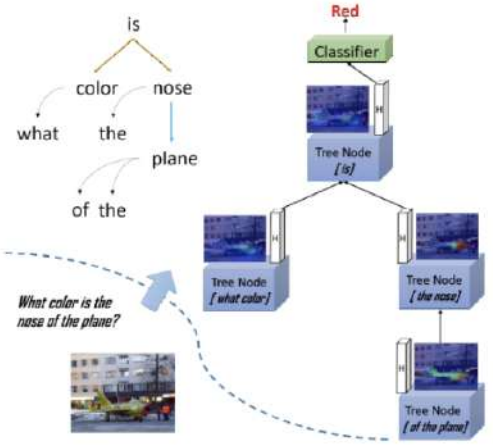
Shintaro Yamamoto

Visual Question Reasoning on General Dependency Tree

Qingxing Cao, Xiaodan Liang, Bailing Li, Guanbin Li, Liang Lin CVPR 2018

概要

VQAの答えだけでなく判断根拠も出力する手法を提案。質問をtree構造に分解し、各nodeに関する情報(例：plane)が画像中のどこに存在するかを示すattention mapを求める。 既に得られているattentionマップ及びhidden stateを更新していくことで、質問の答えとたどり着いていく。 最終的な質問の答えはhidden stateを用いて求める。



新規性・結果・なぜ通ったか？

質問への回答の精度は従来手法と比べて大きく向上されているわけではない。従来の判断根拠を求める研究はルールを人間が設計するもしくはground truthが必要であるのに対してこれらを必要とせずに回答根拠を得ることに成功。

コメント・リンク集

- 論文

[#125]

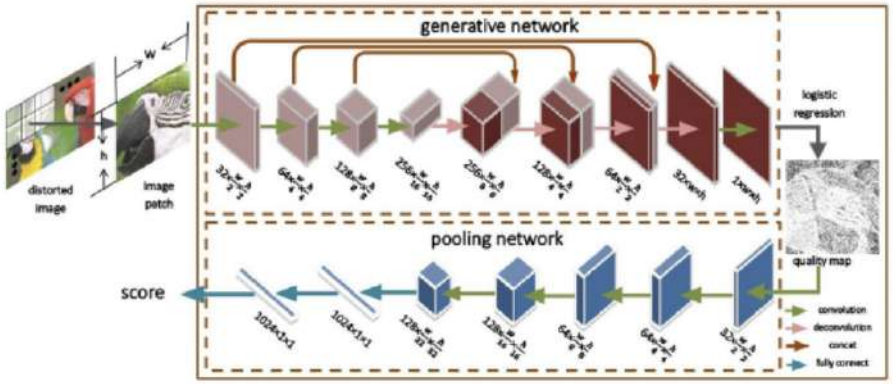
Shintaro Yamamoto

Blind Predicting Similar Quality Map for Image Quality Assessment

Da Pan, Ping Shi, Ming Hou, Zefeng Ying, Sizhe Fu and Yuan Zhang CVPR 2018

概要

画像の品質を評価するためのBlind Predicting Similar Quality Map for IQA(BPSQM)を提案した。CNNを用いた画像の品質評価手法は数多く提案されているが、その大半はブラックボックスとなっている。 本研究は、ピクセル単位の画像の損失度合いを示すquality mapを始めに推定することで、画像圧縮などに伴いどのように画像の品質が低下してるかの可視化を可能とした。 また、qualityマップから画像の損失度合いを表すスコアの算出を行う。



新規性・結果・なぜ通ったか？

従来のquality mapを求める手法は、損失前の画像(reference)が必要なものが大半であり、reference不要なCNNベースの手法はパッチ単位で推定するのみであった。それに対して本研究は、referenceなしでピクセル単位のquality mapを推定することを可能とした。 損失度合いの推定に関しても、referenceなしの手法と比べて精度の向上を実現した。

コメント・リンク集

- 論文

[#126]

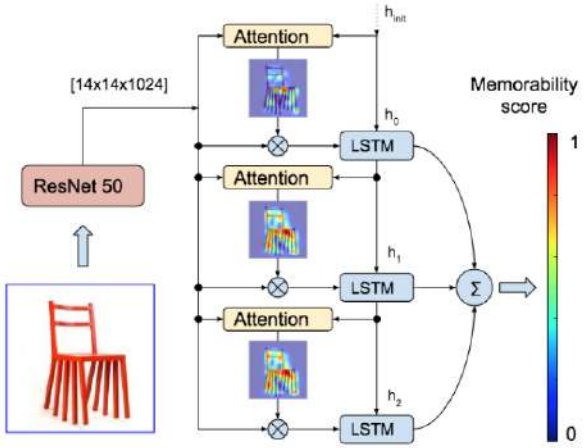
Hirokatsu Kataoka

AMNet: Memorability Estimation with Attention

Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, Paolo Remagnino CVPR 2018

概要

画像中の記憶に残りやすい領域（Memorability）を可視化するネットワークであるAMNet（Attention and Memorability Network?）の提案。ResNet50による特徴表現、LSTMにより実装されたAttention構造の仕組みによりMemorabilityスコアを算出する。アノテーションは従来研究であるLaMem（下記リンク参照）に使用したデータセットであるSUN Memorability（同じく下記参照）を用いて学習を行った。



新規性・結果・なぜ通ったか？

従来法よりも精度が良かった（より人間の記憶の構造に近かった？）ことを示した。これはアテンション構造を用いていることが、より人間の記憶の仕組みにおいて再現性が良かったことを示しているといえる。

コメント・リンク集

記憶の仕組みも人間の直感が必要な高次機能の再現である。このように高次なラベリングが今後は増えてくると思うし、人間のタスクをカバーする意味でも重要になるか？

- 論文
- GitHub
- LaMem
- SUN Memorability

[#127]

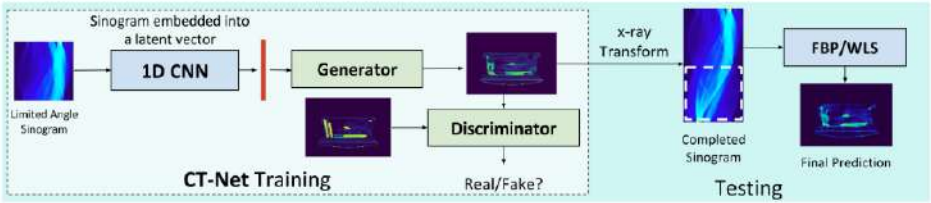
Hirokatsu Kataoka

Lose The Views: Limited Angle CT Reconstruction via Implicit Sinogram Completion

Rushil Anirudh, Hyojin Kim, Jayaraman J. Thiagarajan, K. Aditya Mohan, Kyle Champley, Timo Bremer CVPR 2018

概要

手荷物検査や医療用として用いられるComputed Tomography (CT)画像の復元を、限られた角度のSinogramの入力から行う技術（CTNet）を提案する。CTNetは1D/2D畳み込みで構成され、SinogramからFull-viewのCT画像を復元することができる。図はCTNetの学習とテストを示したものである。学習時にはGAN-likeな手法により構成され、入力から1DCNNにより特徴量を生成、GeneratorがCT画像を復元、DiscriminatorがReal/Fakeを判断することでGeneratorを鍛える。テスト時にはさらにFBP (Filtered Back Projection)/WLS (Weighted Least Squares)なども用いて最終的な結果を得る。



新規性・結果・なぜ通ったか？

角度が限定されたx線画像から、360度のCT画像を生成するというチャレンジングな試みを行ったことが評価された。同課題に対してGAN-likeな手法を提案し、手法的な新規性も打ち出せたことが採択された基準であると考える。PSNRやセグメンテーションベースの方法で評価を行い、従来法よりも優れた手法であることを示した。

コメント・リンク集

CT画像を復元できてしまうのがすごい！

- 論文

[#128]

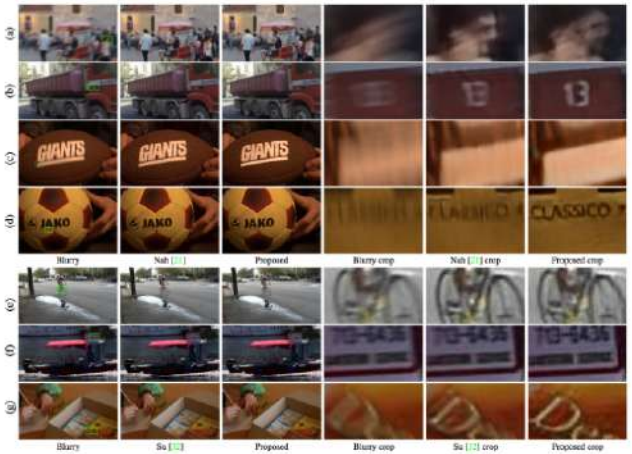
Hirokatsu Kataoka

Learning to Extract a Video Sequence from a Single Motion-Blurred Image

Meiguang Jin, Givi Meishvili, Paolo Favaro CVPR 2018

概要

１枚のブラー画像から時系列フレームを推定して動画像を生成するアプローチを提案。モーションブラーは通常、カメラなどセンサによる露光により発生するが、その分解は非常に困難な問題として扱われていた。本論文では平均化を除去してフレームを時系列方向に並べ、次にDeconvolutionを復元して同問題に取り組む（この問題は通常、Blind Deconvolutionと言われる）。提案法では、深層学習の手法としてこの両者を実現する構造を構築。



新規性・結果・なぜ通ったか？

Blind Deconvolutionの課題を取り扱っているが、さらにここでは単一のブラー画像から動画像を生成するアルゴリズムや深層学習アーキテクチャを提案した。特に、ブラー画像から時系列画像を順次復元するための誤差関数を提案したことが最も大きな新規性である。

コメント・リンク集

もともとあった問題に少し味付けして、新しい問題を作り出すセンスが欲しい。。

- 論文

[#129]

Hirokatsu Kataoka

Learning to Detect Features in Texture Images

Linguang Zhang, Szymon Rusinkiewicz CVPR 2018

概要

テクスチャに対して有効かつスケーラブル、さらに学習可能な局所特徴量を提案する。さらに提案手法は既存のランキングロスやFully-Convolutional Networks (FCN; 全層畳み込みネットワーク)と統合可能である。著者らは、新規の学習誤差関数であるPeakednessという指標を畳み込みマップに対して導入した。画像はテスト画像に対して提案手法を施した結果であり、Repeatableな特徴量（画像の中に再帰的に登場するテクスチャ特徴）が検出されている。

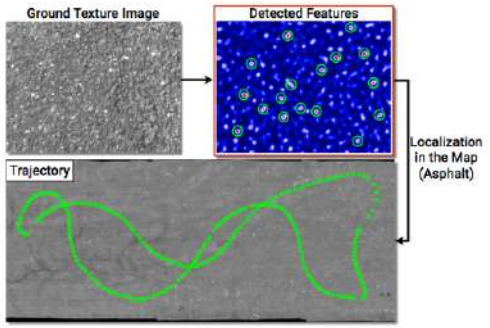


Figure 1. From each test image, our proposed detector extracts highly repeatable features, which can be utilized by Micro-GPS to achieve precise global localization in a pre-built map, such as the asphalt map being shown. Note that Micro-GPS locates each test image *independently* in the map (ignoring temporal coherence).

新規性・結果・なぜ通ったか？

（i）FCN構造によりフルサイズの再帰的なテクスチャパターンを評価することに成功した、（ii）Peakednessという指標を導入し、これを最大化することでテクスチャを評価するための畳み込みマップを洗練化することに成功、という点がもっとも重要な新規性である。実験ではcarpet/asphalt/wood/tile/granite/concrete/coarseといったテクスチャパターンに対して有効であることを示した。

コメント・リンク集

複雑かつ特徴が比較的取りづらいテクスチャの解析は今後さらに重要性を増すと考えられる（道路面のひび割れ調査など）。ここに教師なし学習（Self-Supervision含む）が導入されていくことになると思う。

- 論文
- Project

[#130]

Kota Yoshida

Smart, Sparse Contours to Represent and Edit Images

T.Dekel, C.Gan, D.Krishnan, C.Liu and W.T.Freeman CVPR2018 arXiv:1712.08232

概要

元画像の輪郭情報から画像を再構成する手法を提案.GANをベースとして，入力情報が与えられない領域のテクスチャと細部を合成する.実験では，顔認証システムや人間を対象にして元画像と再構成された画像と区別されないという結果となった.

Figure 6. The source image (a) is reconstructed from different representations kept at the same pixels marked in red (b), using the following methods: (c) Diffusion [13] based solution that propagates RGB values sampled at both sides of each contour pixel. (d) Pix2pix [18] which uses only binary contours as input. (e) Our LFN output using gradient features stored at each contour pixel and (f) our final HFN output.

新規性・結果・なぜ通ったか？

- Pix2pixなどの既存の手法よりも大幅に向上している.
- 2つのネットワークで構成されており，1つ目のネットワークでは，画像全体の構造，色を再構成，2つ目のネットワークでは画像のテクスチャと細部の表現をしている.
- 直感的な操作が可能で，顔のパーツを移動させたり，追加させることもできる.

コメント・リンク集

- 入力情報がない輪郭と輪郭の間の画像部分の再構成にも力を入れてる

- Paper

[#131]

Kodai Nakashima

R-FCN-3000 at 30fps: Decoupling Detection and Classification

Bharat Singh, Hengduo Li, Abhishek Sharma and Larry S. Davis CVPR2018

概要

オブジェクト性検出と分類を分離した物体検出器であるR-FCN-3000を提案した．RoIのための検出スコアを得るために，オブジェクト性検出と分類スコアをかける． R-FCNで提案されたposition-sensitive filterはfine-grained classificationには必要ないというのが基本アイディア． また本論文では，R-FCN-3000はオブジェクト数が増えると性能が向上することが示されている．

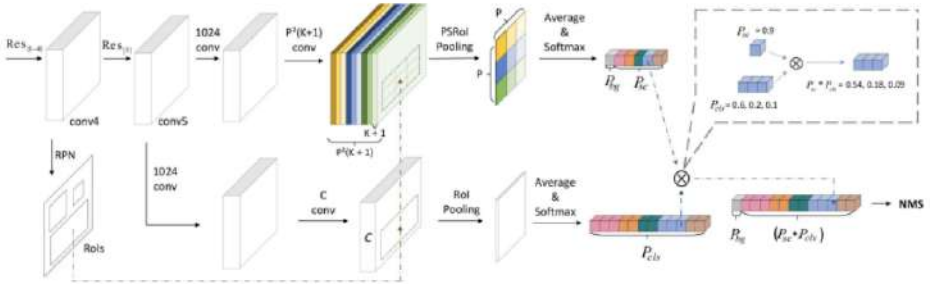


Figure 2. R-FCN-3000 first generates region proposals which are provided as input to a super-class detection branch (like R-FCN) which jointly predicts the detection scores for each super-class (sc). A class-agnostic bounding-box regression step refines the position of each RoI (not shown). To obtain the semantic class, we do not use position-sensitive filters but predict per class scores in a fully convolutional fashion. Finally, we average pool the per-class scores inside the RoI to get the classification probability. The classification probability is multiplied with the super-class detection probability for detecting 3000 classes. When K is 1, the super-class detector predicts objectness.

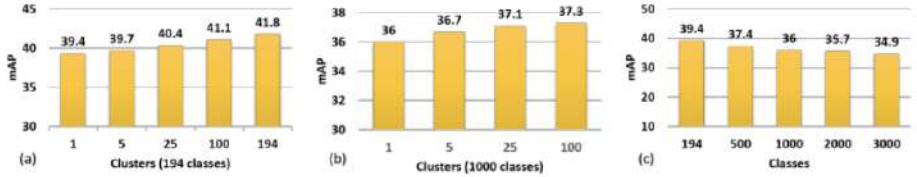


Figure 3. The mAP on the 194 classes in the ImageNet detection set is shown as we vary the number of clusters (super-classes). This is shown for 194 class and 1000 class detectors. We also plot the mAP for different number of classes for an objectness based detector.

Figure 4. The objectness, classification and final detection scores against various transformations such as combinations of scaling and translation are shown. These scores are generated by forward propagating an ideal bounding-box RoI (in green) and a transformed bounding-box RoI (in red) through the R-FCN (objectness) and classification branch of the network. The selectiveness of the detector in terms of objectness is clearly visible against the various transformations that lead to poor detection.

新規性・結果・なぜ通ったか？

ImageNet detection datasetで一秒あたり30枚の画像を処理したところ，mAPが34.9%であった（YOLO9000は18%）．

コメント・リンク集

- 論文URL

[#132]

Kodai Nakashima

Learning to See in the Dark

Chen Chen, Qifeng Chen, Jia Xu and Vladlen Koltun CVPR 2018

概要

暗い環境において，同じシーンを短時間露光で撮影した暗い画像と長時間露光で撮影した明るい画像のrawデータを集めたデータセットを提案した．このデータセットは，5094個の暗い画像のrawデータと424個の明るい画像のrawデータが1対多で対応付けられている． インドアとアウトドアの両方で撮影を行った．

(a) Camera output with ISO 8,000 (b) Camera output with ISO 409,600 (c) Our result from the raw data of (a)

Figure 1. Extreme low-light imaging with a convolutional network. Dark indoor environment. The illuminance at the camera is < 0.1 lux. The Sony α7S II sensor is exposed for 1/30 second. (a) Image produced by the camera with ISO 8,000. (b) Image produced by the camera with ISO 409,600. The image suffers from noise and color bias. (c) Image produced by our convolutional network applied to the raw sensor data from (a).

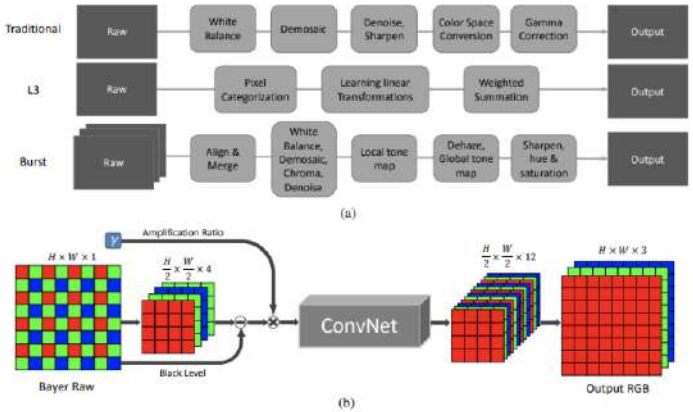


Figure 3. The structure of different image processing pipelines. (a) From top to bottom: a traditional image processing pipeline, the L3 pipeline [18], and a burst imaging pipeline [14]. (b) Our pipeline.

新規性・結果・なぜ通ったか？

このデータセットを用いてFCNをトレーニングし，テストしたところ図に示すような結果が得られた．このネットワークはrawデータを直接扱うため，図に示すように，従来の画像処理パイプラインの多くの代わりになる．

コメント・リンク集

- 論文URL
- github

[#133]

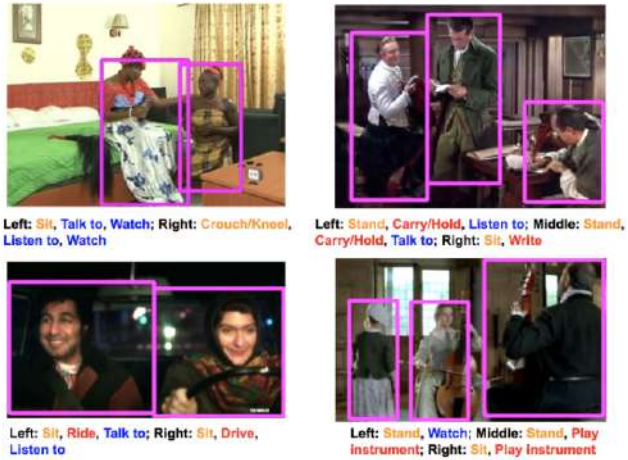
Kensho Hara

AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions

C. Gu et al., CVPR 2018

概要

大規模な新規動画データセットを構築．従来の動画データセットが複合的な行動ラベルを扱うのに対して，このデータセットではStand, Sit, WatchのようなAtomicな行動ラベル (80 classes) を扱う．このようなラベルが1秒間隔で動画中のすべての人にアノテーションされており，しかもBounding Boxまで付いているというのがこのデータセットの強み．80種類ものAtomicな行動ラベルが大規模にしかも密に付いているデータセットは初．加えて，Two-stream I3D & Faster R-CNNというような手法を提案．従来のSpatio-temporal Action Localization用のデータセットではSOTAを達成したものの，このデータセットは15.6% mAPと問題の難しさも主張している．



新規性・結果・なぜ通ったか？

- Bounding Boxまでアノテーションされている初の大規模動画データセットを構築
- 動画中の一部ではなく密にAtomicな行動のラベルがアノテーションされている
- Spatio-temporal Localizationをするためのベンチマークとなる新規手法も提案

コメント・リンク集

- 論文 (arXiv)
- データセット
- ActivityNet Challenge 2018 Task B

[#134]

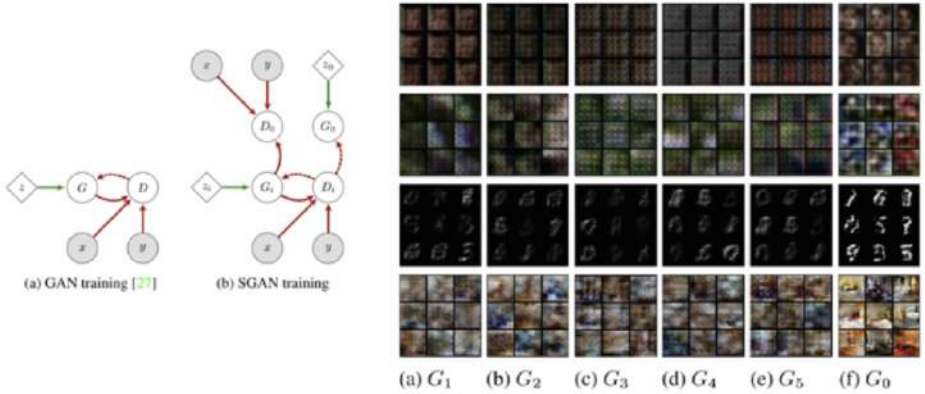
KenichiroWani

SGAN: An Alternative Training of Generative Adversarial Networks

Tatjana Chavdarova, Idiap and EPFL; Francois Fleuret, Idiap Research Institute CVPR2018 1712.02330

概要

General Advesarial Networks(GAN)は現在，コンピュータビジョン分野で広く使われている手法である．しかしながら，複雑な学習をするには時間がかかり，人の手が必要となる．そこでSGANというトレーニングプロセスを検討する．SGANではいくつかの敵対的でローカルなネットワークの組み合わせを独立させて学習させることでグローバルな一対のネットワークの組み合わせを学習することができる．SGANの学習はローカルディスクリミネータとジェネレータによってグローバルディスクリミネータとジェネレータが学習される．



新規性・結果・なぜ通ったか？

adversarial pairs (G1,D1),...,(GN,DN)を学習し， G0はD1,...,DNによって学習， D0はG1,...,GNによって学習させることでグローバルな一対のネットワークを学習する。

コメント・リンク集

- arxiv

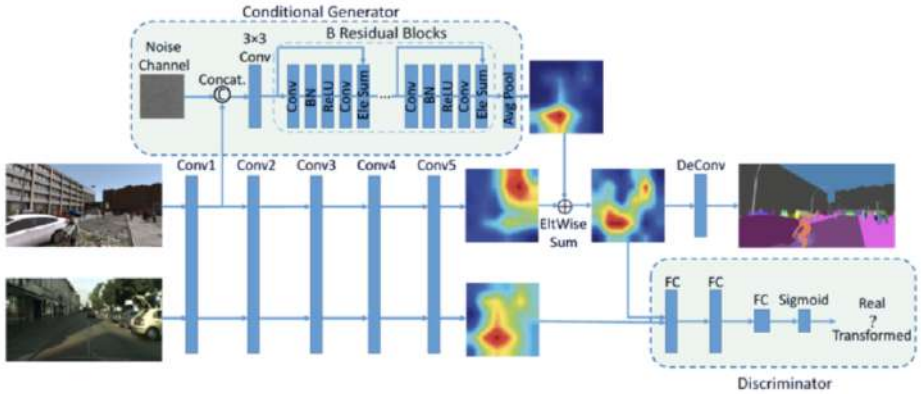
[#135] KotaYoshida

Conditional Generative Adversarial Network for Structured Domain Adaptation

W.Hong, Z.Wang, M.Yang and J.Yuan CVPR2018

概要

コンピュータによって学習用のアノテーションを生成し，実画像のような合成画像として用いることが流行．しかし，ドメインの不一致という問題が起きる．それを解決するために，GANをFCNフレームワークに統合することでSemanticSegmentationのためのドメイン適用のための手法を提案．



新規性・結果・なぜ通ったか？

- 合成画像の特徴を実画像のように変換する条件付きジェネ–レータとディスクリメーターを学習
- ジェネレータは合成画像を実画像のようにディスクリメーターを騙すように学習させることでFCNのパラメータを更新．
- 本手法である実際のラベルを用いずに実験を行い，Cityscapesデータセットの IoU平均が12〜20上回りSoTA．

コメント・リンク集

- FCN＋GANでSemanticSegmentation
- Paper

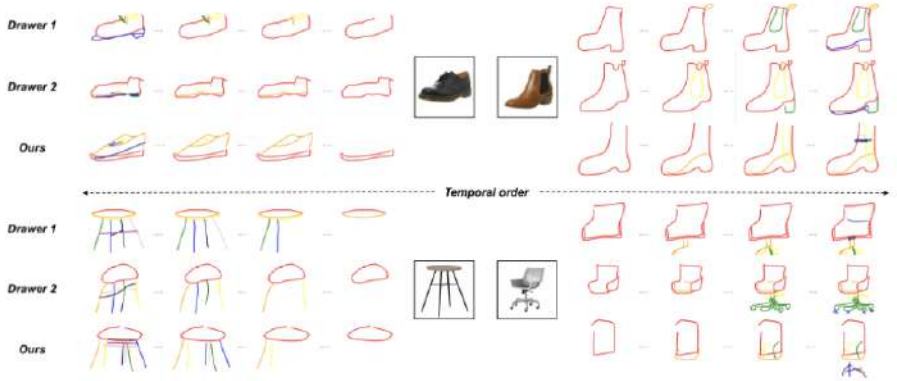
[#136]

Shintaro Yamamoto

Learning to Sketch with Shortcut Cycle Consistency

Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales CVPR 2018

概要



画像からスケッチのストロークを取得する手法の提案。人間が画像からスケッチをすると、同じ画像に対しても様々なバリエーションが生じてしまう。 そこで、教師有学習と教師無学習を組み合わせることによって画像からスケッチの取得を実現する。 教師有学習は、画像からスケッチもしくはスケッチから画像という変換を学習する。 教師無学習は、オートエンコーダのように画像もしくはスケッチを符号化し、元に戻すという処理を学習する。 その際、CycleGANのようにドメイン変換を繰り返すのではなく、符号化したものをそのまま復号化する(Shortcut Cycle)。

新規性・結果・なぜ通ったか？

Pix2pixやCycleGANなどの手法と比較を行い、いずれの手法と比較してもスケッチとして抽象化されつつもセマンティックな特徴を捉えていることを確認した。また、数値評価としてスケッチの認識及び検索タスクを行って評価した。 どちらのタスクにおいても、従来手法と比較して高い精度でスケッチへの変換ができていることを示した。

コメント・リンク集

- 論文

[#137]

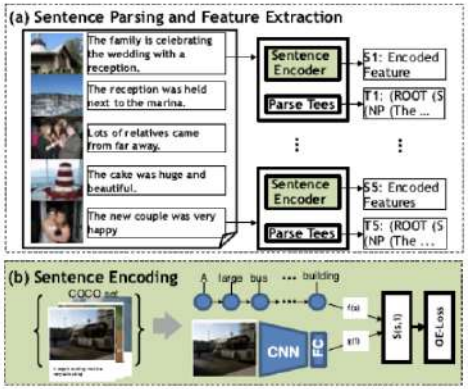
Shintaro Yamamoto

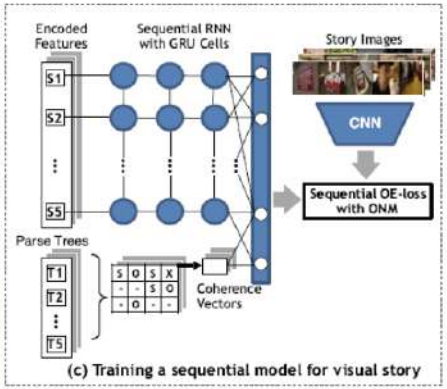
Show Me a Story: Towards Coherent Neural Story Illustration

Hareesh Ravi, Lezi Wang, Carlos M. Muniz, Leonid Sigal, Dimitris N. Metaxas, Mubbasir Kapadia CVPR 2018

概要

複数の文で構成されたテキストの内容を表す画像シークエンスを検索する手法を提案。文章から抽出される特徴と画像から抽出された特徴を対応付けることにより、各文に対して1枚の画像を選択する。 その際、文章特徴はGRUによって前後の文章との関係を含めて抽出する。 また、heやitなどの代名詞が何を指しているかを明らかにするために、テキスト全体としての一貫性を測るcoherence vectorを導入した。





新規性・結果・なぜ通ったか？

ベースラインとなる手法では、文単位で画像の検索を行っているために画像シークエンスとしての一貫性が損なわれてしまう。そこで、GRU及びcoherence vectorによって前後の文で登場した単語などを考慮することが可能となり、テキスト全体を表す画像シークエンスの検索が可能となった。 ユーザースタディにより、ベースライン、coherence vector無し、coherence vector有りの比較を行い、coherence vector有りが最も好まれる結果を得た。 また、画像シークエンスがテキストに合っているかは主観的な評価であるため、saliencyベースの新たな評価指標を提案した。

コメント・リンク集

- 論文URL
- ベースライン

[#138]

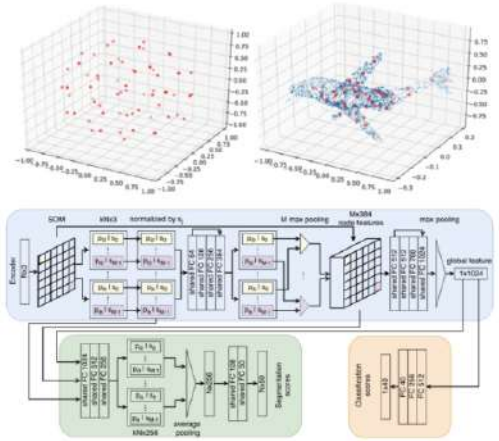
Yoshihiro Fukuhara

SO-Net: Self-Organizing Network for Point Cloud Analysis

Jiaxin Li et al. CVPR 2018

概要

順序構造に対して不変な３次元 Point Cloud のための deep learning アーキテクチャー SO-Net を提案. Self-Organizing Map (SOM) を作ることで点群の空間分布をモデル化し, SOMのノードを用いて階層的な特徴量の抽出を行う. Point Cloud のクラス分類やセグメンテーションなどのタスクを用いた評価実験では, 先行研究と同等以上の結果をより短い学習時間で達成した.



新規性・結果・なぜ通ったか？

- SOM を用いることで Point Cloud を複数の Point Cloud の部分集合に分割し, 各部分集合ごとの特徴量を抽出した後, 全体の特徴量を階層的に抽出する.
- 初期ノードの位置を固定し, 学習を batch 単位で行うことで, SOM の学習が順序構造に対して不変となるようにしている.
- 様々なタスクの事前学習として用いるための Point Cloud の autoencoder を提案.
- ネットワークの構造が単純かつ並列計算可能なため, 先行研究よりも短時間で学習をすることが可能.
- point cloud reconstruction, classification, object part segmentation, shape retrieval などの複数のタスクを用いて評価実験を行った.

コメント・リンク集

- [論文] SO-Net: Self-Organizing Network for Point Cloud Analysis
- [Code] GitHub

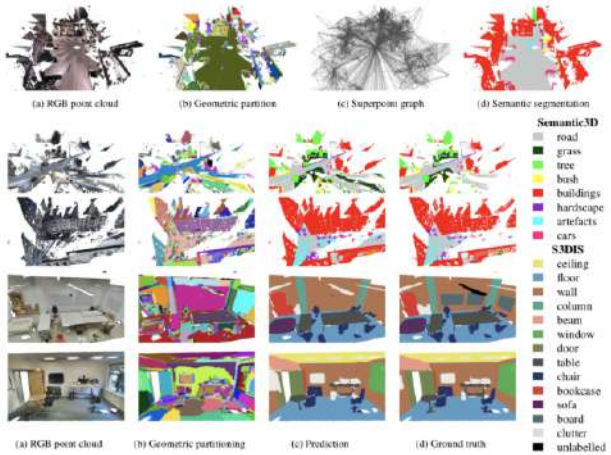
Yoshihiro Fukuhara

Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs

Yoshihiro Fukuhara et al. CVPR 2018

概要

大規模(数百万規模)な point clouds データに対して効率的に Semantic Segmentation を行う研究. まず, point clouds 全体を形状が単純で, 意味的に同じ点が属する部分集合(superpoint）に分類し, superpoint が作るグラフ （SPG）に graph convorution を適用することで segmentation を行う. Semantic3D と S3DIS dataset を用いた評価実験では先行研究よりも良い結果を達成した.



新規性・結果・なぜ通ったか？

- superpoint の構成は先行研究(Guinard+17)で提案された, Global Energy を用いて行う.
- 各 superpoint の特徴量を PointNet を用いて抽出する. (大規模なデータを扱うため, 各 superpoint 内でダウンサンプリングを行っている.)
- 抽出された各 superpoint の特徴量に対して Gated Recurrent Unit (GRU) を用いた graph convorution を適用することで, 各 superpoint のクラス分類を行う.
- Semantic3D と S3DIS dataset を用いた評価実験では, ShapeNet などの先行研究と比較して複数の評価尺度で最も優位な結果を達成した.

コメント・リンク集

- [論文] Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs
- [Code] GitHub

[#140]

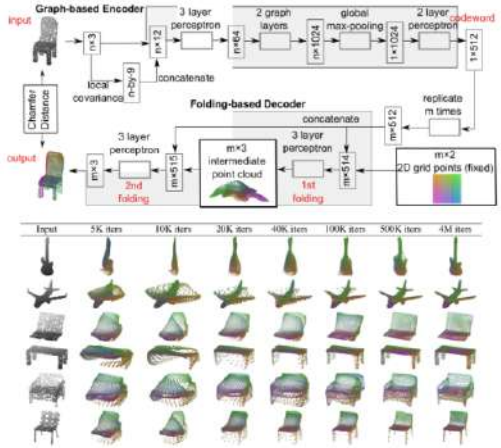
Yoshihiro Fukuhara

FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation

Yaoqing Yang et al. CVPR 2018

概要

3次元点群処理のための autoencoder を提案. Folding という新しい decoding 演算を導入することで, 2次元グリッド上の点から3次元点群の表面上への射影を教師なしで学習した.



新規性・結果・なぜ通ったか？

- 新しい end-to-end な3次元点群処理のための deep autoencoder を提案した.
- 提案手法のdecoderのパラメータ数は既存手法の7%であるが, これで2次元グリッドと任意の3次元点群表面への写像が構成できることを理論的に証明した.
- MN40 や MN10 dataset を用いた classification タスクの評価実験では, 最先端の教師あり手法（Achlioptas+17）などと同等の精度を達成した.

コメント・リンク集

- [論文] FoldingNet: Point Cloud Auto-encoder via Deep Grid Deformation
- [動画] YouTube

[#141]

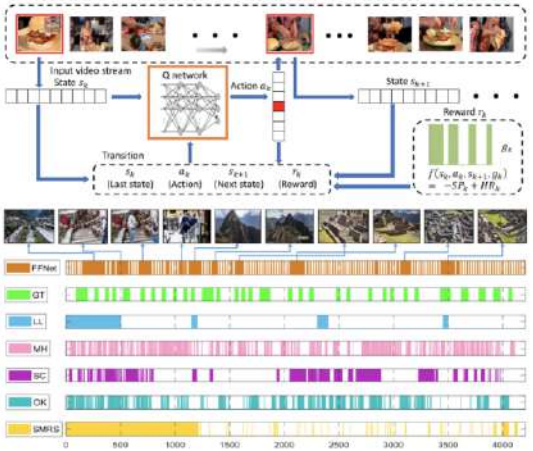
Yoshihiro Fukuhara

FFNet: Video Fast-Forwarding via Reinforcement Learning

Shuyue Lan et al. CVPR 2018

概要

Video Fast-forwarding のタスクを MDP(Markov Decision Process) として定式化し, 強化学習を用いて解く方法を提案. 評価実験では精度と効率の両方に置いて先行研究よりも優れた結果を示した.



新規性・結果・なぜ通ったか？

- Video Fast-forwarding を MDP (Markov Decision Process) として定式化した.
- 現在の Frame の特徴量を状態, スキップする Frame 数を行動として, Q-learningで強化学習を行う.
- 報酬はスキップした Frame の中に重要なものがどの程度含まれていたかに基づいて計算される.
- Tour20 や TVSum dataset を用いた先行研究との比較実験では, 主観評価と定量的評価の両方に置いて最も良い結果となった.(6-20%程度、重要なframeを含んでいる割合が増加)
- 先行研究と比較して80%近く処理するフレーム数を削減し, 効率化することに成功した.

コメント・リンク集

- [論文] FFNet: Video Fast-Forwarding via Reinforcement Learning

[#142]

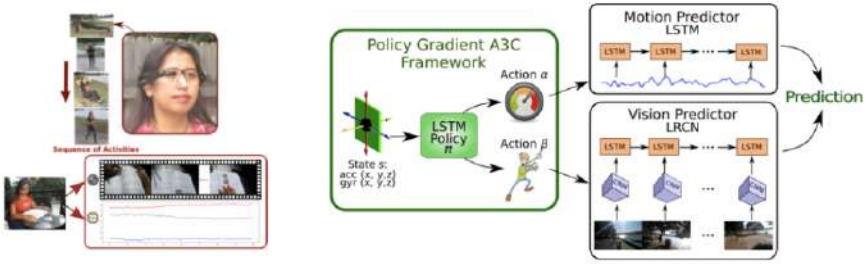
Yoshihiro Fukuhara

Egocentric Activity Recognition on a Budget

Rafael Possas et al. CVPR 2018

概要

ウェアラブルデバイスのような使用可能な電力が限られる状況において, 電力消費と精度を強化学習を用いてバランスするフレームワークを提案. 複数のセンサー情報を用いた行動認識のタスクにおいて, 高精度・高電力消費な predictor と低精度・低電力消費な predictor を強化学習の結果に基づいて適宜切り替えることで少ない消費電力で先行研究と同等の精度を達成した. また, 一人称視点動画行動認識のための新しいデータセットを作成した.



新規性・結果・なぜ通ったか？

- ウェアラブルカメラの情報を用いた高精度・高コストな predictor とモーションセンサーの情報を用いた低精度・低コストな predictor のどちらを使用して推定を行うべきかを A3C の agent が判断する.
- どちらのセンサーの情報を用いても正しい推定結果となるような状況では低精度・低コストな predictor を使用した場合に大きな報酬が得られるように agent の学習を行う.
- 提案手法では報酬についてのパラメータ 1 つを調整する事で精度と消費電力の簡単なトレードオフが可能.
- 一人称視点動画行動認識のための新しいデータセット（DataEgo）を作成.
- Multimodal egocentric dataset を用いた評価実験では従来手法 (Song+16)とほぼ同等の精度を少ない消費電力で達成.

コメント・リンク集

- [論文] Egocentric Activity Recognition on a Budget

[#143]

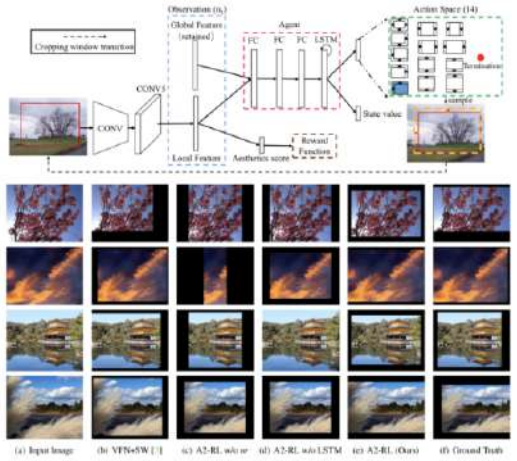
Yoshihiro Fukuhara

A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping

Debang Li et al. CVPR 2018

概要

強化学習 (A3C) を用いて Image cropping を行う手法を提案. 従来の sliding winodow に基づく手法のように膨大な数の cropping 候補を評価する必要がないため, 先行研究よりも短時間で結果の計算が可能. また, 評価実験では精度についても先行研究よりも優位な結果を達成した.



新規性・結果・なぜ通ったか？

- Image cropping を sequential decision-making process として定式化した. (14種類の cropping を action として, Markov 過程としてモデル化.)
- 上記の問題を A3C を用いた強化学習を用いて解いた.
- 報酬については学習済みの View Finding Network (Chen＋2017) を使用.
- 各ステップで候補となる cropping の種類の数が少ないため, 先行研究と比較して非常に短い計算時間で結果を出力することが可能となった.
- Flickr Cropping Dataset, CUHK Image Cropping Dataset, Human Cropping Dataset を用いて行った評価実験ではいずれも先行研究よりも優位な結果を達成した.

コメント・リンク集

- [論文] A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping
- [Code] GitHub

[#144]

Shintaro Yamamoto

Good View Hunting: Learning Photo Composition from Dense View Pairs

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomir Mech, Minh Hoai, Dimitris Samaras CVPR 2018

概要

画像の構図の良し悪しを評価するComparative Photo Compositionデータセットを構築。10800枚の画像から24の構図の画像を作成し、クラウドソーシングによって2つの構図のどちらがいいかをアノテーションした。 また、入力画像をどのようにクロッピングすると良い構図になるかを提示するシステムを構築した。 その際、IOUを評価尺度にすると構図的に評価が低いものも高いスコアになるため、画像を評価するネットワークから得られるスコアを指標とした。

新規性・結果・なぜ通ったか？

従来のデータセットでは画像に対してスコアがついていたのに対して、構図の異なる2枚の画像どちらがいいかを100万ペアアノテーションを行った。構図推薦システムは、ユーザースタディの結果従来手法よりも良いと感じる人が多いことを確認した。 また、計算速度も従来手法と比べはるかに向上した(75FPS+).

コメント・リンク集

- プロジェクトページ

[#145]

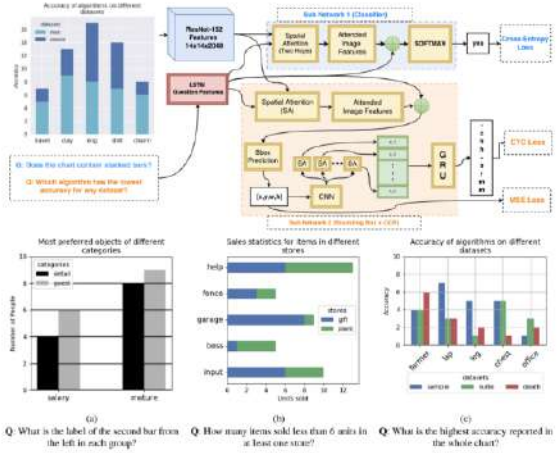
Yue Qiu

DVQA: Understanding Data Visualization via Question Answering

Kushal Kafle, Brian Price, Scott Cohen, Christopher Kanan CVPR 2018 694

概要

- 新規なバーグラフに対して質問回答タスクDVQA及びデータセットの提案．
- バーグラフが情報の一つとしてより豊かな統計的な情報を表現できる．提案手法がバーグラフを対象としたDVQAを提案し，バーグラフの自動的情報抽出と理解を可能にした．
- 大規模なバーグラフQAデータセットDVQAを提案した．DVQAが3Mのグラフ - 質問ペアから構成され，バーグラフに対し３種類の質問(構造理解，データ検索，reasoning)を設定した．また，全部の質問がopen-endedである．
- DVQAタスクにおいて，2種類のネットワーク構造を提案した．①MOM:グラフの局所領域を抽出し文章を生成ことにより回答できる問題を対応するネットワークboundingbox OCR及びグラフの局所領域を抽出せずに回答する一般的な問題を対応するClassifierの二つのサブネットから構成される．どのネットにより回答するかを2クラス分類問題として取り扱っている②SANDY:従来手法SANにダイナミックエンコーディングモデルを用いて，質問文中のchart-specific単語をエンコーディングし，それをベースに直接chart-specificな回答文を生成できる．



新規性・結果・なぜ通ったか？

- 実用性が高い新規なバーグラフに対し質問回答タスクを提案．
- 提案データセットDVQAに対し5種類の従来のVQA手法と提案のMOM,SANDYの比較実験を行った．一般的問題・chart-specific問題の両方に対し提案のSANDYモデルが最も良い精度を達成した．
- 提案のデータセットDVQAがバーグラフの理解と質問文・回答文

コメント・リンク集

- VQAタスクのVを画像からバーグラフに変更し実用性が高い提案である．
- 類似した考えで従来の”V”か“Q”か“A”を同じ処理で別の似た概念に変更する研究をするも面白そう
- 論文

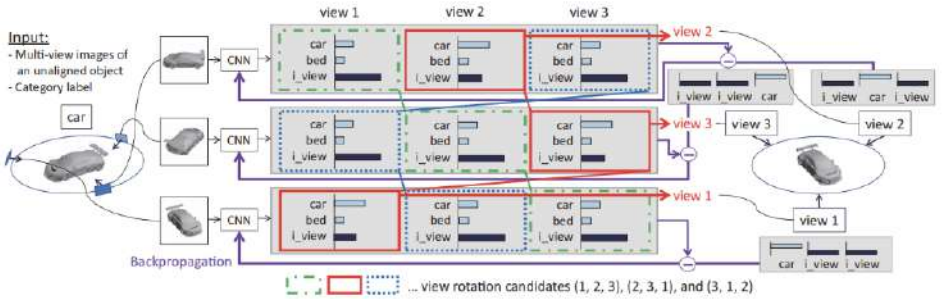
[#146] Yue Qiu

RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints

Asako Kanezaki, Yasuyuki Matsushita, Yoshifumi Nishida CVPR 2018 628

概要

- 物体のマルチ視点の画像からジョイントで3D姿勢推定及び物体認識を行う手法RotationNetの提案.
- 3D MFPにより作成されたマルチ視点画像データセットMIROを提案した．(12classes, 10 instances/class,160viewpoints)
- 物体を観測する視点及び物体のカテゴリをジョイントで推定した方がより良い精度を達成できると指摘し，更にトレーニングする際に物体を観測する視点をlatent variablesとして取り扱い，視点unalignedな学習データセットからunsupervisedで物体の姿勢推定を学習する.
- また，視点-specificな特徴をクラス内だけではなく，異なるクラス間の姿勢アライメントを行う.
- RotationNetのネットワーク構造はマルチ視点の画像から画像ごとにそ全部の視点の確率(その画像がその視点であるか)及び物体カテゴリを予測し，全部の画像から予測した結果から正解ラベルのクラスの確率＊視点の確率の統合を最大化するように学習する.



新規性・結果・なぜ通ったか？

- 物体認識においてはSHREC'17のnormalデータに対し優勝した．また，ModelNet-10,ModelNet-40に対し従来のマルチ視点・ポイントクラウド・ボクセルベースな様々な手法より良い精度を達成.
- 物体姿勢推定において，無監督な方法で従来の監督方法レベルな

コメント・リンク集

- クラス間のViewpoint-specificな特徴を学習することが面白い．可視化手法を加えて学習済みモデルに対しどういうようにアライメントしているのかを知りたい．また，問題定義を詳細的に考える必要がありそう
- 疑問点としては予測したそれぞれの視点の結果の統合は平均をと

[#147]

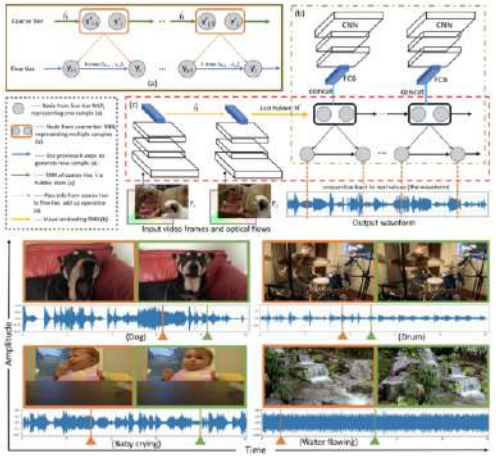
Yue Qiu

Visual to Sound: Generating Natural Sound for Videos in the Wild

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, Tamara Berg CVPR 2018 435

概要

- ビデオからリアルな音声を生成する(waveformな)手法及びビデオ一音声データセットを提案した．
- 人がビジョンとサウンド間の関連性をある程度把握できる．そこで，in-the-wildビデオから音声(waveform型)を自動生成するタスクを提案し，また，このタスクのためのデータセットVEGASを提案した．VEGASはAudioSetデータセットをAMTよりクリーンし，10カテゴリのビデオ及び対応した音声28109ペアから構成される．データセットのビデオの総時間が55時間となる．
- 提案タスクに対応したフレームワークはビデオエンコーダー及び音声ジェネレータから構成される．音声ジェネレータは階層的RNNを用いた．ビデオエンコーダーに対し:①frame-to-frame②sequence-to-sequence③flow-basedの３種類の設計を用いた．3種類モデルの生成結果に対し定量評価及びヒューマンテストを用いて評価し，flow-based構造が最も良い性能とヒューマン評価を達成した．



新規性・結果・なぜ通ったか？

- 従来のビデオから音声を生成する手法はビデオに対し拘束条件を加えている．提案手法は初めてのin-the-wildビデオから音声を生成する手法．
- ビデオから音声を自動生成する手法の応用場面が広い．(VRシステムでの没入感の増強，音声編集作業の自動化，視覚障害の人に視覚体験を聴覚体験として提供)
- ヒューマンテスト (ビデオがリアルかフェクか)に対し，ビデオエンコーダーをflow-basedな構造を用いた場合，平均73.36%の生成

コメント・リンク集

・視覚情報の抽出機に更にコンテンツと物体relationなどを重視したネットワークを用いたら更なる良い結果が得られそう・逆設定として，音声情報からビデオの予測も面白そう

- 論文
- コード

[#148]

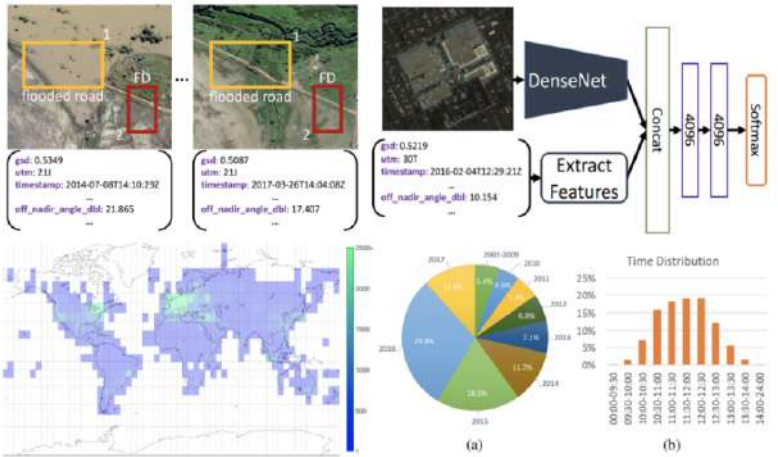
Yue Qiu

Functional Map of the World

Gordon Christie, Neil Fendley, James Wilson, Ryan Mukherjee CVPR 2018 795

概要

- 建物や土地などの機能的目的を予測するタスクに用いられる大規模な衛星画像データセットfMoWの提案(bounding box, 時系列，カテゴリ，メタ情報などのアノテーションがあり)
- データセットの具体的な統計情報は①200以上の国の１,047,691枚画像②63カテゴリ③一枚の画像1つ以上のバウンディングボクス定義④時系列画像が大量に含む．
- このデータセットに対応した新たなタスクを設定した：連続な時系列画像によりバウンディングボクス内の物体を認識する．提案データセットfMoWを用いて5つのネットワーク構造:LSTM-M,CNN-I,CNN-IM,LSTM-I,LSTM-IM(I:画像M:メタ特徴)に対し比較実験を行た．平均F1スコアにおいてLSTM-IMが最も高い精度を示したので，時系列情報及びメタ情報をジョイントでreasoningするアプローチの有効性を証明した



新規性・結果・なぜ通ったか？

- 公開されている最も大規模な衛星画像データセット．
- 異なる国・撮影時間・撮影年代などで撮影された画像から構成され，提案データセットを統計比較などにも用いられる．
- 従来の衛星画像データセットは主にbrief momentsの情報だけをキャプチャーし，メタ情報(ロケーション，時間，太陽角度など)がアノテーションされていない．提案データセットはメタ情報をアノテーションし，様々な応用を可能にした．(例：パーキングエリアの時系列駐車量の統計・影と時間情報によりオブジェクトの高さ推定など)

コメント・リンク集

- 地理情報に関する分析の研究に用いられるデータセット
- 国のバリエーションが豊かなデータセットなので，国ごと上空シーン特徴の比較などにも用いられる
- 論文

[#149]

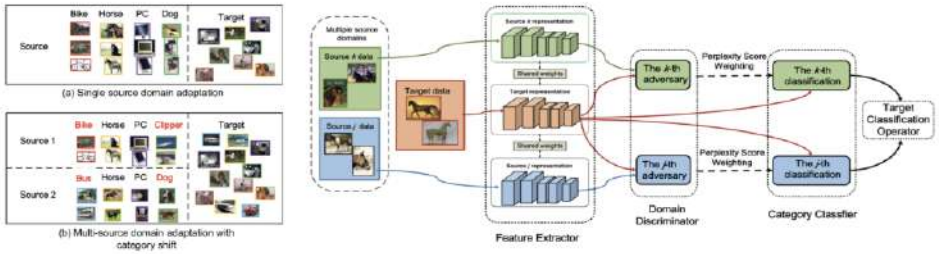
Kazuki Inoue

Deep Cocktail Networks: Multi-source Unsupervised Domain Adaptation with Category Shift

Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, Liang Lin CVPR 2018 Poster

概要

ソースドメインのラベル付きデータセットが複数ある場合のunsupervised domain adaptation(UDA)であるmultiple domain adaptation(MDA)によってターゲットドメインのクラシフィケーションを行う Deep Cocktail Network(DCTN)を提案。MDAではUDAで問題視されるドメインシフトに加えて、 ソースドメインのデータセット間で全てのカテゴリが共有されていないカテゴリシフトが存在する。 DCTNでは、k番目のソースドメインのデータセットとターゲットドメインのデータセットを入力として discriminatorによってperplexity scoreを算出することでどのソースドメインのデータセットの分布に近いかを算出し、 これを全てのソースドメインのデータセットに対して行い、perplexity scoreを重み付けるすることで最終的な識別結果を出力する。



新規性・結果・なぜ通ったか？

- discriminatorによってターゲットドメインがソースドメインのデータセットのうちどのデータの分布に近いかを計算することで、MDAに取り組むDCTNを提案。
- 3つのベンチマークにおいてUDAのstate-of-the-artと比較し他結果、提案手法が最も高い精度を達成。
- カテゴリシフトを解決できているかどうかを確認するために、ターゲットドメイン内でカテゴリの重複あり/なしにおける識別結果を比較したところ、 state-of-the-artと同等以上の精度を達成。

コメント・リンク集

- discriminatorが算出したperplexity scoreによって重み付けをするというシンプルな手法だが、UDAに取り組むstate-of-the-artよりも高い精度を達成している。
- 論文

[#150]

Kazuki Inoue

Unsupervised Correlation Analysis

Yedid Hoshen, Lior Wolf CVPR 2018 Poster

概要

2つのドメインを結合する手法であるCanonical Correlation Analysis(CCA、正準相関分析)を教師なし学習に対して行うUnsupervised Correlation Analysis(UCA)を提案。 既存のCCAは教師あり学習かつ2つのドメインが何らかの対応関係を持っていることを前提としていたが、 UCAは教師なし学習かつ2つのドメインに対応関係がない場合を想定している。 教師あり学習とは異なり、トレーニング時に2つのドメインにおける相関係数を計算することができないため、入力する2つのドメインと、 ネットワークによって射影された潜在変数空間の3つのドメイン間の射影、逆射影がうまくいくように様々なロスをとることで学習を行う。 ロスに対するablationも行なっている。

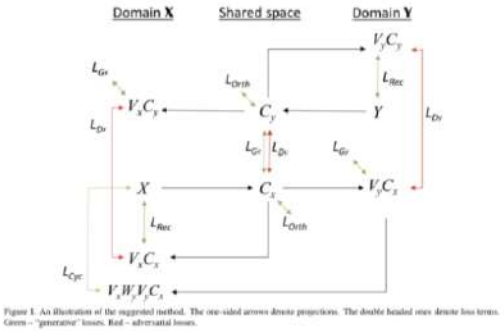


Figure 1. An illustration of the suggested method. The one-sided arrows denote projections. The double headed ones denote loss terms. Green - "generative" losses. Red - adversarial losses.

	Symbol	Description	Computed as:
Variables	$X \in \mathbb{R}^{d_x \times n_x}$	Samples in the first domain $\mathcal{X}$	Input
	$Y \in \mathbb{R}^{d_y \times n_y}$	Samples in the first domain $\mathcal{Y}$	Input
	$C_X \in \mathbb{R}^{d \times n_x}$	X projected to shared space	$W_x X$
	$C_Y \in \mathbb{R}^{d \times n_y}$	Y projected to shared space	$W_y Y$
Networks	$W_x : d_x \to d$	Projection form $\mathcal{X}$ to the shared vector space	
	$W_y : d_y \to d$	Projection form $\mathcal{Y}$ to the shared vector space	
	$V_x : d \to d_x$	Projection form the shared space in $\mathbb{R}^d$ to $\mathcal{X}$	
	$V_y : d \to d_y$	Projection form the shared space in $\mathbb{R}^d$ to $\mathcal{Y}$	
	$D_c : d \to [0, 1]$	The domain confusion discriminator C_X vs. C_Y	
	$D_x : d_x \to [0, 1]$	Discriminator for $V_x C_y$ vs. $V_x C_x$	
	$D_y : d_y \to [0, 1]$	Discriminator for $V_y C_x$ vs. $V_y C_y$	
Const	d_x	Dimensionality of the input domain $\mathcal{X}$	
	d_y	Dimensionality of the input domain $\mathcal{Y}$	
Parameters	d	Dimensionality of the shared space	10
	λ_c	Weight of the loss term L_{G_c}	1
	λ_{cx}	Weight of the losses $L_{G_x}, L_{G_y}, L_{D_x}, L_{D_y}$	1
	λ_{Rec}	Weight of the loss term L_{Rec}	1
	λ_{Orth}	Weight of the loss term L_{Orth}	1
	λ_{Cyc}	Weight of the loss term L_{Cyc}	1
	k	Number of reruns per experiment	100

Table 1. The components of the Unsupervised Component Analysis model

新規性・結果・なぜ通ったか？

- 教師なしかつ2つのドメインに対応関係がない状況におけるCCAの拡張であるUCAを提案。
- 評価尺度として潜在変数空間における相関係数、AUCを用いて以下の5つの状況で実験を行なった。1.MNISTの画像とそのミラー画像、2.MNISTの上半分の画像と下半分の画像、3.鳥の画像とそのキャプション、4.花の画像とそのキャプション、5.Flickerの画像とそれに付随する5つの文章。 関節位置のエラーを測定したところ上記のstate-of-the-artの手法と同等、あるいは上回る精度を達成。
- 教師なし学習の結果をGANと比較しており、全ての実験においてGANよりも高い精度を達成。
- 教師あり学習をUCAで行なった結果も乗せられており、実験3、

コメント・リンク集

- 現状のネットワークを見ると、それぞれのドメインにおける直交性と、それぞれのドメインの射影先が同じ空間になるように様々なロスをとっているだけなので、 もう少しアップデートすることができるかもしれない。
- CCAの特徴であるL_Orthだけを除いた場合に、どれほどの影響が出るのかが気になった。
- 論文

[#151]

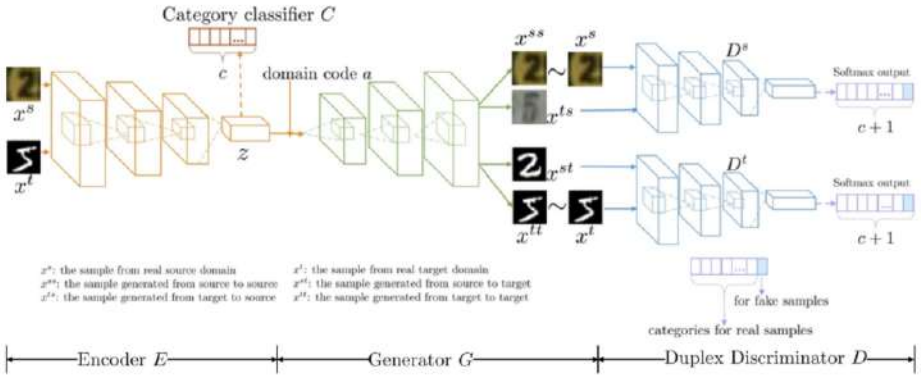
Kazuki Inoue

Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification

Jingya Wang, Xiatian Zhu, Shaogang Gong, Wei Li CVPR 2018 Poster

概要

ラベルなしデータセットにおいてperson re-identification(re-id)を教師なしで行うために、ラベルありデータセットからdomain adaptationを行うTransferable Joint Attribute-Identity Deep Learning(TJ-AIDL)を提案。person re-idとは、街中の監視カメラのような異なる視点、 重複のない領域を撮影された映像内の同一人物を探すことである。 TJ-AIDLにはアイデンティティーを推定するIdentity branch、アトリビュートを推定するAttribute branch、 アトリビュートからアイデンティティーを推定するモジュールであるIdentity Inferred Attirbute(IIA)からなる。 domain adaptationの際には、Attribute branch、IIAの更新のみを行う。



新規性・結果・なぜ通ったか？

- domain adaptationを用いて教師なしでperson re-idを行うために、画像のアトリビュートからアイデンティティーを推定するTJ-AIDLを提案。
- personn re-idのベンチマークである4つのデータセットを使用しており、Rank-1mAPにおいてre-idを教師なしで行うstate-of-the-artよりも高い精度を達成。
- TJ-AIDLにおいてアトリビュート/アイデンティティーのみ学習した際の結果、adaptation有り/無しの結果についても議論しており、提案したTJ-AIDLが最も高い精度となった。

コメント・リンク集

- 論文

[#152]

Kazuki Inoue

Duplex Generative Adversarial Network for Unsupervised Domain Adaptation

Lanqing Hu, Meina Kan, Shiguang Shan, Xilin Chen CVPR 2018 Poster

概要

同一カテゴリのdomain間におけるadaptation, transferをラベル識別と2つのdiscriminatorを用いるネットワークDupGANを提案。target domainにはラベルがない状況である教師なし学習を対象としている。 DupGANはencoderでそれぞれのドメインの潜在変数をエンコードし、generatorでデコードを行い、 2つのdiscriminatorでそれぞれのドメインに対してfake/realとラベルの認識を行う。 結果はdomain transferされた数字画像のラベル認識・生成結果、物体認識の精度において比較を行う。

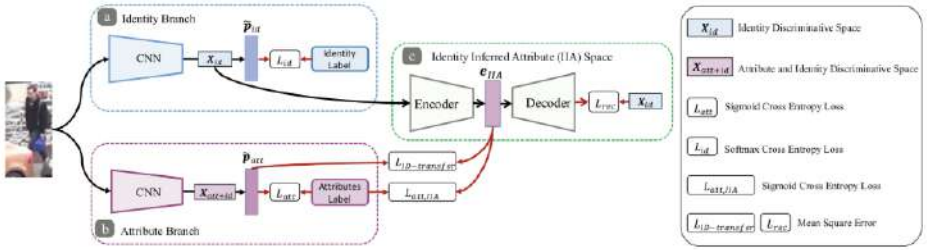


Figure 1. An overview of the proposed Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL).

新規性・結果・なぜ通ったか？

- ラベル認識と2つのdiscriminatorによってdomain adaptaion/transferをおこなうDupGANを提案。
- 既存手法であるDANN、ADDAはadversarial lossを使用してtarget→source のマッピングを行うが、 これらの手法ではマッピングされたtarget domainの分布が歪んでいないことは保証できない。 一方DupGANではラベルの認識を行わせることでカテゴリ構造を保つことができる。また提案手法では画像の生成も可能である。
- state-of-the-artと比較して、数字画像データセットであるMNIST、USPS、SVHN、SVHN-extraそれぞれのデータセット間におけるdomain transferに対するラベル認識の結果、 最も高い精度を達成。またdomain transferによる画像も生成することが可能。
- 31種類のラベル、3つのドメインを持つOffice-31データセットにおける物体認識結果がstate-of-the-artよりも高い精度を達成。

コメント・リンク集

- クラシフィケーション生成された画像ではなくはエンコードされた潜在変数に対して行われている。
- 画像の生成力はそこまで高くなく、実際Office31に対する画像生成は難しかったと主張している。
- 論文

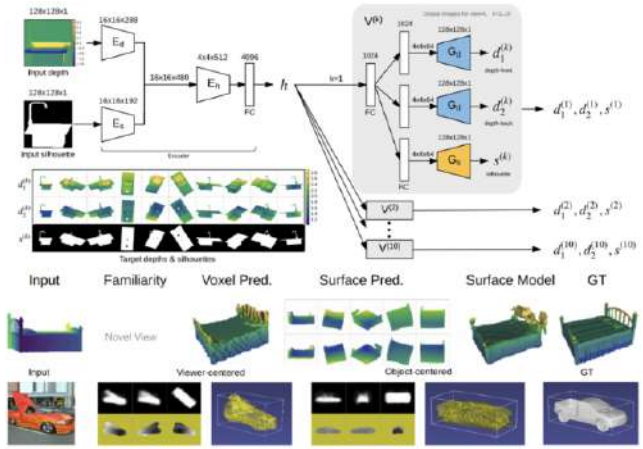
[#153] Yue Qiu

Pixels, voxels, and views: A study of shape representations for single view 3D object shape prediction

Daeyun Shin, Charless Fowlkes, Derek Hoiem CVPR 2018 384

概要

- 1枚の画像から3次元形状を推定するタスクにおいて，異なる形状representation及びcoordinate framesを用いた場合，精度がどのように変化するのかの徹底的比較実験に関する研究．
- 従来形状推定タスクにおいて異なる設計の比較分析の研究がないので，著者達が異なる設計を比較できるフレームワーク及び具体的な実験を行った．
- 比較実験は具体的に，a.RGB画像b.デプス画像からの形状推定タスクにおいて，“①マルチサーフェス画像VS volumetricデータ表示②viewer-centered VS object-centeredな座標”などの設定に対し，定量的及び定性的な比較実験を行った．
- 提案の比較用フレームワークはencoder-decoderベースなネットワークを用いて，decoderに変更を加えることで， マルチサーフェス画像及び volumetricデータの2種類を生成できるようにした．また，coordinate frameをスイッチすることにより，viewer/object centeredを変更できる．



新規性・結果・なぜ通ったか？

- 3次元形状推定タスクにおいて，異なる設定の比較実験を行った．
- 形状representationの設定において，Multi-surfaceの方がvoxelと比べunseenクラスにおいてより良い性能を達成した． Multi-surfaceの方が高い解像度をエンコーディングできるのが理由な可能性があると指摘した．

コメント・リンク集

- 比較をしていない設計(Oct-tree based representationなど)もあるので，そういった構造に対して比較実験を行うのも面白い．
- 3次元あたりの徹底的比較を行って，何らかの結論を出すような研究がまだ少ないので，研究テーマを沢山作れるかも？

[#154]

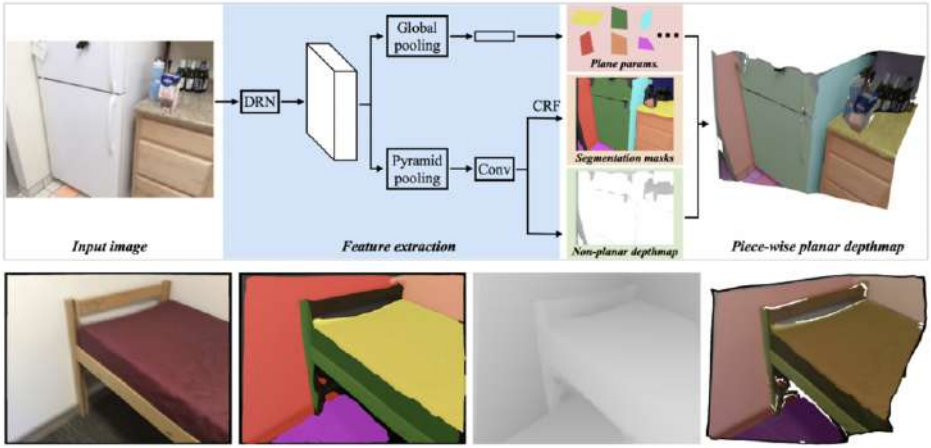
Yue Qiu

PlaneNet: Piece-wise Planar Reconstruction from a Single RGB Image

Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, Yasutaka Furukawa CVPR 2018 336

概要

- 1枚のRGB画像から“piece-wise planar depthmap”を推定するend-to-endなネットワークを提案した．提案手法を用いてRGB画像から平面パラメータ及び平面セグメンテーションマスク及びデプスマップを同時に推定できる．
- 画像からpiece-wiseな平面を検出するタスクはARの応用に一つ重要なタスクとなっている．しかし従来，デプス推定とpiece-wiseな平面検出を同時に行う研究がない．著者達が新たにこのタスク及びタスクに対応できるネットワークを定義した．
- 提案フレームワークは:①DRNs(Dilated Residual Networks)を用いて入力画像から特徴抽出を行う②平面パラメータ推定・non-planarデプスマップ推定・セグメンテーションマスク推定の3つの推定ネットワークを用いる③推定した3つの結果から“piece-wise planar depthmap”を生成する．



新規性・結果・なぜ通ったか？

- 新規な問題定義．実験で提案手法が部屋のレイアウト推定・ARアプリ(テクスチャー編集・バーチャルルーラーなど)に応用できることを指摘した．
- 51,000枚ほどの学習データを作成した．(これが大変そう)
- plane segmentationタスクにおいてNYUデータセットでの精度が従来の三つの手法より優れている(比較している手法は2009年，2009年，2012年の手法だけど。。)
- デプスマップ推定タスクにおいてNYUv2データセットにおいて前述した３つの手法より精度良い

コメント・リンク集

- ARアプリに応用できるところから考えると単純なデプス推定より実用性が高い
- 平面検出も同時に行うので，部屋レイアウト推定に良い精度を達成したのが理解できる．しかし，疑問としては提案手法が平面検出＋デプス推定だけで部屋の幾何構造実際は学習していないので，デプス推定＋平面パーツ検出の従来研究と比べると新規性と技術的の難しさがどこなのかちょっとわからない

[#155]

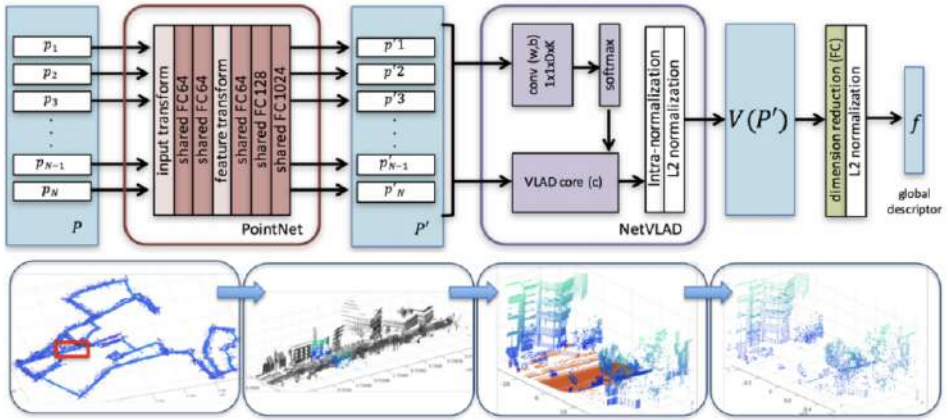
Yue Qiu

PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition

Mikaela Angelina Uy, Gim Hee Lee CVPR 2018 573

概要

- PointNetとNetVLADを用いたポイントクラウドベースな“場所検索”ネットワークPointNetVLAD及びデータセットの提案．
- 従来の自動運転などに用いられる場所検索技術では2次元画像ベースで行われている．しかし，照明条件などに対しロバスト性が低い．ポイントクラウドベースな場所検索が従来良いグローバル特徴抽出機がないため，まだ研究されていない．近年PointNetなどの良いポイントクラウド特徴抽出機が提案され，そこで著者達がPointNetとNetVLADを用いたLiDARで撮ったポイントクラウドをベースとした場所検索手法を提案した．
- 提案データセットの収集過程は:①Oxford RobotCar などのdatasetからフルールートを選択する②フルールートから局所を選択する③選択した局所ポイントクラウドをダウンサンプルと正規処理を行う．また，Oxford RobotCar 以外，3種類の他のデータセットからデータを集めた．
- fixedサイズなポイントクラウドからグローバル特徴を抽出できるPointNet，NetVLADと全結合層をコンバインたend-to-endなグローバル特徴抽出機を構築した．



新規性・結果・なぜ通ったか？

- 新規なポイントクラウドベースな場所検索及び場所検索3次元ポイントクラウドデータセットの提案．
- 従来の2次元画像ベースな場所検索と比べ，提案したポイントクラウドベースな場所検索が照明条件にロバストである．

コメント・リンク集

- PointNet，PointNet++, Kd-networkなどのポイントクラウドデータを扱えるネットワークでポイントクラウドから情報抽出を利用した研究がこれからまだ増えるのかな？

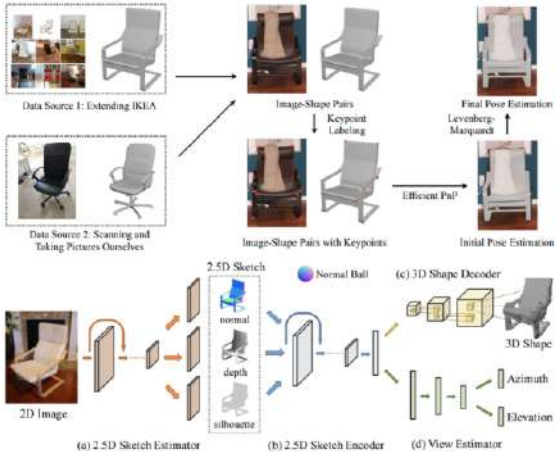
[#156] Yue Qiu

Pix3D: Dataset and Methods for 3D Object Modeling from a Single Image

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Tianfan Xue, Joshua Tenenbaum, William Freeman CVPR 2018 375

概要

- 大規模なピクセルレベルに対応付けられたimage-shape pairsデータセットPix3Dの提案及び画像から同時に三次元形状及び姿勢を推定するネットワークの提案．
- 従来のimage-shape pairsデータセットは①合成データセットを用いる②image-shapeの対応が精密ではない③データセット規模が小さいなどの問題点がある．そこで，著者達が大規模なピクセルレベルに対応付けられたデータセットを提案した．Pix3Dは395個の3次元物体モデル(9カテゴリ)，10069ペアの画像—形状ペアから構成される．画像と形状のペアはピクセルレベルの精密的に対応付けられている．
- データセットの収集段階では:①IKEA及び自撮りで大量な画像—形状ペアを集める②AMTにより画像からキーポイントをアノテーションする③Efficient PnP及びLevenberg-Marquardtを用いて粗い・精密なposeを求める．
- 更に，提案手法は画像から同時に姿勢及び3次元形状を予測できるネットワークを提案した．提案ネットワークはまず画像から2.5Dスケッチを推定し，推定したスケッチをエンコーディングする．また，デコーディングにより3次元形状を推定し，同時にview estimatorネットワークにより姿勢を推定する．



新規性・結果・なぜ通ったか？

- 従来のデータセットではCGモデルで合成されている方が多く，提案のデータセットが実物体を用い，更にピクセルレベルな精密度の画像—形状対応付けアノテーションがある．
- 画像から同時に形状姿勢を推定するフレームワークの定量化結果

コメント・リンク集

- 現在の学習データアノテーション段階でAmazon Mechanical Turkを用いている．Semantic Keypointの自動的検出を用いたら自動化できることはデータセットの更なる拡大化につなぎられそう

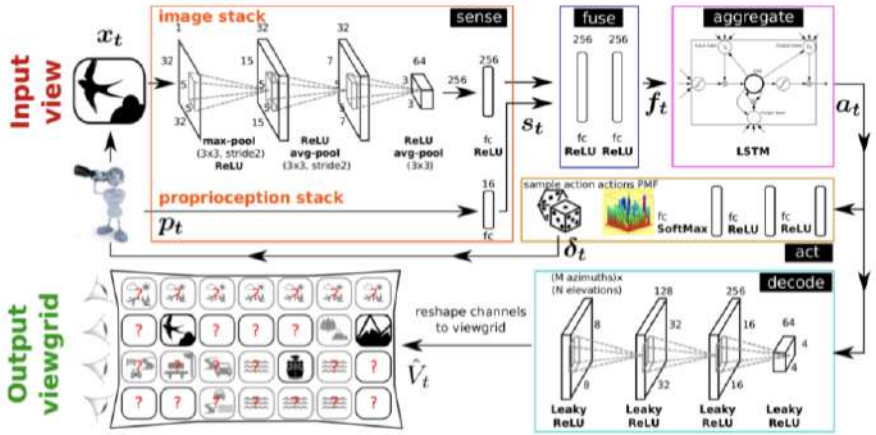
[#157] Yue Qiu

Learning to Look Around: Intelligently Exploring Unseen Environments for Unknown Tasks

Dinesh Jayaraman, Kristen Grauman CVPR 2018 152

概要

- 新規な問題設定“シーンや物体を有効的に観測できる視点を学習する”及びこの問題を対応できる “アクティブ観測補完”ネットワークの提案.
- 従来のCVタスクは主に与えられた観測(画像・ビデオ・ポイントクラウドなど)から視覚性質(クラス分類・検出など)の分析を行う. しかし，リアルな知能はまず環境から目的を達成するための観測を取得することから始まる．また，異なる観測から得られる情報量も異なる．そこで，著者達が“active observation completion”タスクを提案し，未知なシーンかオブジェクトからシーン及び物体のより多く3次元情報が含めた数が限られた観測視点の推定を目標とする.
- 提案手法は強化学習を用いる．RNNベースなネットワークを用いて選択された視点からシーンか物体のパーツ情報を統合する．また，統合されたモデルから推定できるunobserved視点とgt間の誤差をベースにロス関数を設定した.



新規性・結果・なぜ通ったか？

- 学習データを手動でラベリングする必要がないので，大量な学習が行える.
- 提案フレームワークを“シーン”の補完及び“物体モデル”の補完の2種類だいぶ異なったタスクに実験を行い，良い精度を達成したので，”提案した“無監督探索的な”フレームワークを遷移学習でほかのタスクに用いられる.

コメント・リンク集

- Interactive 環境でのVQAタスク(Embodied Question Answering など)は環境から“情報量が豊かな画像”を集めるのが重要の一環なので，提案フレームワークを用いられそう.

- 論文

[#158]

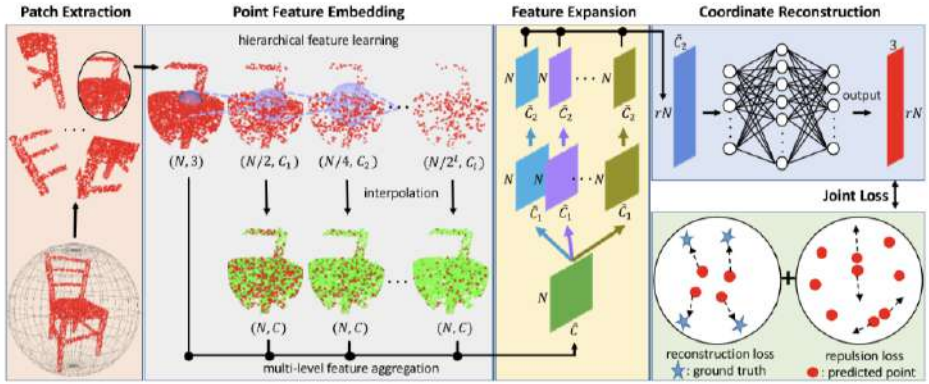
Yue Qiu

PU-Net: Point Cloud Upsampling Network

Lequan Yu, XIANZHI LI, Chi-Wing Fu, Daniel Cohen-Or, Pheng-Ann Heng CVPR 2018 355

概要

- data-drivenなポイントクラウドアップサンプリング手法の提案．スパースなポイントクラウドから，もっとデンスでユニフォームなポイントクラウドを取得できる．
- 従来の2D画像super-resolutionタスクと比べ，3D Upsamplingでは処理対象が空間オーダーとレギュラー構造がないポイントクラウドで，物体の本当のサーフェス(ポイントクラウドのリアル物体)に近づき，点の密度も均等であることがタスクの目標となる．こういったことから，提案手法はポイントクラウドからマルチレベルの特徴を抽出し，更にマルチブランチで特徴を拡張することにより，ポイントクラウドの局所及びグローバルな情報を取得できる．
- 提案ネットワークPU-Netは入力のポイントクラウド(N points)に対し①ポイントクラウドに対し異なるスケールのパッチを抽出し，②パッチからPointNet++を用いたマルチレベルの特徴抽出を行う．③feature expansion構造により特徴を拡張し，④全結合層を用いて出力のポイントクラウド(N＊ｒ points)を生成する．また，物体のサーフェスまでの距離及びポイントクラウドの過密程度を基準に，ジョイントロスを設計した．



新規性・結果・なぜ通ったか？

- 新たな評価指標：“物体のサーフェスまでの距離偏差”及び“ポイントクラウド分布のユニフォーム性”を評価できる指標を提案し，この2つの指標においてSHREC2015データセットに対し従来研究より優れた精度と指摘した．
- Pointnet++を用いてローカル及びグローバル情報抽出を行うの

コメント・リンク集

- 提案手法を更に発展し物体モデルの補完およびアップサンプリング同時にできることを期待される
- Pointnet++を基本構造として使っていることがすごそう

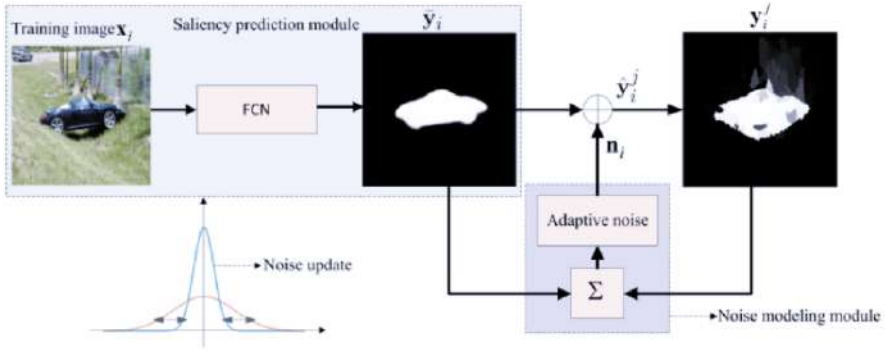
[#159] KotaYoshdia

Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective

J.Zhang, T.Zhang, Y.Daiy, M.Harandi, and R.Hartley CVPR2018 arXiv:1803.10910

概要

深層学習を用いた教師あり学習による顕著性の検出方法は教師データに依存する．そこで，“汎化能力を改善しつつ教師データなしで顕著性マップを学習することは可能か？”という問いに対して，弱いものやのノイズのある教師なし顕著性検出手法によって生成される多数のノイズラベルを学習することによって教師なしで顕著性の検出を行った．



新規性・結果・なぜ通ったか？

- 従来の教師なし顕著性検出に新たな顕著性を推定し，複数のノイズの多い顕著性検出方法から顕著性マップを学習する．
- 我々の深層学を用いた顕著性検出モデルは，人間のアノテーションなしでEnd to Endで学習できとても簡潔である．

結果・リンク集

- 評価実験をしたところ従来の教師なしの顕著性検出方法を大きく上回り，深層学習を用いた顕著性の精度と同等のものとなった．
- Paper

[#160]

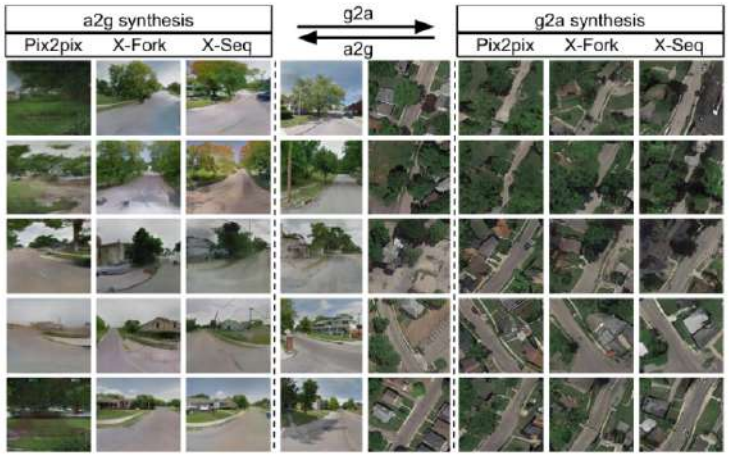
Naofumi Akimoto

Cross-View Image Synthesis using Conditional GANs

Krishna Regmi and Ali Borji CVPR2018

概要

対応する航空写真とストリートビュー写真間の変換を行うcGANを提案．pix2pixによる変換に比べて，オブジェクトの正しいセマンティックスを捉え維持する変換が可能となっている．提案したcGANモデルは２つあり，X-Fork とX-Seq と呼んでいる．出力が変換画像とセグメンテーションマップであることが特徴．Inception Scoreの比較実験をすると，航空写真からストリートビュー方向の変換ではがX-Forkが優れ，逆方向の変換ではX-Seqの生成結果が優れていることがわかった．



256x256の解像度で生成可能．gがストリートビューで，aが航空写真に当たる．

手法

- X-Forkは１つのGeneratorと１つのDiscriminatorから成るシンプルな構成のcGAN．出力は変換後の画像とセグメンテーションマップの２つであることが特徴．
- X-Seqは２つのGeneratorと２つのDiscriminatorから成るcGAN．
  １つ目のGeneratorで変換後の画像を生成．それを元に２つ目のGeneratorでセグメンテーションマップを生成する．
  セグメンテーションマップのGround-Truthには，学習済みのRefineNetを用いた生成結果を使用している．

コメント・リンク集

- 航空写真とストリートビューという劇的に見た目が変わる場合の変換において，どのようなことが問題点となるのか５つ挙げられていたので気になる場合は元論文を参照してください．
- コードやデータは公開予定
- arXiv

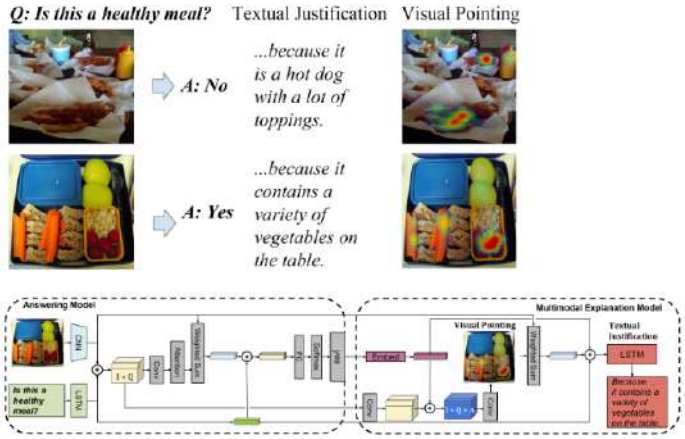
Kensho Hara

Multimodal Explanations: Justifying Decisions and Pointing to the Evidence

D. H. Park et al., CVPR 2018

概要

性能がよく，かつ説明可能なモデルの実現のための新規手法の提案． これまでの説明可能なモデルは視覚的なAttentionのみやテキストの説明のみという単一のmodalだけだったのに対して， この論文では両者を合わせたmulti-modalな説明を出力可能にした． それを行う手法の提案と，学習と評価に使うデータセットを構築したのがこの論文のContribution． データセットはVQAと静止画からのActivity Recognitionのタスクで， 従来あったデータセットに，理由のテキスト説明と視覚的な根拠となった領域のアノテーションを追加して作成． 手法は，まず答えを出力して，それを元に根拠となった理由を出力するという形式のネットワーク構造を採用．



新規性・結果・なぜ通ったか？

- モデルの出力に加えて視覚的，テキストのmulti-modalな根拠説明をする手法を提案
- VQAとActivity Recognitionでそれを評価可能なデータセット（追加アノテーション）を構築

コメント・リンク集

- 論文 (arXiv)
- データセットはまだ公開されていない模様

[#162]

Kazuki Inoue

A Variational U-Net for Conditional Appearance and Shape Generation

Patrick Esser, Ekaterina Sutter, Björn Ommer CVPR 2018 Poster

概要

画像を構成する成分はshape(ジオメトリ、ポーズなど)とappearanceであるという考えのもと、VAEによってappearanceを推定し、 U-Netにshapeを学習させることで入力画像のappearanceとshapeの 片方を保ったままもう一方を変更することが可能なVariational U-Netを提案。 通常のVAEではshape、appearanceの分布を分離することが不可能なため、 VAEに画像とshapeを入力することでappearanceの特徴量を抽出し、U-Netによってshape情報を保つように学習を行う。 shapeとして体のポーズや線画が入力される。トレーニングデータには同一物体に対する様々なバリエーションの画像は必要としない。

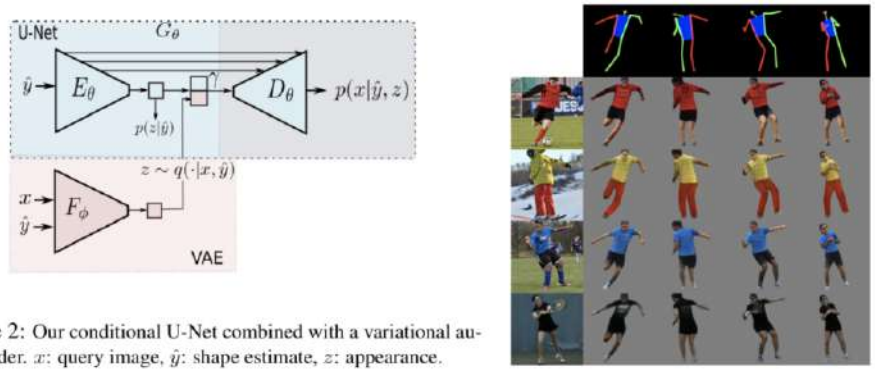


Figure 2: Our conditional U-Net combined with a variational autoencoder. x: query image, $\hat{y}$: shape estimate, z: appearance.

新規性・結果・なぜ通ったか？

- VAEでappearanceを、U-Netでshapeを学習させることで画像に内在する2つの事前分布を別々に学習することができるVarational U-Netを提案。
- コンディションによって画像を編集するpix2pixとポーズをコンディションとして人物画像を編集するPG2と比較を行った。COCO、DeepFashion、Market-1501データセットにおいてSSIMやIS、 関節位置のエラーを測定したところ上記のstate-of-the-artの手法と同等、あるいは上回る精度を達成。

コメント・リンク集

- VAEとU-Netのいいとこ取りをすることで、2つの変数を扱うことが可能になった。
- 論文
- Project page
- GitHub

[#163]

Kazuki Inoue

Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies

Hanbyul Joo, Tomas Simon, Yaser Sheikh CVPR 2018 Oral

概要

表情、体全体の動き、手のジェスチャといった様々なスケールの動きをマーカー無しでキャプチャするdeformation modelである”Frankenstein”と”Adam"を提案。 3Dキャプチャシステムに置いて、画像の解像度と3Dキャプチャシステムの視野はトレードオフであるため、 体の局所的な動きと全体的な動きを同時に捉えことは難しかった。提案手法では顔、両手、両足、 手の指における3Dキーポイントと3D Point Cloudを用いて表情などの 局所的モーションと体全体のモーションをキャプチャすることができるFrankensteinを構築。 また70人のトラッキングデータを用いてFrankensteinモデルを最適化することで、 髪と服を表現することが可能なAdamモデルを提案。結果は既存手法とのトラッキングの精度によって比較している。

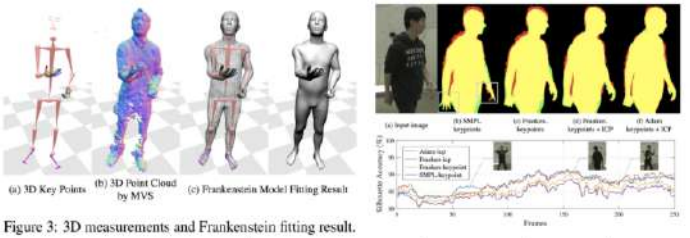


Figure 3: 3D measurements and Frankenstein fitting result.

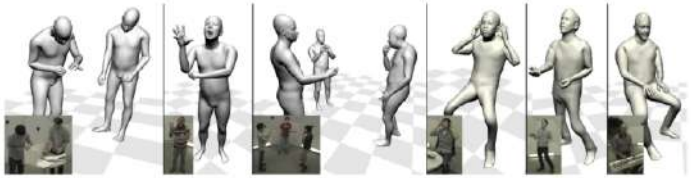


Figure 1: Frankenstein (silver) and Adam (gold). This paper presents a 3D human model capable of concurrently tracking the large-scale posture of the body along with the smaller details of a persons facial expressions and hand gestures.

新規性・結果・なぜ通ったか？

- 表情や手のジェスチャといった局所的なモーションと、体全体の動きを同時にトラッキングすることが可能なdefromation modelを提案。620台のVGAカメラと31台のHDカメラが必要とする。
- state-of-the-artであるSMPLでは顔の表情を表現することは不可能だが、提案手法では可能になっている。
- SMPLとトラッキングにおけるGTとのオーバーラップを計算した結果、SMPLが84.79%であるのに対し提案手法は87.74%となり、提案手法の方が高い精度を達成

コメント・リンク集

- 論文
- Project Page
- Video

[#164]

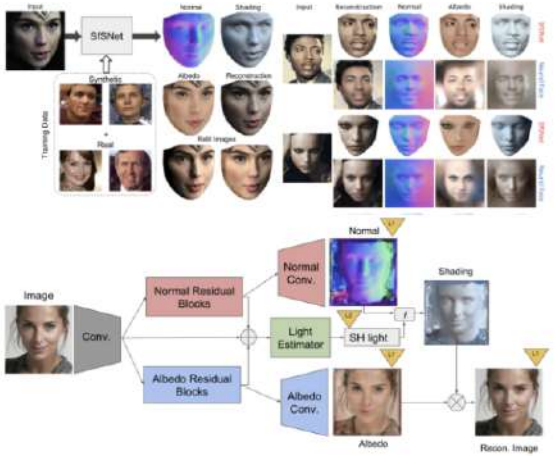
Kazuki Inoue

SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, ; David Jacobs CVPR 2018 Poster

概要

ラベル付き合成顔画像とin-the-wildなラベルなし実顔画像のどちらもトレーニングデータとして使用することで、実顔画像からシェイプ、リフレクタンス、イルミネーションを推定してリコンストラクションをend-to-endに行うSfSNetを提案。 実顔画像に十分なラベルがついているデータセットが存在しない、という問題を解決。 Shape from Shading(SfS)のアイディアに基づき、 低周波成分を合成顔画像から、高周波成分を実顔画像から推定する。リコンストラクションされた画像のL1ロスを取ることで、 トレーニングにおける合成顔画像と実画像の橋渡しが行われる。リコンストラクションにはランバーシアンレンダリングモデルを使用する。



新規性・結果・なぜ通ったか？

- ラベル付きの合成顔画像とラベルなしの実世界顔画像でトレーニングすることで、実世界顔画像の法線、アルベド、シェーディングを推定しインバースレンダリングを行うSfSNetを提案。
- インバースレンダリングによってリコンストラクションされた画像のロスを取ることで、合成顔画像と実世界顔画像の橋渡しを実現。
- インバースレンダリングの見た目がstate-of-the-artよりも良い結果となった。
- 法線・シェーディングの推定精度が、法線・シェーディング単体をそれぞれ推定するstate-of-the-artよりも良い結果となった。

コメント・リンク集

- 画像をリコンストラクションする際によく使われるU-NetではなくResNetを使った理由についても議論されている。
- 論文
- Project Page
- GitHub

[#165]

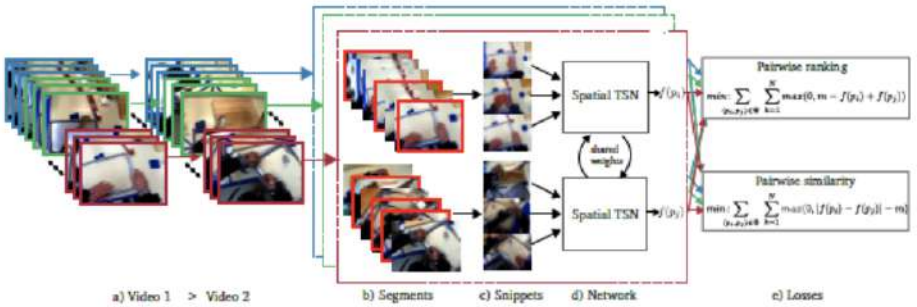
Shintaro Yamamoto

Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination

Hazel Doughty, Dima Damen and Walterio Mayol-Cuevas CVPR 2018

概要

2つの動画から、手術や絵を描くなどの技能がどちらが上かを予測する手法の提案。入力動画をTemporal Segment Networks(リンク参照)によりいくつかのセグメントに分割し，技能評価に用いるフレームを3枚選択する。 技能評価の学習は、2つの動画のどちらが技能が上か、2つの動画の技能が同じであるとき同じであると判定できるかの2つの尺度をロスとして行う。 技能を表すスコアは、Two Stream CNN(リンク参照)によって空間と時間それぞれについてスコアを取得する。



新規性・結果・なぜ通ったか？

手術、ピザ生地をこねる、絵を描く、箸を使うの4つの技能を撮影したデータセットにより実験を行った。そのうち絵を描く、箸を使うは新たにデータセットを構築した。 全てのタスクで70%以上の精度を達成し、箸を使う以外のタスクではベースラインと比べ精度が向上した。

コメント・リンク集

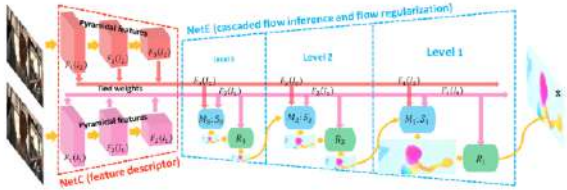
- 論文
- Two Stream CNN
- Temporal Segment Networks

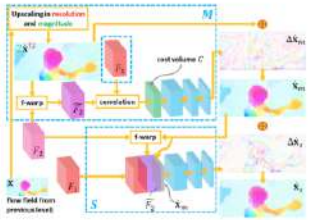
Kensho Hara

LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation

T. Hui et al., CVPR 2018

概要





FlowNet2よりも，性能が良く，モデルサイズが小さく，高速に動作するOptical Flow推定手法を提案． FlowNet2（Feature Warping, Correlation）は性能が良いけどモデルサイズが大きい，
SPyNet（ピラミッド構造を採用）はモデルが小さいけど性能はあまり良くない， ということで，提案手法は両者の良いところを合わせることをしている．2フレームを入力として，各フレームをCNNに入れてピラミッド構造の特徴表現を得る． 一番解像度の低いところから順にFlow推定を繰り返していって洗練化していく． 各Flow推定では軽量な2つのモデルをカスケードさせたりして2フレーム間の大きな移動にも対応しながら， 軽量かつ高速な推定を実現．

新規性・結果・なぜ通ったか？

- 軽量な2つのネットワークをカスケードさせて使うCascaded flow inferenceの提案
- CNNベースのFlow推定にFlow Regularizationを導入
- 高性能，省メモリ，高速な推定を実現

コメント・リンク集

- 論文 (arXiv)
- プロジェクトページ
- コード (GitHub)
- カスケード構造が複雑でなぜこれが良いのか少し納得しにくい
- 実験は各コンポーネントのON/OFFで性能比較がわかりやすい

[#167]

Yuta Matsuzaki

Person Transfer GAN to Bridge Domain Gap for Person Re-Identification

Longhui Wei, Shiliang Zhang, Wen Gao and Qi Tian CVPR2018

概要

Person Re-identification (ReID)のパフォーマンスは大きく向上したが，複雑なシーンや照明の変化、視点や姿勢の変化といった問題の調査は未だなされていない．本稿ではこれらの問題に関する調査を行った．このためにMulti-Scene MultiTime person ReID dataset (MSMT17)を構築した．またドメインギャップがデータ間に存在するため，このドメインギャップを埋めるためのPerson Transfer Generative Adversarial Network (PTGAN)を提案した．実験ではPTGANによってドメインギャップを実質的に狭められることを示した．

Dataset	*MSMT17*	*Duke* [40]	*Market* [38]	*CUHK03* [20]	*CUHK0i* [19]	*VIPeR* [8]	*PRID* [10]	*CAVIAR* [3]
BBoxes	**126,441**	36,411	32,668	28,192	3,884	1,264	1,134	610
Identities	**4,101**	1,812	1,501	1,467	971	632	934	72
Cameras	**15**	8	6	2	10	2	2	2
Detector	**Faster RCNN**	hand	DPM	DPM, hand	hand	hand	hand	hand
Scene	**outdoor, indoor**	outdoor	outdoor	indoor	indoor	outdoor	outdoor	indoor

新規性・結果・なぜ通ったか？

- ReIDを行う際の現実的な問題について網羅的に調査

コメント・リンク集

- 論文

[#168]

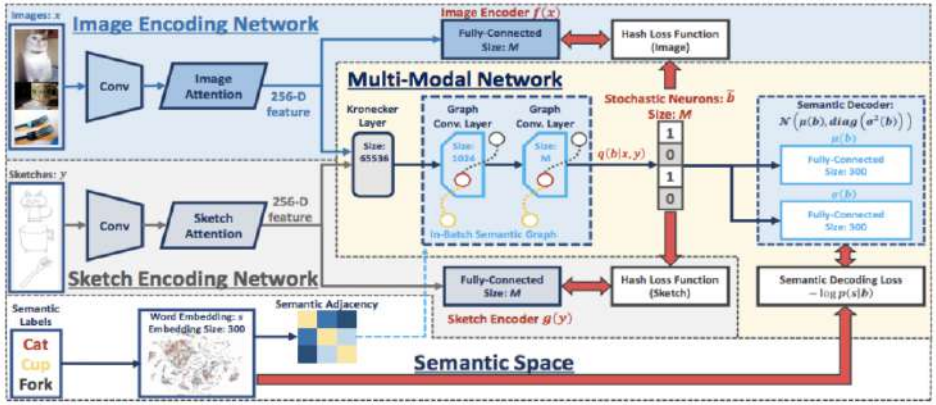
Yuta Matsuzaki

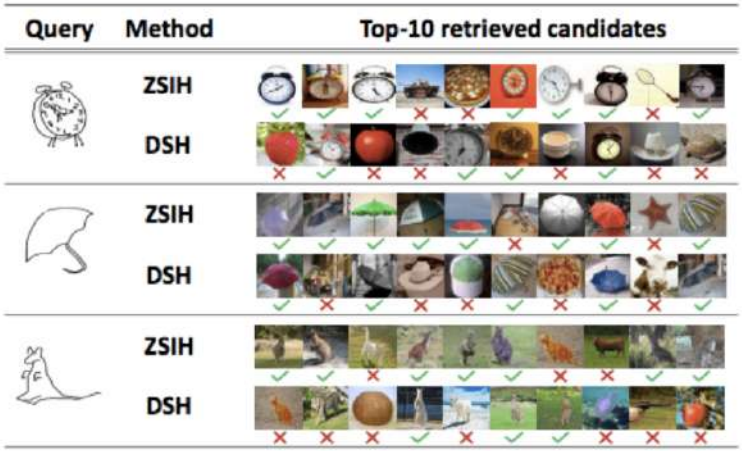
Zero-Shot Sketch-Image Hashing

Yuming Shen, Li Liu, Fumin Shen and Ling Shao CVPR2018

概要

大規模スケッチベース画像検索において，既存の手法では学習中にカテゴリの存在しないスケッチクエリがある場合失敗するという問題がある．本稿ではそのような問題を解決するZero-shot Sketch-image Hashing(ZSIH)モデルを提案した．2つのバイナリエンコーダとデータ間の関係を強化する計3つのネットワークで構成される．重要な点として，Zero-shot検索での意味的な表現を再構成する際に生成的ハッシングスキームを定式化する点である．Zero-shotハッシュ処理を行う初のモデルであり，関連する研究と比較しても著しく精度が向上した．





新規性・結果・なぜ通ったか？

- スケッチイメージハッシングの研究において初のZero-shot
- 意味的な表現を再構成する際に生成的ハッシングスキームを定式化

コメント・リンク集

- 論文

[#169]

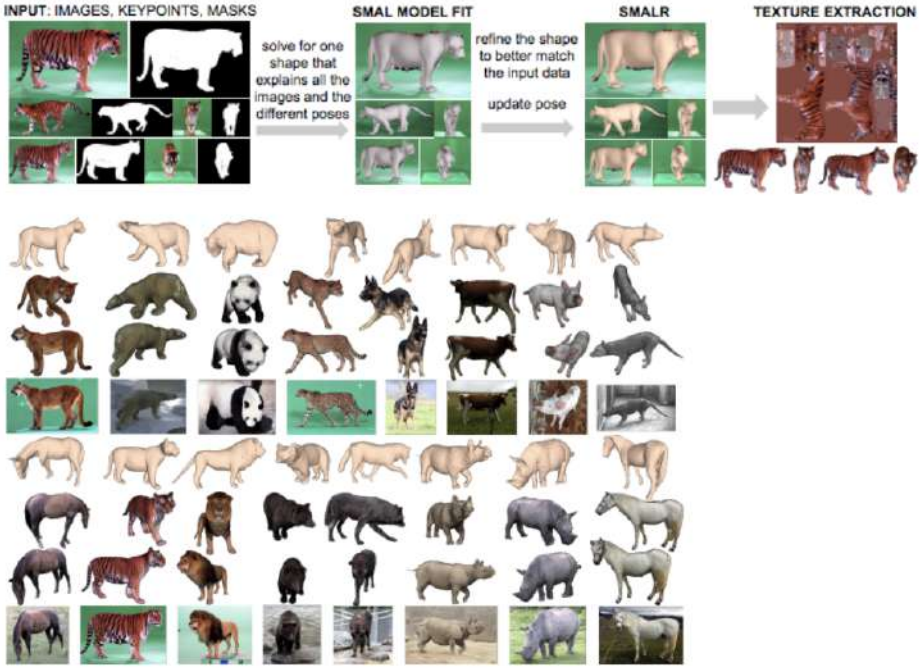
Yuta Matsuzaki

Lions and Tigers and Bears: Capturing Non-Rigid, 3D, Articulated Shape from Images

Silvia Zuffi, Angjoo Kanazawa and Michael J. Black CVPR 2018

概要

3Dスキャンは人間をキャプチャするために設計されており，自然環境での使用や野生動物のスキャンおよびモデリングには不向きという問題がある．この問題を解決する方法として，画像から3Dの形状を取得する方法を提案した．SMALモデルを画像内の動物にフィット，形状が一致するようにモデルの形状を変形(SMALR)，さらに複数の画像においても整合性がとれるよう姿勢を変形させ、詳細な形状を復元する．本手法は，従来の手法に比べ大幅に3D形状を詳細に抽出することを可能にするだけでなく，正確なテクスチャマップを抽出し，絶滅した動物といった新しい種についてもモデル化できることを可能にした．



新規性・結果・なぜ通ったか？

- 3Dスキャンが困難な動物のモデルを構築する方法を提案
- SMALモデルを基として形状を変形させることで，より詳細な3D復元が可能
- 上記手法により，一貫したテクスチャマップの抽出が可能

コメント・リンク集

- 論文

[#170]

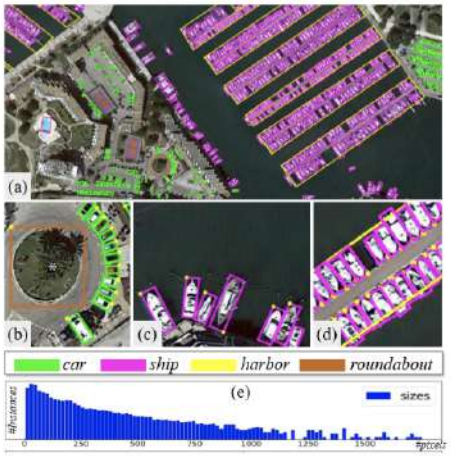
Tetsuya Narita

DOTA: A Large-scale Dataset for Object Detection in Aerial Images

Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, Liangpei Zhang CVPR2018

概要

俯瞰画像から物体検出するためのデータセットを提案．従来のデータセットのものよりも小さい物体が多いデータセットである．各画像は4000×4000ピクセルであり，さまざまな大きさ，向き，形状を示す物体を含む．データセットは15カテゴリに分類されており，188282のインスタンスを含み，それぞれは任意の四角形でラベリングされている．人工衛星での物体検出の基礎構築のために，DOTA上の最先端の物体検出アルゴリズムを評価した．



新規性・結果・なぜ通ったか？

俯瞰画像データセット内のインスタンスは小さいものの割合が高く，細かいものも検出可能人工衛星による物体検出に応用が利く可能性を示唆．

コメント・リンク集

- 論文

[#171]

Kouyou OTSU

Illuminant Spectra-based Source Separation Using Flash Photography

Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, and Aswin C. Sankaranarayanan CVPR2018 752

概要

フラッシュを当てた状態の写真とそうでない写真の2種類を利用して，画像を光源の違いに基づく構成画像へと自動的に分離するアルゴリズムの提案．2つの写真の色情報の違いに基づき，光源に対応するスペクトルや陰影との関係を見出す．従来手法と比較して，光の色合いや陰影を忠実に反映した低ノイズでの分離が可能であることを示した(従来手法(Hsu et.al.)でのSNR:10.13dB 提案手法でのSNR 20.43dB)．また，提案手法が画像のライティングの編集，カラー測光ステレオに有用であることを示した．

(a) Input images | (b) Matrix factorization SNR 16.96 dB | (c) Hsu et al. [27] SNR 10.13 dB | (d) Our results SNR 20.43 dB | (e) Ground truth

新規性・結果・なぜ通ったか？

- 光源分離にカメラのフラッシュを利用（手軽）
- 従来手法を上回る性能．

リンク集

- 論文
- 動画

[#172]

Shusuke Shigenaka

Multi-Label Zero-Shot Learning with Structured Knowledge Graphs

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, Yu-Chiang Frank Wang CVPR 2018

概要

この論文は,各々の入力インスタンスに対して,複数の見えないクラスラベルを予測できるmulti-label learning及びmulti-label zero-shot learning(ML-ZSL)の新しい深層学習の提案した研究． 提案手法は複数のラベル間で人間が関心を持つsemantic knowledgeをグラフの中に組み込むことにより, 情報伝播メカニズムを学習し見えているクラスと見えないクラスの間の相互依存関係をモデル化することに適用できる． 本手法はstate-of-the-artと比較して,同等または改善されたパフォーマンスとして達成をすることができる.

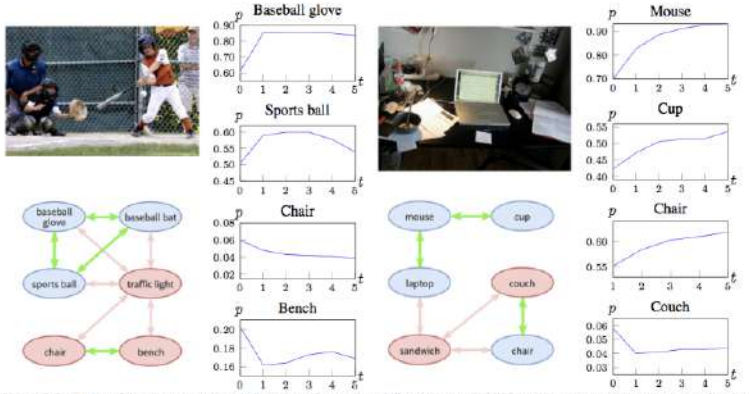


Figure 5. Examples of the constructed knowledge subgraphs and the predicted label probabilities using our proposed method, showing that information propagates across different labels as time step t increases. Note that the blue and red nodes in each subgraph indicate ground truth positive and negative labels, respectively. And, arrows in green or red reflects the corresponding positive or negative relationship.

新規性・結果・なぜ通ったか？

・見た目だけでなく,経験を通して学んだ知識を使って物体を認識・WordNetから観察された知識グラフをend-to-endの学習フレームワークに組み込み,意味空間に電番されるラベル表現と情報を学習 ・NUS-81およびMS-COCOの結果をWSABIE,WARP,Fast0Tag,Logisticsと比べたところ精度について一番高い結果を残した． ・ML-ZSLについてもFast0Tagと比べて高い精度を残している.

リンク集

- 論文

[#173]

Kazuki Inoue

Nonlinear 3D Face Morphable Model

Luan Tran, Xiaoming Liu CVPR 2018 Poster

概要

generatorとdiscriminatorを一つのモデルで表現するIntrospective Neural Network(INN)に対してwasserstein distanceを導入することで、INNと同等の生成能力・識別能力を保ちつつclassifierにおけるCNNの数を20分の1にしたWasserstein INN(WINN)を提案。 生成された画像の比較はDCGAN、INN for generative(INNg)、INNgのclassifierにおけるCNNを一つにしたINNg-singleと行った。 またadversarial exampleに対して頑健な識別精度を達成した。

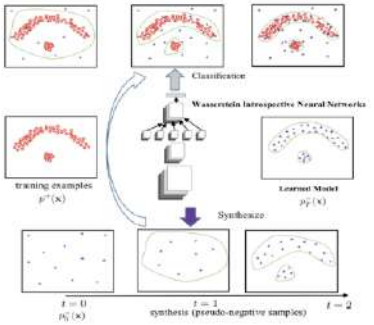


Figure 1: Schematic illustration of Wasserstein introspective neural networks for unsupervised learning. The left figure shows the input examples; the bottom figures show the pseudo-negatives (purple crosses) being progressively synthesized; the top figures show the classification between the given examples (positives) and synthesized pseudo-negatives (negatives). The right figure shows the model learned to approach the target distribution based on the given data.

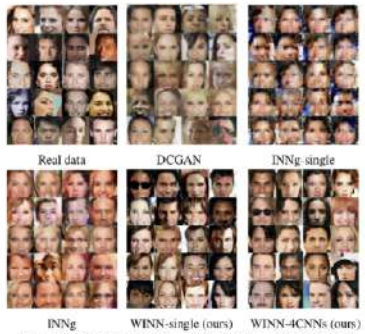


Figure 5: Images generated by various models trained on CelebA.

Method	Adversarial error of **Method** ↓	Correction rate by **Method** ↑	Correction rate by **Baseline** ↓
Baseline vanilla CNN	32.41%	-	-
ICN [21]	19.02%	52.58%	58.68%
WINN-single vanilla (ours)	7.99%	**90.00%**	**46.95%**
Baseline ResNet-32	11.28%	-	-
WINN-single ResNet-32 (ours)	**2.05%**	89.68%	43.69%

新規性・結果・なぜ通ったか？

- INNにwasserstein distanceを導入することで、生成・識別においてINNと同等以上の性能を持ちながら識別器におけるCNNの数が20分の1であるIWNNを提案。
- テクスチャの生成やCelebA・SVHNを学習することで生成された画像はDCGANと比べてはっきりとしており質が高い。
- CIFAR-10の学習によって生成された画像におけるInception scoreはDCGANの方が良い結果となった。
- CNN、ReosNet、ICNと比較して、adversarial exampleに対する誤識別率が低く、 adversarial examples に惑わされずに識別を行うことが可能。

コメント・リンク集

- 論文
- GitHub
- Introspective Neural Networks for Generative Modeling

[#174]

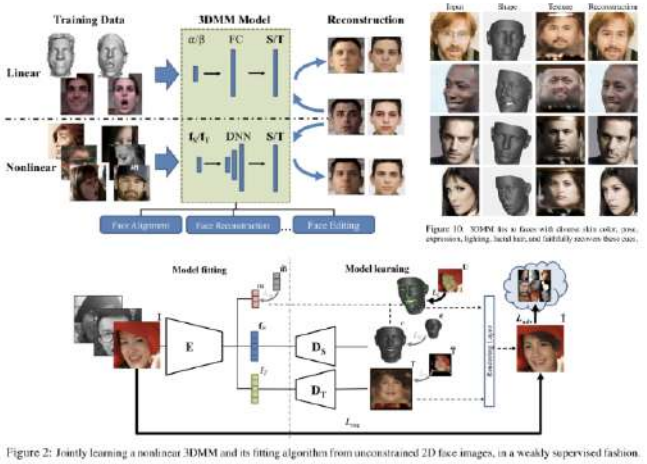
Kazuki Inoue

Nonlinear 3D Face Morphable Model

Luan Tran, Xiaoming Liu CVPR 2018 Poster

概要

3Dスキャンデータを使用せずにin-the-wildな顔画像のみを用いてencoder-decoderによって3D Morphable Model(3DMM)を生成する手法を提案。生成された3DMMを nolinear 3DMMと呼んでいる。 従来のlinear 3DMMは学習のために3Dスキャンデータが必要であり、かつPCAによって次元削減を行うため表現力に乏しいという問題点があった。 提案手法ではencoderによってプロジェクション、シェイプ、テクスチャのパラメタを取得し、decoderによってシェイプ、テクスチャを推定する。 また初期の学習では既存手法によって得られる3DMMのプロジェクションパラメタ、シェイプパラメタとUV空間から得られるテクスチャを擬似的なGTとすることで弱教師学習を行う。



新規性・結果・なぜ通ったか？

- 3Dスキャンデータを使用せずに、in-the-wildな顔画像のみを学習させることで、入力画像から3D Morphalbe Modelを生成する。
- linear 3DMMと比較して、3次元形状、テクスチャの精度が高い。また見た目もGTにより近い。
- 顔のアラインメントにおいてstate-of-the-artよりも高い精度を達成。
- 3次元形状における精度はstate-of-the-artと同等であった。

コメント・リンク集

- 弱教師学習がどれほど影響を持つかが気になった。
- 論文
- Project page

[#175]

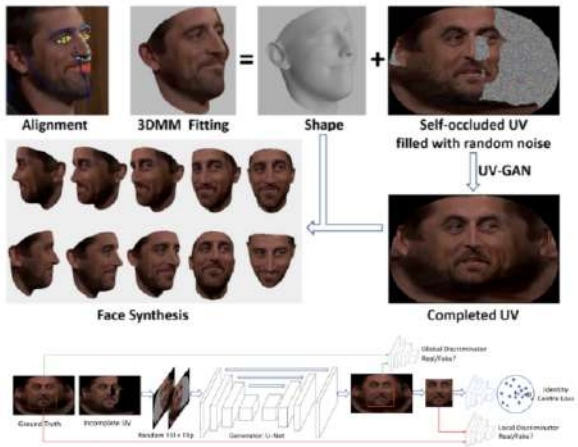
Kazuki Inoue

UV-GAN: Adversarial Facial UV Map Completion for Pose-invariant Face Recognition

Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, Stefanos Zafeiriou CVPR 2018 Poster

概要

in-the-wildな入力顔画像から得られるUVマップの補完をU-Netで行う手法を提案。入力画像に対して3D Morphalbe Modelを適用し不完全なUVマップを取得し、U-Netで補完を行うように学習を行う。discriminatorにはUVマップ全体と顔領域の判定をさせる。 またUVマップの個人性が失われないように、アイデンティティーに関するロスを取る。 1892人のUVマップをもつWildUVデータセットの構築も行った。



新規性・結果・なぜ通ったか？

- in-the-wildな顔画像に対してもリアルかつ精度の高いUVマップの補完を達成。入力されるUVマップが50%欠けていても補完可能。
- 入力画像からUVマップと3D shapeを取得するため、入力画像を任意の顔向きに編集可能。
- 横向き顔画像から生成されたUVマップはPSNR, SSIMにおいて既存手法を上回る精度を達成。
- frontal-profile face verificationにおいてstate-of-the-artを上回る94.05%を達成。
- 1892のアイデンティティーのUVマップをもつ大規模UVマップデータセットであるWildUVデータセットを公開（予定）。

コメント・リンク集

- 論文

[#176]

Ryota Suzuki

LIME: Live Intrinsic Material Estimation

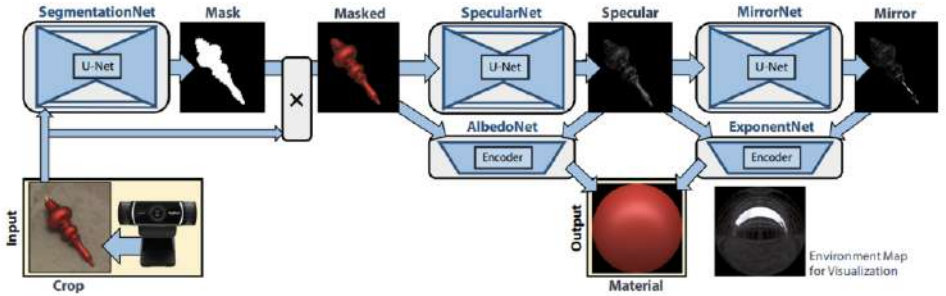
A. Meka, M. Maximov, M. Zollhöfer, A. Chatterjee, H.P. Seidel, C. Richardt and Ch. Theobalt CVPR2018

概要

単RGB画像で，リアルタイムに材質反射特性を推定する手法を提案し，デモシステムを作った．

構造は，主に複数のU-Netからなり，それぞれ前景セグメンテーション，スペキュラー推定，鏡面反射推定を行う．ロス関数も定義．

さらに，形状情報も使えるのなら，低・高周波光源情報の推定も可能．連続撮影時の光源情報の連続性を考慮した時系列統合の枠組みも提案．



新規性・結果・なぜ通ったか？

- 実用的なシチュエーション（リアルタイム，複雑な光源下，連続撮影）で利用可能であることを示している．
- 定性，定量評価を行い，性能の良さを示している．

コメント・リンク集

デモビデオを作り慣れているように見えるあたり，CG勢と思われる．デモも結構評価されているだろうか． アプリケーション枠で評価されるように書いているかもしれない．

- arXiv
- Youtube
- プロジェクトページ

[#177]

Ryota Suzuki

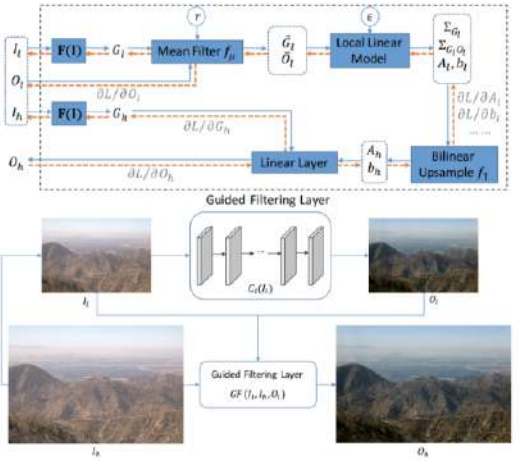
Fast End-to-End Trainable Guided Filter

H. Wu, S. Zheng, J. Zhang, K. Huang CVPR2018

概要

低解像度＋高解像ガイダンスマップを与えると，高解像度画像を効率的（省計算時間，省メモリ）に出力できるGuided Filtering Layerなるものを提案．

GuidedFilterは， 空間的に変化する線形変換行列のグループとして表現でき， CNNに統合可能．つまり，end-to-endで最適化可能な深層ガイデッドフィルタネットワークを構成できる．



新規性・結果・なぜ通ったか？

- Context Aggregation NetworkにGuided Filtering Layerを載せたものを、5つの先進的な画像処理タスクで試したところ，**10〜100倍高速**であり，**SoTA性能**も出た．

コメント・リンク集

かなり省コストになっている．DNN導入可能にするように（エレガントに）定式化し，コストダウンしつつ深層学習できるようにする手法がいくつか見られている．

- arXiv
- GitHub

[#178]

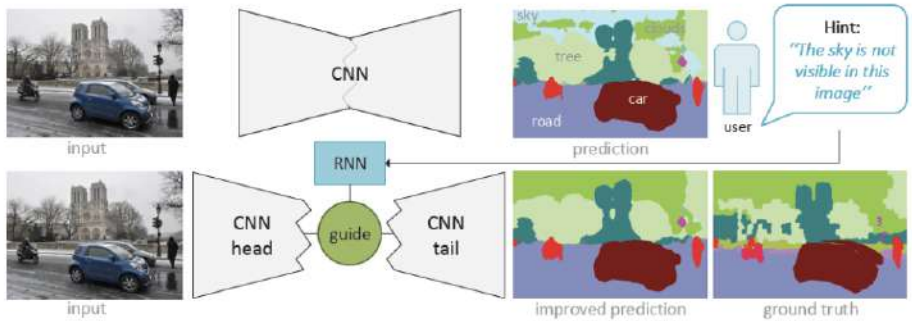
Shintaro Yamamoto

Guide Me: Interacting with Deep Networks

Christian Rupprecht, Iro Laina, Nassir Navab, Gregory D. Hager and Federico Tombari CVPR 2018

概要

CNNにより学習したタスクの出力結果に対して、人間がヒント(例：画像中に空は見えない)を与えていくことで精度向上を図る研究。CNNモデルをheadとtailの2つのパートに分割し、headから得られた特徴マップをヒントによって修正していくことで精度の向上を実現する。 その際、ネットワークの重みを更新するのではなく修正に用いるパラメータを言語情報から推測することで行う。 ネットワークの予測結果とground truthの差分を取り、正しく予測できていない物体の種類や位置を推定することで学習に用いる文章は自動で生成する。



新規性・結果・なぜ通ったか？

セマンティックセグメンテーションにより実験を実施したところ、クラス間違い、物体の一部が欠けている、物体の一部のみが見えるといったケースにおいて精度が向上することを確認した。ヒントを繰り返し与えていくことはノイズとなってしまうためあまり精度が向上しなかった。 従来のディープラーニングは一度学習をしてしまうと得られる出力が固定されてしまうのに対して、人間が介入することで結果を変えるという新しい応用方法を提案している。

コメント・リンク集

- 論文

[#179]

Kazuki Inoue

Face Detector Adaptation without Negative Transfer or Catastrophic Forgetting

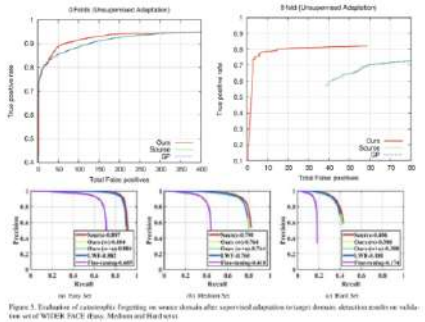
Muhammad Abdullah Jamal, Haoxiang Li, Boqing Gong CVPR 2018 Poster

概要

顔検出におけるターゲットドメインからソースドメインへのadaptationを、negative transferとcatastrophic forgettingの両方を引き起こさずに行う手法を提案。 negative transferとはadaptation後のソースドメインにおける検出精度がadaptation前のソースドメインにおける検出精度に劣ることを指し、 catastorophic forgettingとはadaption後におけるソースドメインの検出精度が著しく下がることを指す。 提案手法では、ソースドメインとターゲットドメインの違いを、ロス関数とDNNの重みの差分で表現し、 この差分がなくなるように学習を行う手法を提案。 またターゲットドメインにface or notのラベルがないという状況も考えて教師あり学習だけでなく教師なし学習、 半教師あり学習の結果についても議論を行った。

$$\mathrm{RES}_t(\widetilde{\mathbf{w}},\widetilde{\theta}) := C\big(y_t, \sigma(\widetilde{\mathbf{w}}^T F(\mathbf{x}_t;\widetilde{\theta}))\big) - C\big(y_t, \sigma(\mathbf{w}^T F(\mathbf{x}_t;\theta))\big), \quad (1)$$

$$\min_{\mathbf{u},\,\widetilde{\theta}} \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \mathbb{E}_t \max_{y_t\in\{0,1\}} \mathrm{RES}_t(\mathbf{w}+\mathbf{u},\widetilde{\theta}), \quad (2)$$



新規性・結果・なぜ通ったか？

- ソースドメインとターゲットドメインの違いを、DNNのロス関数・重みの差分で表現することでadaptationを行った。
- 実験は、CascadeCNN+AFLW(25000 faces), Faster-R CNN+WIDER FACE dataset(393,703 faces, highly labeled)の2つのモデルでソースドメインの学習を行い、ターゲットドメインははFDDB(5171 labeled faces)、COFWで行った。
- 検出結果はターゲットドメインのみを学習した検出器、ソースドメインからターゲットドメインへfine tuningされた検出器、domain adaptaionを行うstate-of-the-artと比較を行った。提案手法はターゲットドメインにおける検出においてもっとも高い精度を達成。 またソースドメインにおける検出においてもターゲットドメインのみを学習した識別器と同等の精度を達成。

コメント・リンク集

- adaptationというより、もはやトレーニングデータセットの事後拡張となっており、後でトレーニングデータを追加したくなった時に有用なのではないだろうか。
- 論文
- Supplementary

[#180]

Kazuki Inoue

Extreme 3D Face Reconstruction: Looking Past Occlusions

Anh Tuấn Trần, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, Gérard Medioni CVPR 2018 Poster

概要S

入力顔画像からバンプマップや視点を推定することで、入力画像からは見えていない側面や、強いオクルージョンがある顔画像からも精度の高い三次元形状を取得する手法を提案。 入力画像から帯域的な情報として三次元の大まかな形と、 局所的な情報としてしわなどのディティールを表現するバンプマップを別々のDNNモデルを使って取得する。 続いてオクルージョンがある場合には、バンプマップが不自然な起伏を持つため深層学習による修正を行う。 最後に顔の対称性を利用して、入力画像からは見えていない側面などをルールベースで復元する。

Figure 1: Results of our method. Detailed, complete 3D reconstructions shown next to their partially occluded input faces.

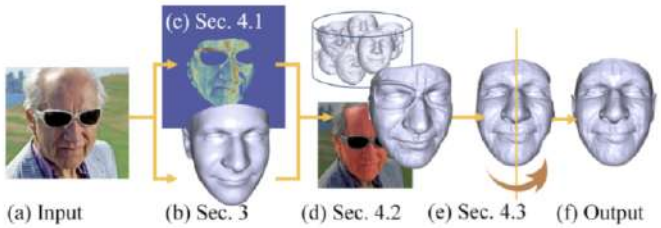


Figure 2: *Method overview.* See related sections for details.

新規性・結果・なぜ通ったか？

- 入力画像から3Dモデル全体を一気に復元するのではなく、帯域的な特徴と局所的な情報を分けて取り扱うことで精度の高い三次元復元を可能にした。
- 結果の評価は復元された三次元形状による個人認証の精度で行っている。画像にオクルージョンがない場合にはstate-of-the-artよりも高い精度を達成。オクルージョンがある場合でも、オクルージョンがない場合よりと比べて2%ほどしか劣らなかった。(state-of-the-artはそもそもオクルージョンを考慮できない。)
- 復元された三次元形状は、既存手法がオクルージョンを考慮することができなかったりシワなどの復元ができていないのに対して、提案手法ではオクルージョンがある場合でもシワなどの詳細な情報を復元できている。

コメント・リンク集

- 帯域的な顔形状の復元やバンプマップの修正などを既存手法に頼っているものの、復元された三次元形状は既存手法に比べて圧倒的なクオリティを持つ。 しかし形状自体のGTとの比較がなかったのが残念。
- 論文
- GitHub

[#181]

Kazuki Inoue

InverseFaceNet: Deep Monocular Inverse Face Rendering

Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, Christian Theobalt CVPR 2018 Poster

概要

実世界の3D顔モデルを使用せず合成された3Dモデルのみで CNNをトレーニングすることで、実世界の顔画像から顔向き、形、表情、リフレクタンス、イルミネーションの3D復元を行う手法を提案。 CNNをトレーニング際の問題点として、実世界の3D顔モデルに対するアノテーションが足りないという問題があった。 これに対して、実世界の顔画像から推定されるパラメタと合成顔から推定されるパラメタに対してself-supervised bootstrappingを行うことで、 トレーニングに使用する合成顔3Dモデルのパラメタの分布を実世界のパラメタの分布に近づくようにトレーニングデータを逐次的に更新を行うことで、 CNNの学習を行った。

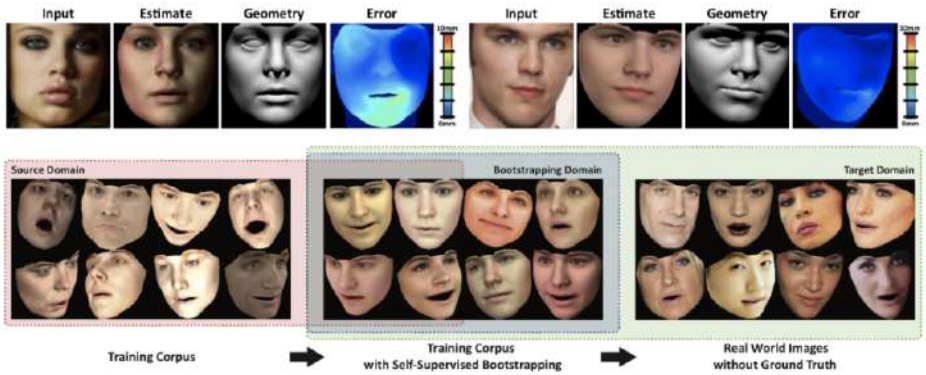


Figure 3. Our approach updates the initial training corpus (left) based on real-world images without available ground truth (right) using a self-supervised bootstrapping approach. The generated new training corpus (middle) better matches the real-world face distribution.

新規性・結果・なぜ通ったか？

- self-supervised bootstrappingを使用することで、実世界のパラメータを再現するように合成顔のデータセットを再構築することで、データセットがないという問題に取り組んだ。
- 既存の学習ベースの手法に比べて、ジオメトリーにおいて最も高い精度を達成。
- 最適化ベースの手法に比べると、パーツのディティールやシワの再現の精度が悪い。
- リミテーションとして、データセットにない顔向きや髪によるオクルージョンを考量することができない。

コメント・リンク集

- 異なるドメインを使ったトレーニングの方法として、GANを使ってcross domainの分布を近づける方法が提案されているなど、トレーニングデータ不足を解決する方法が提案されてきている。
- 論文
- Supplementary

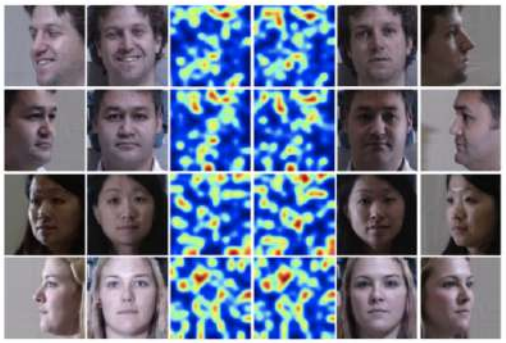
[#182]

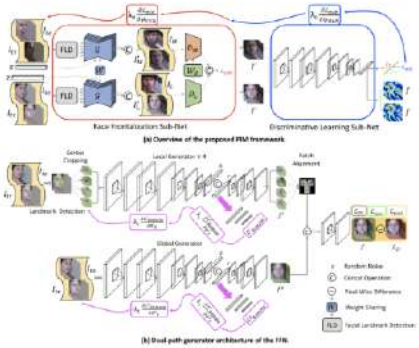
Kazuki Inoue

Towards Pose Invariant Face Recognition in the Wild

Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao CVPR 2018 Poster

概要





様々な照明環境、表情をした横向き顔画像を入力として、正面顔画像を生成することで高い個人認証率を達成するGANベースのPose Invariant Model(PIM)というネットワークを提案。 学習で使用できるトレーニングデータが少ないため、効率的かつ過学習を防ぐために以下のようにPIMを構築。

- 顔全体を生成するgeneratorと両目・鼻・口の4つのパーツを生成するgeneratorを用意。
- 4つのパーツが検出された画像と取得できない画像(横顔画像など)を異なるドメインの画像とみなして、cross-domain adversarial trainingを行うことで、両目・鼻・口を復元。
- 上記のGANを２セット用意し、discriminator同士でlearning to learnを行うことで効率的な学習を行った。

新規性・結果・なぜ通ったか？

- 2つのGANをもつTP-GANやDR-GANは最適化が困難で合ったが、これに対してlearning-to-learnを導入することでこの問題を解決。
- MultiPIE、CFPデータセットにおいて様々な角度の顔画像に対する個人識別においてほぼ全てのケースにおいてstate-of-the-artよりも優れた精度を達成。(唯一Multi-PIEで顔向きが±30°の場合にTP-GANに劣った。)
- 横向き顔画像から生成される正面顔画像において、既存手法ではテクスチャが崩れていたり完全に正面を向いていない場合があっ

コメント・リンク集

- データセットが少ないという根本的な問題に対して、cross-domain adversarial training、learing to learnを行うことで解決しているが、これがデータベースが欠乏している他の問題設定でも解決できるのかを試してみたい。
- 論文

[#183]

Kazuki Inoue

Ring loss: Convex Feature Normalization for Face Recognition

Yutong Zheng, Dipan K. Pal and Marios Savvides CVPR 2018 Poster

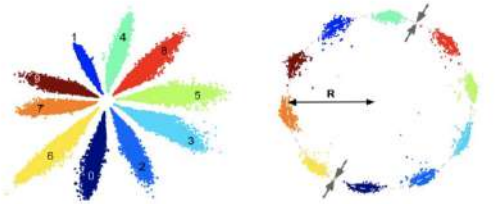
概要

DNNによって得られた特徴量を超球面上に配置するように正規化を行うロス関数であるRing lossを提案。特に教師あり識別問題においてはDNNによる特徴量を正規化することでより精度の高いモデルを構築することができる、 というアイディアもとにRing lossを提案。SoftMaxといった基本的なロス関数と組み合わせることでより高い精度を達成。 実験には様々な識別タスクを行うことができる顔データセットを用いることで、精度の向上を確認した。

Ring loss Definition. Ring loss L_R is defined as

$$L_R = \frac{\lambda}{2m}\sum_{i=1}^{m}(\|\mathcal{F}(\mathbf{x}_i)\|_2 - R)^2 \qquad (4)$$

where $\mathcal{F}(\mathbf{x}_i)$ is the deep network feature for the sample $\mathbf{x}_i$. Here, R is the target norm value which is also learned and λ is the loss weight enforcing a trade-off between the primary loss function. m is the batch-size. The square on the norm



(a) Features trained using Softmax (b) Features trained using Ring loss

新規性・結果・なぜ通ったか？

- SoftMaxとSphereFaceにRing lossを組み合わせることでLFW, IJB-A Janus, Janus CS3, CFP, MegaFaceデータセットにおけるFace verification, identificationにおいて他のロス関数と同等あるいはそれ以上の精度を達成。
- 極端に低解像度の画像におけるface matchingにおいてベースラインの手法を凌駕した。
- 実験ではResNet64を使用。

コメント・リンク集

- 論文

[#184]

Kazuki Inoue

Label Denoising Adversarial Network (LDAN) for Inverse Lighting of Face Images

Hao Zhou, Jin Sun, Yaser Yacoob, David W. Jacobs CVPR 2018 Poster

概要

3Dモデルから実画像へのドメイン変換をGANによって行うことで、単一顔画像から照明パラメタを推定するLabel Denoising Adversarial Network(LDAN)を提案。 人の顔画像に対して照明パラメタ(論文で使用されているのは37次元の球面調和関数)がアノテーションされたデータセットがないため、 3Dモデルを使用してFeature Netと呼ばれるネットワークで中間特徴量を取得し、 中間特徴量からLightning Netを用いて照明パラメタの推定を学習。 続いて人の顔画像に対して、既存手法を用いてノイズが乗った照明パラメタを取得し、 人の顔画像に対してもFeature Netを新しく学習し、 3D モデルから得られた中間特徴量と共にGANに入力することでドメインの変換を行うことでノイズが除去された照明パラメタを取得。

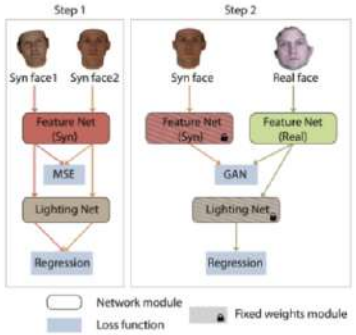
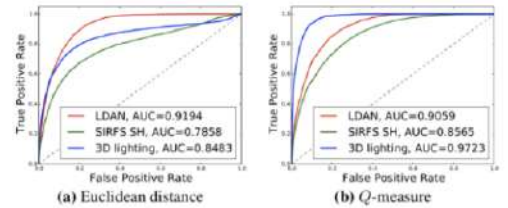


Table 2: Results of ablation study. Standard derivation is shown in the bracket.

	LDAN	LDAN w/o GAN	LDAN w/o Regression
top-1 (%)	**65.87** (±1.2)	60.58 (±0.5)	30.71 (±4.1)
top-2 (%)	**85.17** (±1.1)	81.88 (±0.5)	48.47 (±5.7)
top-3 (%)	**92.46** (±0.5)	90.75 (±0.6)	60.96 (±6.4)



(a) Euclidean distance (b) Q-measure

新規性・結果・なぜ通ったか？

- 単一画像からの照明パラメタの推定という問題に対して、初めて学習ベースの手法を提案。
- 結果の比較は19の照明環境が用意されているMultiPieデータセットで行い、推定されたパラメータに対する識別を行うことで精度を評価。state-of-the-artに比べて識別精度およびユークリッド距離・Q値におけるAUCで最も高い精度を達成。
- 同問題を扱う既存手法が最適化ベースということもあり、既存手法と比べて10万倍のスピードで実行可能。

コメント・リンク集

- GANを使って異なるドメインの特徴量を同じ空間にマップする考え方は既にAdversarial Discriminative Domain Adaptationによって提案されているが、異なる点としては[Eric et al.]はGANのロスしか使っていないが、この方法では写像がうまく行かず、 A→A', B→Bと学習して欲しいところをやA→B', B→A'といった写像を学習してしまう。 これを解消するために、lightning netで得られたパラメータに対するL2ロスを取ることでこれを解消。
- 論文

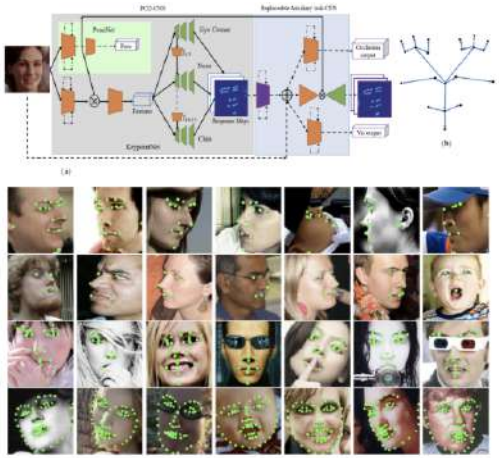
[#185]

Kazuki Inoue

Disentangling 3D Pose in A Dendritic CNN for Unconstrained 2D Face Alignment

Amit Kumar, Rama Chellappa CVPR 2018 Poster

概要



顔向きをコンディションとして与え木構造で表された顔のランドマークを学習させることで、顔のランドマーク推定を行うPose Conditioned Dendritic CNN(PCD-CNN)を提案。 顔のコンディションはPoseNetにより出力された値を使用する。 顔のランドマークを木構造として与えることで、ランドマークの位置関係を利用してCNNを学習させた。 また提案ネットワークはPCD-CNNと通常のCNNの二段階になっており、 後段のCNNをファインチューニングすることでランドマークのポイント数が違うデータセットや顔向き推定などの他のタスクにも適用可能。

新規性・結果・なぜ通ったか？

- ネットワークをPCD-CNNとCNNの二段階で構成することで、異なるランドマークのポイント数や顔向き推定といった他のタスクにも適用可能。
- 顔向きをコンディションとして与えることで推定精度が向上。また、20FPSで実行が可能。
- AFLW, AFWデータセットにおいてランドマークの推定精度がstate-of-the-artよりも高い推定精度を達成。

コメント・リンク集

- 論文

Kazuho Kito

Multi-Image Semantic Matching by Mining Consistent Features

Qianqian Wang, Xiaowei Zhou and Kostas Daniilidis CVPR2018

概要

ノイズを考慮しつつ、数千もの画像セット全てにおいて一致する(信頼できる)特徴を見出すことで、画像間の対応を図るマッチング手法。マッチングはセマンティック性を考慮することができる（目と目、耳先と耳先など）これにより、一貫性がある画像セット内で信頼できる特徴の関係を確立。何千もの画像を処理する場合にスケーラブルな手法。つまりは数に頑健。

新規性・結果・なぜ通ったか？

従来手法では、全てのペアで対応する関係を最適化していたが、本手法では、特徴の選択とラベリングに着目し、信頼度の高い特徴のみを用いた疎なセットのみで識別、マッチングする。

コメント・リンク集

図は中の左が出力結果であり、目は青、耳は黄色、鼻は赤など各特長の意味を理解し、マッチングを成功させている。

- 論文

[#187]

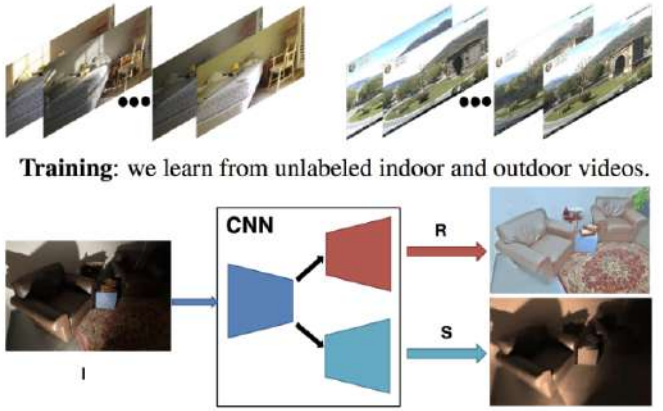
Naofumi Akimoto

Learning Intrinsic Image Decomposition from Watching the World

A. Uthors, B. Uthors and C. Uthors CVPR2018

概要

Intrinsic Image Decompositionのために，時間経過とともに照明が変化するビデオを使ったCNNの学習方法を提案．正解の Intrinsic Imageが不要な点が強みである．学習が完了したモデルは単一画像に対して適用できるよう汎化しており，いくつかのベンチマークに対して良い結果となった．
Contribution：
・データセット（BigTime）の公開．室内，室外両方での照明変化のあるビデオと画像シーケンスのデータセット．
・このGround Truthを含まないデータを使った手法の提案．



Training: we learn from unlabeled indoor and outdoor videos.

Testing: our CNN produces intrinsic images from a single photo.

学習時：ラベル無しで，視点が固定され照明が変化するビデオを学習に利用する．
テスト時：単一画像からintrinsic image decompositionを行う．

手法

最適化ベースのIntrinsic Decomposition手法と，機械学習手法の間に位置する手法と言える．
・U-netに似た構造のCNN．
・Lossの工夫：画像ペア全てを考慮するall-pairs weighted least squares lossとシーケンス全体のピクセル全てを考慮するdense, spatio-temporal smoothness loss．最適化ベースのlossをフィードフォワードネットワークのlossとして利用する．

コメント・リンク集

Intrinsic image decompositionとは，入力された1枚の画像をreflectance画像とshading画像の積に分解する問題のこと．
intrinsic imagesのGround Truthを大規模に揃えることは困難．

- arXiv

[#188]

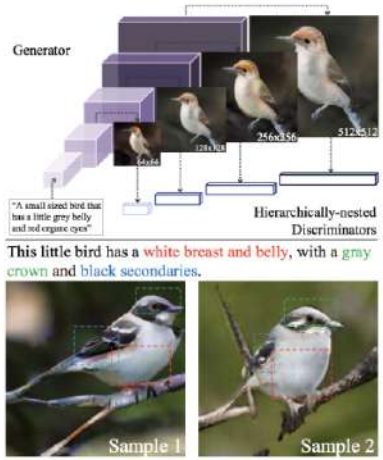
Naofumi Akimoto

Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network

Zizhao Zhang, Yuanpu Xie, Lin Yang CVPR2018

概要

階層的入れ子構造の識別器を使用し，テキストから高解像画像を生成するGANを提案．end-to-endの学習で高解像画像の統計量を直接モデルリングすることが可能な手法．これは，step-by-stepで高解像画像を生成するStackGANとは異なる点である．複数のスケールの中間層に対して階層的入れ子構造の識別器を使用することで中間サイズレベルでの表現に制約を加え，生成器が真の学習データの分布を獲得しやすくする．



手法

新しい構造と，lossの工夫でtext-to-imageのタスクで高解像画像の生成を可能とした．
・hierarchical-nested Discriminatorを使用．
・lossには，pair lossとlocal adversarial lossを使用する．pair lossでは入力テキストと生成画像が一致しているかを評価．local adversarial lossでは生成画像の細部の質を評価する．

コメント・リンク集

- arXiv

[#189]

Naofumi Akimoto

Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images

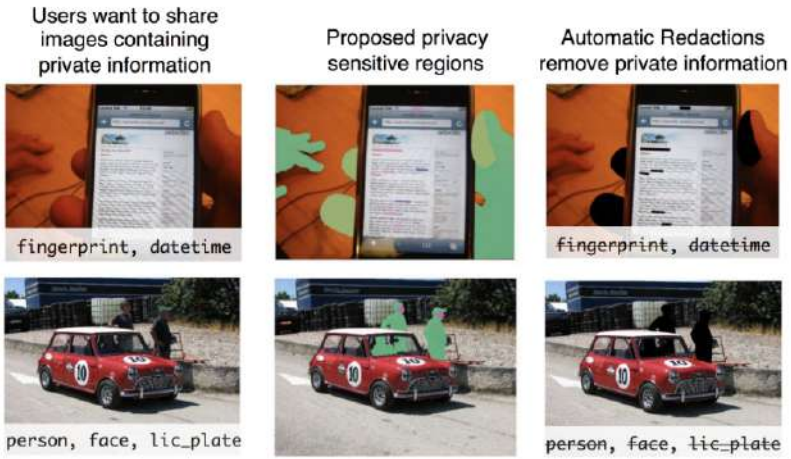
Tribhuvanesh Orekondy, Mario Fritz, Bernt Schiele CVPR2018

概要

プライバシー保護のために画像に含まれる個人的な情報を自動的に改変する手法の提案．プライバシーを守りつつ画像の有用性を保つためのトレードオフが問題となる．有用性を保つためには改変する領域サイズが最小限である必要があり，これをセグメンテーションの問題として取り組む．

Contribution:

- データセットの公開．様々な種類のプライバシーのラベルが，ピクセルレベルとインスタンスレベルで与えられている自然画像の初のデータセット．
- モデルの提案．多様な個人情報を自動的に改変するモデルを提案する．正解のアノテーションに対して83%の正解率を達成した．



指紋，日時，人，顔，ナンバープレートを黒く塗りつぶせている．他にも，住所やメールアドレスのようなテキスト情報や顔や車椅子などの視覚情報，あるいはテキストと視覚情報を合わせたものなど，多様な個人情報に対応するデータセットとモデルを提案．

手法

どのような対象(Textual, Visual, Multimodal)を扱うかで使用するモデルは異なる．
Textualな対象では，Sequence Labelingを使用する．
VisualとMultimodalな対象では，Fully convolutional instance-aware semantic segmentationを使用する．
Nearest Neighborなどのベースライン手法と比較を行なっている．

コメント・リンク集

画像全体を黒く塗ればプライバシーは保護されるが，画像の価値がなくなるので，トレードオフが存在する．
データセットを作った貢献がメイン．プライバシー保護のためのアノテーションを行ったことで，それなりの正解率で個人情報の改変を行えるようになった．

- arXiv

[#190]

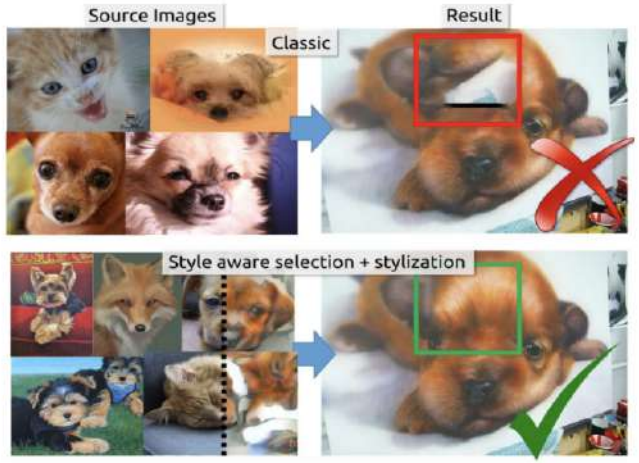
Naofumi Akimoto

Disentangling Structure and Aesthetics for Style-aware Image Completion

Andrew Gilbert, John Collomosse, Hailin Jin, and Brian Price CVPR2018

概要

ノンパラメトリックのInapinting手法を提案．
視覚的な構造とスタイルをdeep embeddingすることで，パッチの検索と選択の際に視覚的なスタイルを考慮することが可能で，さらに，パッチのコンテンツを補完画像のスタイルに合わせるためのneural stylizationが可能となる．この手法は，patch-basedの手法とgenerativeベースの手法の架け橋的な補完手法である．
技術的貢献：
・style-aware optimization
・adaptive stylization



手法

以下の手順で画像補完を行う．
１．スタイルを考慮して穴に埋める候補を検索する
２．補完画像と構造とスタイルが合うパッチをMRFで複数集め，選択する
３．選択されたパッチを補完画像のスタイルに変換する

コメント・リンク集

- 論文pdf

[#191]

Naofumi Akimoto

DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks

Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, Ji˘ri Matas CVPR2018

概要

motion deblurringのためのGAN(DeblurGAN)を提案．structural similarity measureとアピアランスでSoTA．ブラーを除去した画像で物体検出の精度を出すことで，ブラー除去モデルの質を評価するという方法を提案．提案手法は，質だけでなく実行速度も優れており，従来手法の５倍の速さがある．モーションブラーのかかった画像を合成するための方法を紹介し，そのデータセットもコード，モデルとともに公開．

ブレを除去してからYOLOで検出すると精度が良くなることを示している．これをDeblurモデルの指標にすることができると主張．

手法

- loss：WGANによるAdversarial lossとPerceptual loss
- 構造：畳み込み，instance normalization層，ReLU関数から成るResBlockの繰り返しがメインで，出力するときに入力画像を加算するglobal skip connectionを持つ．

コメント・リンク集

最近のGAN手法やテクニックを詰め込んで，新しく作ったデータセットを利用したらSoTAがでたという感じ．テクニカルな貢献はあまりなさそう．

- GitHub
- arXiv

[#192]

Naofumi Akimoto

Learning to Understand Image Blur

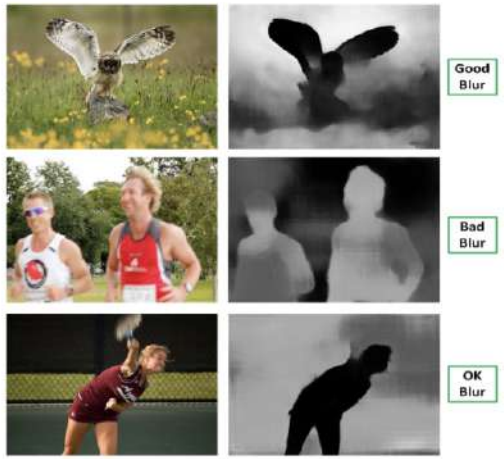
Shanghang Zhang, Xiaohui Shen, Zhe Lin, Radom ír Me˘ch, Joa õ P. Costeira, Jose ´M. F. Moura CVPR2018

概要

ボケ(blur)が望ましいのか否かと，そのボケが写真のクオリティーにどのような影響を与えているのかを，自動的に理解するアルゴリズムは少ない．この論文では，blur mapの推定とこのボケの望ましさの分類を同時に行うフレームワークを提案する．

貢献：

- ボケを検出することと，画像の質という点でボケを理解することを同時に行うのは，おそらく初めての研究．ABC-FuseNetというネットワークを提案．
- １万枚のデータセット（SmartBlur）の公開．ピクセルごとにボケがかかっているか３段階でラベルづけ．さらに，画像ごとにボケの望ましさ(desirability)をラベルづけ．
- SmartBlurと他の公開データセットで実験を行い．blur mapの推定とボケの望ましさの分類がSoTAを超えた．



ボケ具合をピクセルごとに３段階で示し，ボケの望ましさも出力する．

手法

ABC-FuseNetでは，低レベルのボケの推定と高レベルの画像内で重要コンテンツの理解の二つを行う．
A: attention map，FCNである．
B: blur map，Dilated Convolutionとpyramid pooling, Boundary Refinement用の層を使ってblurの推定を行う．
C: content feature map，ResNet-50を使ってコンテンツの特徴を抽出．
ボケの推定はBによって行い，ボケの望ましさの分類はA, B, Cから得られた特徴を用いて行う．ネットワーク全体をEnd-to-endで学習す

コメント・リンク集

ボケを軽減するための研究は多いが，ボケが全て邪魔とは言えない．ボケを効果的に利用することで，写真の印象が良くなることもある．いいボケなのか，悪いボケなのかの判断も必要だというモチベーションがある．

コード，データセットは以下に公開予定

- GitHub

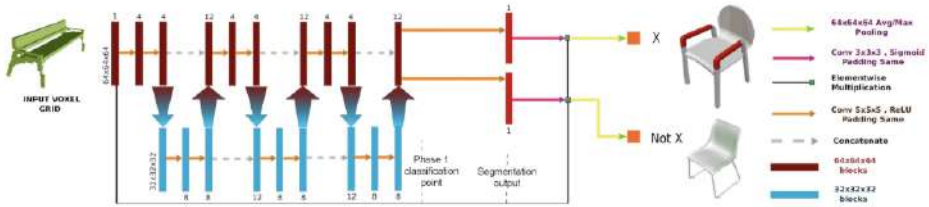
[#193]

Naofumi Akimoto

Tags2Parts: Discovering Semantic Regions from Shape Tags

Sanjeev Muralikrishnan, Vladimir G. Kim, Siddhartha Chaudhuri CVPR2018

概要



指定された形状のタグに強く関係する領域を検出する手法の提案．明示的に領域ごとのラベリングはなく，さらにあらかじめセグメンテーションされていない状況で，形状のタグを与えた時に領域を発見するという問題設定．難しい点は，オブジェクトのタグという弱い教師情報からポイントごとのラベルを細かく出力する必要があること．このために分類とセグメンテーションを同時に行うネットワークを使う．形状ごとのタグからポイントごとの予測を得るためのネットワーク構造（WU-net）を提案したことがメインの貢献．

学習が完了すれば，タグが不明な形状に対しても手法を適用することができる．また，元々Weakly-supervised用に提案しているが，strongly-supervised用としても利用できる手法となった．

手法

U-net風のWU-netを提案．U-netから修正した点は，
・浅いU型の構造を3回くりかし，skip-connectionで密に繋がっている．深いU型1回の場合との結果の違いを図示している．
・セグメンテーションの用の隠れ層にタグ分類用の層を追加．(元々のは，strongly-supervised セグメンテーション用に設計されているので．)

コメント・リンク集

３D形状としてはボクセル表現を使用．64×64×64 cubical gridを入力する．

- arXiv

[#194]

Naofumi Akimoto

Neural 3D Mesh Renderer

Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada CVPR2018

概要

ニューラルネットワークに組み込むことができる3Dメッシュのレンダラーである Neural Renderer を提案。レンダリングの『逆伝播』と呼ばれる処理をニューラルネットワークに適した形に定義し直した．そしてこのレンダラーを
・一枚の画像からの3Dメッシュの再構成（ボクセルベースの再構成との比較あり）
・画像から3Dへのスタイル転移と3D版ディープドリーム
に応用できることを示した．

2D-to-3Dスタイルトランスファーの例

方法

従来のままでレンダリングの操作が処理の途中にあると逆伝播が行えない状態であるので，レンダリングのための勾配を定義することでニューラルネットワークの中にレンダリング操作を加えても学習を行えるようにした．

コメント・リンク集

- プロジェクトサイト
- GitHub
- ３Dの形式には様々ある（ポイントクラウド，ボクセル，メッシュなど）が，3Dメッシュは効率的で表現能力が高く直感的な形式だそう．

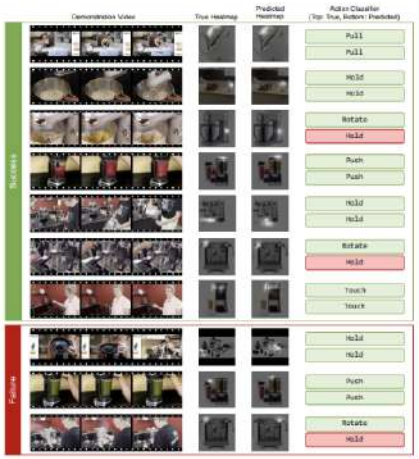
KazuhoKito

Demo2Vec: Reasoning Object Affordances from Online Videos

Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese and Joseph J. Lim CVPR2018 1387

概要

商品などのデモンストレーションの映像の特徴を通してその商品などのアフォーダンスを推論する研究．映像から埋め込みベクトルを抜き出すことで，ヒートマップと行動のラベルとして特定のもののアフォーダンスを予測するDemo2Vecモデルを提案．また，YouTubeの製品レビュー動画を集め，ラベリングすることでOnline Product Review detaset for Affordande(OPRA)を構築．



新規性・結果・なぜ通ったか？

アフォーダンスのヒートマップと行動のラベルの予測に関し，RNNの基準よりよいパフォーマンスを達成

コメント・リンク集

YouTubeで公開されている動画では，Demo2Vecを用いてある物体のデモ動画からSawyer robotのEnd Effectorを予測したヒートマップの地点に移動するように制御させている様子を見ることができる．

- 論文
- ProjectPage
- YouTube

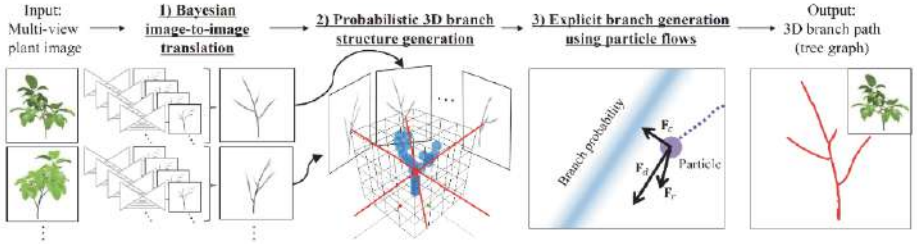
[#196]

Goshi Sasaki

Probabilistic Plant Modeling via Multi-View Image-to-Image Translation

Takahiro Isokane, Fumio Okura, Ayaka Ide, Yasuyuki Matsushita, Yasushi Yagi CVPR 2018 368

概要



葉に隠れていても３次元の枝構造を多視点画像から推測できるようにした。多視点からの植物画像を入力として枝構造の２次元確率マップをdropoutを取り入れたPix2Pixで推測して、それらから３次元の確率構造を作成した。最後にpartical flowシュミレーションによって明確な３次元の枝構造を生成した。

新規性・結果・なぜ通ったか？

葉や他の枝によって隠れてしまっていても枝構造を生成できるようにした。ベイジアンPix2Pixを利用することで植物の３次元構造をより正確に表せるようにした。

コメント・リンク集

- 論文

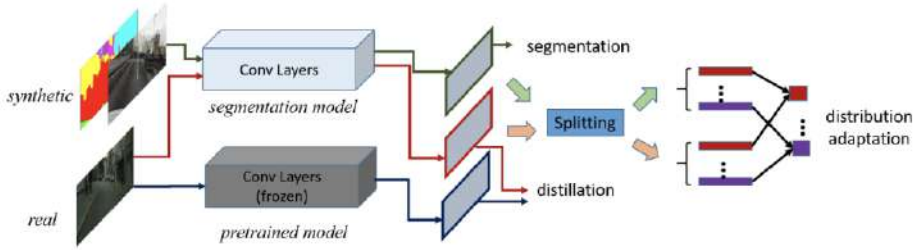
Hiroaki Aizawa

ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes

Yuhua Chen, Wen Li, Luc Van Gool CVPR2018

概要

synthetic-to-realな変換を行う際に、1)モデルがsyntheticにoverfitするstyleの側面と、2)syntheticとrealの分布の違いの側面から発生する2つの問題があることに著者らは着目している。解決するために、前者はtarget guided distillation、後者はspatial-aware adaptationという手法を提案し、それを組み合わせた Reality Oriented ADaptation Network(ROAD-Net)を考案。GTAV/SYNTHIA - Cityscapesの適合タスクで評価し、sotaのsemantic segmentationモデルの汎化性能を向上したことを確認。



新規性・結果・なぜ通ったか？

- Semantic SegmentationへのDomain Adaptationの適用が新しい。
- 結果もまたNonAdaptなPSPNetからmIoUが約11.6%向上している。

コメント・リンク集

- Learning to Adapt Structured Output Space for Semantic Segmentationと目的と対象が似通っている。どちらもクラス分類で得られる特徴(ImageNetで学習されたpretrain model)がsegmentationでは有効ではないという主張であり、これをもとにそれぞれmulti-scaleな手法と、distillationによる手法と異なるアプローチをとっているのが興味深い。
- spatial-aware adaptationはPatchGANと似通っており同様の性質を持つ？
- arxiv

[#198]

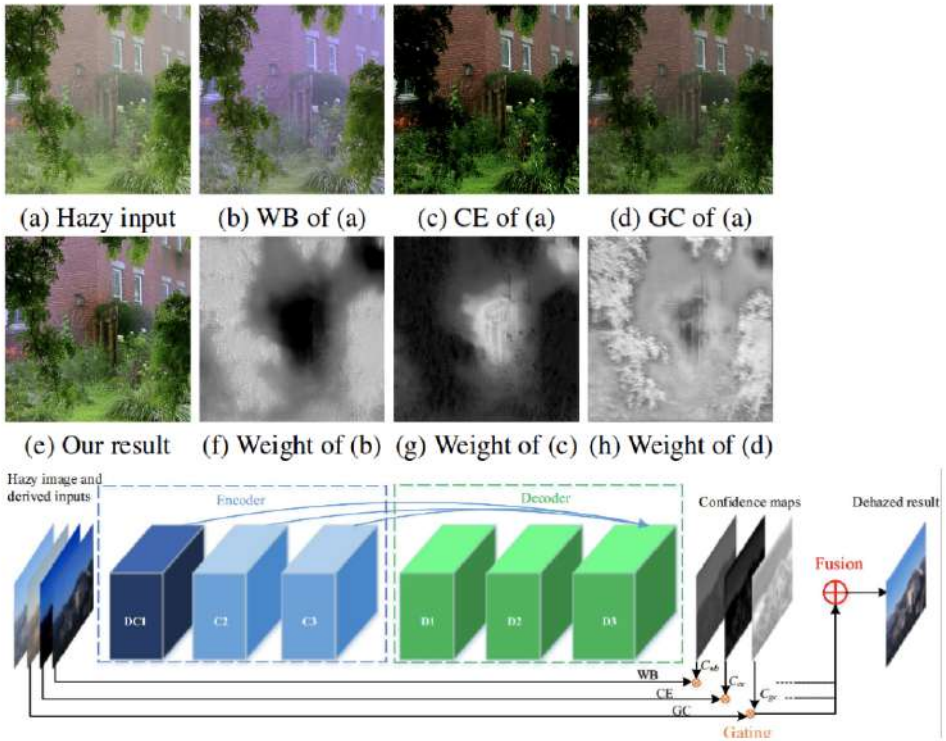
Kazuma Asano

Gated Fusion Network for Single Image Dehazing

Wenqi Ren Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, Ming-Hsuan Yang CVPR2018 404

概要

霧がかかった画像(hazy input)から更に３つの入力，White balanced input，Contrast enhanced input，Gamma corrected inputを計算して導出し，これらの異なる入力間の外観差に基づきピクセル単位のConfidence Mapを計算する．これらを学習することで鮮明な画像を生成するMulti-scale Gated Fusion Network(GFN)を開発した．



新規性・結果・なぜ通ったか？

従来手法と比較し，実装や再現が容易であり，また出力結果もPSNR，SSIMともに従来手法より高い評価となっている．

コメント・リンク集

- arXiv
- プロジェクトページ

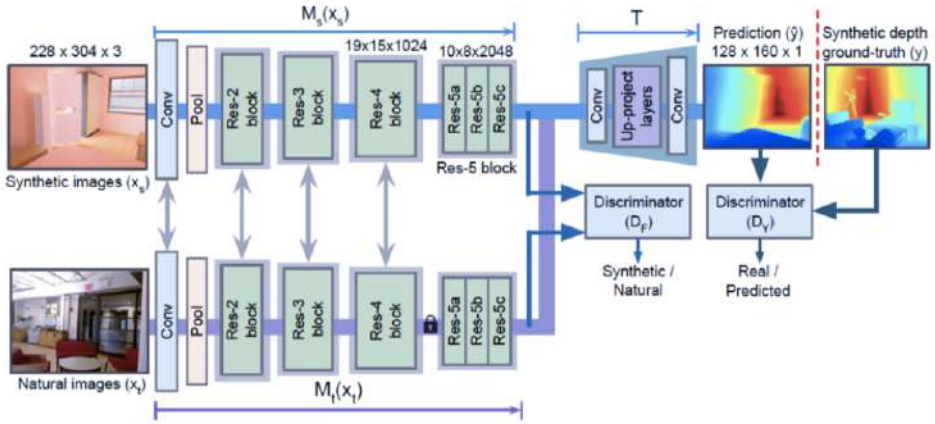
KotaYoshida

AdaDepth: Unsupervised Content Congruent Adaptation for Depth Estimation

J.Nath, K.Phani, K.Uppala, A.Pahuja and R.V.Babu CVPR2018 arXiv:1803.01599

概要

教師あり深層学習による手法は単眼カメラ画像における深さ推定に対して良い結果を出している．しかし．grand truthを得るためにはノイズに影響され，コストもかかる．合成データセットを用いた場合の深度推定では固有のドメインにしか対応していなく，自然なシーンに対して対応するのが難しいと言われる．この問題に対応するため，Adversalな学習と対応したターゲットの明確な一貫性をかすこと事によりAdaDepthを提案．



新規性・結果・なぜ通ったか？

- 高次元の構造化エンコーダ表現に作用する，教師なしの敵対的適応設定AdaDepthを提案．
- 新規の特徴を再構成する正則化フレームワークを使用して適応表現にコンテンツ一貫性を課すことでモード崩壊の問題に取り組んだ．
- 最小限の教師データでの自然シーンの深度推定タスクにおいてSoTAを達成．

コメント・リンク集

- Paper

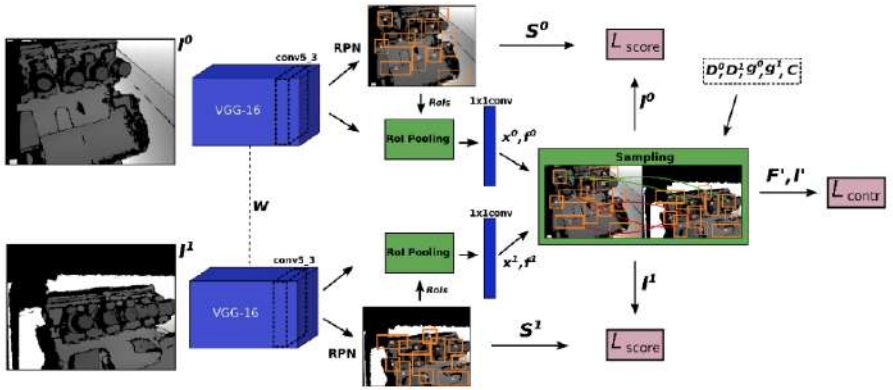
[#200] Goshi Sasaki

End-to-end learning of keypoint detector and descriptor for pose invariant 3D matching

Georgios Georgakis, Srikrishna Karanam,Ziyan Wu,Jan Ernst,Jana Kosecka CVPR 2018 227

概要

End-to-Endで3次元空間における特徴点の抽出とマッチングを行う手法を提案した。2つの距離画像を入力とし、VGG-16 を利用した Faster R-CNNを基本構造としている。 ２つの距離画像からそれぞれVGG−16を利用して特徴マップを作成し、RPNにより領域候補を推定して、ROIプーリング層、全結合層を経て特徴量ベクトルを作り出す。最終的にcontrastive lossを利用して得られた特徴量間の対応関係を求めた。



新規性・結果・なぜ通ったか？

初めてEnd-to-Endで3次元マッチングを行えるようにした。ノイズ環境下においてキーポイントマッチングで従来手法のHarris3D +FPFHなどよりも10%以上高い精度を出した。

コメント・リンク集

- 論文

[#201]

Naofumi Akimoto

AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He CVPR2018

概要

アテンションドリブン，複数ステージでのRefineによって，テキストから詳細な画像を生成するGANを提案．CUBデータセットとCOCOデータセットでinception scoreがstate of the artを超えた．生成画像の特定の位置をワードレベルで条件付けしていることを示した．

貢献：
・Attentional Generative Adversarial NetworkとDeep Attentional Multimodal Similarity Model(DAMSM)の提案．
・実験でstate-of-the-art GAN modelsを超えたことを示す．
・ワードレベルで自動的に生成画像の一部をアテンションするのは初である．

手法

・Attentional Generative Networkはセンテンスの特徴から始めて段階的に画像を高精細にしていくネットワークで，途中にアテンションレイヤーからのワード特徴を入力して条件付けする．
・各解像度に対してそれぞれDiscriminatorがある．
・最終的な解像度になったあと，Image Encoderにて局所的な画像特徴量とし，ワード特徴量とDAMSMにて比較することで，生成画像の細部がどれくらい単語に忠実であるか評価する．

コメント・リンク集

・StackGANの著者も共著にいる．
・アテンションにより生成箇所を局所に向けることで，COCOのような複雑なシーンでも対応できるようになっている．

- arXiv

[#202]

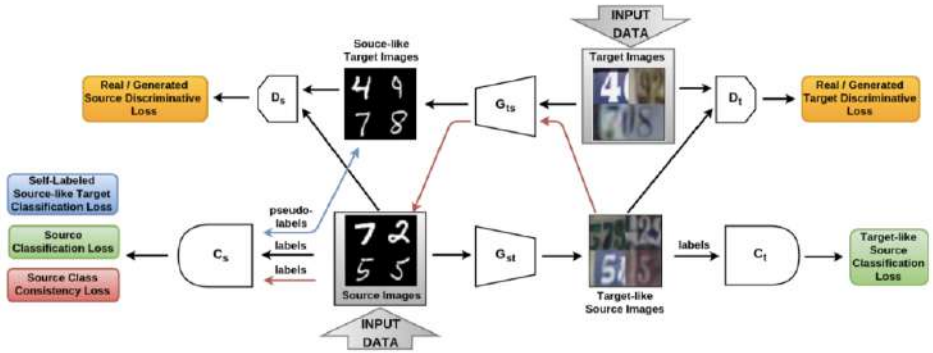
Naofumi Akimoto

From source to target and back: Symmetric Bi-Directional Adaptive GAN

Paolo Russo, Fabio M. Carlucci, Tatiana Tommasi and Barbara Caputo

概要

SBADA-GANの提案．（Symmetric Bi-Directional ADAptive Generative Adversarial Network）
unsupervised cross domain classificationにフォーカス．
ラベルが与えられるSourceのサンプルを利用して，最終的にはTargetの分類問題を解く．SourceのサンプルをTargetのドメインに(Image-to-Imageの)マッピングをし，同時に逆方向も行う．分類器の学習に利用するのは，Sourceサンプル，TargetをSource風にしたもの，SourceをTarget風にしてさらにSource風に戻した３種類を使う．それぞれにラベルもしくは擬似ラベルを付与して学習する．テスト時はTargetサンプルのクラスを予測したいので，Target用の分類器と，TargetサンプルをSource風にしてから入力するSource用の分類器の２つを使用する．



手法

- セルフラベリングの使用．Source用の分類器に制約を課す
- class consistency lossの導入．Generatorとともに利用することで両方向のドメイン変換がお互いに影響し合うようになる．安定性と質向上の効果．最終的な目標である分類問題を解くことに有効．
- 例えばSource側のDiscriminatorは，RealサンプルとしてSource画像を使い，FakeサンプルとしてTarget画像をSource画像風にGeneratorでドメイン変換した画像を使う．
- （問題設定的に）Source側の分類器にはクラスラベルによる学習ができる．
- SourceとTargetの双方向のサンプル生成のための二つadversarial

コメント・リンク集

- arXiv

[#203]

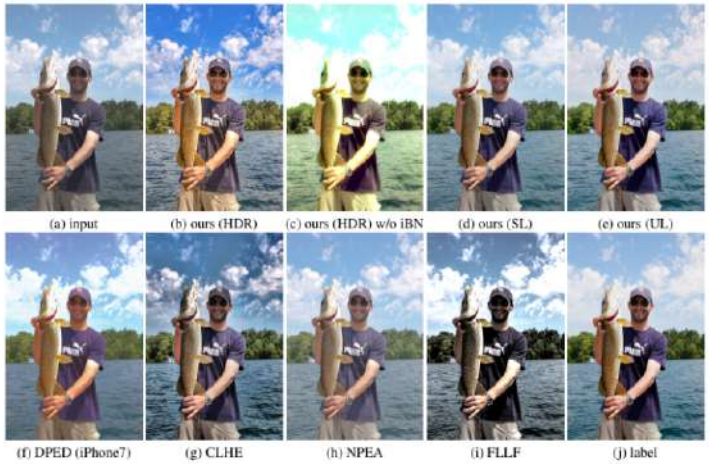
Naofumi Akimoto

Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, Yung-Yu Chuang CVPR2018

概要

学習ベースで画像のエンハンスメントを行う手法の提案．入力として「良い」写真のセットを使う．このセットに含まれる特色を持つように変換することが「エンハンスメント」に繋がると定義する．エンハンスメント問題をimage-to-imageの問題として扱い，提案手法は「良い」写真のセットの中で共通の特色を発見することを狙っている．普通の写真のドメインを「良い」写真のドメインに変換すれば良いとし，（CycleGANのような）２方向GANを以下の３つの工夫とともに利用する．



CONTRIBUTION

- global featureを使ったU-netの利用．これがシーンの状況，照明条件，対象のタイプの情報を捉える．
- WGANのためのadaptive weighting schemeを提案．収束を早める．
- individual batch normalization layersの利用．Generatorは入力データの分布により適応するようになる．

コメント・リンク集

- Flickerのレタッチされた写真を利用するなどしている．
- Adobeがプロ写真家一人一人のレタッチ方法を再現するという機能を実装するのも近いかもしれない．
- ハイダイナミックレンジの写真にしたらエンハンスされていると思っている節がある．
- 論文

[#204]

Naofumi Akimoto

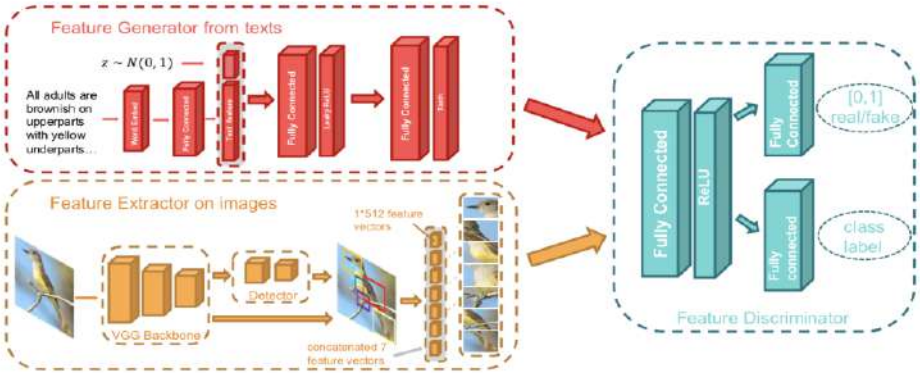
Imagine it for me: Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, and Ahmed Elgammal CVPR2018

概要

Wikipediaのようにノイズの多いテキストからzero-shot learningを行うためのGAN用いる方法を提案．GANを使ってテキストが表現するオブジェクトのビジュアル的な特徴を生成する．オブジェクトのクラスごとに特徴を近い位置にembeddingできれば良い．これができれば後は教師あり手法で分類を行えることになる．
コントリビューション：

- zero-shot learningにおいてUnseenであるクラスのテキスト記述からvisual featureを生成することで，zero-shot learningを従来の分類問題にしてしまう．generative adversarial approach for ZSL (GAZSL)．
- ノイズを抑制するためのFC層と埋め込み後のクラス識別性を高めるvisual pivot regularizationの提案．
- zero-shot recognition, generalized zero-shot learning, and zero-shot retrievalという複数のタスクでstate-of-the-art手法を超えた．



左上段がFakeデータを作るストリーム．左下段がRealデータを作るストリーム．

手法

Unseenクラスについてのノイズを含むテキスト記述を入力とし，このクラスのvisual featureを生成するGANを提案．テキストから生成されるvisual featureをFakeデータとし，真の画像から得られるvisual featureをRealデータとしてGANを学習．

- テキストのembedding後，FC層で次元圧縮をし，ノイズの影響を

コメント・リンク集

- arXiv

[#205]

Naofumi Akimoto

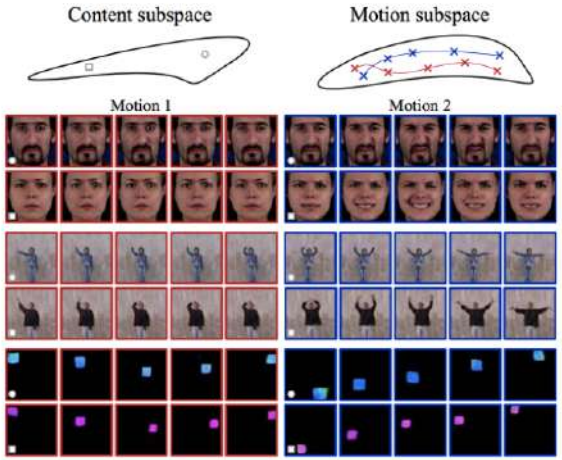
MoCoGAN: Decomposing Motion and Content for Video Generation

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, Jan Kautz CVPR2018

概要

教師不要でコンテンツとモーションという要素に分解し，ビデオを生成するGANを提案．コンテンツを固定しモーションのみ変化させることや，逆も可能．広範囲の実験を行い，量と質ともにSoTAであることを確認．人の服装とモーションの分離や，顔のアイデンティティーと表情の分離が可能であることを示している．

Contribution:・ノイズからビデオを生成する，条件なしでのビデオ生成GANの提案．・従来手法では不可能である，コンテンツとモーションのコントロールが可能なこと ・従来のSoTA手法との比較



手法

- GAN．
- ランダムベクトルのシーケンスをビデオフレームのシーケンスにマッピングするGenerator．ランダムベクトルの一部はコンテンツ，もう一部はモーションを指定するもの．
- コンテンツの部分空間はガウス分布でモデル化．モーションの部分空間はRNNでモデル化．
- Generatorは一つのフレーム分をベクトルからフレームにマップする働きだけなので，モーションを決めるのは連続するベクトルを生成するRNN部分となる．
- 1枚のフレームを入力とするDiscriminatorと連続した数フレームを入力とするDiscriminatorを使うGAN構造を新たに提案．

コメント・リンク集

- ビデオはコンテンツとモーションに分けられるという前提（prior）からスタート
- arXiv

[#206]

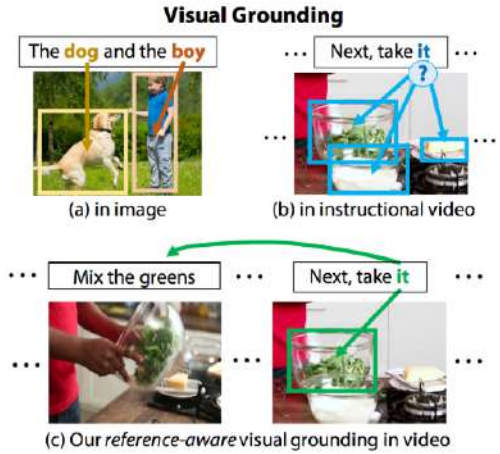
Hirokatsu Kataoka

Finding “It”: Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos

De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, Juan Carlos Niebles CVPR 2018

概要

言語的な文脈の中で指示語からそれが何であるかを特定する問題（Visual Grounding;「それを取ってください」の「それ」を動画中から探索するなど）を扱う論文である。この問題に対してMIL（Multiple Instance Learning）を参考にした弱教師付き学習であるReference-aware MIL（RA-MIL）を用いて解決する。



新規性・結果・なぜ通ったか？

画像に対するVisual Groundingが空間的な関係性を捉えるのに対して、Visual Groundingは時間的な関係性を捉える課題である。YouCookII/RoboWatch datasetにて処理を行った結果、弱教師付き学習であるRA-MILを適用するとVisual Groundingに対して精度向上することを明らかにした。

コメント・リンク集

Language and Visionの課題はすでに動画にまで及んでいる。Visual Groundingのみならず、新規問題設定を試みた論文として精読してもよいかも？それと視覚と言語のサーベイ論文は読んでみたい

- 論文
- 著者
- 視覚と言語のサーベイ論文

[#207]

Hirokatsu Kataoka

Practical Block-wise Neural Network Architecture Generation

Zhao Zhong, Junjie Yan, Wei Wu, Jing Shao, Cheng-Lin Liu CVPR 2018

概要

ブロック単位でのアーキテクチャ生成手法であるBlockQNNを提案。Q学習（Q-Learning）を参考にして高精度なニューラルネットを探索的（ここではEpsilon-Greedy Exploration Strategyと呼称）に生成する。基本的には生成したブロックを積み上げることによりアーキテクチャを生成するが、早期棄却の枠組みも設けることで探索を効率化している。

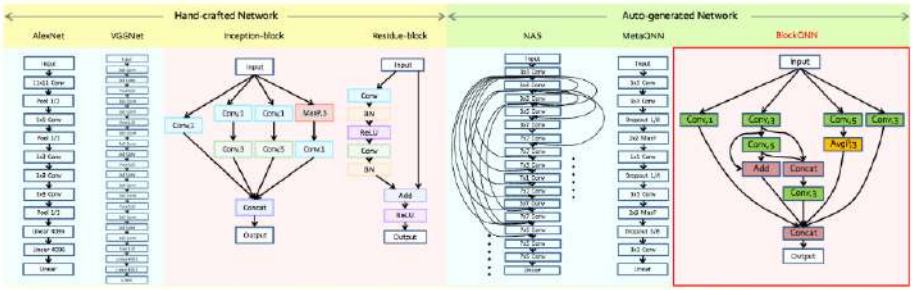


Figure 1. The proposed **BlockQNN** (right in red box) compared with the hand-crafted networks marked in yellow and the existing auto-generated networks in green. Automatically generating the `plain` networks [2, 37] marked in blue need large computational costs on searching optimal layer types and hyperparameters for each single layer, while the `block`-wise network heavily reduces the cost to search structures only for one block. The entire network is then constructed by stacking the generated blocks. Similar `block` concept has been demonstrated its superiority in hand-crafted networks, such as inception-block and residue-block marked in red.

新規性・結果・なぜ通ったか？

ブロック単位でニューラルネットのアーキテクチャを探索するBlockQNNを提案した。同枠組みはHand-craftedなアーキテクチャに近い精度を出しており（CIFAR-10のtop-1エラー率で3.54）、探索空間を削減（32GPUを3日間使用するのみ！）、さらに生成した構造はCIFARのみならずImageNetでも同様に高精度を出すことを明らかにした。ネットワーク構造の探索問題においてブロックに着目し、性能を向上させると同時に同様の枠組みを複数のデータセットにて成功させる枠組みを提案したことが、CVPRに採択された基準である。

コメント・リンク集

ここから数年で、practicalなGPU数（8GPUや4GPUなど）、1日以内の探索で解決するようになると予想される（し、してくれないと一般の研究者/企業が参入できない）。

- 論文

[#208]

Hirokatsu Kataoka

Residual Dense Network for Image Super-Resolution

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, Yun Fu CVPR 2018

概要

低解像画像から高解像画像（SR; super-resolution image）を復元するための研究で、DenseNet（論文中の参考文献7）を参考にしたResidual Dense Networks (RDN)を提案して同課題にとりくんだ。異なる劣化特徴をとらえたモデルであること、連続的メモリ構造（Contiguous Memory Mechanism）やコネクションを効果的にするResidual Dense Blockを提案したこと、Global Feature Fusionにより各階層から総合的な特徴表現、を行い高解像画像を復元した。DenseNetで提案されているDense Blockと比較すると、提案のResidual Dense Blockは入力チャネルからもスキップコネクションが導入されているため、よりSRの問題設定に沿ったモデルになったと言える。

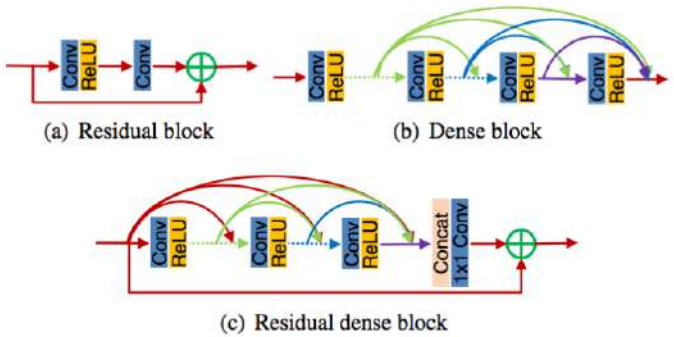


Figure 1. Comparison of prior network structures (a,b) and our residual dense block (c). (a) Residual block in MDSR [17]. (b) Dense block in SRDenseNet [31]. (c) Our residual dense block.

新規性・結果・なぜ通ったか？

高解像画像を復元するための改善として、DenseNetを改良したRDNを提案した。Dense Blockを置き換え、より問題に特化したResidual Dense Blockを適用。実験で使用した全てのデータセット（Set5, Set14, B100, Urban100, Manga109）の全てのスケール（x2, x3, x4）にて従来手法よりも良好なAverage PSNR/SSIMを記録した。結果画像はGitHubのページなどを参照されたい。

コメント・リンク集

課題の肝をつかんで、従来提案されている効果的な手法を改善できるセンスを磨きたい。

- 論文
- GitHub

[#209]

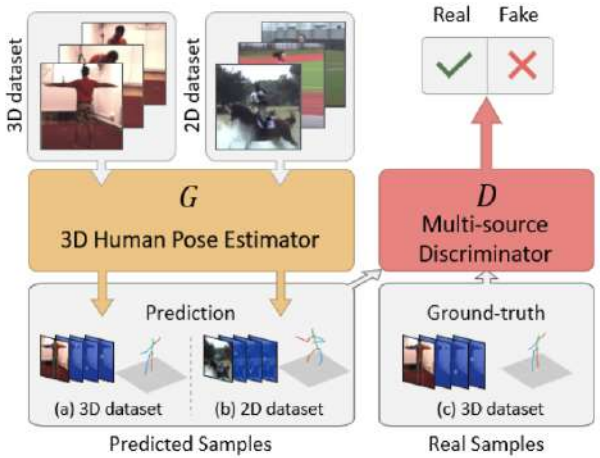
Hirokatsu Kataoka

Three Dimension Human Pose Estimation in the Wild by Adversarial Learning

Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, Xiaogang Wang CVPR 2018

概要

現在でもチャレンジングな課題として位置付けられる人物に対する3次元姿勢推定に関する研究で、Adversarial Learning (AL)を用いて学習を実施。問題設定としては「多量の」2次元姿勢アノテーション+「少量の」3次元姿勢アノテーションを使用することで、新規環境にて3次元姿勢推定を実行することである。本論文で提案するALではG（生成器）として、2D/3Dのデータセットからそれぞれ2D/3Dの姿勢を推定、実際のデータセットからアノテーションを参照（リアル）して、生成されたものか、データセットのアノテーションなのかを判断（D; 識別器）させることで学習する。G側の姿勢推定ではHourglassによるConv-Deconvモデルを採用、D側には3つの対象ドメイン（オリジナルDB、関節間の相対的位置、2D姿勢位置と距離情報）を入れ込んだMulti-Source Discriminatorを適用する。



新規性・結果・なぜ通ったか？

GANに端を発する敵対的学習を用いて、3次元姿勢に関するアノテーションが少ない場合でもドメイン依存をすることなく3次元姿勢推定を可能にする技術を提案した。また、もう一つの新規性としてドメインに関する事前知識を識別器に入れ込んでおくmulti-source discriminatorについても提案した。

コメント・リンク集

少量のラベル付きデータが用意できていれば、ドメイン関係なく推定ができるという好例である。データとアノテーションに関連するのはCG/敵対的学習/教師なし/ドメイン適応などで、これらは現在のCVにおいても重要技術。少なくともお金がないとクラウドソーシングでデータが集められないという構図を変えたいと思っている。

- 論文
- 著者

[#210]

Hirokatsu Kataoka

Gesture Recognition: Focus on the Hands

Pradyumna Narayana, J. Ross Beveridge, Bruce A. Draper CVPR 2018

概要

手部領域に着目してチャネルを追加することにより、ジェスチャ認識自体の精度を高めていくという取り組み。従来型のマルチチャネル（rgb, depth, flow）のネットワークでは限定的な領域を評価して特徴評価を行なっていたが、提案のFOANetでは注目領域（global, right hand, left hand）に対して分割されたチャネルの特徴を用いて特徴評価を行い識別を実施する。図に示すアーキテクチャがFOANetである。FOANetでは12のチャネルを別々に処理・統合し、統合を行うネットワークを通り抜けて識別を実施する。

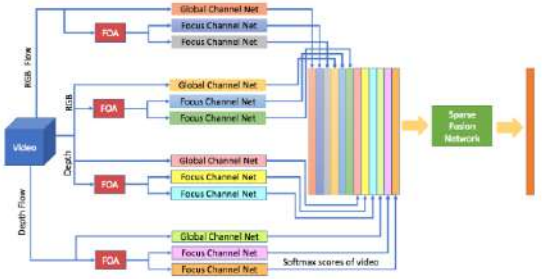


Figure 2. The FOANet Network Architecture. The architecture consists of a separate channel for every focus region (global, left hand, right hand) and modality (RGB, depth, RGB flow and depth flow). FOA module is used to detect hands. The video level softmax scores from 12 channels are stacked together. Sparse fusion combines softmax scores according to the gesture type.

新規性・結果・なぜ通ったか？

手部領域に着目し、よりよい特徴量として追加できないか検討した、とういアイディア自体が面白い。また、ChaLearn IsoGD datasetの精度を従来の67.71%から82.07まで引き上げたのと、同じようにNVIDIA datasetに対しても83.8%から91.28%に引き上げた。

コメント・リンク集

あまりメジャーに使用されているDBではないが、重要課題を見つけてアプローチする研究は今後さらに必要になってくる？一番最初に問題を解いた人ではないが、二番目に研究をして実利用まで一気に近づけられる人も重宝される。

- 論文

[#211]

Ryota Suzuki

Direct Shape Regression Networks for End-to-End Face Alignment

X. Miao, X. Zhen, V. Athitsos, X. Liu, C. Deng and H. Huang CVPR2018

概要

顔のアライメントにおいて，Direct shape regression networkを提案．いくつかの新しい構造を組み合わせている．(1)二重Conv，(2)フーリエ特徴プーリング，(3)線形低ランク学習．顔画像-顔形状間の高い非線形関係性（初期化への強い依存性，ランドマーク相関導出の失敗）の問題を解決する．

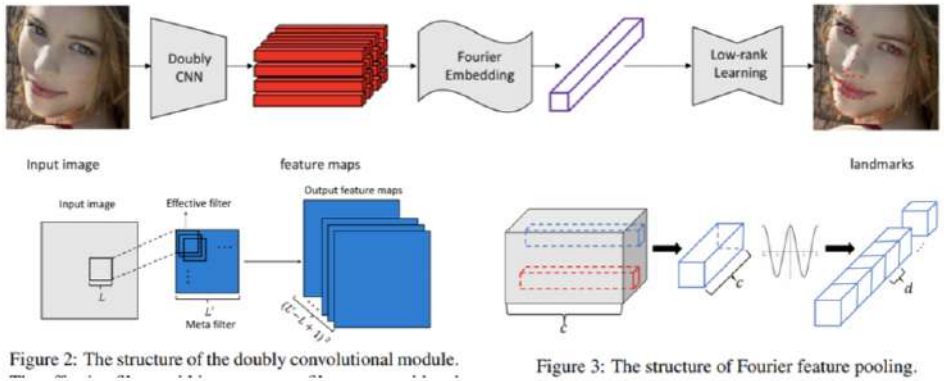


Figure 2: The structure of the doubly convolutional module.

Figure 3: The structure of Fourier feature pooling.

新規性・結果・なぜ通ったか？

- 複数の新しい構造の定義
- いくつかのケースでSoTAを超える性能．

コメント・リンク集

- 論文

[#212]

Ryota Suzuki

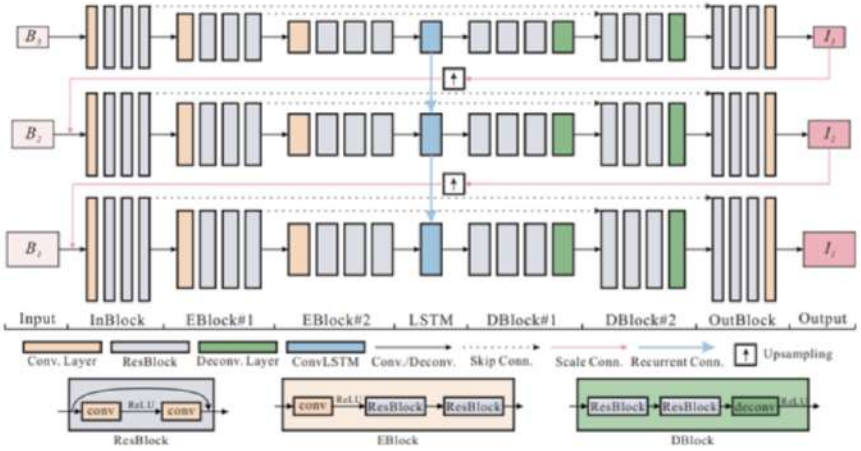
Scale-recurrent Network for Deep Image Deblurring

X. Tao, H. Gao, Y. Wang, X. Shen, J. Wang, J. Jia CVPR2018

概要

coarse-to-filneに単画像デブラーリングする，Scale-recurrent Network (SRN-DeblurNet)を提案．

構造的には，(1)入出力がピラミッド画像， (2)中間はUnet， (3)最終層の出力を第1層に注入（Recurrent）し，ピラミッド画像の枚数分実行．



新規性・結果・なぜ通ったか？

- シンプルでパラメータ数が少ない．
- SoTAを超える性能．例もすごいきれいになっているように見える．

コメント・リンク集

見た目明らかにきれいになっていると，やはり評価したくなる．

- arXiv

[#213]

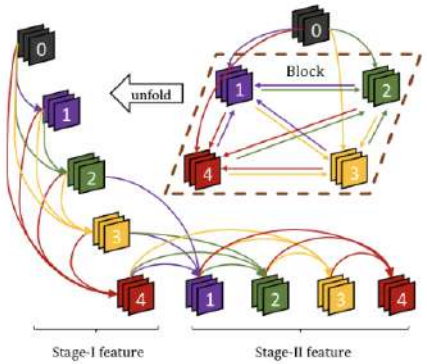
Tomoyuki Suzuki

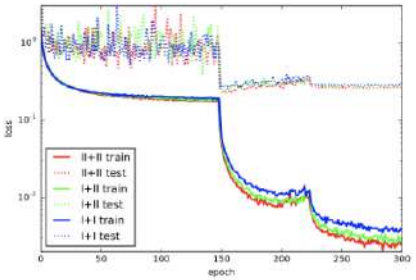
Convolutional Neural Networks with Alternately Updated Clique

Yibo Yang et al., CVPR 2018

概要

従来のCNNの構造では基本的に決められた方向へのみのforwardを行うのに対して、すべてのレイヤー間で結合を持つClique blockで構成されるClique Netの提案。CIFAR-10でSoTA、その他ImangeNetやSVHNでも少ないパラメータでSoTAに匹敵する精度を記録。





手法・なぜ通ったか？

Clique blockでは以下のような処理が行われる。

- 畳み込み層によってすべての層を共通の特徴マップで初期化。
- ある層に対して、他のすべての層から畳み込み結合した値で更新。これを各層に対して順次行い、すべての層で更新したら1つのStageが終了。
- 上記を決められたStage数行う。畳み込み結合の重みはStage間で共有する。

DenseNetの拡張に近い構造のため妥当性があり、実際に精度が出ている点が強い。

コメント・リンク集

- 論文

Tomoyuki Suzuki

Geometry Guided Convolutional Neural Networks for Self-Supervised Video Representation Learning

Chuang Gan et al., CVPR 2018

概要

合成画像のペア間のフローと教師ラベルのない実画像のペア間のデプスを推定することによってシーン認識、行動認識のための表現学習を行う研究。フロー推定を行ったのち、デプス推定にfine-tuningし、さらに目的となるタスクにfine-tuningする。 直感的には、低レベルな特徴が獲得されそうだが、行動認識などの高次な問題設定でも効果を発揮した。

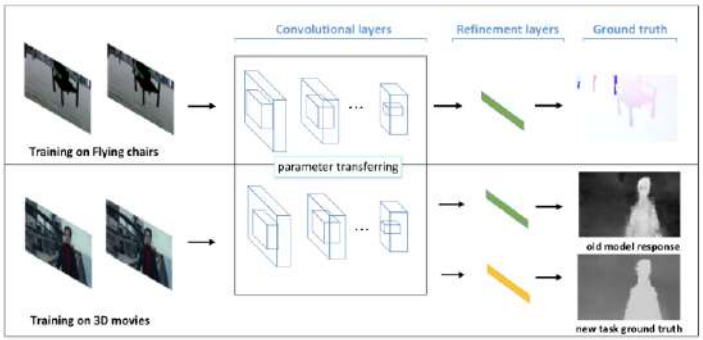


Figure 2. The framework of our proposed geometry guided CNN. We firstly use the synthetic images to train a CNN, and then use the 3D movies to further update the network.

手法・なぜ通ったか？

多段にfine-tuningするため、初期の問題設定によって獲得した特徴が失われてしまう可能性があるので、２段目のfine-tuning時にはfine-tuning前の出力結果への蒸留を同時に行う。ImageNetのpretrainingとも行動認識において補間的な関係がある。表現学習自体での使用データが少ないのに関わらず高い精度向上が実験的に示されたことが大きなcontributionだと考えられる。

コメント・リンク集

特徴のforgetを防ぐ手法は、複数のタスクで学習済みモデルを作成する際に、その順番が重要となるような状況で有用だと思われる。既存手法との比較においては今回は+αのデータを利用している点はフェアではないと感じた。 また、目的のタスクへのfine-tuningの際のフレームペアの選び方などの詳細な設定が記されていなかった。主に精度評価のみで、高次なタスクでうまくいく考察が少なく、疑問もあった。

- 論文

[#215]

Ryota Suzuki

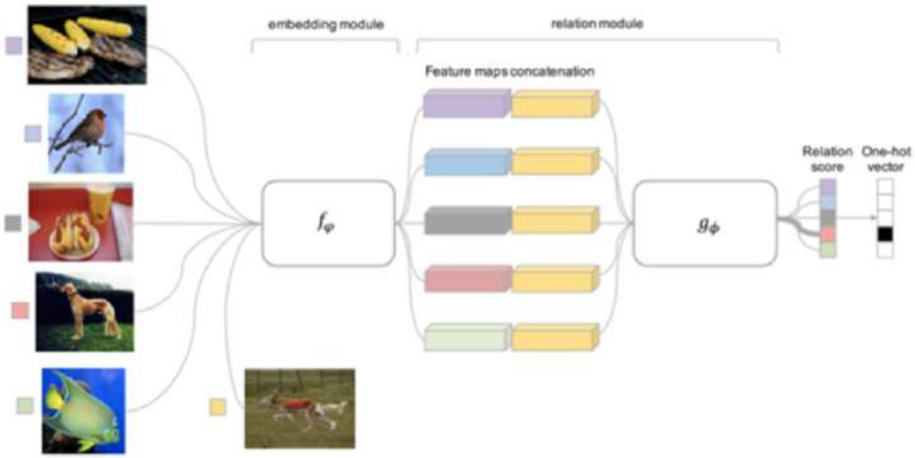
Learning to Compare: Relation Network for Few-Shot Learning

F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr, T.M. Hospedales CVPR2018

概要

メタ学習を用いたFew-shot learningの新しい枠組み，Relation Networkの提案．一度学習されれば，ネットワークのアップデートの必要なしに新しいクラスの画像分類ができるようになる．

1エピソードにおける少数の画像の比較によって距離メトリックを学習するメタラーニングを行う．少数の新クラスの代表画像群とクエリ画像の関連性スコアの比較により，追加学習なしに新クラス画像分類が行える．



新規性・結果・なぜ通ったか？

- 再学習しなくても，データさえ用意しておけば未知のクラスも分類可能な画像分類器ができる．
- Zero-shot learningにも拡張可能．
- シンプルで，高速に動作し，拡張性も高い．

コメント・リンク集

テスト時も少数のデータを用意しておけば，という考え方はイマドキ感がある．

- arXiv
- GitHub

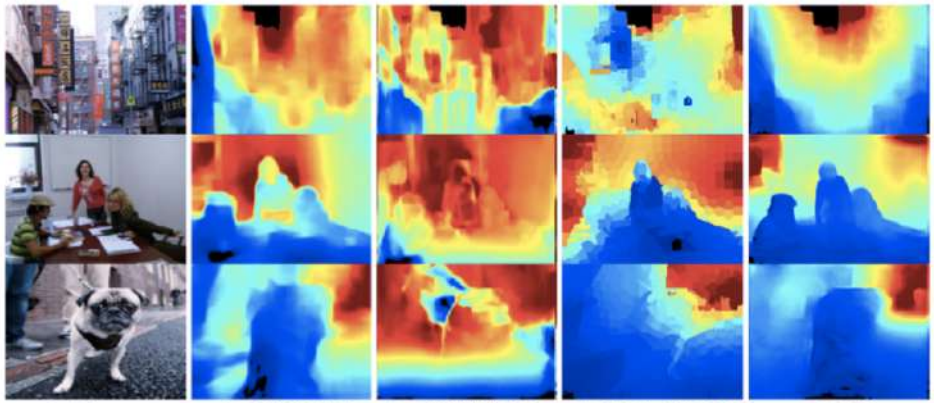
[#216] KotaYoshida

MegaDepth: Learning Single-View Depth Prediction from Internet Photos

Z.Li and N.Snavely CVPR2018 arXiv:1804.00607

概要

画像における深度予測はCV分野において基本的なタスクである．既存の手法は学習データによる制約が伴う．今回提案する手法では，インターネットの画像をデータセットとするMVSの手法を改良し，既存の3D reconstructionとsemantic ラベルを組みわせて大規模な深度予測モデルであるMegaDepthを提案．



(a) Image (b) NYU [6] (c) KITTI [13] (d) Make3D [22] (e) MD

新規性・結果・なぜ通ったか？

- セマンティックセグメンテーションを用いた順序による深度関係を自動で拡張
- MegaDepthが強力なモデルであることを示すために膨大なインターネット画像を使い検証

コメント・リンク集

- 深度予測にsemantic ラベルを取り入れることで精度が向上．
- semanticラベルを用いており，複雑背景における物体検出にも応用可能かも！！
- Paper

[#217]

Kazuki Inoue

Real-Time Rotation-Invariant Face Detection with Progressive Calibration Networks

FXuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, Xilin Chen CVPR 2018 Poster

概要

リアルタイムで顔の回転に頑健な顔検出を行うProgressive Calibration Network(PCN)を提案。PCNは3つのステージで構成されており、それぞれのステージでは検出された領域を0° or 180°回転させる、 0° or 90° or -90°回転させる、頭が上にくるように顔を回転させる、という処理をそれぞれ行う。 また各ステージ共通で検出された領域が顔であるか顔でないかという識別を行う。第1,2ステージで粗く回転を行うことで第3ステージにおける回転量と、 各ステージにおける顔識別の学習が容易になったことで、高精度かつリアルタイムに顔検出を行うことが可能となった。

新規性・結果・なぜ通ったか？

- 従来手法であるデータオーギュメンテーション、角度の値域を分割してそれぞれの検出器を学習させる方法、角度の回転角を推定する流手法では、どれもネットワークが大きくなりすぎるためにリアルタイムでの実行が難しかった。
- 解像度が40x40以上の顔を検出。
- state-of-the-artの手法と比べて同等の精度を達成し、かつGPUを使用した際の実行スピードは4.2倍となった。

コメント・リンク集

- GitHub with Demos
- 論文

[#218]

Kazuki Inoue

Partially Shared Multi-Task Convolutional Neural Network with Local Constraint for Face Attribute Learning

Jiajiong Cao, Yingming Li, Zhongfei Zhang CVPR 2018 Poster

概要

顔のアトリビュート推定に有効なネットワークであるPS-MCNN/-LCを提案。従来手法のMCNNでは、類似度の高いアトリビュートの識別率を高めるために、 類似度の高いアトリビュートのごとにグループを形成し、MCNNの高い層では各グループごとにCNNを形成して学習を行なっていた。 そのため低い層で得られていた特徴量が消失するという問題が起きていた。 これを解決するために、MCNNに対して各レベルで得られた特徴量を教諭するShared Netを導入したPS-MCNNを提案。 また同一人物において推定されたアトリビュート同士のロスをとるPS-MCNN-LCも提案した。 ネットワークの構築に関する議論も行なっている。

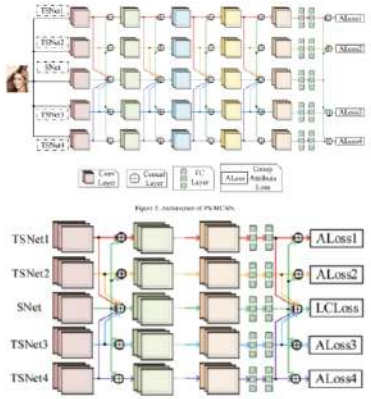


Figure 3. The architecture of PS-MCNN-LC, which shares the architecture of PS-MCNN but is different at the loss function.

新規性・結果・なぜ通ったか？

- 同一人物において推定されたアトリビュート同士のロスをとることで、アトリビュートの空間を限定することが可能となるという考えのもとPS-MCNN-LCを提案している。
- state-of-the-artに比べて、CelebAデータセットではPS-MCNN-LCが40種全てのアトリビュートにおいて最も高い精度を達成、LFWAデータセットではPS-MCNN/-LCを合わせて37種において最も高い精度を達成。

コメント・リンク集

- 精度が上がったことはもちろんだが、既存研究であるMCNNのリミテーションを正確に見抜いてネット枠を改善している点が採択につながったと考えられる。
- 論文

[#219]

Kazuki Inoue

Deep Semantic Face Deblurring

Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, Ming-Hsuan Yang CVPR 2018 Poster

概要

顔に対してセマンティックセグメンテーション(face sparsing)を利用することで、モーションブラーが加えられた正面顔画像に対するCNNベースのデブラーリング手法を提案。 face sparsingによって顔のパーツの位置関係や形といった情報を利用することができると主張。 また学習の際には様々なカーネルサイズによるブラー画像を同時に与えるのではなく、 小さなカーネルサイズのブラー画像から順々に学習させるincremental trainingことでデブラーリング精度を向上させた。

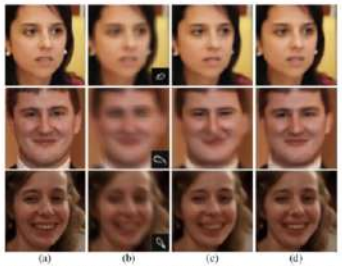
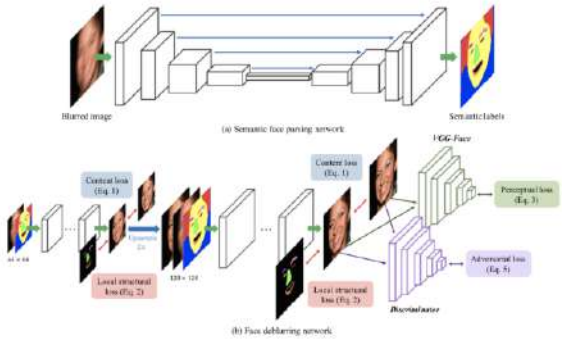


Figure 1. **Face deblurring results**. We exploit the semantic information of face within an end-to-end deep CNN for face image deblurring. (a) Ground truth images (b) Blurred images (c) Ours w/o semantics (d) Ours w/ semantics.



新規性・結果・なぜ通ったか？

- ブラー画像はランダムな3D cameraの軌道によって与えられ、カーネルサイズは13x13~27x27までを学習させた。
- ロスとしてデブラーリング画像のL1 loss, face parsing画像のL1 loss, adversarial loss, CNNの特徴量マップのL2 ロスを使用。
- tate-of-the-artに比べてデブラーリング画像とソース画像のPSNR、SSIM、顔の検出率、個人認証の精度においてもっとも良い精度を達成し、それぞれ約5%, 5%, 28%, 4%向上した。
- state-of-the-artに比べて実行スピードが約44%向上した。

コメント・リンク集

- 学習データを少しずつ変化させて、順々に最適化を行うincremental trainingは、学習データをパラメトリックに変化可能な他の問題に対しても有用なトレーニング方法だと思われる。
- 論文

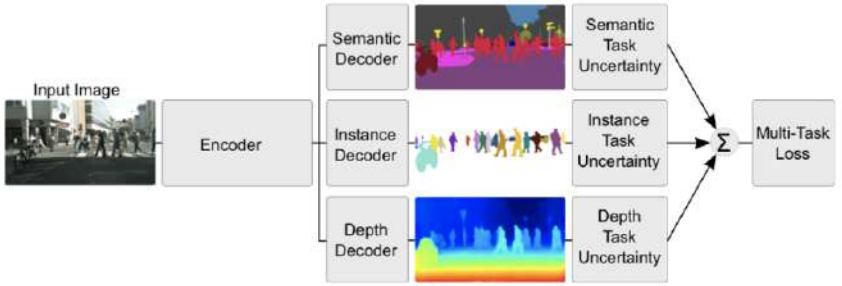
Tomoyuki Suzuki

Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics

Alex Kendall et al., CVPR 2018

概要

学習時のタスクごとの重みによって精度がかなり変化する。そこでNNのマルチタスクモデルにおいて各出力を分布表現にし、その同時確率を最尤推定するように学習することで結果的にタスクごとの不確実性を考慮した重み付けを損失関数に課す。実験ではSemantic Segmentation, Instance Segmentation, Depth estimationのマルチタスク学習を行い、等しい重みや手動での重み設計時よりも良い結果となった。



手法・なぜ通ったか？

モデルから各タスクに対して不確実性を表す値を同時に出力させる。回帰タスクの場合はこれが分散を表し、最終的には回帰出力値を平均とするガウス分布として表現する。識別タスクについては不確実性が分布の温度パラメータとして扱われる。これらの同時確率を最尤推定すると、通常の損失に対してタスクごとに適応的に重み付けされた損失を最適化していることになる。理論的にも妥当であり、精度向上は大きくチューニングの手間が省けるという点でかなり便利である。

コメント・リンク集

簡単な実装でハイパーパラメータが減るという点でかなり有用に感じた。様々なマルチタスクで行った訳ではないのでこの手法の汎用性がきになる。結局、識別の場合は通常でも不確実性は考慮しているので、本質的に新しいのは回帰の場合である。

- 論文

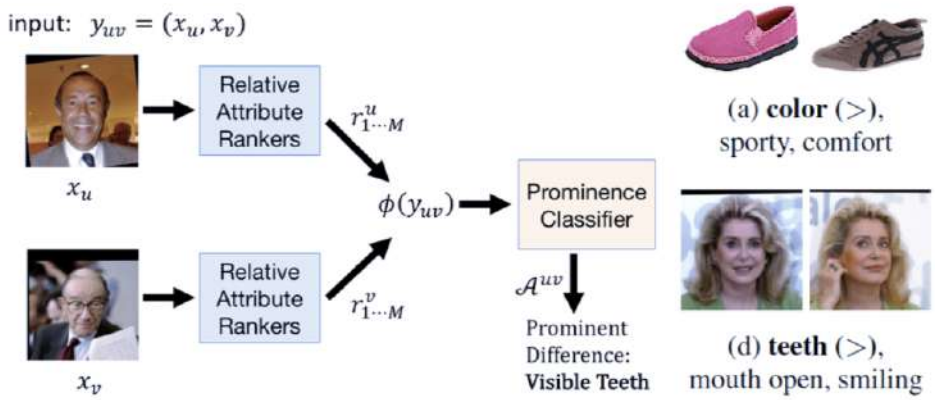
[#222] KotaYoshida

Compare and Contrast: Learning Prominent Visual Differences

S.Chen and K.Grauman CVPR2018 arXiv:1804.00112

概要

2つの画像間で最も顕著な違いは表せられるがその他の細かい違いは示されないことが多い．それに対して，より多くの違いによって画像を比較できるようなモデルの構築をした．また，そのモデルを使って，UT-Zap50K shoesとthe LFW10のデータセットを用いて評価したところSoTAであった．構築したモデルを画像記述と画像検索に導入し，拡張を図った．



新規性・結果・なぜ通ったか？

- 画像中から目立つ部分をアノーテーションで収集し，ランク付けすることでモデルの構築．
- UT-Zap50K shoes（靴）とthe LFW10（顔）のデータセットを用いて評価．
- 画像記述と画像検索のタスクに応用し，拡張を図る

コメント・リンク集

画像説明文に応用できればキャプショニングの幅を広げられそう．

- Paper

[#223]

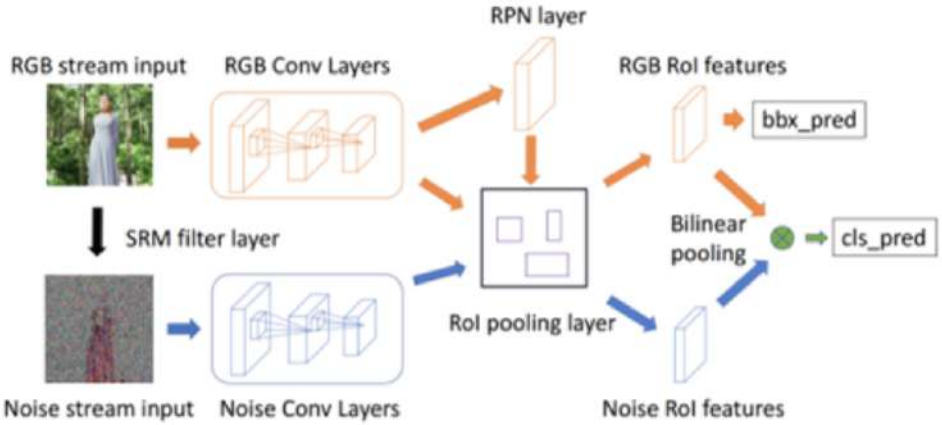
Ryota Suzuki

Learning Rich Features for Image Manipulation Detection

P. Zhou, X. Han, V.I. Morariu and L.S. Davis CVPR2018

概要

画像修正検出．修正箇所をちゃんと注目すべきで，リッチな特徴の学習が必要．修正後画像から修正領域を検出するtwo-stream Faster R-CNNを提案． RGB stream：コントラスト差，不自然境界とかを捉える．Noise stream：ノイズの非一貫性を捉える．Steganalysis Rich Modelでとれたノイズ特徴に基づく． そして，両者のバイリニアプーリングで共起性を捉える．



新規性・結果・なぜ通ったか？

- 修正箇所のノイズ感の差を見るアイデアは昔にあったが，それを導入したという温故知新．
- 実験によりリサイズや圧縮に対するロバスト性におけるSOTAを確認．

コメント・リンク集

- arXiv
- 過去のノイズ感の差を使った画像加工領域検出の例

[#224]

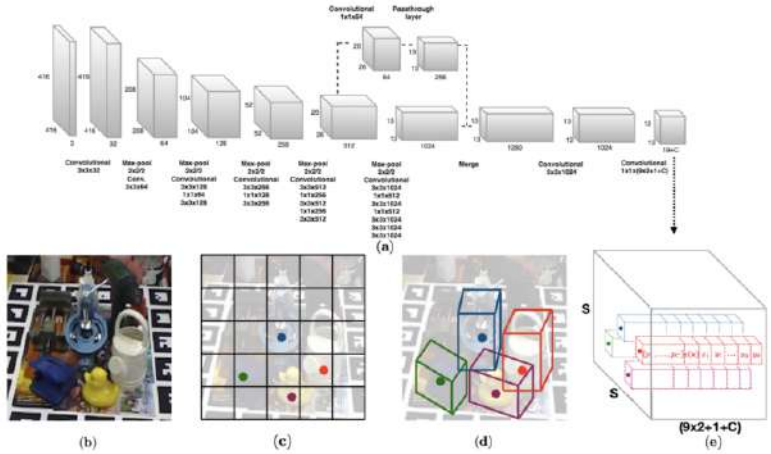
Yoshihiro Fukuhara

Real-Time Seamless Single Shot 6D Object Pose Prediction

Bugra Tekin et al. CVPR 2018

概要

1枚のRGB画像から物体の6次元姿勢を推定する研究. CNN を用いた単一のネットワーク (YOLO v2 ベース) で RGB 画像から物体の 3D bounding box を直接推定する. post-process 無しで高精度な姿勢推定が可能なため, 実時間（従来手法の約５倍速）で従来手法と同程度の推定精度を達成した.



新規性・結果・なぜ通ったか？

- ネットワークはRGB画像1枚の入力に対して, 各物体の制御点（3D bounding box 8点 と centroid 1点）の位置, カテゴリー, 推定の確信度を出力する.
- 推定された物体の9つの制御点の位置に対して PnP 問題を解くことで6次元姿勢を推定する.
- 物体の bounding box の情報から学習を行うので物体の詳細な3次元モデルが必要無い. また, テクスチャーが殆ど無い物体に対しても適用が可能.
- 物体が複数あった場合でも PnP 以外の部分の計算量は増えないので, 物体数に関わらず計算時間はほぼ一定.(従来手法の SSD-6D は線型に増加.)
- LINEMOD や OCCLUSION データセットを用いた評価実験では従

コメント・リンク集

- [論文] Real-Time Seamless Single Shot 6D Object Pose Prediction
- [著者HP] Bugra Tekin

[#225]

Yoshihiro Fukuhara

Video Captioning via Hierarchical Reinforcement Learning

Xin Wang et al. CVPR 2018

概要

Video captioning のための階層型強化学習フレームワークを提案. Caption を複数のセグメントに分割し, High-level の Manager Module が各セグメントのコンテキストをデザインし, Low-level の Worker Modeule が単語を生成することで順次セグメントを作成する. 提案手法は MSR-VTT データセット を用いた評価実験で既存手法よりも複数の評価尺度で良い結果となった. また, video captioning のための新しい大規模データセットを公開.

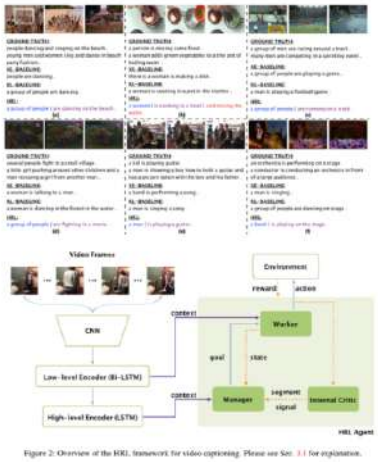


Figure 2: Overview of the HRL framework for video captioning. Please see Sec. 3.1 for explanation.

新規性・結果・なぜ通ったか？

- Video captioning の問題を強化学習の問題として定式化し直し, 効率的に学習をすることができる階層型強化学習手法を提案した.
- High-level の Manager Module が目標を達成するために必要なゴールを設定し, Low-level の Worker Modeule がゴールを達成するための基本行動を行う. また, Internal Critic がゴールが達成されたかどうかの評価を行う.
- Action recognition や segmentation で主に用いられている Charades データセットをもとにvideo captioning のための新しい大規模データセットを作成. 既存の MSR-VTT データセットよりも詳細で長い caption が与えられている.
- MSR-VTT データセットを用いた評価実験では, 既存手法（Mean-Pooling, Soft-Attention, S2VT等）と比較して複数の評価尺度で最も良い結果を得た

コメント・リンク集

- [論文] Video Captioning via Hierarchical Reinforcement Learning
- [著者HP] Xin Wang

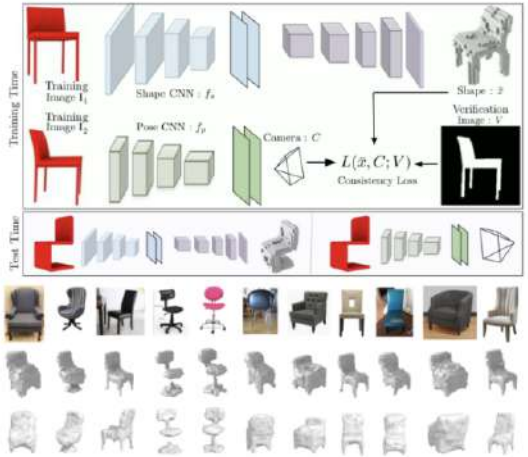
Yoshihiro Fukuhara

Multi-view Consistency as Supervisory Signal for Learning Shape and Pose Prediction

Shubham Tulsiani et al. CVPR 2018

概要

１枚のRGB画像から物体の形状とカメラ姿勢の両方を推定する研究. 異なる視点から見たときの一貫性(具体的には物体の輪郭または深度情報の一貫性)を教師情報として用いるため, 従来手法と異なり学習時に物体の３次元形状と姿勢のいずれについても直接の教師データも必要としない.



新規性・結果・なぜ通ったか？

- 物体の形状とカメラ姿勢の両方を推定するタスクに置いて, 直接の教師データを用いずに学習する方法を提案した.
- 学習時の入力は同一の物体を異なる位置から撮影したRGB画像２枚と２枚目の画像の物体の Mask または Depth 画像.
- １枚目の画像から３次元形状, ２枚目の画像からカメラ姿勢をそれぞれ推定し, 推定された形状を推定された姿勢から見た時に, 与えられたマスク画像と同じ結果が得られるように学習を行う.
- ShapeNet データセットを用いた評価実験では, 直接の教師あり学習を行った手法とほぼ同等の結果であった.

コメント・リンク集

- [論文] Multi-view Consistency as Supervisory Signal for Learning Shape and Pose Prediction
- [Project Page]
- [Code]

[#227]

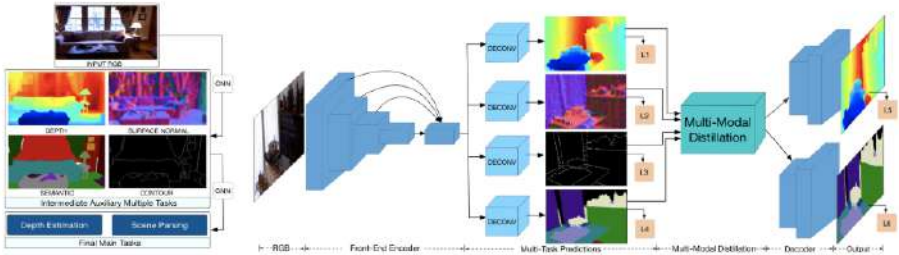
Tomoyuki Suzuki

PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing

Dan Xu et al., CVPR 2018

概要

CNNに対して中間的に法線方向推定と輪郭推定も加えることで最終的にdepth推定とscene parsingの精度を向上させる。法線方向と輪郭についてはdepthとscene parsingのラベルから計算可能であるので追加にアノテーションする必要はない。 NYUD-v2とCityscapesにおいてSoTA。



手法・なぜ通ったか？

中間的に推定した結果を元に最終的な目的タスクを出力するが、その中間出力として3つのパターンを考えた(タスクをに分けずconcat, タスクごとにconcat, attention機構を取り入れたconcat)。attention機構を取り入れたconcatが最も良い結果となった。シンプルな手法だが、実験結果が良いので評価されたと考えられる。

コメント・リンク集

「distillation」という言葉を用いているが、生徒モデルと教師モデルがあるようなdistillation手法は使われておらず、単に複数の中間タスクからのMulti-modalな情報の統合に対してその言葉が使用されている。 単に通常のマルチタスク推定に中間タスクを導入したのみでかなりシンプルな印象。

- 論文

[#228]

Hirokatsu Kataoka

Convolutional Sequence to Sequence Model for Human Dynamics

Chen Li, Zhen Zhang, Wee Sun Lee, Gim Hee Lee CVPR 2018

概要

時空間的な特徴を捉えて、長期のモーション予測を行う研究である（ここではいかに最初の限られた情報量のみでシーケンスを推定できるかどうかについて検証を行なっている）。この課題に対し、Convolutional Long-term Encoderを用いてより長期的な隠れ変数をデコーダにより推定する。このエンコーダ-デコーダ構造にて短期〜より長期的な変数の予測を可能にする。本手法では主にRNNベースのSequence-to-SequenceなモデルにConvolutionalな要素を加えたことが技術的発展であると主張。

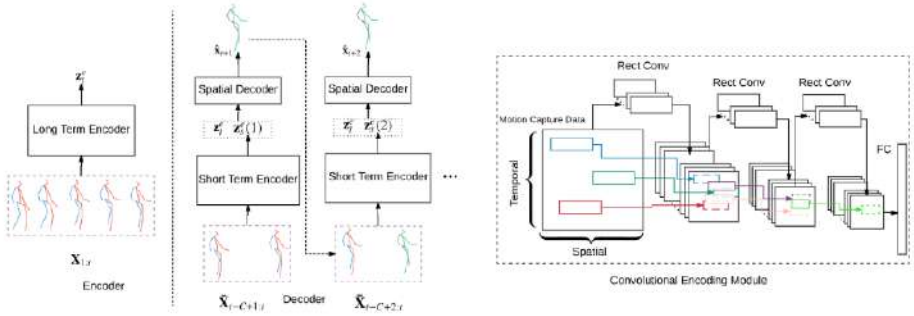


Figure 2. An illustration of our network architecture.

新規性・結果・なぜ通ったか？

より長期の（といっても数秒間のシーケンス？）人物モーション予測（ここでは人物姿勢位置を予測）を実現したことが課題設定として大きい。手法としてはConvolutional Long-term Encoderやその抽象化された特徴をデコーダにより長期隠れ変数を推定。
Human3.6MやCMU Motion Capture datasetにて高い精度を実現した。

コメント・リンク集

Short-termからLong-term（Short-term: 〜3秒、Long-term: 5秒〜; 明確な定義はなされていないが。。）の行動/姿勢の予測はまだまだ未解決だし、何を予測するかに関しての定義づけ自体の整備も曖昧なままである。まだまだ参入の余地が残されているように見える。

- 論文
- GitHub

[#229]

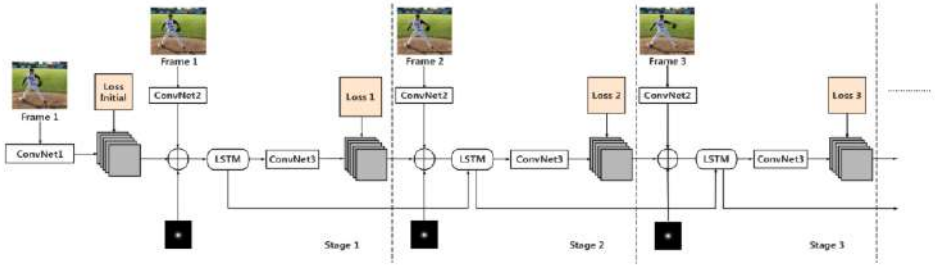
Hirokatsu Kataoka

LSTM Pose Machines

Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, Liang Lin CVPR 2018

概要

Convolutional Pose Machine (CPM)のCNN部分を再帰的ネットであるLSTM (Long-short term memory)により置き換えた人物姿勢推定手法。時系列的に連続するフレーム（e.g. t, t+1, t+2）の入力に対して処理を実行し姿勢を推定する。CPMとは基本となるアーキテクチャの考え方（multi-stage algorithm）は同様であるが、それぞれのステージ間でパラメータを共有している点で異なる。



新規性・結果・なぜ通ったか？

CPMと同じmulti-stageの姿勢推定学習を、LSTMの構造にて実現したことが技術的なポイントである。さらに、CPMとは異なりステージ間でパラメータを共有することで精度向上が見られたと説明。Penn Action datasetやJHMDB datasetにて最高精度を叩き出した。JHMDBにて93.6@PCK(=0.2)、Penn Actionにて97.7@PCK(=0.2)を記録。さらに、各フレーム時のメモリチャンネルの挙動も可視化し、どのような際に成功するか/失敗するかを明らかにした。複雑姿勢（複雑背景？）の際にはエッジに着目していて、姿勢推定が成功する際にはピンポイントで関節位置を回帰する傾向にある。処理速度の面においても本論文の技術では25.6msで動作した（CPMは48.4ms）。

コメント・リンク集

アーキテクチャ自体上手くいったモデルを異なるアプローチ（この場合はConvolutionalなモデルをRecurrentに変更）で実行して、どういった改善ができるかを試しているところが面白い。LSTMの場合にはステージ間でパラメータを共有するところがポイントであった。その上で精度を向上している点が実行力に優れていると言える。

- 論文
- 著者
- GitHub
- YouTube

[#230]

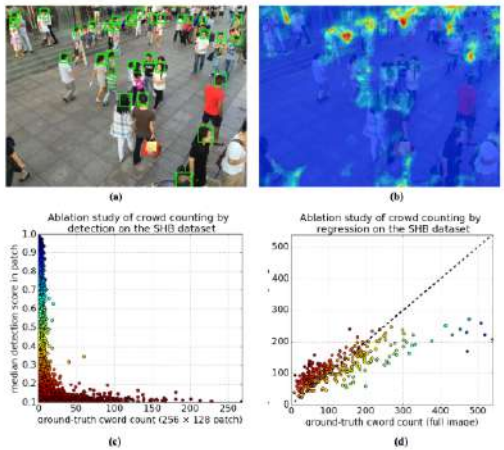
Hirokatsu Kataoka

DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation

Jiang Liu, Chenqiang Gao, Deyu Meng, Alexander G. Hauptmann CVPR 2018

概要

混雑時の人数カウントにおける問題点を解決するため、End-to-Endで学習可能なDecideNet（DEteCtIon and Density Estimation Network）を提案する。混雑時の人数カウントでは、従来（1）人物検出では認識ミスによる過不足によりカウントを誤ってしまう、（2）回帰ベースの手法では人物が存在しない領域が蓄積されると実際のカウントよりも多く集計されてしまう、という問題が存在した。DecideNetでは検出ベース/回帰ベースを別々に行い、それらの結果を総合してカウントを行うという点で従来法を解決していると言える。実験では本論文で提案のDecideNetが混雑時の人数カウントにおいてもっとも優れた精度を達成したと主張。検出/回帰の手法としてはFaster R-CNN/RegNetを適用している。



新規性・結果・なぜ通ったか？

3つのベンチマーク（Mall, ShanghaiTech PartB, WorldExpo10 dataset）においてState-of-the-artな精度を達成すると同時に、混雑時の人数カウントの問題と異なるアプローチを同時実行して相補的なアプローチDecideNetを提案したことが採択された大きな理由である。

コメント・リンク集

異なる複数のアプローチを統合して最高精度を達成するためには、その分野における積み重ねと実装力が必要である。論文の書き方と合わせて鍛えていくことで毎回難関国際会議に突破できる力がつくと思われる。

- 論文
- Project
- GitHub

[#231]

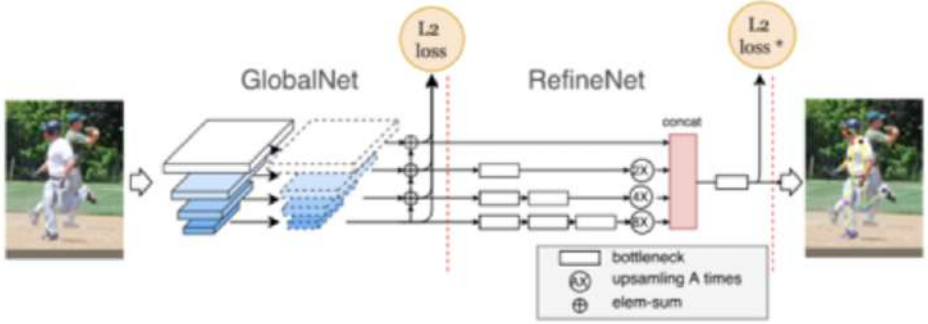
Ryota Suzuki

Cascaded Pyramid Network for Multi-Person Pose Estimation

Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun CVPR2018

概要

複数人ポーズ推定には，キーポイントの半／全遮蔽や，複雑な背景といった要素(hard keypoints)が問題になる．Cascaded Pyramid Networkを提案． hard keypointに対応するためのもの．2つの構造からなる．



- GlobalNet
  ピラミッド構造をしていて，遮蔽などの無いシンプルなキーポイントの検出として作用する．この時点ではhard性にはあまり対応していない．
- RefineNet
  hard keypointを考慮した層． GlobalNetのピラミッドな特徴を拾って，ResNetのBottleneckにかける． ここで，何もしないとシンプルキーポイントだけ見てしまうので，損失関数の計算時，online hard keypoints miningする． テスト時のロスを参考にオンラインでhard keypointを選択，選んだキーポイントのものだけバックプロパゲーションにまわすという作業．

新規性・結果・なぜ通ったか？

- 新規ネットワーク構造の提案
- MS COCO keypoint benchmarkにてSOTA
- 実験を結構頑張っている様子．online hard keypoint miningの有無に関する議論などある．

コメント・リンク集

online hard keypoint miningについて実装可能なレベルでは詳しく書いてなかった．コード読めということか．

- arXiv
- GitHub

[#232]

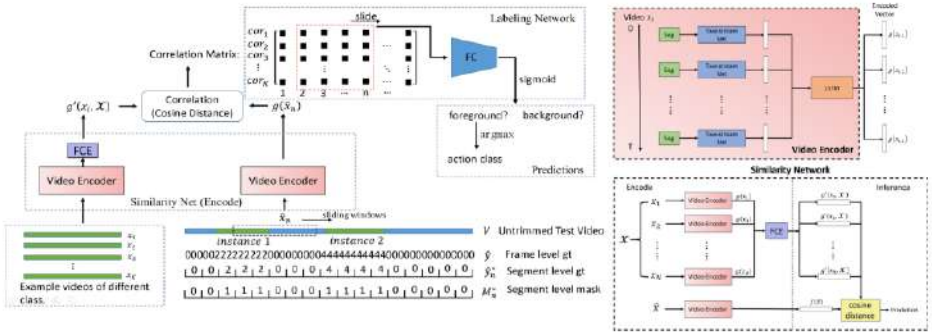
Kensho Hara

One-shot Action Localization by Learning Sequence Matching Network

H. Yang et al., CVPR 2018

概要

ある長い動画中から指定した対象動画と同じActionを探してくるOne-shot Action Localizationの研究． Matching Networkという手法がベースになっていて，それを動画のAction Localizationに応用． 基本的には動画をEncoding (Video Encoder) して， 類似度を計算 (Similarity Network) して，ラベリング (Labeling Network)．長い方の動画はSliding Windowで分割 (Proposals) して，Proposalsと指定動画の間で類似度を計算． Encoderは動画でよくやられるTwo-stream CNNとLSTMを利用． 学習はMeta Learningの形式で定式化され，End-to-Endで学習可能．



新規性・結果・なぜ通ったか？

- Deep時代になってからほとんどやられていなかったOne-shot Action Localization (Action search)
- ProposalsのEncoding，類似度計算，ラベリングと3つすべてが微分可能でEnd-to-Endで学習可能
- 普通のTemporal Action LocalizationのSOTA手法よりもOne-shotの設定では高い性能を実現

コメント・リンク集

- 論文（著者ページ）
- やっている事自体は至って普通のアプローチに感じる
- End-to-End, Meta Learningと今風の形で実現できているのが評価されているのかな

[#233]

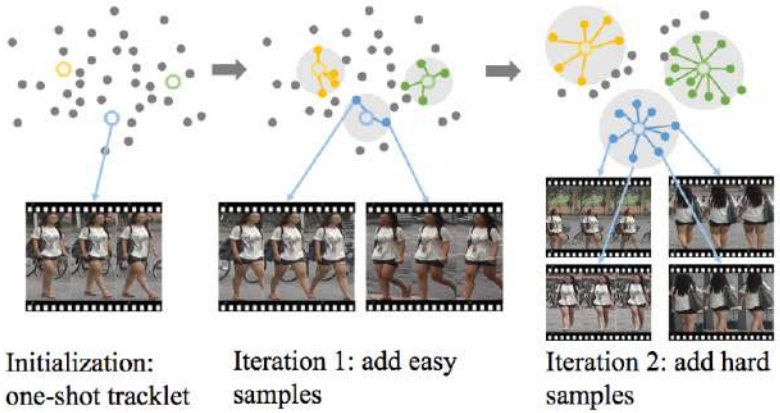
Hirokatsu Kataoka

Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning

Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, Yi Yang CVPR 2018

概要

ワンショット学習（One-shot Learning）により動画像における人物再同定（person re-identification）を実行する論文。ラベルなしのtracklets（人物から抽出した動線）が容易かつ事前に手に入ることから、このtrackletsを徐々に改善しつつ人物同定率を高めるようにCNNを学習していく手法を提案する。本論文での学習では、最初にひとつのラベルを用いて初期化したあと、（1）信頼度の高い少量のサンプル（簡単なサンプル）に対して擬似ラベルを付与、（2）擬似ラベルを含めたラベルを元にカテゴリを更新してより難しいサンプルも取り込む、を繰り返して学習を行う。実験的に擬似ラベルを選択する方法についても議論している。



新規性・結果・なぜ通ったか？

正解ラベルが付与されたある画像一枚を準備するだけで擬似ラベルを推定して徐々に学習を進めていくワンショット学習を提案した。人物再同定の問題においては有効な解決策であることを示したことがCVPRに採択された基準である。ワンショット学習によりrank-1の精度が21.46@MARS dataset、16.53@DukeMTMC-VideoReID datasetであり、コードも公開されている。

コメント・リンク集

ワンショットのラベルと信頼できる擬似ラベルから徐々に概念を獲得するのはうまいやり方。あらゆる枠組みで用いることができそう。

- 論文
- Project
- GitHub
- 著者

[#234]

Hirokatsu Kataoka

PoseTrack: A Benchmark for Human Pose Estimation and Tracking

Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, Bernt Schiele CVPR 2018

概要

動画シーケンスにおいて2D姿勢推定のベンチマークを提供する。本論文で提案するベンチマークでは特に、人物の重なりを含む混雑シーン、密なアノテーションを提供する。さらに右の画像で示すようにドメイン依存していない多様な（diverse）シーンを捉えつつ姿勢アノテーション数でも有数、1画像に対する複数人物/ビデオに対するラベルづけにも対応している。トータルでは23,000画像に対して153,615人の姿勢アノテーションを行なった。チャレンジとしては単一フレームに対する姿勢推定（single-frame pose estimation）、ビデオに対する姿勢推定（pose estimation in videos）、姿勢トラッキング（pose tracking）を提供し、評価用サーバも提供する。同DBに対するベンチマーキングではOpenPoseにも導入されているPAFを改良したML-LAB（引用52）がトップ（70.3@mAP）、Mask R-CNNをベースにしたProTracker（引用11）は64.1@mAPであった。

Dataset	# Poses	Multi-person	Video-labeled poses	Data type
LSP [25]	2,000			sports (8 act.)
LSP Extended [26]	10,000			sports (11 act.)
MPII Single Person [1]	26,429			diverse (491 act.)
FLIC [38]	5,003			feature movies
FashionPose [9]	7,305			fashion blogs
We are family [10]	3,131	✓		group photos
MPII Multi-Person [1]	14,993	✓		diverse (491 act.)
MS COCO Keypoints [28]	105,698	✓		diverse
Penn Action [51]	159,633		✓	sports (15 act.)
JHMDB [23]	31,838		✓	diverse (21 act.)
YouTube Pose [6]	5,000		✓	diverse
Video Pose 2.0 [39]	1,286		✓	TV series
Multi-Person PoseTrack [22]	16,219	✓	✓	diverse
Proposed	**153,615**	✓	✓	**diverse**

新規性・結果・なぜ通ったか？

大規模かつ静止画ではなく動画に対する人物姿勢データセットを構築し、さらには評価サーバを提供、さらに最先端手法に関するベンチマーキングを行なっていることが新規性およびCVPRに通った理由であると考える。

コメント・リンク集

データセットの比較図に多様なドメインから収集（diverse）と書かれているが、これらをすべて統合すると相当な量のデータになるのでは？（だれかやってそう）もしくはドメインを合わせれば学習の効果がありそう。

- 論文
- Project
- 著者

[#235]

Hirokatsu Kataoka

Camera Style Adaptation for Person Re-identification

Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, Yi Yang CVPR 2018

概要

Person Re-ID（人物再同定）は異なるカメラ間で同一人物を対応づける問題設定であり、画像の質や形式が異なるため非常に困難である。本論文ではカメラ間のスタイル変換を行うことでカメラに依存せず安定して認識できる特徴抽出（camera-invariant descriptor subspace）を行い、人物再同定の問題を高度に解決することを目的とする。この問題に対してCycleGANを適用することでカメラ間の特徴変換を捉えた上で、データ拡張を行う。存在するノイズへの対策として有効と思われる正則化:Label Smooth Regularization (LSR)を適用する。LSRを使用する場合では学習データに対するオーバーフィッティングが見られず、有効な手法であることが判明した。

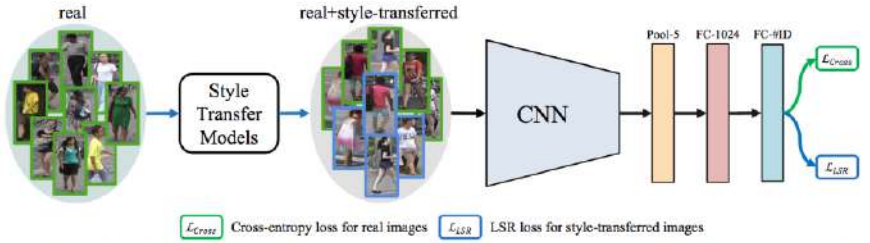


Figure 3. The pipeline of our method. The camera-aware style transfer models are learned from the real training data between different cameras. For each real image, we can utilize the trained transfer model to generate images which fit the style of target cameras. Subsequently, real images (green boxes) and style-transferred images (blue boxes) are combined to train the re-ID CNN. The cross-entropy loss and the label smooth regularization (LSR) loss are applied to real images and style-transferred images, respectively.

新規性・結果・なぜ通ったか？

CycleGANによるカメラ間のスタイル変換を実現してデータ拡張、LSRによりノイズへの対応を行いオーバーフィッティングを回避していることが新規性である。また、人物再同定においてその高い精度（Market-1501のrank-1にて89.49%、DukeMTMC-reIDのrank-1にて78.32%）を実現している。さらに、LSRを用いることでベースラインからの精度向上が見られる。

コメント・リンク集

CycleGANが学習データを増やすという意味でも取り上げられている。例えば東大井上氏の論文もCycleGANを用いてデータ拡張を行い、スタイルの異なる画像に変換することでデータアノテーションの労力を削減している。

- 論文
- Project
- 著者
- CycleGAN
- 東大井上氏論文

[#236]

Hirokatsu Kataoka

Dense 3D Regression for Hand Pose Estimation

Chengde Wan, Thomas Probst, Luc Van Gool, Angela Yao CVPR 2018

概要

単眼距離画像から簡易的かつ効果的に3次元手部姿勢推定を実施する技術について提案する。従来の3D手部姿勢回帰の手法と比較して、本論文ではピクセルごとの（pixel-wise）解析を可能とする。手法としては2D/3Dの関節点を返却するカスケード型の多タスクネットワーク（multi-task network cascades）を提案し、End-to-Endでの学習を行う。その後MeanShiftによりピクセルごとの姿勢位置を推定する。

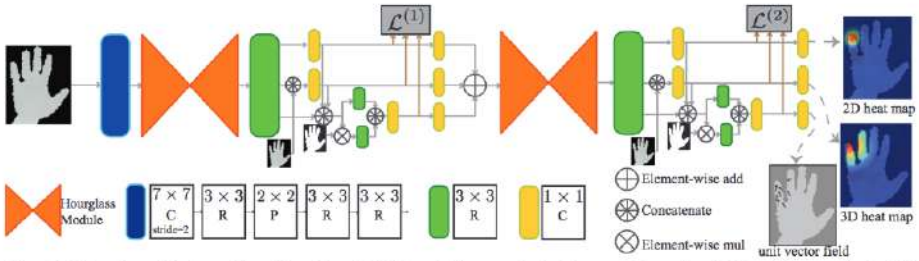


Figure 1. **Network architecture.** The abbreviations C, P, R stands for convolution layer, pooling and residual module respectively. We choose 128*128 as size of input depth map and 32*32 as the input and output resolution of hourglass module[23] with 128 feature channels in each layer. In this paper, we use 2 stacks due to real-time performance constraint. The network estimate 2D,3D heat maps and unit vector field for each joint, we only show the pinky tip point here. Figure is best viewed in colour.

新規性・結果・なぜ通ったか？

従来のほとんどの手法では関節レベルの手部姿勢推定であったのに対して、本論文で提供する技術はピクセルベースの3D手部姿勢推定であることが新規性である。ピクセルごとの回帰はノンパラメトリックな手法を構築した。MSRA/NYU hand datasetにてすべての従来手法よりも高い精度で手部姿勢推定を実行した。また、ICVL hand datasetでは（頭打ちになっていると思われる）論文5には及ばなかったが、接近した精度を叩き出すことに成功した。

コメント・リンク集

HandTrackingも激戦であるが、本論文ではピクセルベースの回帰とState-of-the-art（最高精度）という強みを活かして論文を通している。

- 論文
- GitHub

[#237]

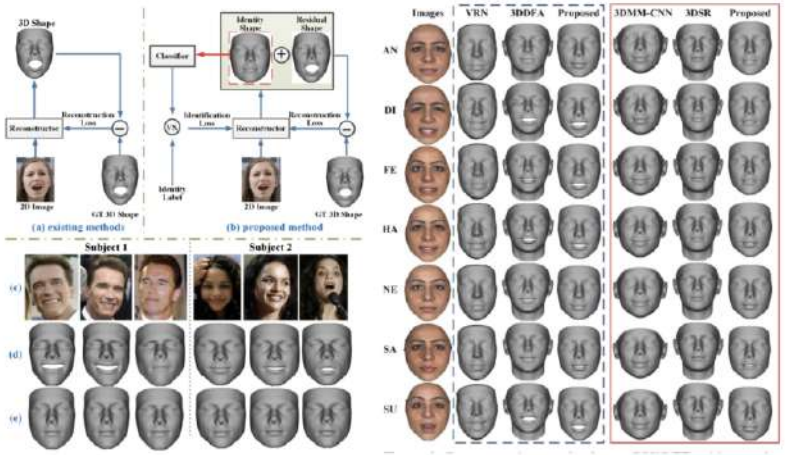
Kazuki Inoue

Disentangling Features in 3D Face Shapes for Joint Face Reconstruction and Recognition

Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, Xiaoming Liu CVPR 2018 Poster

概要

顔画像からshapeの三次元復元を行う際に、画像から個人性(顔の形など)を反映した3Dモデルと、個人性以外(表情など)を反映した3Dモデルをencoderで別々に生成しdecoderで三次元復元を行う手法を提案。 生成された顔のshapeは三次元復元におけるstate-of-the-artよりも高い精度を達成し、 また生成されたshapeによる顔認証においても多くの既存手法より高い精度を達成した。



新規性・結果・なぜ通ったか？

- 従来の三次元復元の手法では顔のディティールは再現するものの、アラインメントなどの個人性の再現が完全ではなかった。提案手法では個人性を反映したモデルとそうでないモデルを分離して学習させることで、この問題を解決した。
- 様々なデータセットにおいて、生成された顔の3D shapeはstate-of-the-artに比べて最も低いaccuracyを達成。
- 生成された3D shapeにおけるランドマークなどのaccuracyにおいてももっとも低い値を獲得。
- 生成された3D shapeによる個人認証においても、多くの既存手法よりも高い精度となった。

コメント・リンク集

- disentangleのファクターとして個人性を選んだのはあくまで人間であって、今後の発展ではもっと優秀なファクターを深層学習が導き出してくれるかもしれない。
- 論文

[#238]

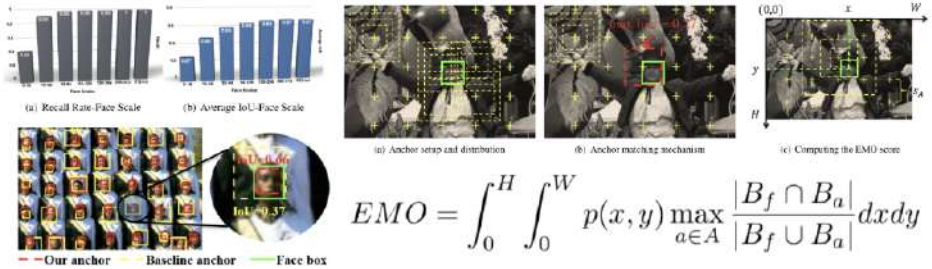
Kazuki Inoue

Seeing Small Faces from Robust Anchor's Perspective

Chenchen Zhu, Ran Tao, Khoa Luu, Marios Savvides CVPR 2018 Poster

概要

アンカーベースで画像中の小さな顔に対する検出精度を向上させる手法を提案。アンカーベースの手法では画像中に等間隔で並べられた点(アンカー)を中心とした矩形によって物体を検出する。 アンカーによる検出精度を評価する数値としてExpected Max Overlapping(EMO) scoreを提案し、 EMOを深層学習に学習させることで、小さな顔(16X16)に対する検出精度を向上した。



$$EMO = \int_0^H \int_0^W p(x,y) \max_{a \in A} \frac{|B_f \cap B_a|}{|B_f \cup B_a|} dxdy$$

新規性・結果・なぜ通ったか？

- 従来のアンカーベースの手法ではIoUを学習させていたため、解像度が16x16などの小物体に対する学習が困難であったが、EOM scoreを学習させることで小物体の検出精度が大きく向上。
- 従来のアンカーベースの手法よりも検出精度が向上、特に小さな顔に対する検出精度が大きく向上したが、実行時におけるスピードは従来手法と同程度。

コメント・リンク集

- 論文

[#239]

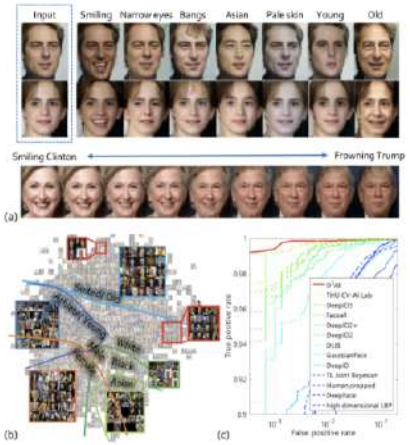
Kazuki Inoue

Exploring Disentangled Feature Representation Beyond Face Identification

Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, Xiaogang Wang CVPR 2018 Poster

概要

顔に関するタスクに汎用的な特徴量を得ることができるDistilling and Dispelling Autoencoder(D2AE)を提案。Encoderによって顔から個人性を表現する特徴量(性別など)と個人性を排除した特徴量(表情など)を抽出する。 取得された特徴量により、個人識別、アトリビュートの識別、顔のアトリビュート編集、顔の生成を行うことができる。



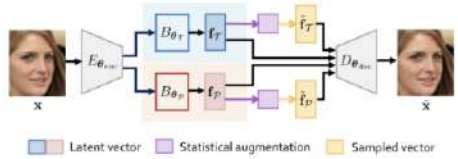
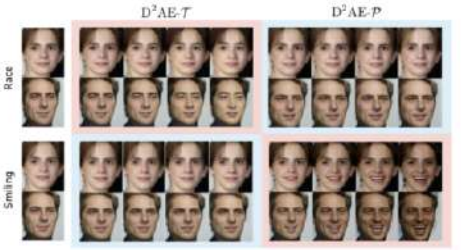


Figure 2. The Distilling and Dispelling Autoencoder model.



新規性・結果・なぜ通ったか？

- Encoderによって顔から個人性を表現する特徴量と個人性を排除した特徴量を抽出することで、これらの特徴量により様々な顔のタスクを行うことが可能となった。
- LFWデータセットにおける個人識別でaccuracyが約99.0%、TPRが約98.0%であり、既存手法と同等の精度を達成。
- LFWA、CelebAデータセットにおける顔のアトリビュート認識は83.16%となり、アトリビュートを学習していないにも関わらず、アトリビュートを学習した既存手法と同等の精度を達成した。
- 顔のアトリビュートの編集、アトリビュートを保ったアイデンティティーの転写といった編集が可能。

コメント・リンク集

- このネットワークを用いて他の物質の個人性を抽出して何が出てくるのか興味がある。例えば顔の代わりに魚を学習させて、鯛ごとの個人性、マグロごとの個人性を抜き出してみるなど。
- 論文

[#240]

Ryota Suzuki

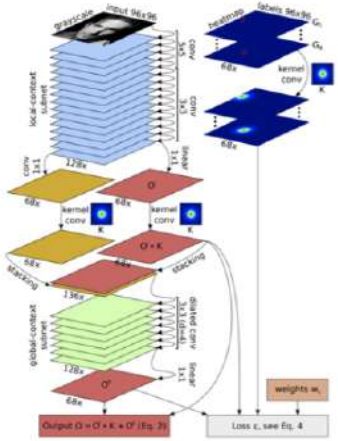
Robust Facial Landmark Detection via a Fully-Convolutional Local-Global Context Network

D. Merget, M. Rock and R. Gerhard CVPR2018

概要

FCNの中にKernel convolutionを暗黙的に入れ込み，大域的特徴情報を残すというアイデアを提案．Conv層で局所特徴を取り，KernelConvでそれをブラーにかけ，DilatedConv層で大局的特徴をリファインするという構造．

特に解像度に独立・きっちりROIがとれない・要複数検出対応・要遮蔽対応な顔ランドマーク検出タスクに有効．KernelConvによって勾配平滑化と過学習抑制が働き収束しやすくなる． アウトライア弾きのために，事前処理ステップにおいて，ネットワーク出力をシンプルなPCAベース2D形状モデルにフィットしておく．



新規性・結果・なぜ通ったか？

- 従来は階層構造やプーリング，統計モデルへのフィッティングで対応していたところを，FCNに直に大域的特徴を入れ込むようにした．
- 構造単純化により，学習パラメータが少なくなる．
- 顔ランドマーク検出に適用してみて，いくつかのSOTAな手法より良い性能を出した．

コメント・リンク集

- 論文

[#241]

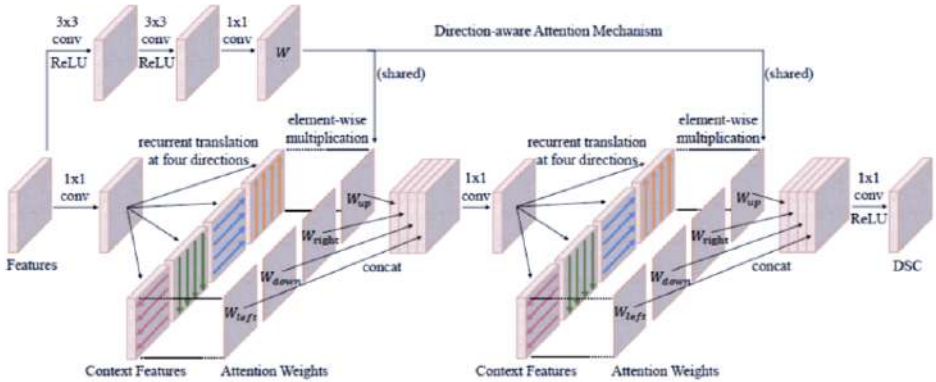
KotaYoshida

Direction-aware Spatial Context Features for Shadow Detection

X.Hu, L.Zhu, C.W.Fu, J.Qin, and P.A.Heng CVPR2018 arXiv:1712.04142

概要

影の周りには様々な背景があり，セマンティクスを理解しなければならないため，影の検出は基本的のようで困難である．それに対して，方向認識の方法で画像のコンテキストを解析することで影検出手法を提案する．空間のRNN内のコンテキスト特徴が密集している箇所にアテンションを導入することで方向認識の手法を定式化する．97%の検出精度と38%のバランスエラー率の低減を実現．



新規性・結果・なぜ通ったか？

- 空間的なRNNに対してアテンション機構を設計しdirection-aware spatial context (DSC)モジュールを構築することで方向認識の方法で空間的なコンテキストを学習．
- 重み付き交差エントロピー損失が影と影でない領域における検出精度のバランスが取れるように設計．

コメント・リンク集

影の検出だけでなく，顕著性検出およびセマンティックセグメンテーションなどの他のアプリケーションで使用する事もできそう．

- Paper

[#242]

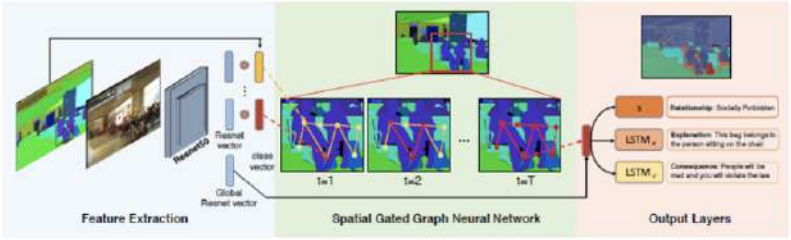
Kazuho Kito

Learning to Act Properly: Predicting and Explaining Affordances from Images

Ching-Yao Chuang, Jiaman Li, Antonio Torralba and Sanja Fidler CVPR2018

概要

現実の多様な場面での環境の物体に対するアフォーダンスの推定する研究。ADE20kを基にしたADE-Affordanceというデータセットの提案。このデータセットはリビングなどの屋内から、道路や動物園などの屋外まで幅広いタイプの画像とそのannotationで構成。また、画像中の物体に対してアフォーダンスの推理を行うための，画像からcontextual informationを伝えるGraph Neural Networksの提案。



新規性・結果・なぜ通ったか？

・ある場面の状況下での適切でない行動の理由について身体的や社会的な観点から説明・画像上のある物体に対してだけでなくその場面を全体としてとらえてアフォーダンスの推論を行っている． ・物体間の依存関係をモデル化することでアフォーダンスとその説明を生成

コメント・リンク集

- 論文
- Project Page

[#243]

KotaYoshida

Discriminability objective for training descriptive captions

R.Luo, B.Price, Scott Cohen and G.Shakhnarovich CVPR2018 arXiv:1803.04376

概要

現在のキャプショニング方法は，2つの異なる画像であるにも関わらず，同じキャプションを生成してしまうなどの弁別性にかけている．それに対して，学習の際に画像とキャプションの一致度を直接関連付けるLossを組み込むことによって他のキャプションよりも弁別性のあるキャプションを生成している．

Human: a large jetliner taking off from an airport runway
ATTN+CIDER: a large airplane is flying in the sky
Ours: a large airplane taking off from runway

Human: a jet airplane flying above the clouds in the distance
ATTN+CIDER: a large airplane is flying in the sky
Ours: a plane flying in the sky with a cloudy sky

新規性・結果・なぜ通ったか？

機械翻訳の評価指標であるBLEU，METEOR，ROUGE，CIDErやSPICEにおいても既存のキャプショニング手法よりも高いスコアを示している．

コメント・リンク集

これにより，同じような画像に対するバリエーションが増え，ユニークなイメージキャプショニングの幅が広がった!!

- Paper

[#244]

Kazuki Inoue

A Face-to-Face Neural Conversation Model

Hang Chu, Daiqing Li, Sanja Fidler CVPR 2018 Poster

概要

入力された会話文に対して、その返答と適切な顔のジェスチャーを生成する手法。映画データセットを元にトレーニングデータセットを構築。 RNNに対してディスクリミネータの出力を報酬とした強化学習を行った。

source text	source face sequence	true target text	text only [12, 27]	text+face
we went to the hickory stick, we had a drink, two drinks. she doesn't know where he is.		and then? and then i went home alone. i don't know where he is.	we drank a bottle of champagne. i'm sorry.	and then i went to bed. i don't know where she is.
and he sleeps only one hour a night. a night that marked the opening of a new chapter. i hope you're not a hothead like sonny.		he's a great man. in world history. he's a good kid.	he sleeps in the same bed. for the future. he's got a lot of something.	he's a good man. in the history of the world. he's a good kid.
i guess they was worried they wouldn't find a vein in my arm. oh, he's so cute.		what's that number? oh, my god.	what's that? oh, my god	i don't think so. he's so cute.
can you hear me? i'm still here. scott. stop. so i don't really remember, yeah. i can't feel my legs.		i'm here. scott. stop. yeah, right. stupid. i can't feel my legs.	i'm sorry. yeah, yeah, yeah. and i can't breathe.	what the f*** are you doing here? well, you know what? i'm sorry. it's too much.

新規性・結果・なぜ通ったか？

- 入力は会話文のみ、あるいは動画。動画が入力の場合には同じテキストでも発話者の表情によって出力される返答文が変化する。
- 出力が会話文だけの場合よりも、同時に顔のジェスチャを生成した方が生成された会話文がよりGTの会話文に近くなったことを主張。
- データセットは250種類の映画データセットMovieQAにおいて単一人物が写っているシーンにおいて顔向、ジェスチャカテゴリ、タイムスタンプを取得することで構築した。
- 生成された返答文の妥当性を評価するためにamazon mechanical turkを実施。GANを導入したことで返答文の多様性、妥当性がstate-of-the-artの手法に勝った。
- このモデルで学習したボットとリアルタイムで会話することも可能

コメント・リンク集

- デモを見るとまだ返答文自体には違和感があるが、顔のジェスチャがつくことで会話している気分になる。ボットのモデルが謎のおじさん。
- 論文
- Project page

[#245]

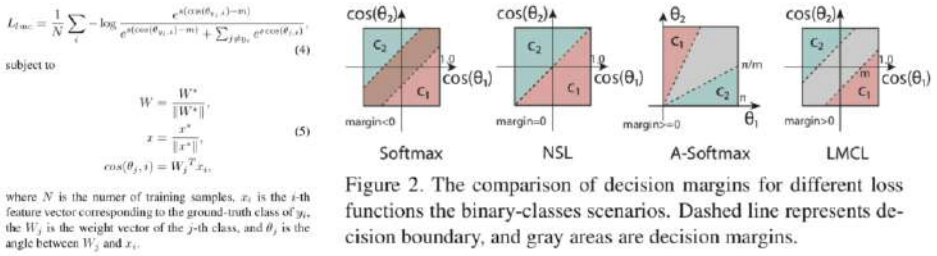
Kazuki Inoue

CosFace: Large Margin Cosine Loss for Deep Face Recognition

Syed Zulqarnain Gilani, Ajamal Mian CVPR 2018 Poster

概要

顔認識のための新たなロス関数としてソフトマックス関数をベースとしたLarge Margin Cosine Loss(LMCL)を提案した研究。LMCLはソフトマックス関数の指数部分を重みベクトルWと特徴量ベクトルxの内積においてWとxのノルムを1とし、定数mを引いた関数。 認識タスクでは異なるクラスタ間の距離を遠く、同じクラスタ間の距離を近くする、という基本的な考えがある。 LMCLはこの考えを元に上記のようにL2正則化を施すことで、Wとxのノルムに左右されることなくWとxの角度空間においてクラスタの分離を行う。



where N is the numer of training samples, x_i is the i-th feature vector corresponding to the ground-truth class of y_i, the W_j is the weight vector of the j-th class, and θ_j is the angle between W_j and x_i.

Figure 2. The comparison of decision margins for different loss functions the binary-classes scenarios. Dashed line represents decision boundary, and gray areas are decision margins.

新規性・結果・なぜ通ったか？

- ソフトマックス関数において重みベクトルの大きさ、入力特徴量のノルムを除外することで、cosの影響を最大限に大きくしWとxの角度空間におけるマージンの最大化を提案。
- face identification(この人はAさんであるか？)、face verification(この人は女性であるか？)の多くのタスクにおいて,ソフトマックス関数由来のロス関数、state-of-the-artの手法よりも良い精度となった。

コメント・リンク集

- 汎用的な認識タスクに使用できそうだが、顔認識に限定したのはデータセットや既存研究との比較のため？
- 論文

[#246]

Kazuki Inoue

Sparse Photometric 3D Face Reconstruction Guided by Morphable Models

Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Cen Wang, jingyi Yu CVPR 2018 Poster

概要

異なる位置の点光源１個によって照らされた５枚の正面顔画像から高品質な３次元顔形状を最適化によって復元する研究。被写体の正面に5つのLED点光源が配置されいている照明環境で撮影を行う。
入力画像に対して3D morphable modelを適用することで簡易的な3次元顔形状を生成し、法線マップ組み合わせることで点光源の位置をピクセル単位で推定する。 またセマンティックセグメンテーションを行うことで体毛が生えいてる領域とそうでない領域に分割し、体毛が生えている領域にはフィルタ処理を行うことでノイズを除去する。

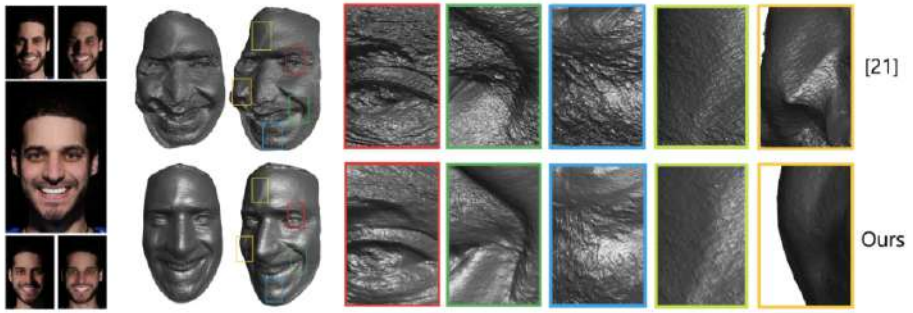


Figure 8: Reconstruction results of [21] (top row) vs. ours (bottom row). [21] causes large deformations and high noise when using a sparse set of images. Our approach is able to faithfully reconstruct face geometry without deformation and at the same time recover fine details.

新規性・結果・なぜ通ったか？

- 顔画像からいきなり光源位置を推定するのではなく、一度morphalbe モデルに生成することで推定精度が大きく向上。
- 3Dスキャンなどの大掛かりな装置を必要としない。
- 顔の小じわ、毛穴、まつ毛なども再現するほど高品質な3次元顔形状を復元。

コメント・リンク集

- 推定された光源位置自体の精度結果を見てみたかった。
- 配置する点光源の位置については特に言及がなかったが、配置による影響の比較結果がみてみたかった。
- 論文

[#247]

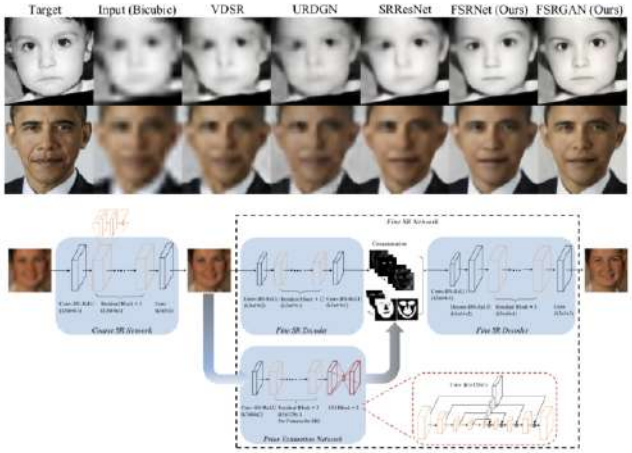
Kazuki Inoue

FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors

Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, Jian Yang CVPR 2018 Poster

概要

顔の超解像度化を学習させる際にランドマーク、パーツの位置推定を同時に行うネットワーク(FSR Net)を提案した研究。同ネットワークをベースにFSR GANも提案。 また生成された高解像度画像に対する評価尺度として生成画像とGTにおけるランドマークのNRMSE、顔パーツに対するセマンティックセグメンテーション画像(parsing)に対するPSNR、SSIM、MSEを提案。 GANベースの手法では高精細な画像が生成されるがPSNR、SSIMが低くなり、MSEをロスとしたネットワークではPSNR、SSIMは高いがボケた画像になってしまう、というジレンマから上記の評価尺度を導入。



新規性・結果・なぜ通ったか？

- 入力画像は16x16の様々な顔むきの画像、出力は128x128に超解像度化された画像。
- state-of-the-artの手法よりもSSIM、PSNRが高く、また新たな評価尺度として提案したランドマーク、face parsingの位置推定も既存手法よりも高い精度となった。
- 新たに提案した評価指標自体の妥当性は、FSR GANとFSR Netを比べた際に、FSR Netの方がボケた画像を生成したにも関わらずSSIM、PSNRが高く、一方でFSR GANの方がランドマーク、face parsingの推定精度が高かったことを根拠に主張している。

コメント・リンク集

- 比較画像において既存手法の画像があまりにもボケているため、既存手法のコントリビューションを確かめるという意味でも調査が必要と感じた。
- 論文
- GitHub

[#248]

Hirokatsu Kataoka

2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning

Diogo C. Luvizon, David Picard, Hedi Tabia CVPR 2018

概要

相互に関連性がある2D/3D姿勢推定+人物行動認識を多タスク学習（Multi-task Learning）により最適化した論文である。それぞれで学習を行ったときよりも高い精度を実現することを明らかにし、複数のデータセットにてState-of-the-artな性能を叩き出した。2Dと3Dの姿勢推定、人物行動の特徴量が相補的に補完し合い特徴学習をより高度にしている？

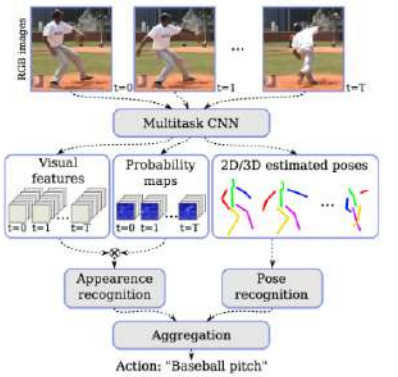


Figure 1. The proposed multitask approach for pose estimation and action recognition. Our method provides 2D/3D pose estimation from single images or frame sequences. Pose and visual information are used to predict actions in a unified framework.

新規性・結果・なぜ通ったか？

姿勢推定（しかも3D姿勢推定も含めて）や人物行動認識を単一の枠組みで解決、さらには多タスク学習により別々に学習したときよりも高い精度でふたつの問題を解決した。さらに複数のベンチマーク（姿勢推定：Human3.6M, MPII/行動認識：PennAction, NTU）にて最高精度も叩き出したことが採択の理由である。

コメント・リンク集

動画シーケンスから姿勢と行動を同時出力する、ありそうでなかった研究である。先にやったもの勝ちだが、高度な最適化を実施し特に最高精度を出すのは難しい。CVPRではState-of-the-artとなるかどうかがひとつの採点基準でもある（が、全てではない）ため、実装力をつけておくに越したことはない。

- 論文
- Youtube

[#249]

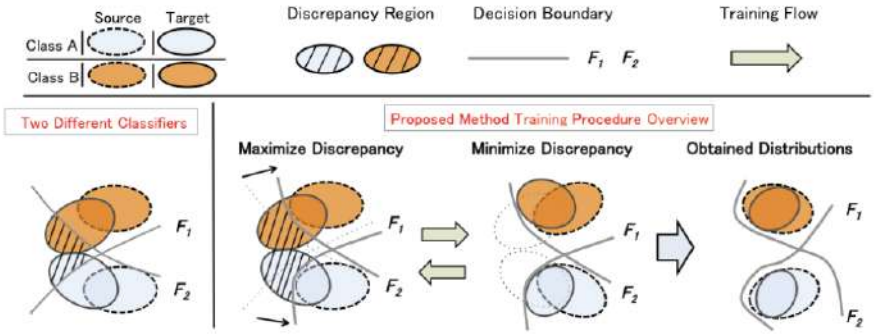
Tomoyuki Suzuki

Maximum Classifier Discrepancy for Unsupervised Domain Adaptation

Kuniaki Saito et al. CVPR 2018

概要

目的のタスクに特化した２つの分離境界を利用したドメイン適応手法。従来の埋め込み空間においてドメイン間の分布を単に近づける方法に対して、あるタスクと解くための分離境界を考慮して適応を行う。この枠組みでの適応はtargetでの損失の上界を下げる埋め込み空間への写像を求める作業と類似している。さまざまなドメイン適応のベンチマークにおいてSoTA。



手法・なぜ通ったか？

Source(S)で学習を行った二つの識別境界を作成する。その識別器がTarget(T)で異なる判断を行ったサンプル(discrepancy)はSの分布とは乖離している領域であると考えられる。以下のような敵対的な適応を行う。(1) TにおけるDiscrepancyが増加するよう識別境界を学習。(2) Discrepancyが減少するように埋め込み空間を学習。(3)Sでの識別は常にうまくいくよう学習。 識別境界を考慮した適応という新規性、理論的な背景、論文の明快さ、精度としての結果が揃っている。

コメント・リンク集

アイデアの面白さと同時に論文が非常にわかりやすかった。識別境界はあくまで埋め込み関数を適化するために得たものなので、この枠組みで得られる最終的なもの以外(得られた埋め込み空間上で新たに学習したもの)でもうまくいくのではないかと感じた。

- 論文

[#250]

Naofumi Akimoto

Generative Non-Rigid Shape Completion with Graph Convolutional Autoencoders

Or Litany, Alex Bronstein, Michael Bronstein, Ameesh Makadia CVPR2018

概要

非剛体的な変形を伴う３Dオブジェクトの形状補完．部分的な形状補完のための学習ベースの手法としてgraph-convolutionを含むVAEを提案した．推論時には，既知の部分的な入力データに合う形状を生成できる変数を潜在空間で探すように最適化する．結果として人体と顔の合成データ，リアルなスキャンデータに対する補完が可能であることを示した．

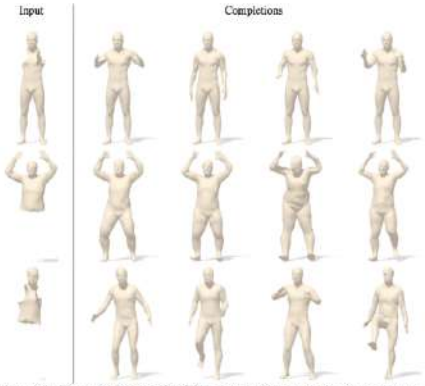


Figure 4. **Completion variability.** When large contiguous regions (e.g. limbs) are missing, the solution to shape completion is not unique. Shown here are different reconstructions with our method obtained using random initializations.

従来手法よりも優れている点

- 訓練中に部分的な形状を見る必要なしに，任意スタイルで一部として切り出されたデータを扱えること
- 人間以外にも，任意の種類の３Dデータに適用できる手法であること
- 形状補完はデータに適合する解が複数ある問題であり，複数のもっともらしい解を生成し，この問題に対応できること

コメント・リンク集

- arXiv

[#251]

Naofumi Akimoto

Eye In-Painting with Exemplar Generative Adversarial Networks

Brian Dolhansky, Cristian Canton Ferrer CVPR2018

概要，新規性

eye-Inpaintingを行う手法．顔のようなそれぞれ固有の特徴を持つ画像においてのInpaintingで，従来のDNNによる手法は新しい顔を生成するなどidentityを保たなかった．exemplar informationを利用するconditional GAN（ExGANs）を提案．参照画像やperceptual codeというidentifying information（exemplar information）をGANの複数の箇所で利用することで，perceptualに優れ，identityを反映した結果を生成することができた．identifying informationをGANの複数の箇所で利用することが新しい．さらに，将来の比較のためにEye-Inpaintingのタスクの新しいベンチマークとデータセットを用意した．

手法概要

cGANの一種．参照画像のIdentityを符号化するネットワークと，Generator，Discriminatorから成る．identifying informationを生成に利用するだけでなく，DiscriminatorやPerceptual lossの算出にも利用している．参照画像をベースにした場合と符号をベースにした場合にアプローチを分けている．

コメント・リンク集

- arXiv

Naofumi Akimoto

Logo Synthesis and Manipulation with Clustered Generative Adversarial Networks

Alexander Sage, Eirikur Agustsson, Radu Timofte, Luc Van Gool CVPR2018

概要

特徴ベクトルのクラスタリングでGANの入力ベクトルを作成する学習方法で，ロゴの生成と操作が可能とした．ロゴのデータは高マルチモーダルのデータであり，従来のSoTAではmode collapseを起こしてしまうが，提案する学習方法では多様なロゴを生成する．
iWGANをCIFER-10で学習するとき，提案する学習方法によって，Inception scoreでSoTA達成．Contribution:

- 600k以上のロゴを収集してデータセットを構築
- マルチモーダルなロゴデータでのGANの学習方法
- 潜在空間の探索によって，インタラクティブなロゴ生成

上段はデータセットから．下段が生成結果．

手法

Clustered GAN Trainingと読んでいる．GANのネットワークは，DCGANとimproved Wasserstein GAN with gradi- ent penalty (iWGAN)を利用．オートエンコーダーの中間特徴ベクトルもしくは，Resnetの特徴ベクトルをクラスタリングして，Generatorの入力ベクトルとする．このクラスタリングでセマンティックに意味のあるクラスタを形成し，GANの学習を向上させることが可能．

コメント・リンク集

- データセット
- ロゴ・ジェネレーター・インターフェースも用意されている．スライダーを動かして，生成結果を操作できる
- arXiv

[#253]

Naofumi Akimoto

Multi-Agent Diverse Generative Adversarial Networks

Arnab Ghosh, Viveka Kulharia, et al. CVPR2018

概要

多様で意味のあるサンプルを生成可能な，複数のGeneratorと１つのDiscriminatorから成るGAN(MAD-GAN)を提案．一つのGeneratorが一つの構成要素を担当する混合モデルとしてはたらく．いくつかの従来のGAN手法と比較実験を行い，MAD-GANは多様なモードを獲得できることを確認．さらに，理論的な分析も行っている．

それぞれの行が異なるGeneratorによって生成した結果．行はそのGeneratorにランダムなノイズzを入力して生成した結果．マルチビューなデータセットから異なるモードを異なるGeneratorが学習していることを確認できる．

手法

- Multi-agent GAN．複数のGeneratorと１つのDiscriminatorで構成．
- Generator同士は，最終層以外は重みを共有している．
- 複数のGeneratorの生成サンプルと真のサンプルをDに入力し，Discriminatorは，FakeとRealの判別だけではなくて，そのFakeの生成サンプルを与えるGeneratorがどれであるかも予測する．これによって，複数のモードがある時，個別のモードに対してそれぞれのGeneratorを振り分けるようにDiscriminatorが学習する．

コメント・リンク集

- image-to-image変換,multi-view生成， face generationなど多数の実験を行っている．
- 展望は，MAD-GANでは複数のGeneratorを使うことになるが，いくつのGeneratorが必要なのかを推定できるようにすること．
- arXiv

[#254]

Naofumi Akimoto

SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis

Wengling Chen, James Hays CVPR2018

概要

スケッチから写真を生成する手法の提案．50のカテゴリの写真を生成することができる．スケッチに対して，自動でデータ拡張をする方法を示し，その拡張方法がタスクに有効であることを示す．さらに追加の目的関数と新しいネットワーク構造も提案．マルチスケールの入力画像を入れることで情報の流れを向上させている．結果はまだphotorealisticとは言えないが，従来手法よりリアルでinception scoreの高い結果を得た．

手法

- データ拡張の方法として，エッジ検出などのいくつかの処理を組み合わせている．
- ネットワーク構造はU-net構造だが，各ブロックで入力画像で条件付けを行うのが特徴．以前の層で抽出された特徴マップと比べ新しい特徴量を入力画像から選択的に抽出するための内部マスクを学習するため，Masked Residual Unitというブロックモジュールを導入した．（DCGAN, CRN, ResNetとの比較がある）

コメント・リンク集

- GeneratorにもDiscriminatorにも途中で画像やラベルの情報をinjectionする方法が増えている印象．
- sketchから似ている写真を検索してくるという方法がこれまでよく研究されていた．今回は，スケッチから新しく写真を生成する（質はまだ低い）
- arXiv

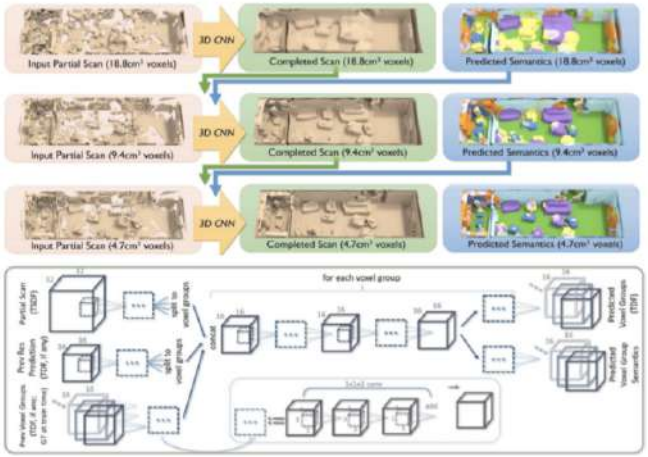
[#255] Yue Qiu

ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans

Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Juergen Sturm, Matthias Nießner CVPR 2018 584

概要

- 部分的なシーンの3Dデータからシーンの幾何及びボクセルごとのセマンティック情報をコンプリートする手法ScanCompleteを提案した．
- 従来，シーンの3次元情報を完全に収集するのが非常に困難，シーンの３次元のデータの膨大さや形状情報のバリエーションの多さは従来のシーン補完に対して困難な問題設定である．そういったため，シーンのコンプリートでは出力の質が低いという問題点がある(contentsとして応用するレベルではない)．こういった困難を解決するため，提案手法は①trainとtestデータの入力解像度を異なる値に設定し， testの場合シーンのサイズの変化を対応できるようにする．②coarse-to-fineなfully convolution 3DCNNを用いて，グローバルなシーンの構造特徴および精密な局所的補間をできるようにする．



新規性・結果・なぜ通ったか？

- 異なる入力シーンのサイズを自由に対応できる（最大70×60×3m くらいまでできる）
- 従来の手法：3D-EPN,SSCNetなどの従来手法と比べ，scene completion, semantic labeling両方精度がSOTA
- 出力結果が3D Contentsとして応用できるレベル

コメント・リンク集

- 論文

[#256]

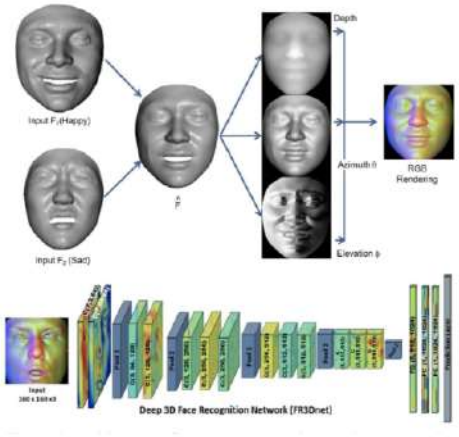
Kazuki Inoue

Learning from Millions of 3D Scans for Large-scale 3D Face Recognition

Syed Zulqarnain Gilani, Ajamal Mian CVPR 2018 Poster

概要

大規模3D顔データセットを構築し、そのデータによってトレーニングされたCNNが高い3D顔認識精度を持つことを示した論文。従来の3D顔データセットはデータ数が少なく、最も多いND-2006でも888アイデンティティー・13540種類のみであったが、本論文で構築されたトレーニング用データセットはおよそ10万アイデンティティー・310万種類。 このトレーニングデータを用いてCNNを学習させることで、認識精度は98.74%となりstate-of-the-artよりも優っていることを確認した。 また既存の3D顔データセットをマージすることで、1853アイデンティティー・31K種類のテスト用3D顔データセットを構築した。



新規性・結果・なぜ通ったか？

- トレーニング用の3D顔データは1000人の3Dスキャンデータに対して、変形に要するエネルギーがもっとまた商用ソフトを使用すること300種類の顔のうち顔の形状・表情が似ている顔を合成して生成。も高くなる顔のペアを合成して生成。また商用ソフトを使用すること300種類の顔のうち顔の形状・表情が似ている顔を合成して生成。 前者は別の顔を識別するため、後者は似た顔を識別する目的で用意されたデータである。 生成された顔に対して水平方向、垂直方向から15度ずつ撮影することで、計100,005アイデンティティー・3,169,275種類の3D顔データを生成。
- 既存の3D顔認識・2D顔認識手法に対してオープン・クローズドテスト両方における精度を比較したところ、提案モデルがもっとも良い精度となった。

コメント・リンク集

- 論文

Naofumi Akimoto

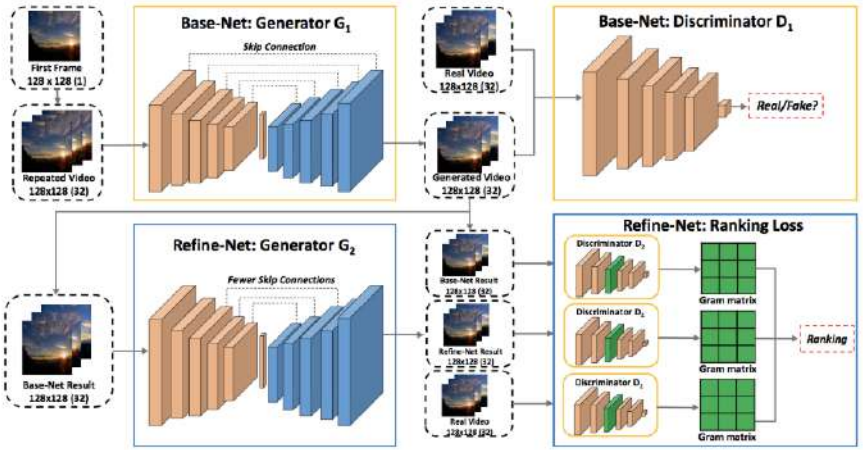
Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, Jiebo Luo CVPR2018

概要

高解像(128x128)のリアルタイムなタイムラプス動画の生成をするGANを提案．最初のフレームを与えると，近未来のフレームを生成する．新規性としては，

- タイムラプスデータセットを作成
- タイムラプス向きの近未来予測ネットワークを提案（Multi-stage Dynamic Generative Adversarial Network (MD-GAN) ）
- モーションのモデリングにGram matrixを導入し，実世界ビデオのモーションを模倣するためのadversarial ranking lossを提案



手法

corse-to-fineの２ステージアプローチのGAN．ステージを分けた狙いとしては，１ステージ目でコンテンツの生成を行い，２ステージ目でモーションのモデリングを行うこと．１ステージ目のU-net風のネットワークでは3D convolutions と deconvolutions を含んでいる．

２ステージ目のDiscriminatorとして，モーションパターンをモデル化するためにGram matrix使って，adversarial ranking lossを算出する．1ステージの出力ビデオ，2ステージ目の出力ビデオ，真のビデオからランキングをとる．

コメント・リンク集

- arXiv

タイムラプス用のGANが初めて提案されたことが評価されたのかなという印象．定量的な評価はメインがPreference Opinion Scoreで，他はMSE, PSNR and SSIM．

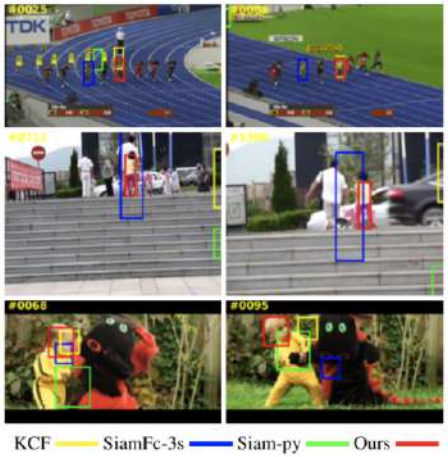
Yoshihiro Fukuhara

Hyperparameter Optimization for Tracking with Continuous Deep Q-Learning

Xingping Dong et al. CVPR 2018

概要

Object Tracking 手法において用いられる複数の Hyperparameter を強化学習によって各シークエンス毎に最適化する手法を提案. Hyperparameter の選択を Action, Tracking の精度の良さを Reward として, Normalized Advantage Functions (NAF) を用いた強化学習を行なっている. また, Heuristic を導入することで, 学習の遅さの問題を緩和した.



新規性・結果・なぜ通ったか？

- Object Tracking における Hyperparameter の最適化問題を強化学習の問題として定式化した.
- 上記の問題を既存の強化学習手法である NAF　(連続な行動が取れるように拡張された Q 学習の手法) を用いて解いた.
- 強化学習を適用した際に, 状態空間の次元の多さなどに由来する学習速度の遅さを huristic を導入することで緩和した.
- OTB-2013 や VOT-2015 などのデータセットを用いて既存研究 (Siam-py等)と比較. 同程度の速度で, 正確性とロバスト性の両方に置いて既存手法を上回った.

コメント・リンク集

- [論文] Hyperparameter Optimization for Tracking with Continuous Deep Q-Learning
- [関連論文 : NAF] Continuous Deep Q-Learning with Model-based Acceleration

[#259]

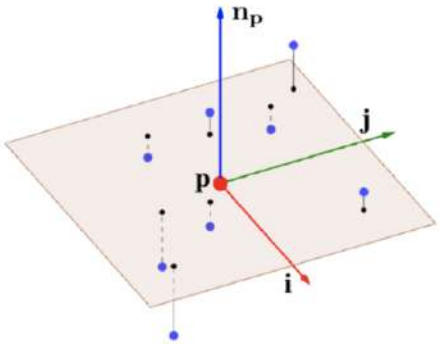
Yoshihiro Fukuhara

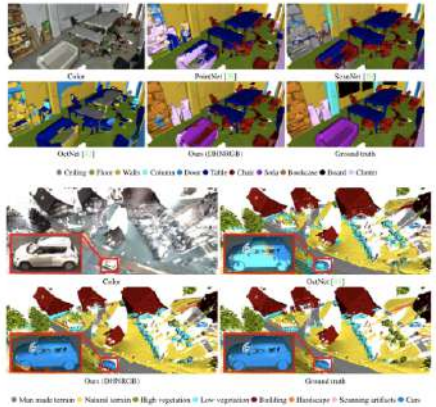
Tangent Convolutions for Dense Prediction in 3D

Maxim Tatarchenko et al. CVPR 2018

概要

３次元データを扱う新しい convolutional の方法 "Tangent Convolution" を提案. 全ての点の近傍点を仮想的な接平面上に射影し, 接平面上で畳み込みを行う. 接平面は法線ベクトルが計算できれば構成する事ができるため, 複数のデータ形式に対して同様に適用が可能. また, 事前計算を行う事によって大規模なデータベースに対しても効率的に計算を行う事が可能となった.





新規性・結果・なぜ通ったか？

- 入力データの形式は法線ベクトルを近似的に求められるもの (point clouds, meshes, dpolygon soup) であればなんでも良い.
- 事前計算を行う事によって大規模なデータ（数百万オーダーの点群）も効率的に扱う事ができる.
- 提案手法の有効性を示すために Tangent Convolution を用いたネットワークを Semantic 3D Scene Segmentation のタスクに置いて既存手法 (PointNet, ScanNet, OctNet) と比較し, 複数の評価尺度に置いて最も良い精度となった.

コメント・リンク集

- [論文]

[#260] Yue Qiu

Im2Pano3D: Extrapolating 360 Structure and Semantics Beyond the Field of View

Shuran Song, Andy Zeng, Angel Chang, Manolis Savva, Silvio Savarese, Thomas Funkhouser CVPR 2018 466

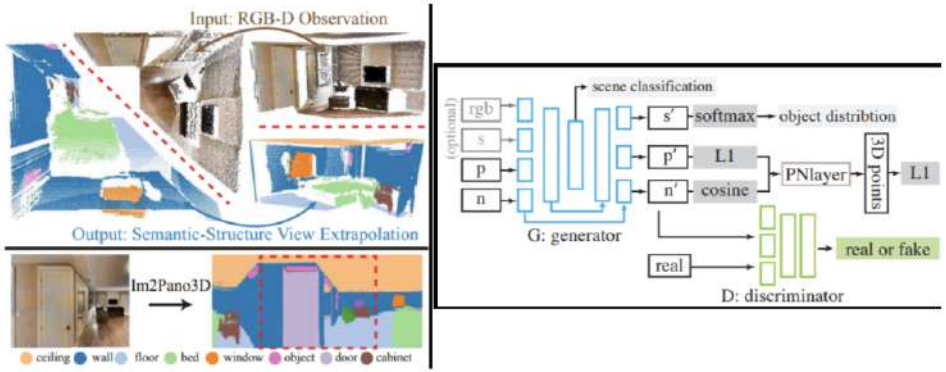
概要

・部分的に観測されたシーン(RGB-D)から，full sceneの構造及びセマンティックラベルを推定する新規な問題設定”semantic-structure view extrapolation”及びフレームワークを提案した．

・従来のview extrapolationは画像のboundryの色情報しか行わず，シーンのセマンティック構造に対してextrapolationを行う研究がない．そこで，この論文で，著者達がsemantic-structure view extrapolationを提案し，50%以下のシーンの観測データから構造及びセマンティックをextrapolation予測する．

・提案フレームワークは：①一枚のマルチチャンネルpanorama画像でシーンの情報(RGB，構造，セマンティック)を表示する；②3次元構造をデプスのような詳細な三次元情報を用いずに，3次元平面方程式で表示する．③マルチロス関数(ピクセルレベル，グローバルコンテキスト)を用いる．

・提案フレームワークの考え方は入力と出力を一枚のマルチチャンネルpanorama画像として表示し，encoder-decoderにより，欠損した入力からfullなpanorama画像を出力する．



新規性・結果・なぜ通ったか？

・CG データセットSUNCG及びリアルシーンデータセットMatterport3Dを用いて従来手法よりシーンの構造及びセマンティックの予測が優位．

・一枚のマルチチャンネルpanorama画像でシーンの情報を表示し，シーンの情報を固定なサイズにできるので，2次元畳み込みを

コメント・リンク集

・マルチチャンネルpanorama画像でシーンの情報を保存するところが賢い

・提案フレームワークは構造的に理解しやすい，実装してみたい

- 論文

[#261]

Tomoyuki Suzuki

Deep Image Prior

Dmitry Ulyanov et al. CVPR 2018

概要

「CNNは理論上任意の関数を近似できるが、その構造自体に汎化性能をあげるようなPriorが含まれている」という考えのもと、ランダム初期化されたCNNを用いて高いレベルの画像復元、ノイズ除去などを行った。また、CNNのPrior をさらに裏付けるものとして、自然画像を復元するより、ノイズ画像を復元する学習の方がiteration数がかかることも示された。

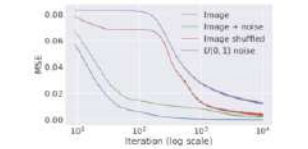


Figure 2: Learning curves for the reconstruction task using: a natural image, the same plus i.i.d. noise, the same randomly scrambled, and white noise. Naturally-looking images result in much faster convergence, whereas noise is rejected.

$$\min_{\theta} \|f_\theta(z) - x_0\|^2$$

Figure 6: **Region inpainting.** Our method is able to successfully inpaint large regions. Despite using no learning, results are comparable to [16] which does. The choice of hyper-parameters is important (for example (d) demonstrates sensitivity to the learning rate), but a good setting works well for all images.

(a) Original image (b) Corrupted image (c) Shepard net works [16] (d) Deep Image Prior

手法・なぜ通ったか？

ノイズ画像 z をencoder-decoderモデルに入力して、生成された画像を欠損画像にMSEで近づけるように学習するだけである。注意点として、完全に学習仕切ってしまうと欠損画像と同じものが出るだけなので、学習をある程度のiterationで止めると、復元されたような画像が得られる。また、CNNのPrior をさらに裏付けるものとして、自然画像を復元するより、ノイズ画像を復元する学習の方がiteration数がかかることも示された。着眼点や面白い実験方法に加え結果も伴っている研究

コメント・リンク集

畳み込み処理×SGDの異常なまでの汎化性能を実験的に裏付けていると思われ非常に面白い。逆にCNNのPriorの苦手なところとして、Adversarial exampleやGANのチェッカーボード現象も関係してそう。畳み込み処理の派生(Deformable convなど)でのpriorの検証も気になる。

- 論文

[#262]

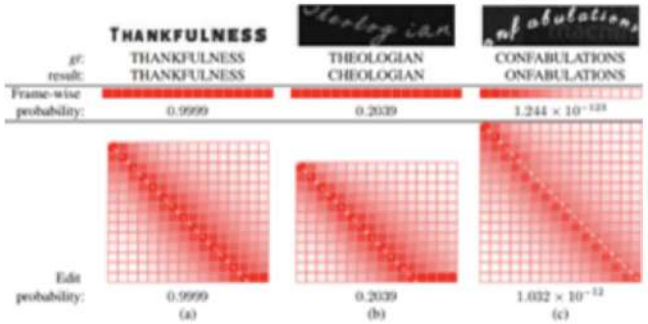
Ryota Suzuki

Edit Probability for Scene Text Recognition

F. Bai, Z. Cheng, Y. Niu, S. Pu and S. Zhou CVPR2018

概要

OCRのstate-of-the-artな手法として，encoder-decoderで文字カテゴリごとのAttentionを取ってからテキスト認識をするvisual attentionベーステキスト認識があるが， ある文字がよく見えなかったり1文字でも複数ピークが出てしまったりする問題はある． GTとの差を取るとして，エンコード後の文字列で比較する編集距離を取ることが考えらえるが， 本稿ではVAで出る尤度分布で比較する，編集確率（Edit Probablity）を提案する． これにより，字抜けや余分な字を拾ってしまうような誤認識に強い文字認識を実現可能．



新規性・結果・なぜ通ったか？

- Attentionベーステキスト認識においてstate-of-the-artな性能．
- まさに正統進化といえる．

コメント・リンク集

正統進化を，他のラボが，1年未満に行ってしまっているあたり，CV分野の流れの早さがうかがえる．

- arXiv
- Visual attention models for scene text recognition（**ICDAR2017**）

[#263] Yue Qiu

iVQA: Inverse Visual Question Answering

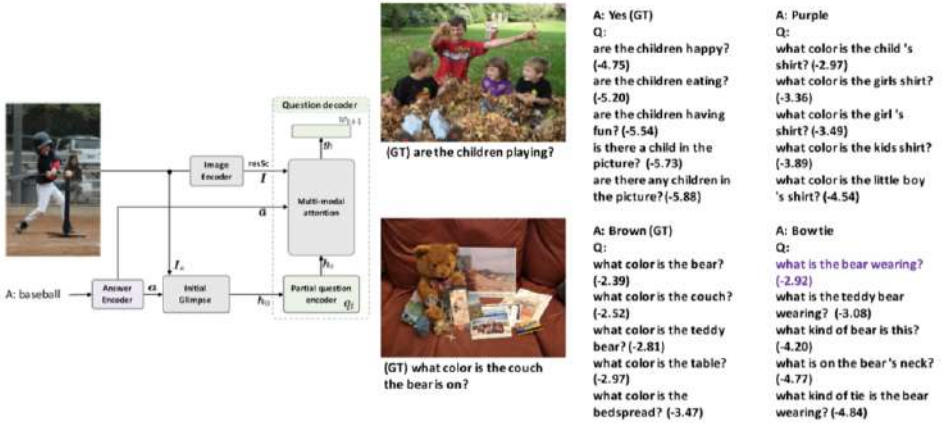
Feng Liu, Tao Xiang, Timothy Hospedales, Wankou Yang, Changyin Sun CVPR 2018 1199

概要

・VQA問題の逆問題iVQA設定及びモデルを提案し (画像及び回答文から，質問文を生成する)，更に iVQAもVQAと同じく“視覚-言語”の理解のベンチマック問題設定になれると指摘した．

・iVQAタスクに用いられるmulti-modal dynamic inferenceなフレームワークを提案した．提案フレームワークは回答文を生成する段階で，“回答文”，“生成した部分的な質問文”によって導かれ動的に画像attentionを調整できる．

・更に，回答文の従来の自然言語的評価に， ランキングベースなiVQAタスクの回答文を評価できる指標を提案した．その指標により，などの面を評価できる．



新規性・結果・なぜ通ったか？

・近年，従来のVQAの成功がデータセットバイアス及び質問文からの情報理解，画像の内容に対する理解がまだVQAにおいて深く利用されていないことが指摘された．そこで，画像と回答文から質問文を予測する問題設定iVQAを提案した， iVQAタスクにおいてはVQAと比べ，①画像内容の理解の要求が高い，②また回答文が常に短いので，質問文と比べよりスパースな情報抽出しかできないため，回答文に頼りすぎることにならない．③モデルの推定及びreasoning能力が更に必要である．

・提案フレームワークの各パーツ(dynamic attention, multi-modal inferenceなど)の有効性に関してAblation　studyを詳しく行った．説得力がある．

コメント・リンク集

・VQAの問題点を深く理解した上での新規問題設定．

・Dynamic attention mapsの可視化分析により問題文を生成する段階で，動的に関連する画像領域にattentionすることを指摘した．

・新奇な考え方・詳しい分析実験・論文の理解しやすさなどが非常に良い

- 論文

[#264]

Ryota Suzuki

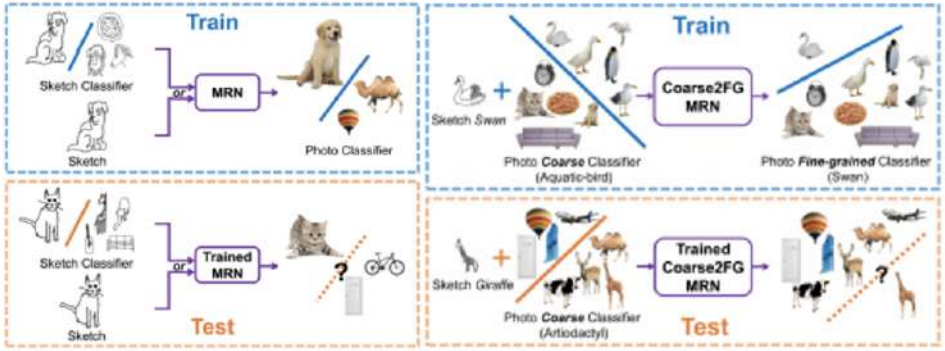
Sketch-a-Classifier: Sketch-based Photo Classifier Generation

C. Hu, D. Li, Y. Song, T. and T.M. Hospedales CVPR2018

概要

手書き画像から，書いたものの判別をする画像分類器を出力するメタ学習の提案．学習していない手書きカテゴリでも，そのカテゴリの画像分類器が出力される．3つの枠組みが作れる． (1)スケッチ画像カテゴリ分類モデルを入力 (2)スケッチ画像を入力 (3)コースなリアル画像分類モデル＋スケッチ画像を入力

枠組みとしては，Model Regression Networkによる．論文では，SVMパラメータの学習を行っている．



新規性・結果・なぜ通ったか？

- 多様性がある．作ったモデルの性質がよく把握されている
- 知識転用の新しい形が見える

コメント・リンク集

- arXiv
- プロジェクトページ

[#265]

Ryota Suzuki

ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing

C. Lin, E. Yumer, O. Wang, E. Shechtman and S. Lucey CVPR2018

概要

画像合成の際に，背景に対して位置やサイズ感などが正しくなるように幾何的変換を求め，修正を加えてくれるGANを提案．たとえば，家具が適切な場所に置かれたり，メガネが適切に掛けられたりする．

構造的には複数のSpatial Transformer Networkをジェネレータとして組み込んでいることが特徴．複数のSTNにおける，反復画像ワーピング（画像変形方法の一つ）と逐次学習を導入している．

新規性・結果・なぜ通ったか？

- 画像変換が得られるので，間接的に高解像度画像に適用可能
- ナイーブな単ジェネレータよりも高性能．
- 大きな差には弱い．奇抜なデザインのものや，大きな移動

コメント・リンク集

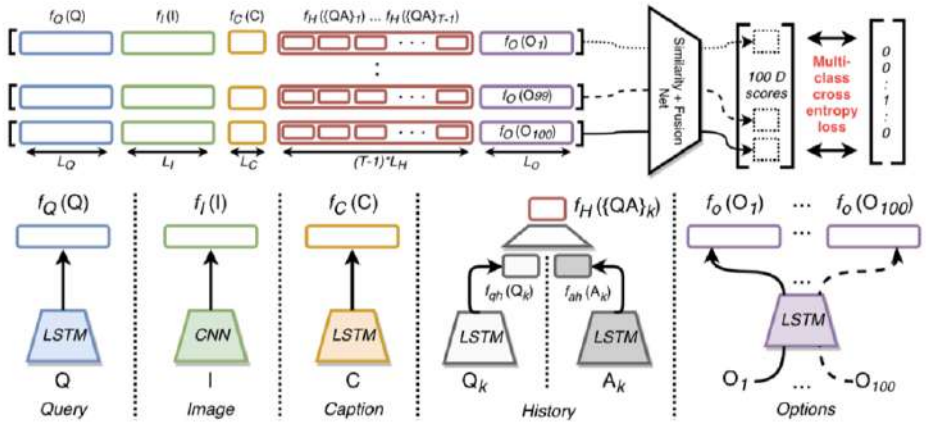
- arXiv
- プロジェクトページ

[#266] Yue Qiu

Two can play this Game: Visual Dialog with Discriminative Visual Question Generation and Visual Question Answering

Unnat Jain, Lana Lazebnik, Alex Schwing CVPR 2018 705

概要



・Visual Dialogタスクに用いられる質問の回答文と質問文を両方予測できるネットワークを提案した．

・提案フレームワークは100個の回答文(質問文)から正解を予測する(discriminative). 提案フレームワークは質問文，画像，キャプション，QA履歴，選択などの情報をsimilarity+Fusionネットにより100次元のベクトルを生成し，正解ラベルとのcross-entropy誤差を求める．

・また，従来Visual Dialogの質問文を評価する指標がない，著者達が質問文を評価できる“VisDial-Q evaluation protocol”を提案した．提案protocolは質問文を100個に固定し，予測した質問文がどれくらい通常の人により提出される可能性が高いかにより評価を行っている．

新規性・結果・なぜ通ったか？

・同じネットワークで質問文と回答文を両方予測できる．

・質問文を評価できる指標の提案．

・Discriminative VQAタスクにおいて， VisDial評価指標は従来手法(HRE, MN, HCIAE-D-NP-ATT)より良い性能を達成した．

・VQGタスクにおいて，提案した評価指標“VisDial-Q evaluation protocol”により55.17% recall@5 と 9.32 mean rankを達成した．

コメント・リンク集

- 論文

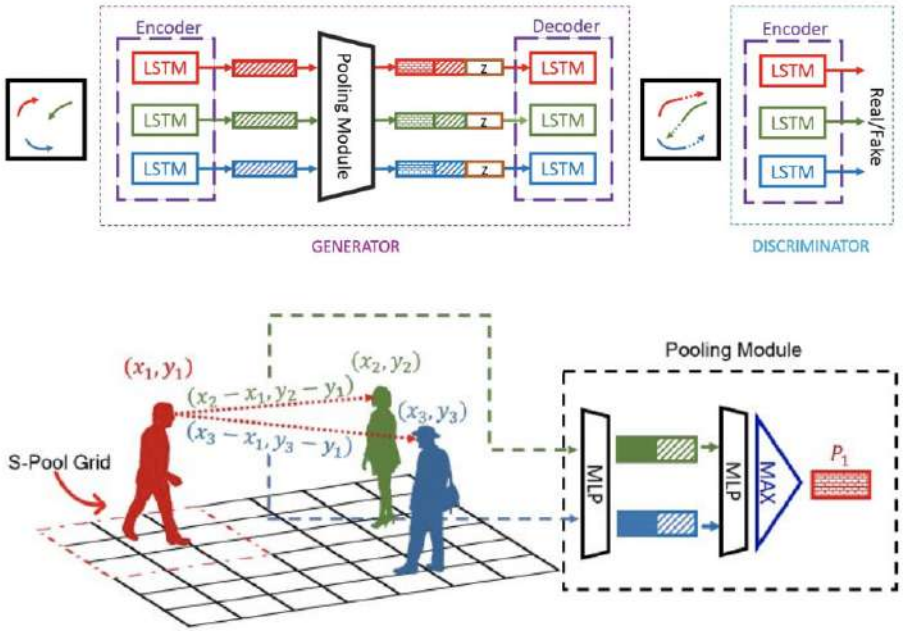
Ryosuke Araki

Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese and Alexandre Alahi CVPR2018 234

概要

人や自律移動プラットフォームが，移動している人を避けるにはいくつかの経路が考えられる．本手法は，人間の経路予測にシーケンス予測とGANを組み合わせたツールを用いて，複数の経路予測を行う．Recurrent sequence-to-sequence modelは，複数の人の間で情報を集約するための新しいプーリング手法を用いて，観測者の行動を予測する．そして，GANを用いてもっともらしい行動をいくつか予測する．予測された経路はDiscriminatorへ入力され，Fake/Real判別をしGANを訓練していく．



新規性・結果・なぜ通ったか？

Generatorでは，複数の人が同時にどう動くか予測するために，Encoderの各LSTMの出力をまとめるプーリングモジュールを導入した．Discriminatorは，経路そのものがFake（人として社会的にあり得ない行動）またはReal（あり得る行動）を判断する．ETHやHOTELなどのデータセットを用いて評価実験を行った．12ステップ後のAverage Displacement Error（全ての時間での真値と予測値の誤差）は0.58（Social LSTM: 0.72），Final Displacement Error（最終目的とでの真値と予測値の誤差）1.18（Social LSTM: 1.54）となった．

コメント・リンク集

GANを使う手法は多く出てきているが，これは面白い応用方法だと思った．Discriminatorをどうやって学習していくかが肝になりそう．

- arXiv

[#268]

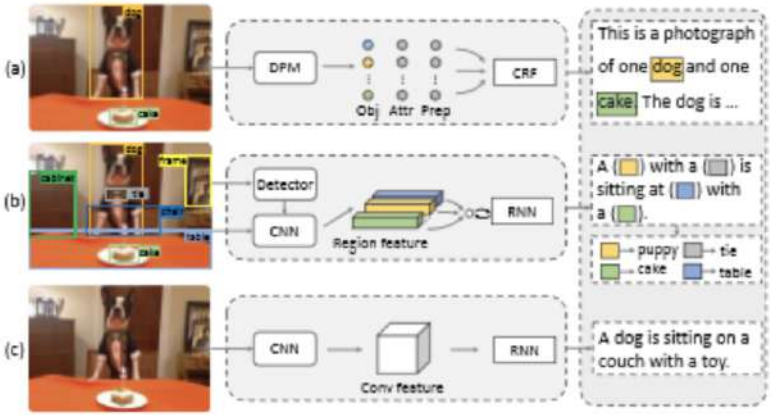
Takumu Ikeya

Neural Baby Talk

Jiasen Lu, et al. CVPR 2018

概要

画像内で検出した物体から文章を生成するイメージキャプショニングタスクを行うための新たなフレームワークの構築を行った．単語が格納されるスロットを文章内に生成し，生成したスロットを満たすように検出した物体を当てはめていくことでキャプションを行う．



新規性・結果・なぜ通ったか？

検出された物体の名称が入るスロットを最初に生成し，生成したスロットを満たしていくことでキャプションを行う手法が新しい．

イメージキャプショニングタスクにおいてFlickr30KとCOCOデータセットでSOTAを達成した．

コメント・リンク集

- 論文
- GITHUB

Naofumi Akimoto

Attentive Generative Adversarial Network for Raindrop Removal from a Single Image

Rui Qian, Robby T. Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu CVPR2018

概要

- 写真から雨粒を除去する手法の提案
- このタスクが難しいのは，
    1. どの領域が，雨粒によって隠されているか不明なこと
    2. 雨粒に隠された背景側の情報がないこと
- GAN，LSTMを利用
- Generatorは，Attentive-Reccurent networkとContextual Autoencoderから構成
- はじめにAttentive-Reccurent networkでattention mapを生成 次にContextual Autoencoderで，mapと入力画像から雨粒除去後の画像を生成　attention mapは，Discriminatorの中間出力とMSE lossを取る際にも利用
- visual attentionという情報によって，
    1. Generatorでは雨粒の領域と，周辺の構造にアテンションをより向けることができる
    2. Discriminatorは復元した領域をより局所的に評価を行える

新規性

- GeneratorとDiscriminatorの両方でvisual attentionを利用するようにしたこと
- 自作の1119枚の雨粒ありと無しのペア画像を用意し学習に利用

コメント・リンク集

- arxiv

Naofumi Akimoto

Deformable GANs for Pose-based Human Image Generation

Aliaksandr Siarohin, Enver Sangineto, Ste ́phane Lathuilie`re, and Nicu Sebe CVPR2018

概要

与えられたポーズ情報を条件として人物画像を生成するタスクを扱う．任意ポーズへの変形タスクで発生する，（服などの）変換前のピクセルと変換後のピクセルの対応が不整列である問題に対応するために，deformable skip connectionを対案する． 従来手法と比べ，条件画像の服の色・テクスチャを保存して別ポーズの画像を生成できている． 人物画像の生成に限らず，キーポイントを与えることのできる不整列のオブジェクトであれば，この手法が適用できると著者らは考えている．

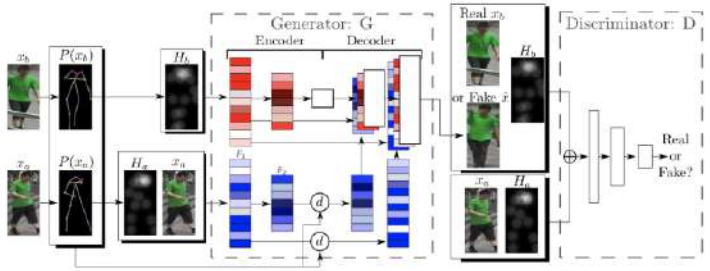


Figure 2: A schematic representation of our network architectures. For the sake of clarity, in this figure we depict $P(\cdot)$ as a skeleton and each tensor H as the average of its component matrices H_j $(1 \leq j \leq k)$. The white rectangles in the decoder represent the feature maps directly obtained using up-convolutional filters applied to the previous-layer maps. The reddish rectangles represent the feature maps "shuttled" by the skip connections from the H_b stream. Finally, blueish rectangles represent the deformed tensors $d(F)$ "shuttled" by the deformable skip connections from the (x_a, H_a) stream.

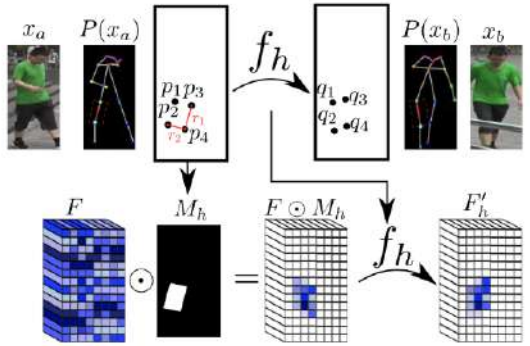


Figure 3: For each specific body part, an affine transformation f_h is computed. This transformation is used to "move" the feature-map content corresponding to that body part.

手法

U-net likeのEncoder-Decoder, GANdeformable skip connectionについて． 変換前後の両方のポーズ情報が既知なので，キーポイント周辺のピクセルが変換前から変換後にどこへ移動するか知ることができる．したがって，キーポイントの座標からアフィン変換を求

リンク集

- arXiv
- プログラム

[#271]

Yue Qiu

VizWiz Grand Challenge: Answering Visual Questions from Blind People

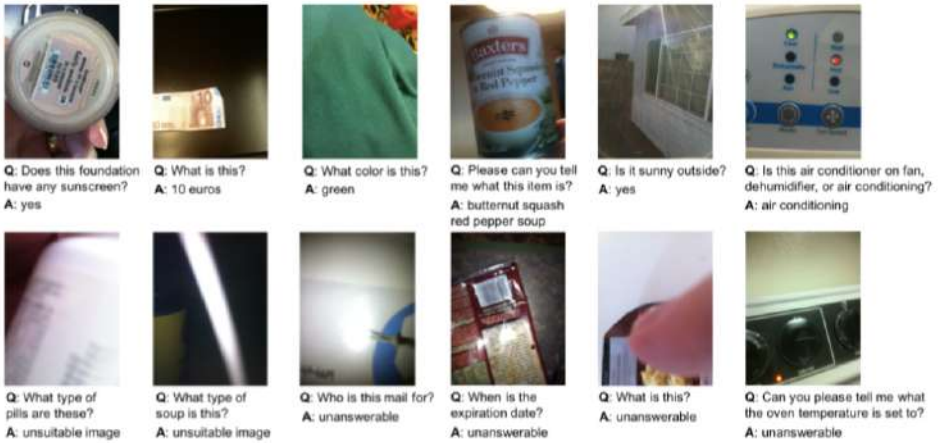
Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo CVPR 2018 491

概要

・盲人に集められたVQAタスクのデータセットVizWiz（画像と音声質問文）を提案した．VizWizが31,000枚の盲人が携帯により撮影し，画像ごとに画像を撮影した盲人が提出した音声質問文一つ付き．質問文ごとに，10個の回答文がアノテーションされている．

・従来のVQAデータセットほぼ人工設定により作成された方が多く，また現実環境の盲人ユーザを対象に“goal oriented”なVQAデータセット未だにない．そこで，盲人がカメラにより周囲環境を撮影し，環境を理解することを目的にして，盲人ユーザにより集められた画像及び質問文のデータセットを構築した．

・ 盲人ユーザにより撮影されたのでVizWizは画像の質が良くなく，又質問文が音声情報なので，はっきり発音が取れない場合などの問題点がある．提案データセットで現状のVQAモデルで検証した結果，性能が従来のデータセットで検証した性能より劣るので，VizWizが将来的の盲人のためのVQA応用に新たな挑戦を提出した．



新規性・結果

・初めての盲人により撮影及び質問したVQAデータセット．

・従来のVQAデータセットと比べ，もっと画像の周りの環境に関する質問文が多い．

・従来のVQAデータセットとの質問文の詳細的な特徴比べも行っている．

リンク集

・盲人のためのVQAシステム構築に有力なデータセット．

- 論文

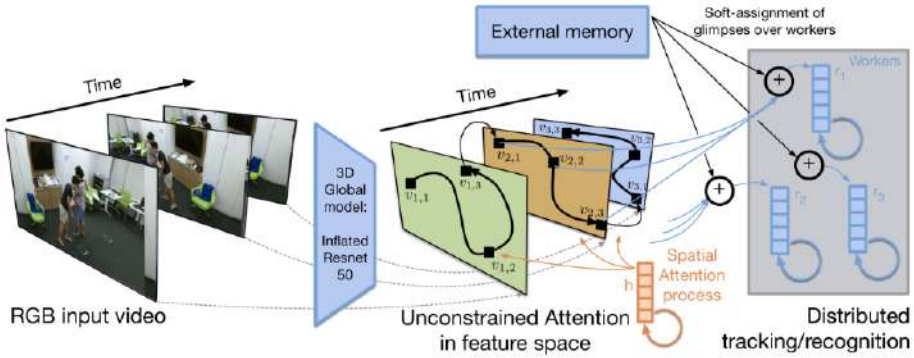
Kensho Hara

Glimpse Clouds: Human Activity Recognition from Unstructured Feature Points

F. Baradel et al., CVPR 2018

概要

RNNベースの行動認識を提案．学習はRGB-Dを使うが，テスト時にはRGBのみを使うという設定．テスト時にRGB-Dが使えてPose情報が使えればそれを使えばいいが，それが使えないときもあるからそれに変わる手法を提案するという主張．Poseでの間接位置に代わって，Attentionベースでフレーム中から重要な局所要素 (Glimpse) を抽出＆トラッキング．Glimpseの集合に基いて行動を認識するというフレームワーク．Glimpseの抽出やトラッキングはそれぞれRNNベースで行う手法になっている．



新規性・結果・なぜ通ったか？

- 姿勢の代わりに別の局所要素を使うフレームワークを提案
- Attention, External Memoryといった流行り?の要素が詰め込んである
- RGB-D行動認識データセットにおいてRGBのみの利用でSOTAを達成

コメント・リンク集

- 論文（著者版）
- 論文 (Long-ver., arXiv)
- 動画 (YouTube)
- 姿勢ベースの行動認識を姿勢を使わずにやるような話に近い印象

[#273]

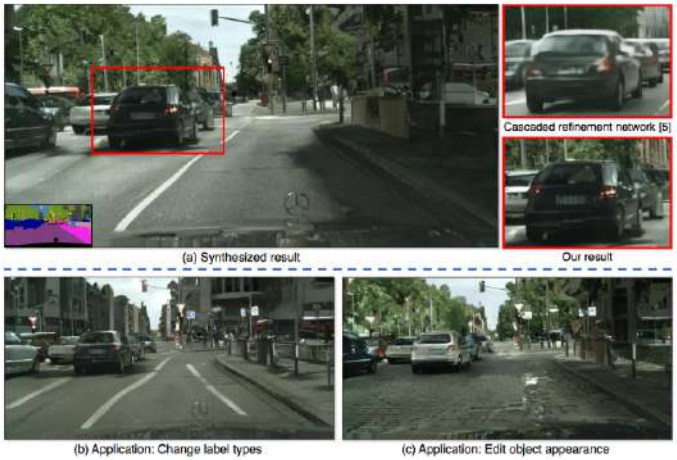
Hirokatsu Kataoka

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, Bryan Catanzaro CVPR 2018 (oral)

概要

GANの枠組みにてセマンティックラベルからの高精細画像（HD-Image）生成に関する研究。意味ラベルからリアルな画像を生成するのみならず、インタラクティブな操作で画像生成をコントロールすることも可能。Residual blocksにより構成されるエンコーダ/デコーダ構造を（入力をスケールが異なる画像として）入れ子構造にしデコーダ直前の中間層で統合して画像生成を実行する。さらに、ラベルのみならずインスタンスレベルの特徴量を用いることで写実性が向上したと主張（論文中図4では物体境界面あたりに出ているボケが綺麗になっている）。



新規性・結果・なぜ通ったか？

従来法より、見た目の画像生成が明らかに良くなり、高画質の画像を対象にしても画像生成ができるようになった。従来手法（pix2pix（論文中文献21）, CRN（論文中文献5））さらに、インタラクティブな操作により生成画像を所望の結果に近づけることができる。動画像を見れば従来手法よりも鮮明になっていることは明らかであり、アーキテクチャや生成に関する知見も得ている。CVPRでoralになるための準備やプレゼンが論文中にも書かれていると感じた。やはりNVIDIAはずるいと言われるくらいの計算機環境が揃っているのではないか。

コメント・リンク集

これはもう、学習画像として使えるのでは？（すでにだれか使って精度検証しているのでは？）

- 論文
- Project
- GitHub
- arXivTimes
- YouTube

[#274]

Ryota Suzuki

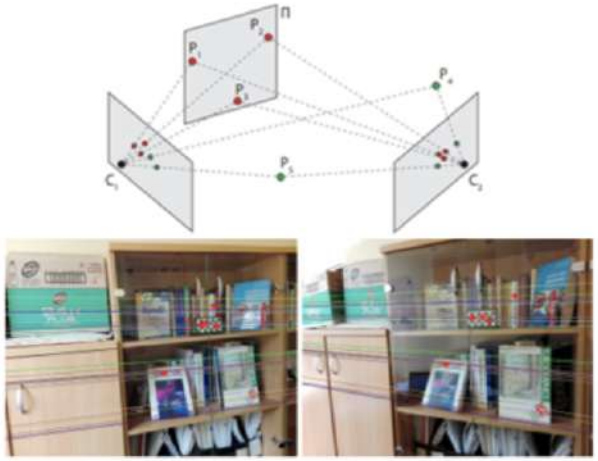
Five-point Fundamental Matrix Estimation for Uncalibrated Cameras

D. Barath CVPR2018

概要

2つの未キャリブレーションカメラにおいて，**5点のみ**で基礎行列を推定する手法を提案．

回転不変な特徴点（SIFT等）を使う．3点は平面にあれば，他2点はどこでも可能．グラフカットRANSACのようなロバスト対応点推定と組み合わせれば，state-of-the-artな性能が出る．



新規性・結果・なぜ通ったか？

通常，7点や8点取るアルゴリズムが用いられるが，リーズナブルな制約で，少ない情報のみでキャリブレーションできるのはうれしい．例えば図のようにキャリブレーションボードを小さくできたりする． 大変有用な研究成果．

コメント・リンク集

- arXiv

[#275]

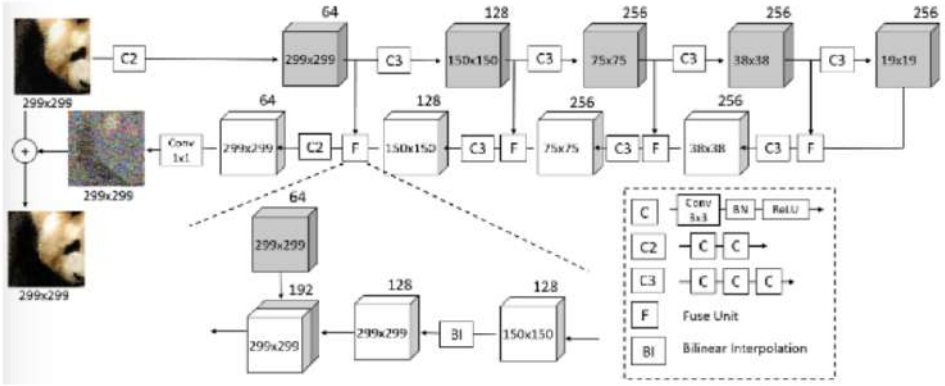
Kodai Nakashima

Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Jun Zhu and Xiaolin Hu CVPR 2018

概要

画像分類におけるadrversarial attackの防御手法として, high-level representation guided denoiser (HGD) を提案.target model (メインの処理を担うネットワーク) への前処理段階で用いる. HGDは, マルチスケールインフォメーションを得るためU-netの構造を使い, トレーニングするための損失関数として, 元画像とノイズの乗った画像をそれぞれ入力したときの出力差を用いる. 右図に提案手法の詳細を示す.



新規性・結果・なぜ通ったか？

pixel-levelの損失関数を課した従来のdenoiserと比べ, より良い結果が得られた.

state-of-the-artな防御手法であるensemble adversarial trainingと比べ, 3つのメリットがある.

1. target modelがwhite-boxとblack-boxの両方に対してよりロバスト.
2. 大規模データセットでの学習が簡単.
3. 他のtarget modelへ使い回すことが可能.

コメント・リンク集

- 論文URL

[#276]

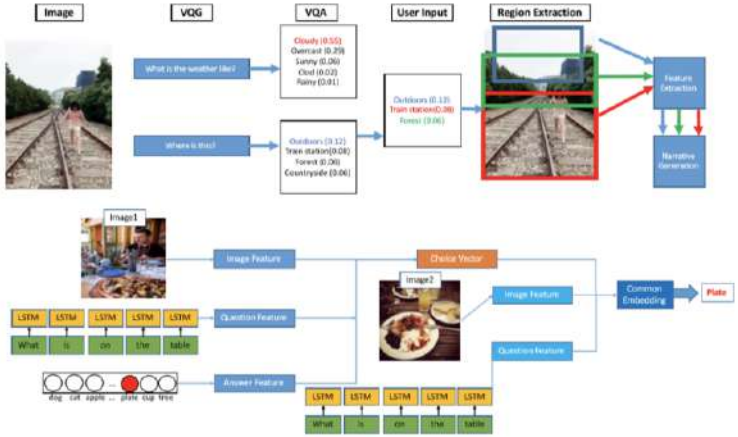
Yue Qiu

Customized Image Narrative Generation via Interactive Visual Question Generation and Answering

Andrew Shin, Yoshitaka Ushiku, Tatsuya Harada CVPR 2018 1224

概要

・新規の“Customized画像説明文生成”タスクを提案した．また，インタラクティブにユーザに自動的に画像に関する質問をし，回答文を収集できるような仕組みを提案した．・従来の画像説明文生成タスクにおいて，異なるユーザの性質や画像の注目領域などにより，多様な説明文を生成できることが検討されていない．このような性質に応じて，多様な質問文を生成できる仕組み及びユーザとインターアクションしユーザの個性的な回答文を収集しユーザの特徴を学習することにより，Customizedで画像説明文を生成できる仕組みを提案した． ・提案仕組みは具体的に：①画像から self Q&A modelにより，画像中のマルチリジョンを注目し(attention構造を利用した)質問文を生成し，VQAモデルにより回答する(マルチ回答がある質問文だけを保留)；② ①により生成できた質問文をユーザに提示し，回答させる；③画像リジョン・質問文・回答文の統合した画像説明文を生成する． ・画像リジョン・質問文・ユーザ特有な回答文からchoice vectorを抽出し，このベクトルを利用してほかの画像が入力された場合，ユーザの個性的な画像説明文を生成できる．



新規性・結果

・新規な問題設定“Customized画像説明文生成”・提案手法により，画像からより多様でユーザの個性を含んだ説明文を生成できる． ・Automatic Image Narrative Generationにおいて，従来のデータセットCOCO, SIND, DenseCapなどと比べ”diversity”,”interesting”,”naturalness”,”expressivity”などの指標に対しパフォーマンスが良い ・ Interactive Image Narrative Generationにおいて，ヒューマンテストで良い評価を達成した．

リンク集

・ユーザの個性を学習できる仕組みは応用場面が広そう

- 論文

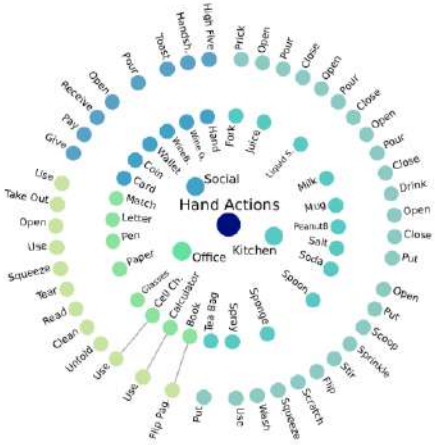
Kensho Hara

First-Person Hand Action Benchmark with RGB-D Videos and 3D Hand Pose Annotations

G Garcia-Hernando et al., CVPR 2018

概要

一人称視点動画 (RGB-D) データセットの提供．手（21点の3D間接位置）と物体（6D姿勢）の情報に加えて，45クラスの行動ラベルが付けられている．データ数は1175シーケンス，10万フレーム．手の3D姿勢と行動ラベルが付いている一人称視点動画データセットはこれまでになかった．実験では従来手法やLSTMによるベースライン手法を合わせて18個を比較した結果が議論されており，手の姿勢情報を使う手法が高い性能を示す傾向があることが確認されている．



新規性・結果・なぜ通ったか？

- 手の3D姿勢を使った行動認識のためのデータセットを提供．
- RGB, Depth, Poseといった様々な特徴を用いる各手法が詳細に議論されている．
- 一番良い手法で78%程度の認識率．

コメント・リンク集

- 論文 (arXiv)
- 動画 (YouTube)

[#278]

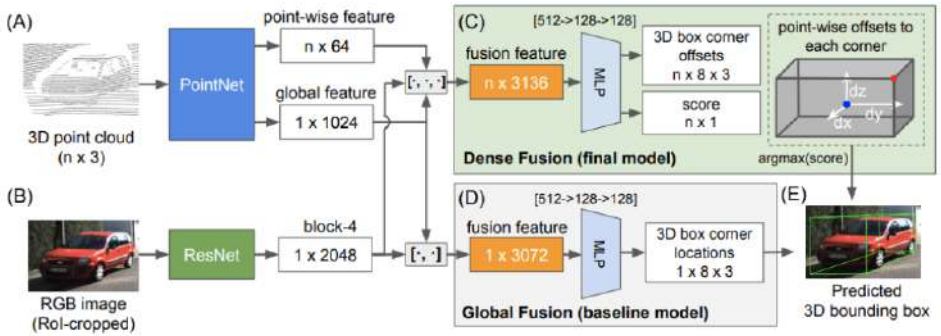
Yue Qiu

PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation

Danfei Xu, dragomir Anguelov, Ashesh Jain CVPR 2018 50

概要

・画像と点群情報を利用した3D物体検出のフレームワークPointFusionを提案した．・従来のマルチセンサーの情報を利用した3D物体検出は前処理が必要、マルチセンサーを異なるパイプラインで処理し，他のセンサーのコンテキストをうまく利用できないなどの問題点がある．PointFusionは①異なるネットワーク構造を用いて画像(CNN)と点群情報(PointNet)を直接処理し，②デンスフュージョンネットワーク構造を提案し，画像と点群の抽出情報を統合しより精密な3D物体検出を行う． ・2種類のデンスフュージョンネットワークを提案した．①画像情報及びPointNetにより抽出したグローバル情報を統合し， 3Dボックスのコーナー位置を推定する．②画像情報及びPointNetにより抽出したグローバル情報、ポイントフィーチャーを統合し， 3Dボックスのオフセット及びconfidence scoresを予測する．最後の2つの結果を統合し，最終的な結果を予測する



新規性・結果

・点群データの前処理が必要無し．・対応できるデータの形式が広い，室外環境と室内環境両方対応できる． ・多様な三次元センサーのデータを対応できる．(RGB-D, LiDar, Radar,…) ・KITTI, SUN-RGBDデータセットにおいてstate-of-the-artな結果

リンク集

・室内・外環境両方対応できるので、応用場面が広そう・将来的にend-to-endに実現できたら更に良い

- 論文

[#279]

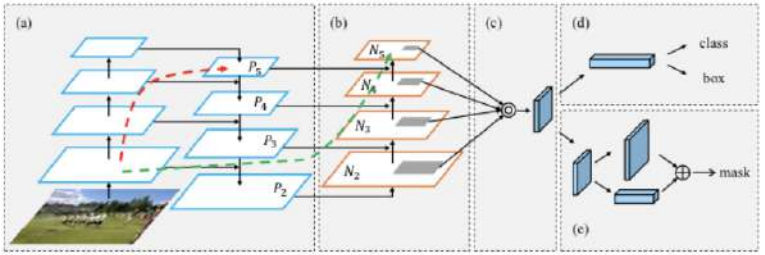
Hiroshi Fukui

Path Aggregation Network for Instance Segmentation

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, Jiaya Jia CVPR2018, arXive:1803.01534 912

概要

Feature Pyramid Network(FPN)ベースのMask R-CNNに，下位層の特徴マップを上位層に伝播させるPath Aggregation Networkを提案．インスタンスセグメンテーションの傾向として，上位層では物体全体に強く反応するが，下位層では物体の局所的な領域に強く反応する． そのため，Path Aggregation Networkでは，上位層と下位層の特徴マップを用いることで，インスタンスセグメンテーションの精度を向上させている． Path Aggregation Networkは，COCOのベンチマークで2位の性能を達成しており，CityscapeとMVDでも高い性能を達成している．



新規性・結果・なぜ通ったか？

Path Aggregation Networkの構造は右図のようなシンプルな構造になっている．(a)の部分はFPNと同様の構造となっており，FPNの特徴マップから(b)で新しい特徴マップを作成する． ここで，(a)と(b)では，緑線と赤線のように短距離と長距離のショートカットを導入する． これにより，下位層の特徴を上位層に伝播することが可能である．

コメント・リンク集

- 論文リンク

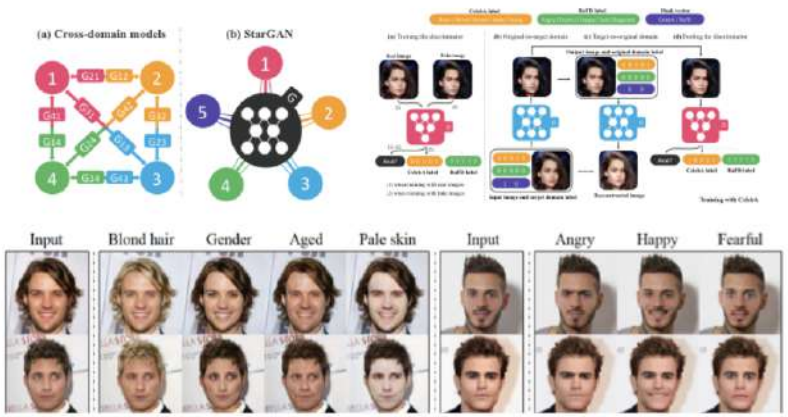
[#280] Hiroshi Fukui

StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo CVPR2018, arXive:1711.09020 872

概要

1つのネットワークでマルチドメイン対応の画像変換が可能なStarGANを提案．pix2pixやCycleGANの場合，左上図のように特定の1つのドメイン変換しかできないため，複数のドメイン変換をする時には各ドメインを変換するネットワークをそれぞれ構築しなければいけない． StarGANでは，入力する条件とロス設計を適切に設計することで，シンプルなネットワークで多ドメインな画像変換を実現している． 実験では，顔属性のCelebAと表情のRaFD Datasetを使用し，2つのデータセットでGANを学習して下図のような多様な顔画像変換を可能にしている．



新規性・結果・なぜ通ったか？

StarGANの構造は，右上図のようになっている．ここで，入力はそれぞれのドメインの画像がランダムに入力される． まず，real imageとfake imageでDiscriminatorを学習する． そして，次にGeneratorを学習する． Generatorは，生成したい顔画像の条件とreal imageを入力して，画像変換する． ここで，変換した画像はDiscriminatorに入力される． 変換した顔画像はCycleGANのようにreal imageを再変換する． 定義するロスは，一般的なAdversarial Loss，ドメインを認識するロス，real imageと再変換したimageのL1 Lossである． また，複数のデータセットを学習するために，各データセットのラベルとデータセットの情報が格納されたMask vectorを導入している． これにより，多ドメインかつ複数データセットに対応したGANを構築できている．

コメント・リンク集

多ドメインかつ複数データセットに対応したGAN．変換するドメインの数に依存しないので，非常に用途が広がりそう．

- 論文リンク
- コードリンク

[#281]

Hirokatsu Kataoka

Semi-parametric Image Synthesis

Xiaojuan Qi, Qifeng Chen, Jiaya Jia, Vladlen Koltun CVPR 2018 (oral)

概要

意味ラベル（Semantic Layout）から写真のようにリアルな画像をSemi-parametricな手法にて生成する。Semi-parametricはNon-parametricとParametricの強みを相補的に適用する手法である。セマンティックセグメンテーションのアノテーションとその対応する画像をペアとした外的なメモリにより対応関係を学習、Canvasとしてその順番や境界面を初期ステップとして出力する。次にCanvasと意味ラベルを入力としてConv-Deconv構造のネットワークにより写真のようにリアルな画像を出力とする。

新規性・結果・なぜ通ったか？

Cityscapes, NYU, ADE20Kデータセットとセマンティックセグメンテーションに関するラベルが付与されていれば学習/テストが可能であり、同データセットにて従来法よりもさらにリアルな画像を生成するに至った。図には従来法（Chen and Koltun, ICCV 2017）との比較があり、従来法ではエッジ付近にボケが生じているが、提案法ではボケを相殺してさらに光の度合いまでもリアルに復元できている。

コメント・リンク集

意味ラベルから写真を復元することに成功した。今後、さらに生成するアピアランスや配置をコントロールする手法が登場すれば、学習データを無限に増やすことができたり、作りたい写真を再構成することが可能になる。

- 論文
- Project
- GitHub
- Video

[#282]

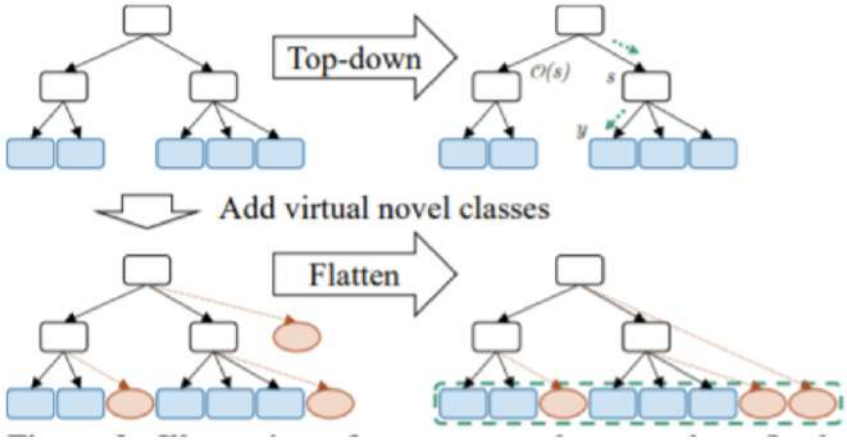
Yue Qiu

Hierarchical Novelty Detection for Visual Object Recognition

Kibok Lee, Kimin Lee, Kyle Min, Yuting Zhang, Jinwoo Shin, Honglak Lee CVPR 2018 131

概要

・最も近いスーパークラスを予測することにより階層的新規(novelty)物体識別及び検出のフレームワークを提案した．・従来，新規なunseen物体識別は”known”と"unknown"に回帰する問題として対応されている．この論文で，物体のクラスを階層的に取り扱い，unseen物体の最も近いスーパークラスを求める．提案フレームワークによりgeneralized zero-shot learningタスクに用いられる階層的エンベディングを得られる． ・2種類の階層的な新規(novelty)物体検出構造を提案した．①top-down構造ではconfidence-calibrated classifierにより物体を分布の一致性が高いスーパークラスに分類する．②flatten構造では階層的分類構造の全体を用いずにerror aggregationを避ける単一的なclassifierを用いる．また，①と②を組み合わせすることにより，階層的検出精度を向上できることを示した．



新規性・結果

・従来のクローズデータセットを用いた物体検出と比べ，提案手法はオープンデータセットを対応できる．・generalized zero-shot learningタスクで提案フレームワークを用いられる ・ ImageNet, AwA2, CUBなどのデータセットで階層的新規(novelty)物体識別においてベースラインより高い精度を達成した．

リンク集

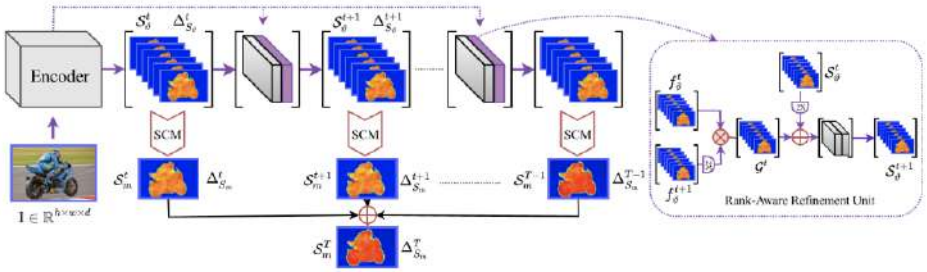
- 論文

[#283] Yue Qiu

Revisiting Salient Object Detection: Simultaneous Detection, Ranking, and Subitizing of Multiple Salient Objects

Md Amirul Islam, Mahmoud Kalash, Neil D. B. Bruce CVPR 2018 892

概要



・マルチsalientオブジェクトおよびそれぞれのsalientランキングを同時に検出するネットワークを提案した．・従来のsalientオブジェクトタスクに，salientランキングは観測者によって異なる結果が出る性質があるため，オブジェクトのsalientランキングについてまだ検討されていない．この文章でsalientランキングを有効的に得られるネットワークを提案した．またsalientランキング手法の評価方法も提案した．・具体的なネットワーク構造はまずencoderネットワークにより粗末な相対salientスタックを生成し，そしてStacked Convolutional Module (SCM)により粗末なsaliency mapを生成する．またrank-awareでstage-wiseなネットワークによりsalientスタックをリファインする．ヒュージョンレイヤーにより各stageのsaliency mapを統合する．

新規性・結果

・saliency ランキングの提案・AUC, max　F-measure, median F-measure, average F-measure,MAE, and SORなどの 評価方法により，state-of-the-artなsalientオブジェクト検出性能を達成した．

リンク集

- 論文

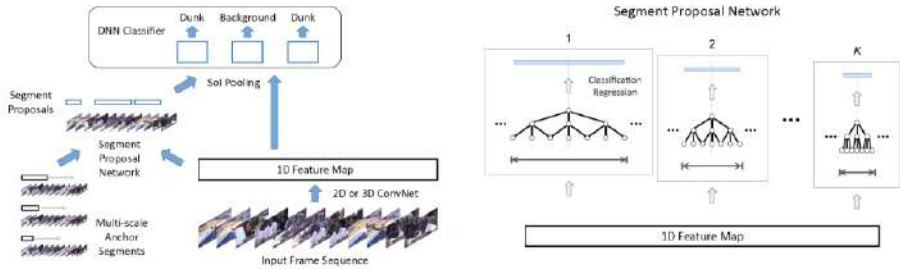
[#284] Kensho Hara

Rethinking the Faster R-CNN Architecture for Temporal Action Localization

Y. Chao et al., CVPR 2018

概要

動画中の行動のラベル，開始・終了時刻を推定するTemporal Action Localizationの研究． Faster R-CNNによる物体検出をベースにLocalizationをする． ここで，スケールのバリエーションが非常に大きい，前後の行動などのコンテキストが重要， RGBとFlowをどう統合するか，といった3点の検討が重要としてこれらに取り組んでいる． 提案手法であるTAL-Netのポイントとしては， アンカーごとに適切なスケールの受容野を持つ異なるCNNを組み合わせて利用している点． 各問題に対する設計がそれぞれ精度向上に寄与している点を実験から確認し， THUMOS'14でのSOTAを達成．



新規性・結果・なぜ通ったか？

- 行動の時間スケールについての検討をちゃんと行った点は新規性がある
- 提案手法の各要素についての実験がされていて，それぞれによる精度向上を確認できている

コメント・リンク集

- 論文 (arXiv)
- 目新しいアイデアはないように思うが，問題点に対する解法を検討してかっちりと評価している
- この辺りのスケールの話は大事そうなのにこれまで意外とちゃんとやられてきてなかったところ

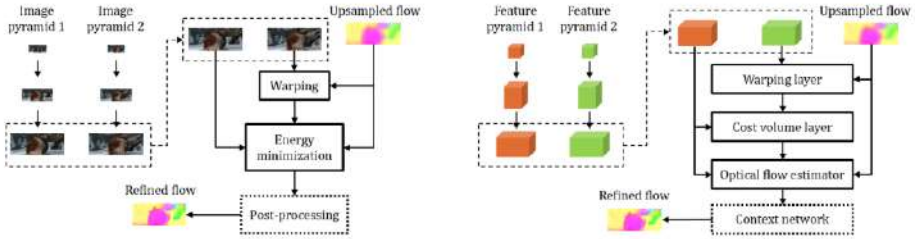
[#285]

Hirokatsu Kataoka

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, Jan Kautz CVPR 2018 (oral)

概要



コンパクトかつ効果的なオプティカルフロー推定を実現するPWC-Netを提案する。ピラミッド構造かつ学習可能な階層的処理、射影（Warping）、コストボリュームにより設計され、軽量化しながら高精度なフロー推定を実現している。図は従来法（左図）と提案法（右図）のアーキテクチャの概略を示している。従来は画像のピラミッド構造により全てのサイズを階層的にオプティカルフローの射影や最適化を行い、最後に後処理をしていたが、提案法のPWCNetではあるひとつの階層内で後処理を行い、コンテキストを考慮したネットワーク（ContextNetwork; Dilated Convによる、各階層のオプティカルフローを入力するとそれらを総合的に解釈して最良のオプティカルフローを出力する）を通り抜けることで出力する。間には{Warping, Cont Volume, Optical flow}を行う層により構成される。

新規性・結果・なぜ通ったか？

従来法であるFlowNet2よりも17分の1の軽量化モデルでありながら、MPI Sintel final pass/KITTI 2015 BenchmarkにてState-of-the-art、Sintel 1024x436の解像度にて35fpsで動作する。

コメント・リンク集

- オプティカルフロー/距離画像の推定など、RGBの入力から異なるチャンネルを出力する課題が登場して本論文のように精度向上やコンパクト化、処理速度向上が著しい。ただし、出力したオプティカルフローや距離画像の出力自体の正当性を保証するような評価方法が必要？特に、異なるドメイン（ドイツの道路データで学習して日本の道路データでテストするなど）での適応とその性能保証は欲しいところ。
- （さすがNVIDIA！？）実験量がとても多く見える。Table1~7までびっしり実験結果が埋められている。
- 論文

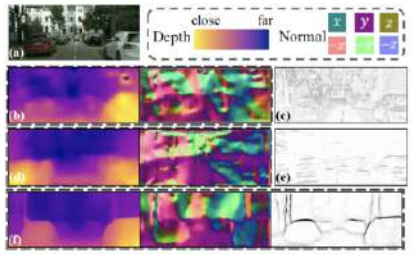
[#286]

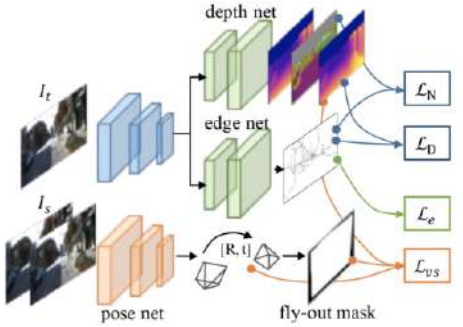
Kensho Hara

LEGO: Learning Edge with Geometry all at Once by Watching Videos

Z. Yang et al., CVPR 2018

概要





ラベルなし動画からの3次元幾何 (Depth, Normal) の推定．従来研究のものだと画素ごとの誤差で最適化していたのでボケた幾何構造推定になっていたのが問題と主張．提案手法はエッジと3次元幾何を同時に推定して最適化することで，左図 (f) のような正確な幾何構造を推定可能にした．ベースは従来手法同様で，カメラ姿勢を推定し，それに基づくWarping結果と元のフレームとの間の誤差をとって最適化．これに，エッジ推定と3D-ASAP (as smooth as possible in 3D) Priorを導入したところがポイント．3D-ASAPはある2点間の間にエッジがなければその2点は同一平面上にあるという仮定に基づく提案手法．

新規性・結果・なぜ通ったか？

- 3次元幾何とエッジ推定を同時にする手法の提案
- 3D-ASAP Priorの定式化とそれによる精度向上を実現
- KITTIやCityScapesでのSOTAを達成

コメント・リンク集

- 論文 (arXiv)
- 結果動画 (YouTube)

[#287]

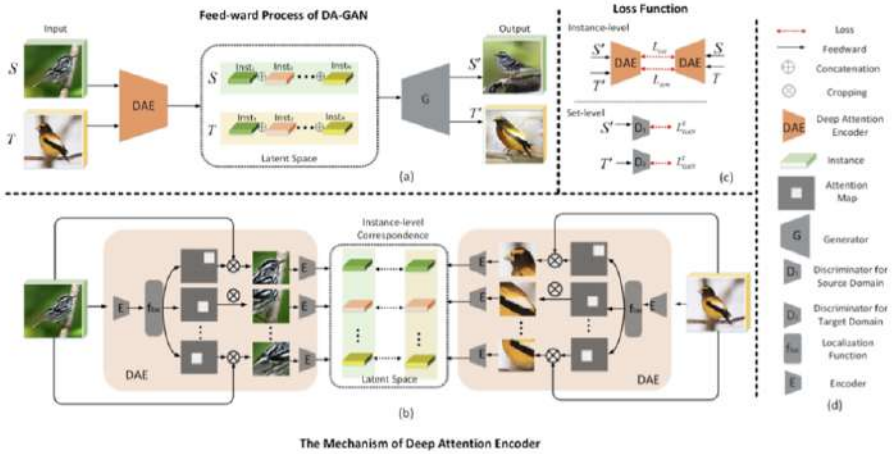
Yue Qiu

DA-GAN: Instance-level Image Translation by Deep Attention Generative Adversarial Network

Shuang Ma, Jianlong Fu, Chang Chen, Tao Mei CVPR 2018 695

概要

・無監督インスタンスレベルのattentionを用いたImage Translationフレームワークを提案した．・従来の無監督Image Translationではセットレベルで実現され，物体パーツレベルの対応ができないため，従来手法より生成した物体画像が幾何や意味的な情報のリアル性が低い場合がある．それと比べ，提案フレームワークは①物体をはattentionを用いた高構造化latent空間に変換し，このlatent空間によりインスタンスレベルなImage Translationを可能にした．②さらに，source samplesとtranslated samplesをセマンティック的に対応させるconsistency lossを提案した．



The Mechanism of Deep Attention Encoder

新規性・結果

・初めてattentionをGANに導入したと宣言・MNIST , CUB-200-2011, SVHN , FaceScrub and AnimePlanet 1などのデータセットを用いて実験を行い，ドメンadaption，テキスト-画像合成，ポーズモーフィング，顔 - アニメーション化などのタスクにおいて，state-of-the-artな精度を達成した．

リンク集

・attentionをGANに導入し，さらに精密で構造化した画像生成ができるので，様々なアプリで応用できそう

- 論文

[#288]

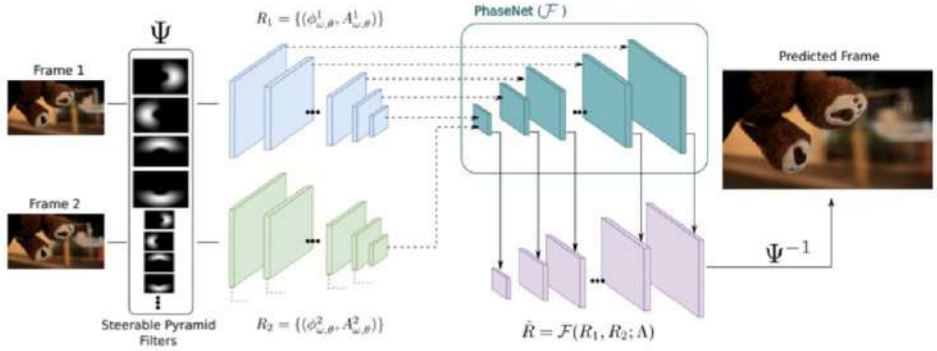
Munetaka Minoguchi

PhaseNet for Video Frame Interpolation

Simone Meyer, et al. 1804.00884

概要

様々なシーンに頑健かつ、大きな動きにも対処しながらビデオフレームの補間を行うPhaseNetの提案。中間のフレームにおける位相と階層構造を推定するnnのデコーダを搭載。これにより、既存の位相ベースの手法よりも広範囲に渡る動きに対応。



新規性

既存のビデオフレーム補間アプローチは、フレーム間において密な対応付けが必要であり、照明変化や被写体ブレに頑健でない。カーネルに依存した深層学習ベースの手法でもある程度緩和することはできるが不十分。ピクセル単位の位相ベースの手法ならば上手くいくことが実装されている。位相ベースでnnを用いた手法を提案。

結果・リンク集

位相のlossとノルムを組み合わせることで、チャレンジングなシーンでも視覚的に綺麗な画像を生成できる。

- 論文

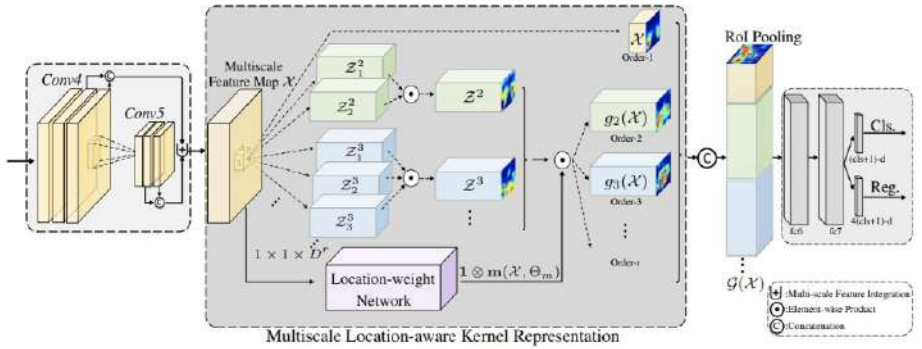
Ryosuke Araki

Multi-scale Location-aware Kernel Representation for Object Detection

Hao Wang, Qilong Wang, Mingqi Gao, Peihua Li and Wangmeng Zuo CVPR2018 153

概要

物体検出時に特徴量の高次の統計量（high-order statistics）を獲得するためのMulti-scale Location-aware Kernel Representation（MLKP)を提案する．MLKPはSSDで用いるような，複数解像度の特徴マップを結合したマルチスケール特徴マップを用いて効果的に計算できる．マルチスケール特徴マップをMLKPに入力すると，畳み込みと要素ごとの積算を行いr次の表現Z^rを得る．このとき，location-weight networkは各位置の寄与度を学習する．その後，各次の表現を重みつき結合し，RoI Poolingへ入力する．



新規性・結果・なぜ通ったか？

最近の分類メソッドでよく用いられる高次統計量を物体検出器の高精度化に用いる手法である．Faster R-CNNにMLKPを統合することで，Faster R-CNNよりも精度が4.9%(mAP, VOC2007），4.7%（mAP, VOC2012），5.0%（MSCOCO）向上した．DSSDやR-FCNと比較しても同等もしくはそれ以上の性能である．

コメント・リンク集

流行りのマルチスケール手法をR-CNNに昇華した感じ．R-CNNベースの手法もまだまだ煮詰める余地は十分ある．

- arXiv
- コードpy-faster-rcnnをベースにされている．マルチGPU版もあり

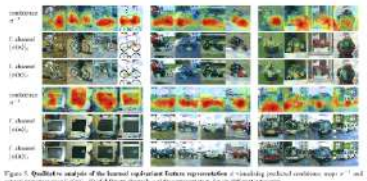
Tomoyuki Suzuki

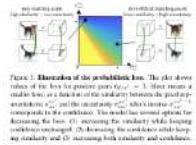
Self-supervised Learning of Geometrically Stable Features Through Probabilistic Introspection

David Novotny et al. CVPR 2018

概要

幾何学変換を利用したGeometrically Stable な特徴表現の獲得手法。オリジナル画像とそれに幾何学変換を施した画像を同じCNNに学習し、中間特徴マップ上で対応するpixelでの特徴量の類似度が高くなるように学習する。キーポイントマッチングなどの問題設定で教師あり学習以上の効果を発揮。Pixelによってはマッチングが困難ば場合も存在するため、不確実性を考慮した学習を提案。





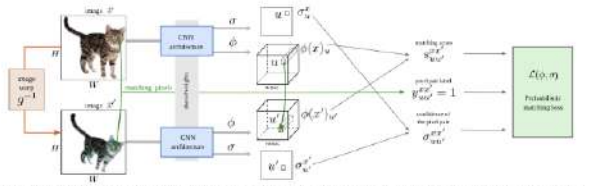


Figure 2. **Overview of our approach**. Image x is warped into image x' using the transformation g^{-1}. Pairs of pixels and their labels (encoding whether they match or not according to g^{-1}) are used together with a probabilistic matching loss to train our architecture that predicts i) a dense image feature $\phi(x)$ and ii) a pixel level confidence value $\sigma(x)$.

手法・新規性

ペアとなる画像を同じNNに入力し、各pixel ペアの類似度と、不確実性を表す値を算出。不確実性を考慮した損失関数を定義することで、結果的にNNはマッチング可能かつ対応するpixelに関しては高い類似度と低い不確実性を、マッチングが困難なものに関しては高い不確実性を算出するように学習される。

メモ・リンク

定義された距離尺度において対象に直接近づける枠組みが多い通常の類似度学習と異なり、連続値である類似度を確率変数とすることで、不確実性を考慮するのは興味深い。しかし、定式化としては論文内のものよりも、不確実性利用してモデルが類似度の分布を算出しているという定式化にした方がわかりやすいのではないかと思った。

- 論文

[#291]

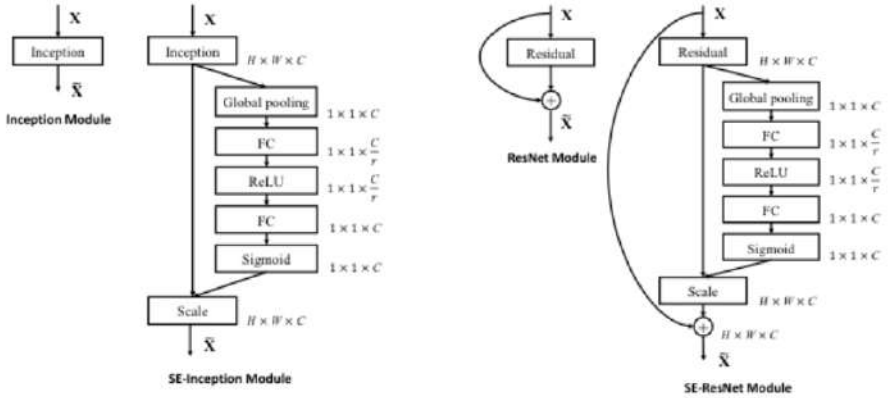
Hiroshi Fukui

Squeeze-and-Excitation Networks

Jie Hu, Li Shen, Gang Sun CVPR2018, arXive:1709.01507 891

概要

Residualモジュール, Inceptionモジュールに対してAttention機構を導入したネットワーク．Squeeze-and-Excitation Networks(SENet)では，生成される特徴マップのチャンネルに対してAttentionを導入している． SENetは，ImageNetでstate-of-the-artな性能を達成している．(現在1位) また，Place Datasetでも高い性能を達成している．



新規性・結果・なぜ通ったか？

SENetには，右図のように2つのモジュールが提案されている．SE Inception moduleは，VGGやAlexNet等の順伝播ネットワークで使われるSEモジュール． SE Residual moduleは，ResNet系のネットワークに使われるSEモジュールである． 基本的には，Global Average Poolingを施した後に，全結合層を何層か通してチャンネル毎のAttentionを生成する． この構造は，ResNet等の様々なネットワークモデルにも適応できる．

コメント・リンク集

Attention機構を導入した物体認識法．最近，物体認識にもAttentionが流行し始めているので，その先駆けな手法になりそう．学習モデルもGitHub上で公開．

- 論文リンク
- コードリンク

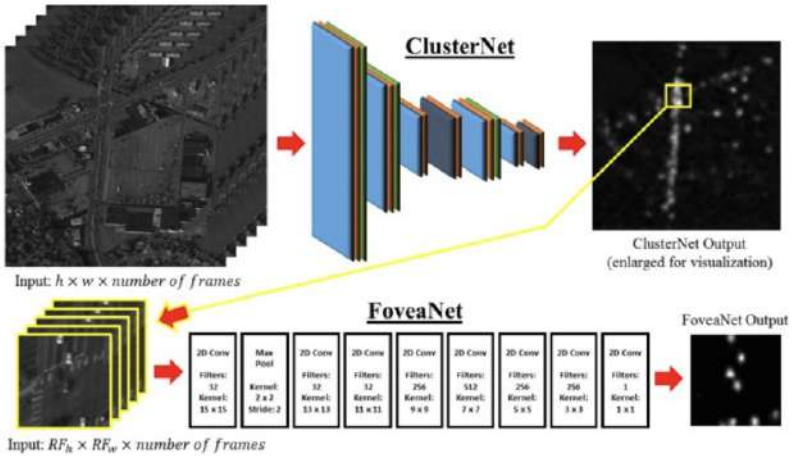
Munetaka Minoguchi

ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information

Rodney LaLonde, Dong Zhang, Mubarak Shah 1704.02694

概要

1平方キロメートル以上の広範囲の領域を撮影できるWide Area Motion Imagery(WAMI)の映像から、車などの小さい物体を検出する手法の提案。まず、ClusterNetでビデオフレームから、CNNを使って動きと外観情報を結合し、regions of objects of interest(ROOBI)を出力。次に、FoceaNetによって、ヒートマップ推定を介して、ROOBI内の物体の重心位置を推定する。



新規性

WAMIを使った従来の物体検出は、アピアランスベースの分類器であまり精度が出ず、背景差分やフレーム間差分などの動き情報に依存しがち。Fast R-CNNなどにおけるこれらの問題を検証し、効率的かつ効果的な新たな2ステージCNNを提案。

結果・リンク集

広範囲の情報から数百の物体を同時に検出する。他の手法では扱えない停止車両なども検出できる。

- 論文
- WAMI

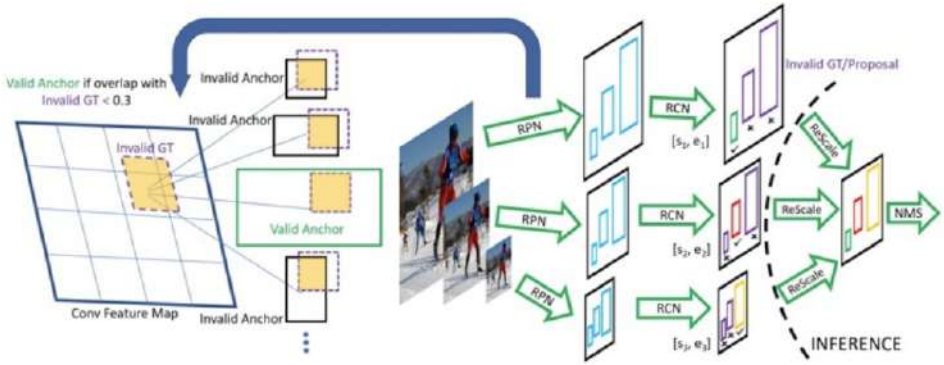
Munetaka Minoguchi

An Analysis of Scale Invariance in Object Detection – SNIP

Bharat Singh, Larry S. Davis 1711.08189

概要

極端なスケール変化に頑健な物体検出手法であるSNIPの提案。物体検出において、大きな物体と小さな物体をそれぞれ検出することは困難。そこで、学習時に異なるサイズの物体における勾配を、選択して逆伝播する。物体の幅広いスペクトルに対処し、ドメインシフトを低減する。ピラミッド型のネットワークとなっており、end-to-end学習可能。



新規性

まず、現代の物体検出手法の欠点として、スケール変化について解析している。小さい物体を検出するために“アップサンプリング画像が必要か”などを、ImageNetを使ってパフォーマンスを評価。これらの解析に基づいてSNIPを開発。

コメント・リンク集

COCO2017 challengeにおける、最優秀学生応募作品。

- 論文
- コード

[#294]

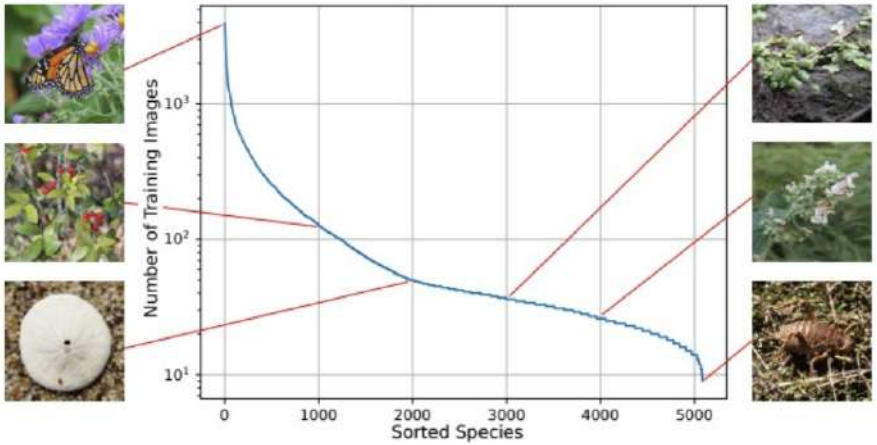
Munetaka Minoguchi

The iNaturalist Species Classification and Detection Dataset

Grant Van Horn, el al. 1707.06642

概要

自然界にける、“写真に写り易さ”を考慮した画像分類・検出タスク用データセットの提案。5000種類以上の植物や動物からの85万9000の画像で構成。世界各地の多種多様な種やシチュエーションで撮影され、様々なカメラタイプで収集することで画質の変化し、クラスの均衡が大きい。



新規性

従来の画像分類・検出用データセットでは、カテゴリごとに画像数が統一されている傾向にある。しかし，写真に収め易い種と、そうでない種があるため、自然界はとても不均衡。この差に着目し、現実世界の状況に近い状況で分類・検出に挑戦するデータセットを提案した。

コメント・リンク集

やはり既存の手法では精度を出すのは難しそう。このデータセットで精度を出すチャレンジングな研究をするのはアリ。

- 論文

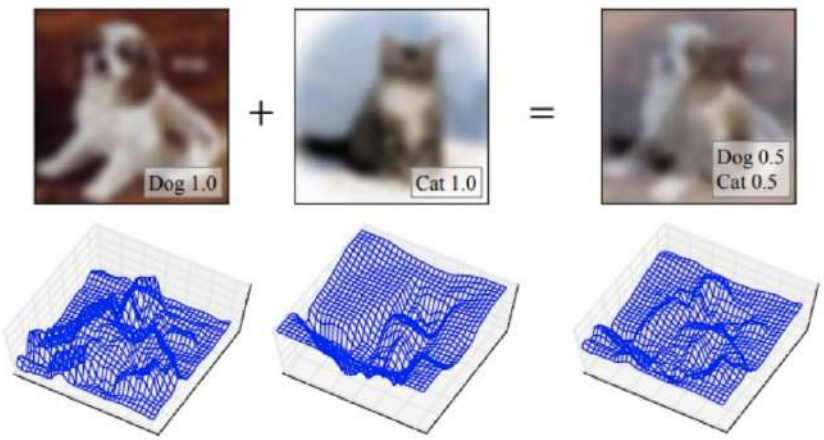
Munetaka Minoguchi

Between-class Learning for Image Classification

Yuji Tokozume, Yoshitaka Ushiku and Tatsuya Harada 1711.10284

概要

Between-Class learning(BC learn)という画像分類タスクにおける新学習方法の提案。まず、異なるクラスの2枚の画像をランダムな比率で混合したbetween-class imageを作成。そして、画像を波形として扱うためにミキシングを行う。混合画像をモデルに入力し、学習することで混合した比率を出力する。これにより、特徴分布の形状に制約をかけることができるため、汎化性能が向上する。



新規性

もともとは、混合できるデジタル音声のために開発された手法。CNNは“画像を波形として扱っている”という説から、本手法を提案。2つの画像を混合する意味に疑問はあるが、実際にパフォーマンスが向上している。

結果・リンク集

混合とミキシングの提案手法によって分類精度が向上。画像の混合にどんな意味があるのかを解明してほしい。

- 論文

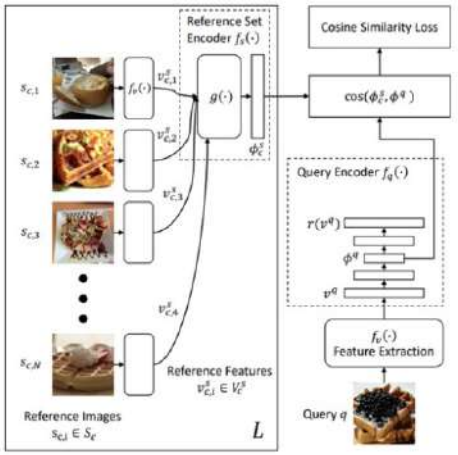
Munetaka Minoguchi

CleanNet: Transfer Learning for Scalable Image Classifier Training with Label Noise

Kuang-Huei Lee, Xiaodong He, Lei Zhang and Linjun Yang CVPR 2018

概要

ラベルノイズを使って、画像分類モデルを学習するCleanNetの提案。人間による“ラベルノイズの低減”という作業を低減する。事前知識として人の手で分類されたクラスの一部の情報だけを使い、ラベルノイズを他のクラスに移すことができる。また、CleanNetとCNNによるクラス分類ネットワークを1つのフレームワークとして統合。ラベルノイズ検出タスクと、統合した画像分類タスクの両方で、ノイジーなデータセットを使って精度検証。



新規性

人間がラベルのアノテーションをすると時間がかかり、学習はスケーラブルじゃない。逆に人間に頼らない手法はスケーラブルだが、有効性が低い。少し人間に頼って、あとは自動的にノイズ除去をするというハイブリットな手法。

結果・リンク集

弱教師付き学習と比較して、ノイズを41%低減。画像分類タスクにおいて、47%パフォーマンスが向上。

- 論文

[#297]

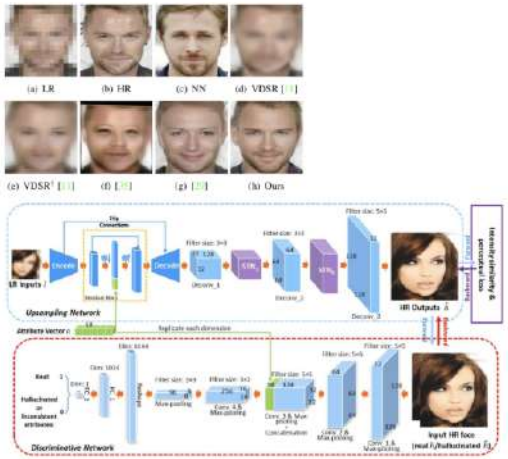
Kazuki Inoue

Super-Resolving Very Low-Resolution Face Images with Supplementary Attributes

Xin Yu, Basura Fernando, Richard Hartley, Faith Porikli CVPR 2018 Poster

概要

顔画像のアトリビュートを使用することでGTとなる高解像度画像(HR)を使用せずに低解像度画像(LR)を超解像度化する研究。LRとともに顔のアトリビュートも入力として使用することで超解像化における曖昧さを解決。 ネットワークの大枠はGANを採用。 ジェネレータにおいてLRをauto encoderに噛ませる際にエンコードされた特徴量にアトリビュートを付け足してでコードを行う。 ディスクリミネータはGTのHR画像なら1を、ジェネレータによる画像or画像にアトリビュートが含まれていないと判断した際には0を返す。



新規性・結果・なぜ通ったか？

- 入力は16x16画像、出力は入力画像が128x128に超解像度化された画像。
- PSNR、SSIMを評価指標として既存手法と比べたところもっとも良い精度を得た。
- 既存手法で入力されたLRに対して一意的なHRのみしか出力することができなかった。一方提案手法では入力するアトリビュートに伴って出力するHRの見た目を変更することが可能。

コメント・リンク集

- トレーニングで使用したデータセットはCelebAであり、使用したアトリビュートはCelebAに付属する40種類のうちからgender, ageなど18種類。
- 論文

[#298]

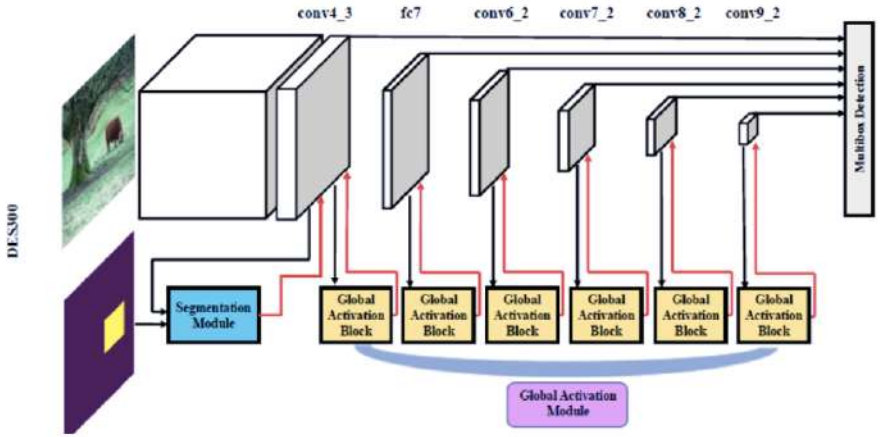
KotaYoshida

Single-Shot Object Detection with Enriched Semantics

Z.Zhang, S.Qiao, C.Xie, W.Shen, B.Wang and A.L.Yuille CVPR2018 arXiv:1712.00433

概要

Detection with Enriched Semantics (DES)というシングルショットオブジェクト検出器を提案．セマンティックセグメンテーションブランチとオブジェクト検出ブランチで構成. セマンティックセグメンテーションブランチとグローバルアクティベーションモジュールによってオブジェクト検出の特徴であるセマンティクスを向上． 既存のSSDなどのシングルショット検出器よりも速度と精度が向上．



新規性・結果・なぜ通ったか？

- セマンティックセグメンテーションブランチに高レベルのオブジェクト特徴のためのオブジェクト検出特徴チャンネルとオブジェクトクラスとの意味的関係を学習するためのグローバルアクティベーションブロックを加える．
- 一般的なシングルショット検出器と比較して大幅に検出精度が向上，
- Titan Xp GPU1台で、31.7 FPSを達成し、R-FCNやResNetベースのSSDよりも高速.

コメント・リンク集

- Paper

[#299]

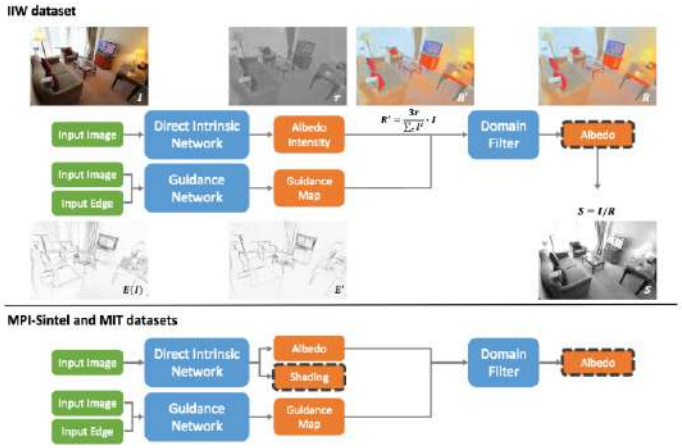
Hirokatsu Kataoka

Revisiting Deep Intrinsic Image Decompositions

Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, David Wipf CVPR 2018 oral

概要

光の反射やシェーディングなどを再計算することで自然画像の分解と再構成（Image Decomposition）を行う問題設定である。従来型の事前情報を陽に与えるフィルタリング手法とは異なり、深層学習による提案手法では（十分なラベル付きデータが存在すれば）画像の内的な情報を効果的に捉えて画像の再構成をより自然に行うことができると主張。この問題を解決するために、２種類のカテゴリに関する問い ー（１）詳細なラベル付きデータ（２）弱教師付き学習により比較的多様なラベル付きデータを学習ー を解決することができる。これにより学習データには詳細なラベル付けを行わず弱い事前知識（Loose Prior Knowledge）のみで大量のサンプルを準備することができる。手法面において、最初は荒く光の反射（Albedo）やシェーディングを推定し、次いでエッジやテクスチャ等を推定できるようにフィルタリングを学習する。



新規性・結果・なぜ通ったか？

主要な画像再構成のベンチマークにおいて全てState-of-the-artの（最先端の）結果を達成した。さらに、従来まではデータセットに対してアドホックである（と思われる）が、本論文にて提供するデータや手法はよりオープンかつリアルな問題に対して汎用的に使用できる。弱い事前知識のみでリアルデータを学習できるようにしたことも新規性として挙げられる。CVPRの査読を突破できた理由として、State-of-the-artな精度を全てのデータにて達成したことや、その学習法/アーキテクチャの提案にあると考える。

コメント・リンク集

光の反射（Albedo）や陰影（shading）を同時に推定できる技術はよりリアルな画像の生成には重要技術なのでどんどん進んで欲しい。

- 論文
- MIT Intrinsic Images Dataset
- MPI Sintel Flow Dataset
- Intrinsic Images in the Wild

[#300]

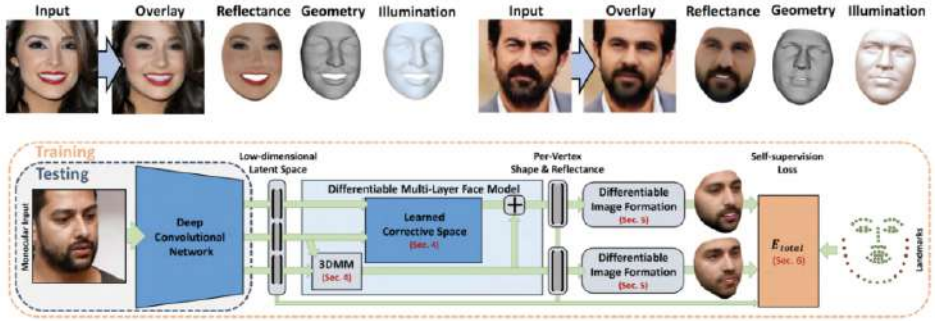
Kazuki Inoue

Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Perez, Christian Theobalt CVPR 2018 Oral

概要

単眼顔画像からリフレクタンス、ジオメトリー、照明情報を推定する研究。トレーニングデータには上記の情報のアノテーションを必要とせず、3D Morpahlbe Modelを使用することで高品質な3Dパラメトリックモデルを生成。 3D Morpahlbe Modelを使用することで高品質な3Dパラメトリックモデルを生成。 テスト時には250Hz以上で実行することができる。



新規性・結果・なぜ通ったか？

- 大量のアノテーションが必要という既存手法の問題点を解決
- 様々な表情に対応することができ、口髭や化粧も再現することが可能。
- 既存のラーニングベースの手法と比較した結果、同等の実行時間でより精度の高いリコンストラクションが可能となった。最適化ベースの手法と比較すると10%ほど精度は落ちるものの、最適化ベースの手法では実行時間が120secかかるが提案手法では4msで実行可能。

コメント・リンク集

- 目元やおでこの皺の再現には至っていない
- 論文
- Project page

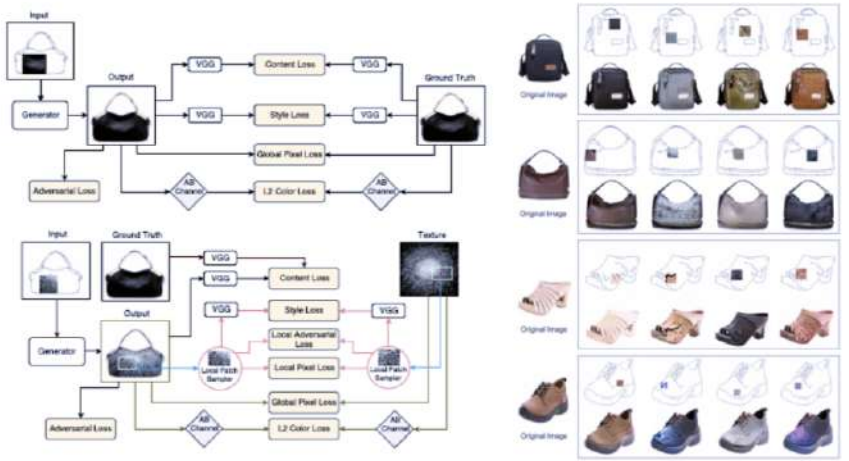
[#301] KotaYoshida

TextureGAN: Controlling Deep Image Synthesis with Texture Patches

W.Xian, P.Sangkloy, V. Agrawal, A.Raj, J.Lu, C.Fang, F.Yu and J.Hays CVPR2018 arXiv:1706.02823

概要

ユーザが色，スケッチ，テクスチャから深層画像合成を行うTextureGANを提案．既存手法では，カラーやスケッチによる制御を行っているが今回の手法ではユーザがテクスチャパチをスケッチ上に配置することによってテクスチャによる制御を実現．



新規性・結果・なぜ通ったか？

- 深層画像合成における細かいテクスチャ制御の妥当性を初めて実証
- ユーザが特定のテクスチャをスケッチの境界に「ドラック＆ドロップ」するテクスチャインタフェースの提案.
- 生成ネットワークで既存のオブジェクトに見られないテキスチャであった場合でも扱うようにする局所テクスチャロスを定義.

結果・リンク集

- TextureGANをローカルテクスチャで制約することにより，テクスチャとスケッチベースの画像合成の効果を実証.
- 別のテクスチャデータベースから抽出されたテクスチャから生成されたスケッチを用いて実験を行い、提案アルゴリズムがユーザコントロールに忠実な妥当な画像を生成されることを確認.
- Paper

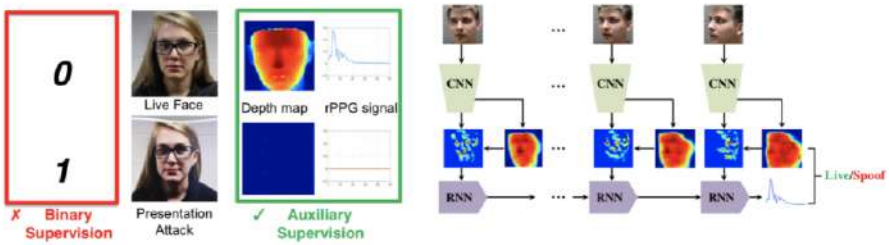
[#302]

Kazuki Inoue

Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision

Yaojie Liu, Amin Jourabloo, Xiaoming Liu CVPR2018 SPOTLIGHT

概要



入力された動画が生身の人間によるものか、あるいはそれ以外のspoofing（撮影された動画や顔のプリントなど）を判定する研究。空間的な情報として顔のデプスマップ、時間的な情報としてrPPG（信号のパルス信号）。CNN-RNNモデルを使用しCNNでデプスマップと顔の特徴量マップを、RNNは各時刻でCNNによって推定された顔の特徴量マップを入力としてrPPGを推定する。既存研究では様々なパターンのspoofingがあるにも関わらずCNNによるバイナリの識別問題として捉えていたため、CNNの広すぎる空間を学習してしまい結果的に過学習をしてしまっていた。提案手法では補助的な情報としてデプスマップ、rPPGを使用することで識別精度を向上した。更に165の被写体に対して様々な照明環境、ポーズ、表情、顔むきごとの動画を収集し、anti-spoofingのためのSiWデータベースを構築した。

新規性・結果・なぜCVPRに通ったか？

- 提案手法では既存研究のようにバイナリの識別問題とはとらえず、デプスマップとrPPGを使用することで学習したパターンのspoofingを確実に検出できることを目的とした。
- 既存研究とAPCER、BPCER、ACER、HTER値における比較を行なった結果、提案手法優位な結果となった。識別精度は約72%、state-of-the-artの研究では約34%。
- 165の被写体に対して様々な照明環境、ポーズ、表情、顔むきごとの動画を収集し、anti-spoofingのためのSiWデータベースを構築。

コメント・リンク集

- 論文

[#303]

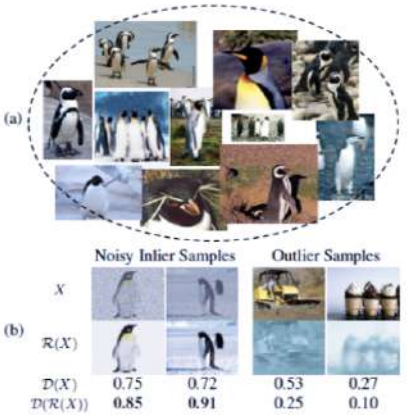
KotaYoshida

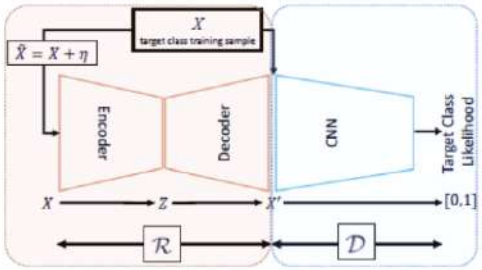
Adversarially Learned One-Class Classifier for Novelty Detection

M.Sabokrou, M.Khalooei, M.Fathy and E.Adeli CVPR2018 arXiv:1802.09088

概要

1クラス分類の際のノベリティ検出のために2段階のネットワークを構築．1つのネットワークはノベリティの検出をし，もう1つでは，inlierを強化しoutlierを歪ませる． 画像と動画で検証．





新規性・結果・なぜ通ったか？

- 1クラス分類のためのend to endネットワークを導入したもの
- GANを用いた手法では学習後に片方のモデルのみが使われるが，今回の手法ではテストの際に両方のモデルを掛け合わせることで効率化を図る

結果・リンク集

- inlierとoutlierの分類は元のクラスのサンプルの決定よりも優れている．
- ノベリティクラスのサンプルが無くても学習し，動画や画像の異常を検知でき，様々なアプリケーションで高いパフォーマンスを示す．
- Paper

[#304]

Hirokatsu Kataoka

Feature Space Transfer for Data Augmentation

Bo Liu, Mandar Dixit, Roland Kwitt, Nuno Vasconcelos CVPR 2018

概要

画像空間上ではなく、特徴空間上でデータ拡張（Data Augmentation）を行う研究である。この課題に対して著者らは特徴空間上で物体姿勢/見え方のバリエーションを多様体として考慮するFeature Transfer Network (FATTEN)を提案。従来の特徴空間上でのデータ拡張とは異なり、提案法であるFATTENはEnd-to-Endでの学習が可能であり、より効果的にデータ拡張を実行可能である。同ネットワークは姿勢やカテゴリの多タスク学習により学習を行う。図は直感的な特徴空間上での挙動を示したもので、Pose/Appearanceにおける特徴空間の動線を把握した上でデータ拡張を行うことができる。One-/Few-shot学習でも効果を発揮し、特にOne-shotでは他を大きく離して優れていることを示した。

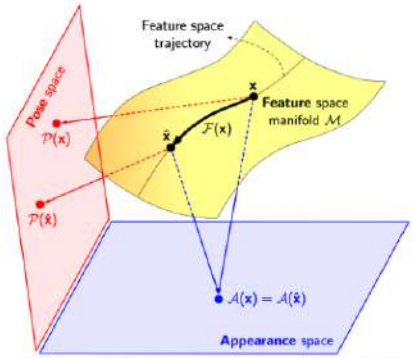


Figure 1. Schematic illustration of *feature space transfer* for variations in *pose*. The input feature $\mathbf{x}$ and transferred feature $\hat{\mathbf{x}}$ are projected to the same point in *appearance space*, but have different mapping points in *pose space*.

新規性・結果・なぜ通ったか？

新規性としては複数の属性（ここでは姿勢・アピアランス）を同時に考慮しながら特徴空間上でデータ拡張を行える点が新規性としてあげられ、さらに関連研究と異なるのはEnd-to-Endで学習できる点も優れている。直感的にはビューポイントの違いとそれに対応するアピアランスを拡張する形で特徴学習ができていると言える。FATTENを適用しModelNet/SUN-RGBDのデータセットにてデータ拡張を行った結果、はっきりとした精度向上を確認した。

コメント・リンク集

RotationNetとの比較や統合（RotationNet+FATTEN）が気になる。もともとこの論文で扱っている問題に対して精度が高いRotationNetに本論文のデータ拡張手法を使用するとさらに大きく精度向上するのでは？

- 論文
- RotationNet

[#305]

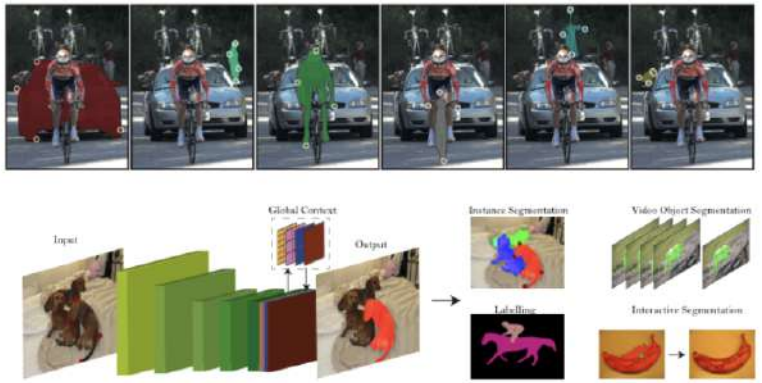
Hiroshi Fukui

Deep Extreme Cut: From Extreme Points to Object Segmentation

Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, Luc Van Gool CVPR2018, arXiv:1711.09081 88

概要

Extreme pointを用いた物体セマンティックセグメンテーション法．このExtreme pointは，セグメンテーションの上端，下端，右端，左端を使用している． 4つのExtreme pointは，物体の大まかな形状の情報を取り込みながらCNNを学習することができる． Pascal VOC, COCO, DAVIS2016, DAVIS2017, Grabcutで評価し，どのベンチマークにおいても高い性能を示している． また，セマンティックセグメンテーションのアノテーションツールとして応用できることも示している．



新規性・結果・なぜ通ったか？

使用するネットワークは，ResNet101をBackboneにしたDeepLab-v2である．提案手法のDeep Extreme Cutでは，Extreme pointを有効的に学習するために，点にガウシガウシアンを施してヒートマップを作成し，そのヒートマップを入力画像のチャンネルに追加している． この学習方法は，様々なタスクのセグメンテーションに有効であり，セマンティックセグメンテーション，動画のセグメンテーション，インスタンスセグメンテーション，インタラクションセグメンテーションに応用することができる． また，セグメンテーションのアノテーションツールにも応用でき，従来のアノテーションコストを10分の1まで削減できていることを示している．

コメント・リンク集

- 論文リンク
- プロジェクト＆コードリンク

[#306]

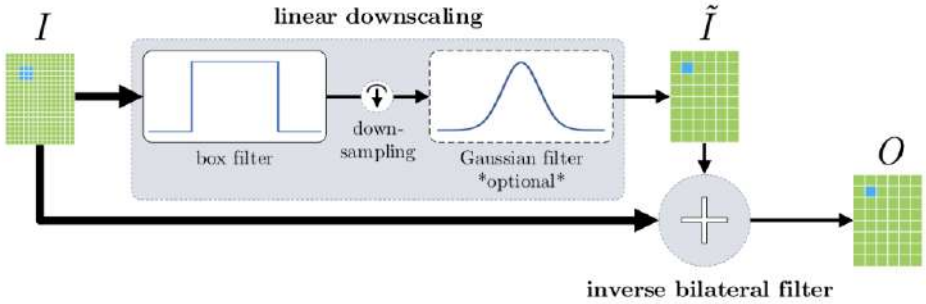
Hirokatsu Kataoka

Detail-Preserving Pooling in Deep Networks

Faraz Saeedan, Nicolas Weber, Michael Goesele, Stefan Roth CVPR 2018

概要

徐々にダウンサイジングしながらも詳細な情報は保持するという問題設定を解決するDNN、特に微分可能なプーリング手法であるDetail-Preserving Pooling（DPP）を提案する。同ネットワークでは隠れ層にて徐々にダウンスケールを行う。図にはフローチャートが示されている。このように線形ダウンスケーリングを施した画像に対して、出力が情報量をできる限り失わないように学習できるプーリングを提案することで任意の畳み込みネットに対して性能向上を見込める手法とした。



新規性・結果・なぜ通ったか？

データセットにより最良なプーリングの手法が異なるという欠点を解決するべくDPPを提案した。また、グラフィクスの分野にて提案されているDPID（文献31）を参考にして微分可能（学習可能）なプーリング手法を提案した。このようにして作成されたプーリングはあらゆるネットワークに対し有効にフィットし、（max/average poolingなどより）精度向上を保証すると主張した。例として単純にResNet-101のアーキテクチャのプーリングを置き換えてもCIFAR10にてエラー率が下がっている。このように学習可能であり、汎用的に使用できて高精度が期待できるプーリング手法を提案したことが採択された理由であると考える。

コメント・リンク集

本手法が汎用的に使用できるのであれば、早い段階でDLフレームワーク（e.g. PyTorch, TensorFlow）などに実装されて使用されるかも？実装面の難しさがどの程度あるか次第か。

- 論文
- Project
- GitHub

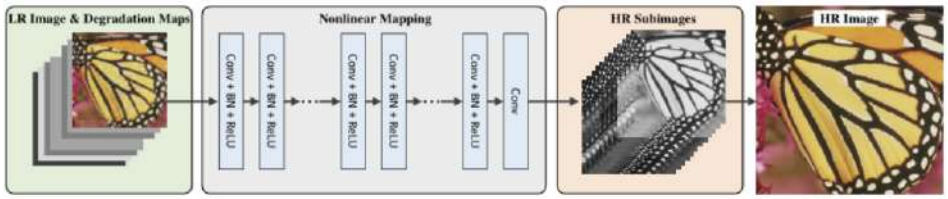
kodai nakashima

Learning a Single Convolutional Super-Resolution Network for Multiple Degradations

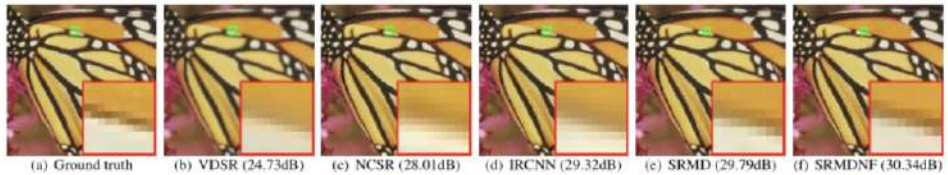
Kai Zhang, Wangmeng Zuo and Lei Zhang CVPR2018

概要

従来の単一画像の超解像手法では，低解像度の画像は，高解像度の画像からのバイキュービック的にダウンサンプリングされたものであるという仮定を置いている．そのため，この仮定に従わない場合，性能が低下する．さらに，複数の劣化に対処するスケーラビリティーも欠けている．本論文ではこれらの問題に対処するため，畳み込み超解像ネットーワークに低解像度画像とdegradation map（ブラーカーネルとノイズレベルから作成）を入力する方法を提案している．



Degradation Settings			VDSR [23]	NCSR [10]	IRCNN [35]	DnCNN [34]+SRMDNF	SRMD	SRMDNF
Kernel Width	Down-sampler	Noise Level	PSNR (×2/×3/×4)					
0.2	Bicubic	0	37.56/33.67/31.35	– /23.82/–	37.43/33.39/31.02	–	37.53/33.86/31.59	37.79/34.12/31.96
0.2	Bicubic	15	26.02/25.40/24.70	–	32.60/30.08/28.35	32.47/30.07/28.31	32.76/30.43/28.79	–
0.2	Bicubic	50	16.02/15.72/15.46	–	28.20/26.25/24.95	28.20/26.27/24.93	28.51/26.48/25.18	–
1.3	Bicubic	0	30.57/30.24/29.72	– /21.81/–	36.01/33.33/31.01	–	37.04/33.77/31.56	37.45/34.16/31.99
1.3	Bicubic	15	24.82/24.70/24.30	–	29.96/28.68/27.71	27.68/28.78/27.71	30.98/29.43/28.21	–
1.3	Bicubic	50	15.89/15.68/15.43	–	26.69/25.20/24.42	24.35/25.19/24.39	27.43/25.82/24.77	–
2.6	Bicubic	0	26.37/26.31/26.28	– /21.46/–	32.07/31.09/30.06	–	33.24/32.59/31.20	34.12/33.02/31.77
2.6	Bicubic	15	23.09/23.07/22.98	–	26.44/25.67/24.36	– /21.33/23.85	28.48/27.55/26.82	–
2.6	Bicubic	50	15.58/15.43/15.23	–	22.98/22.16/21.43	– /19.03/21.15	25.85/24.75/23.98	–
1.6	Direct	0	– /30.54/ –	– /33.02/ –	– /33.38/ –	–	– /33.74/ –	– /34.01/ –



新規性・結果

畳み込み超解像ネットワークにブラーカーネルやノイズレベルも入力しようとすると，低解像度画像とのサイズの違いによりネットワークの設計が困難になる．本論文では，dimensionality stretcing strategyを導入することによりこの問題を解決した点が新しい．

劣化されたSet5などのデータセットに対して，従来法や提案手法を適用し，PSNRとSSIMにより評価した結果，提案手法が最も良い結果を示した．

リンク集

- 論文URL

[#308]

Kazuki Inoue

Super-FAN: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with GANs

Adrian Bulat, Georgios Tzimiropoulos CVPR2018 SPOTLIGHT

概要

任意の向きの低解像度顔画像に対して超解像度化する研究。生成された超解像度画像に対してランドマーク推定を同時に行うことで画像の精度が良くなることを主張。顔画像の高解像度化の際にランドマークを特定することは有用であることはすでに示されていたが、低解像度かつ任意の顔向きの際にはランドマークを使用して高解像度化することが難しかった。提案手法ではGANによって低解像度顔画像から超解像度化された顔画像を生成し、生成された顔画像に対してランドマークのヒートマップを推定を推定することでネットワークの学習を行う。

Figure 1: A few examples of visual results produced by our system on real-world low resolution faces from WiderFace.

新規性・結果・なぜCVPRに通ったか？

- 解像度はそれぞれ入力画像が16x16、出力画像が64x64
- 生成された顔画像の評価指標としてPSNR、SSIMを、ランドマーク推定の評価指標としてAUCを使用し、 顔向きが30・60・90度の顔画像に対してどちらも既存研究より良い顔画像を生成することが可能となった。
- トレーニングの際に複数のロス関数を提案しているが、各ロス関数ごとの結果に関しても議論を行っている。

コメント・リンク集

- 論文

[#309]

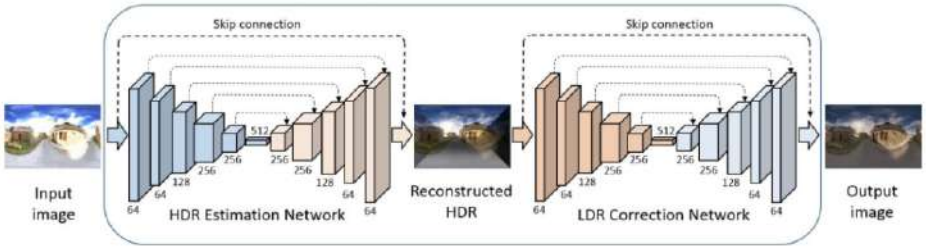
Kazuki Inoue

Image Correction via Deep Reciprocating HDR Transfromation

Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, Rynson W.H.Lau CVPR 2018

概要

入力されたLDR画像に対する露光量の調節をend-to-endに行う研究。２つのU-Netを使用し、LDR画像からHDR画像の推定と、推定されたHDR画像からLDRドメインへの変換、という２つ学習によって実現する。LDR画像に内包されている問題として、露光量が少ない箇所ではピクセルが黒く塗りつぶされてしまい、実際のシーンにおける色の推定が難しいという問題がある。そこで、LDR画像から一度HDR画像を生成することで、塗りつぶされた領域を修復する。



(b) Deep Reciprocating HDR Transformation (DRHT) pipeline

新規性・結果・なぜCVPRに通ったか？

- 入力LDR画像の露光量が多い部分や少ない部分に対しても適切な画像修復が可能となった。
- 同様の問題を扱う最新手法と比較した結果、提案手法優位な結果となった。主な理由としてはHDR画像からLDR画像へ変換する際に画像の局所的な詳細情報を保てていることをあげている。
- 定量評価として画像の質を表す数値であるPSNR、SSIM、FSIM、Q-scoeによる評価を行った。

コメント・リンク集

- 論文
- Project page

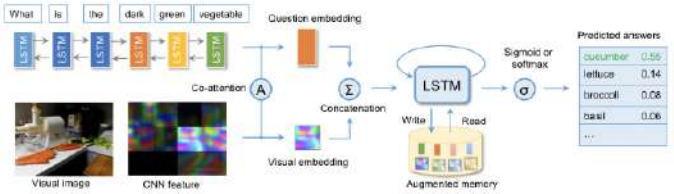
Hiroshi Fukui

Visual Question Answering with Memory-Augmented Networks

Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, Ian Reid CVPR2018, arXive: 1707.04968 875

概要

学習サンプルに少ないような質問に対しても回答ができるような手法を提案．ベースはMemory-Augmented Network (One-shot learningを導入したMemory Network)であり，記憶ブロックとAttentionの機能により，稀に発生する質問に対しても正確に回答をすることができる． VQA benchmark datasetとCOCOのVQAタスクで評価し，高い性能を示している．



Q: What is the dark green vegetable?
A: Cucumber (Ours) A: Broccoli [21] A: Lettuce [2]

新規性・結果・なぜ通ったか？

この手法の大まかな構造はMemory-Augmented Networkになっており，特徴抽出部分が質問文と画像特徴である．画像特徴はVGGやResNetの特徴マップを使用しており，質問文はLSTMの特徴ベクトルを使用している． この2つの特徴ベクトルは結合され，質問と画像特徴の2つのAttentionがそれぞれ与えられてAugmented memoryに格納される． そして，Augmented memoryを用いて最終的な回答が出力される． 提案手法では，右下図のように，稀に存在する困難な質問に対しても正確な回答を得ることができる．

コメント・リンク集

- 論文リンク

[#311]

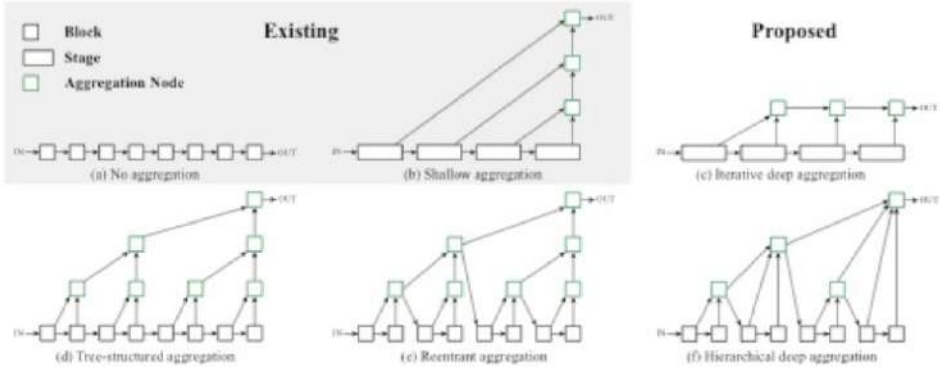
Hiroshi Fukui

Deep Layer Aggregation

Fisher Yu, Dequan Wang, Evan Shelhamer, Trevor Darrell CVPR2018, arXive: 1707.06484 272

概要

Deep Neural Networkにおける，層間の結合に関して様々な検討を行った論文．従来のネットワーク(ResNet, DenseNet, FCN, U-Net等)のスキップ結合は，”浅い”結合しか適用されていなかった． この論文では，より”深い” 結合をネットワークに取り入れ，少パラメータかつ高精度なネットワークモデルを構築している． 画像分類をはじめ，様々な認識タスクで実験を行い，高精度化を実現している．



新規性・結果・なぜ通ったか？

この論文では，右図のような4つのモデルを検討している(c~f)．(c)のようにシンプルに特定の層を集約して連鎖的に入力していくモデルから，(d~f)のように様々な層を集約して連鎖的に集約していくモデルを検討しており，上位層と下位層の層を効率的に伝播することで，認識精度を向上させている． また，(c)と(f)のモデルを組み合わせることで，より性能を向上させることも可能である． 画像分類，Fine-grained Recognition，物体検出，セマンティックセグメンテーションで実験を行っており，全ての認識タスクにおいて高い性能を示している．

コメント・リンク集

Deep CNNの次期モデルを検討しているような論文．結局，画像分類，検出，セグメンテーションではスキップ結合が重要であることを再確認できる．

- 論文リンク

[#312]

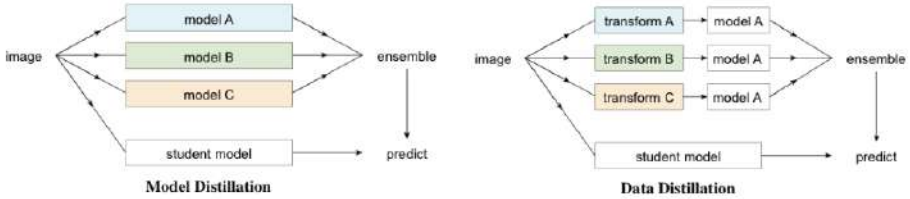
Hiroshi Fukui

Data Distillation: Towards Omni-Supervised Learning

Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, Kaiming He CVPR2018, arXive: 1712.04440 536

概要

ラベル付きとラベルなしデータを用いることで画像認識の精度を向上させるData Distillationを提案．この手法では，self-trainingとHinton先生のKnowledge distributionをベースに提案されている．この手法は，インターネット上のラベルなしデータを大量に学習できる． この論文では，Mask R-CNNによる人のKeypoint検出と，FPNをbackboneにしたFaster R-CNNによる物体検出で高精度化を実現している．(COCOをラベル付き，Sports-1M statistic framesとCOCO2017unlabel imagesをラベルなしデータとして使用．)



新規性・結果・なぜ通ったか？

一般的なラベルなしデータを扱うModel Distillationとは異なり，Data Distillationは1つのteacher modelとstudent modelを用いる．構造としては，1つの画像を複数の単純な変形を加え，それぞれの認識結果を得る． そして，それぞれの認識結果を統合し，統合した認識結果をラベルとしてstudent modelを学習する． ここで，学習に使用するラベルは”soft”なラベルではなく，”hard”なラベル．COCOをベースに実験をしており，ラベルなしデータを併用することで人のKeypoint検出と物体検出で高精度化を実現している．

コメント・リンク集

シンプルかつ少量データの学習にも応用できるできるので，今後これをベースにした手法が増えそう．

- 論文リンク

[#313]

Hirokatsu Kataoka

Actor and Observer: Joint Modeling of First and Third-Person Videos

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, Karteek Alahari CVPR 2018 (spotlight)

概要

一人称（First Person View; 頭部にカメラを装着して撮影）かつ三人称（Third Person View; 環境に設置したカメラから撮影）の視点から人物行動や操作している物体を撮影したデータセットCharades-Egoを提供する。一人称/三人称視点は互いに対応付けされており、実に157の行動カテゴリ、112人の実演、4,000の動画ペア、全8,000動画を保有するデータベースの構築に成功した。手法の側面ではTripletによる弱教師付き学習（Weakly-supervised Learning）により一人称/三人称から抽出した複数の特徴量を評価する枠組みActorObserverNetを提案する。さらには、三人称から一人称視点への知識転換（Transferring Knowledge）をZero-shot行動認識の枠組みで実行する。

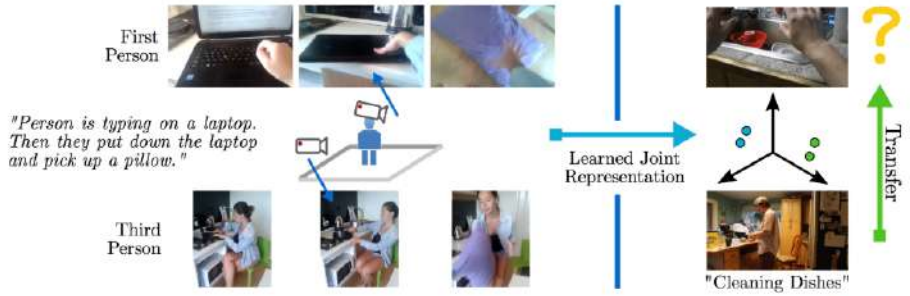


Figure 1: We explore how to reason jointly about first and third-person for understanding human actions. We collect paired data of first and third-person actions sharing the same script. Our model learns a representation from the relationship between these two modalities. We demonstrate multiple applications of this research direction, for example, transferring knowledge from the observer's to the actor's perspective.

新規性・結果・なぜ通ったか？

一人称/三人称は従来独立に撮影されて、それぞれのデータベースを構築して来たが、ここでは同時解析することにより行動に関するより詳細な考察（e.g. 間接的に行動を観察した方が良い vs. 操作している物体で行動を認識する方が良い）を行えるようにした。また、弱教師付き学習により特徴学習できるActorObserverNetを提案した。CVPRに通った理由はなんといってもデータベース（とそのベンチマーキング）、弱教師付き学習によるものである。

コメント・リンク集

Hollywood in HomesのようにAMT（クラウドソーシング）にてユーザがフリーで使用を許可した動画を収集するのはアリにしている。公開してフリーにしても良い人だけの動画を効率良く集める仕組みが今後流行ってくるか？（ただ日本だと難しいかも？）データベースに対するベンチマーキングは若干少ない印象を受けるが、データベースの意義自体が優れているため査読を突破したと思われる。

- 論文
- YouTube
- 著者
- Project/Database

[#314]

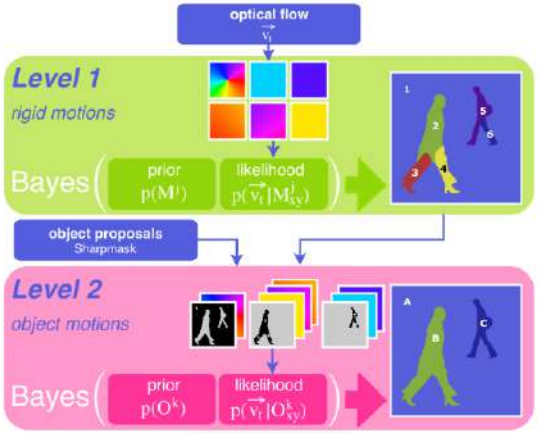
Hirokatsu Kataoka

The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation

Pia Bideau et al. CVPR 2018

概要

モーションセグメンテーションの問題を扱う。従来のモーションセグメンテーションは幾何的制約を設けることで効果的に動作をセグメントして来たが、高次なセグメントに失敗していた。一方でCNNについては従来方とは逆の特性があった。この両者の特性を活かして、両者にとって良いところどり（The Best of Both Worlds）することでモーションセグメンテーションの性能を向上させた。手法は図に示すようにオプティカルフローを用いた剛体の動き推定（Perspective Projection Constraints）、変形可能でより複雑な物体形状を推定できるようCNNによるセマンティックセグメンテーションを実行。物体のモーションモデルを形成するために、SharpMask（論文中文献35）による物体候補も導入し物体に関する知識を導入した。



新規性・結果・なぜ通ったか？

クラシカルなフローによる剛体モーション推定とCNNによる物体セグメンテーションを統合、両者の良い部分を引き出しているところが評価に値した。アブストラクト/図１が非常にわかりやすくこの２つで問題設定を把握できるところもグッド。

コメント・リンク集

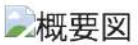
- 論文
- UMASS CV Lab.
- SupplementaryMaterial

Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present

Xi.Cheny, L.Mazx, W.Jiangzx, J.Yaoy and W.Liuz CVPR2018 arXiv:1803.11439

概要

encorder/decorderモデルにhiden stateと過去のhiden stateを再構成することによって隣接するhiden stateの接続を強化するためのARNetを導入.



従来手法問題点

- 従来のRNNのtrainとinferenceの間にはexposure biasと呼ばれる相違が存在する.
- decorderはの入力に依存する演算子を用いて，キャプション生成する.

結果・リンク集

- RNNにおけるtransition dynamicsの正則化を助け，シーケンス予測の不一致の緩和が見られた.
- ソースコードキャプション，イメージキャプションの両方で精度の向上が見られた.
- Paper
- github

[#316]

Hiroshi Fukui

Repulsion Loss : Detecting Pedestrian in a Crowd

Xinlong Wang, Tete Xiau, Yuning Jiang, Shuai Shao, Jian Sun and Chunhua Shen CVPR2018, arXive:1711.07752 1005

概要

群衆に頑健な歩行者検出法を提案．Faster R-CNNで群衆を検出したとき，歩行者同士の間にBounding Boxが出現しやすい． これは，Bounding Box回帰の誤差を算出する時に誤差を最小にしようとして歩行者同士の間にBounding Boxが発生してしまう． この現象を解決するために，新たにRepulsion Lossを導入し，群衆に対しても高精度な歩行者検出を実現している．

新規性・結果・なぜ通ったか？

Repulsion Lossの中身は， L1 smooth lossをベースにしたL_RepGTとL_RepBoxから構成されている．L_RepGTは，targetの歩行者付近から最も近いGTとの誤差を示しており，targetと最も近いGTにBounding Boxが検出されると誤差が大きくなるように誤差が設計されている． L_RepBoxは，複数のBounding Boxが特定の箇所に集中するように誤差を設定している． L_RepBoxの目的は，NMSの割合の影響を減らすためである． 歩行者検出のCaltech, CityPerson(Cityscape)でstate-of-the-artな性能を出しており，Pascal VOCにおいても有効であることを示している．

コメント・リンク集

歩行者検出のベンチマークにおいて非常に高い性能を示しており，ResNetベースのFaster R-CNNに対してDilated Conv.を導入する等のちょっとしたテクニックも色々導入されている．

- 論文リンク

[#317]

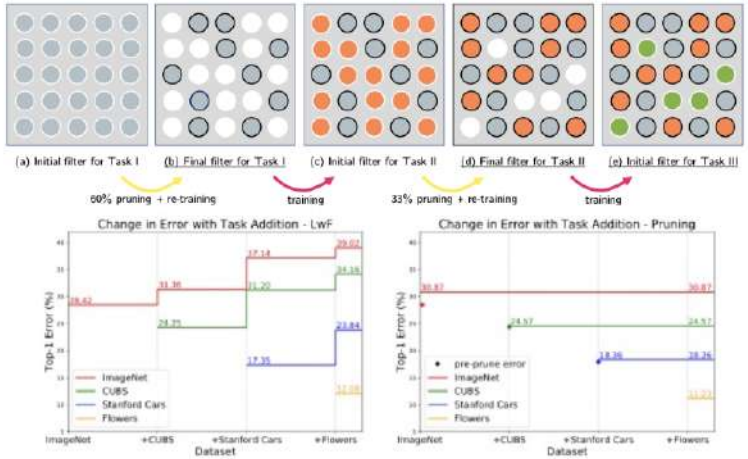
Hiroshi Fukui

PackNet : Adding Multiple Tasks to a Single Network by Iterative Pruning

Arun Mallya, Svetlana Lazebnik CVPR2018, arXive:1711.05769 1004

概要

複数のデータセットを1つのネットワークで学習する場合，通常は過去に学習したデータセットは段々と精度が低下していく．これは，全てのパラメータに対して更新するため，過去に学習したデータセットの特徴を抽出できなくなっていくのが原因である． この論文で着目していることは，大規模なネットワークは特定のパラメータは学習をサボる傾向があるところであり，このサボっているパラメータを使って効率よく学習させて複数のデータセットを学習させている．



新規性・結果・なぜ通ったか？

手法自体は非常にシンプルであり，特定のパラメータをプルーリング(右上図の白領域)して再学習する．そして，プルーリングしたパラメータのプルーリングを解放してパラメータをアップデートする． 特定のタスク(データセット)を学習した後は同じ要領でまたプルーリングと再学習を行う． 特定のパラメータを特定のタスクに割り当てるような学習をすることで，複数タスクに対応している． 結果としては，右図のようにタスクが追加されても性能がほとんど低下していない．

コメント・リンク集

単純な手法でありながら，非常に強力な手法．図2のインパクトがすごかった．様々な応用にも繋げれそう(Transfer Learning, Domain Adaptation等)

- 論文リンク
- コードリンク

[#318]

Hiroshi Fukui

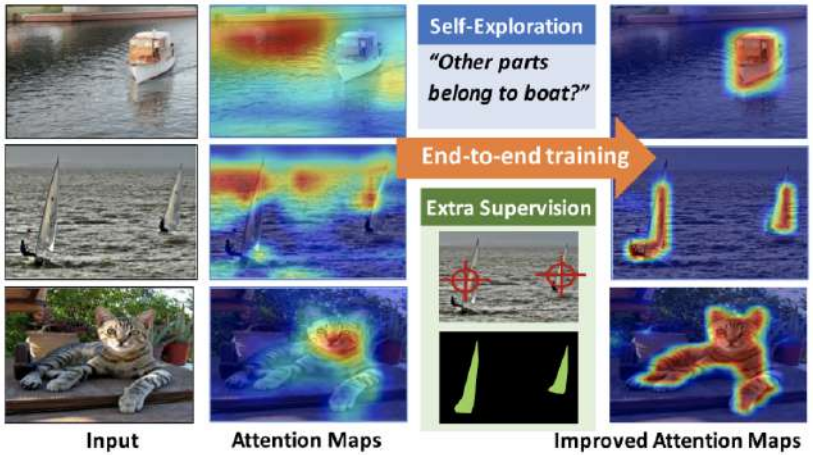
Tell Me Where to Look : Guided Attention Inference Network

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Yun Fu CVPR2018, arXive: 1802.10171 1247

概要

弱教師あり学習で得られる物体のローカライゼーションを高精度にする研究．方法としては2つ提案しており，

1. GAPのローカライゼーションを用いて物体の領域と背景の領域を明示的に学習させる方法と，
2. セマンティックセグメンテーションのラベルを用いて物体の詳細な領域を学習させる方法がある．セマンティックセグメンテーションと視覚的解釈に対する評価をしており，どちらのタスクも高い性能を示している．



新規性・結果・なぜ通ったか？

1)の方法では，2streamなCNNをベースにしており，入力はそれぞれ通常の画像と，GAPのローカライゼーションから物体領域を排除した画像を入力する．この処理により，物体と背景を明示的に学習できる． そして，セマンティックセグメンテーションでは， 1)のネットワークに加えて，セマンティックセグメンテーションのラベルと出力したAttention mapとの誤差を算出させることで，Attention mapを最適化させる． Pascal VOCのweakly-supervisedによるセマンティックセグメンテーションのタスクで評価し，高い性能を示している． また，発生するAttention mapの領域に対してオリジナルのデータセットを作成して評価している．

コメント・リンク集

- 論文リンク

[#319]

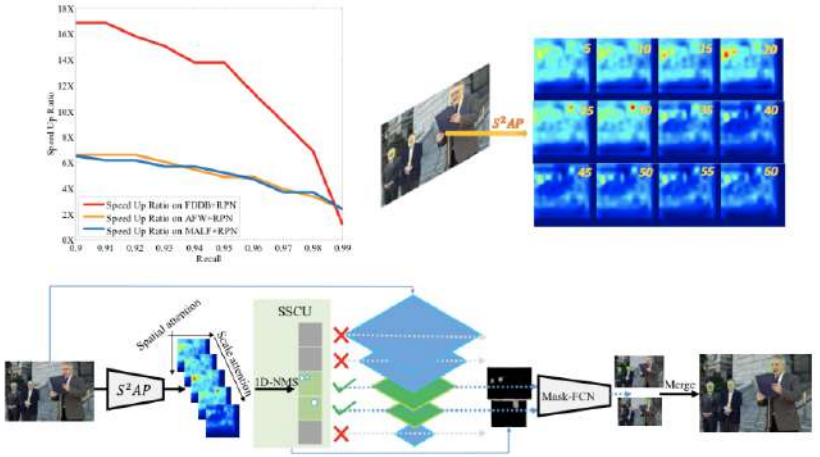
Hiroshi Fukui

Beyond Trade-off: Accelerate FCN-based Face Detector with Higher Accuracy

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, Yun Fu CVPR2018, arXive: 1802.10171 1003

概要

一般的な顔(物体)検出法(Faster R-CNN, FPN, SSD, YOLO等)は，Backboneな部分がFCNベースで構築されているため，各ピクセルを密に畳み込んで検出結果を出力する．しかし，顔検出では背景領域を大量に含んでおり，検出に必要な領域はごく僅かである． 本論文では，顔検出を効率化するために，2つのAttentionを適応して高速化を試みており，左上図のように高い性能を維持しつつ，4倍以上の高速化を実現している．



新規性・結果・なぜ通ったか？

本手法で適応しているAttentionは，右上図のようなspatial attentionとscale attentionである．spatial attentionは2次元上における顔の位置を示しており，scale attentionは出力されたスケールピラミッドから最適な特徴マップをAttentionで表現している．spatial attentionは2次元の位置のattentionから探索する領域を制限するために使用し，scale attentionは探索するスケールピラミッドを制限するために使用する． ネットワークは下図のようになっており，2つのAttentionにより背景と判定された領域は，マスクされた状態で後段のMask FCNに入力される． AFW, FDDB, MALFでstate-of-the-artな性能かつ，高速な検出が可能(最速で14.2ms)．

コメント・リンク集

Attentionを計算コスト削減に適応した物体検出法．顔検出や車載系の物体検出等の背景領域を多く含む問題設定では非常に効果的に使えそうな手法． (COCO, VOCではあまりコストに対しては言及していない)

- 論文リンク

[#320]

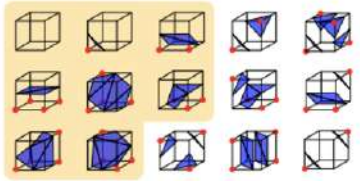
Ryota Suzuki

Deep Marching Cubes: Learning Explicit Surface Representations

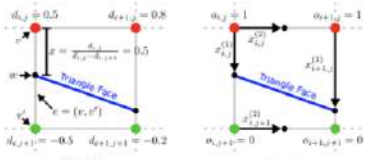
Y. Liao, S. Donné and A. Geiger CVPR2018

概要

既存の学習ベースの3D面推定方法は，End-to-Endでの学習ができないが，本研究では，end-to-endでの学習を可能にした．3D面推定手法の一つのマーチングキューブは微分不可．そこで，代替の微分可能定式化を行い，これを3DNNの最終層として追加する． また，疎な点群で学習が行えるようにロス関数群を提案． サブボクセル精度での3D形状を推定可能であることを確認した． 本モデルは形状エンコーダ・推論と組み合わせられる柔軟さがある．



マーチングキューブで使われる
ボクセル面表現



通常　提案

新規性・結果・なぜ通ったか？

End-to-endで行われたものはない．適用範囲が広そう．

コメント・リンク集

- 論文

[#321]

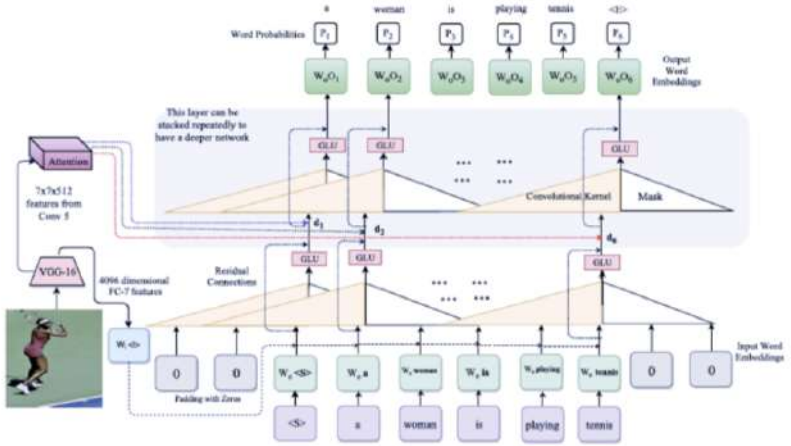
KotaYoshida

Convolutional Image Captioning

J.Aneja, A.Deshpande and A.Schwing CVPR2018 arXiv:1711.09151v1

概要

近年，条件付き画像生成や機械翻訳において畳み込みニューラルネットの功績は大きい，これを画像キャプションに応用してみた．ベースラインであるLSTMモデルと同等の精度を示し，パラメータ数ごとの学習時間の短縮をすることができた．



従来手法の問題提起

- RNNは学習プロセスが逐次的
- LSTM，RNNは画像の分類精度が低い

結果・リンク集

- RNNとCNNのアプローチを分析し，CNNを用いたアプローチは出力確率分布のエントロピーの増大，単語予測精度の向上，消失勾配の影響の低下を示すことができた．
- 論文
- github

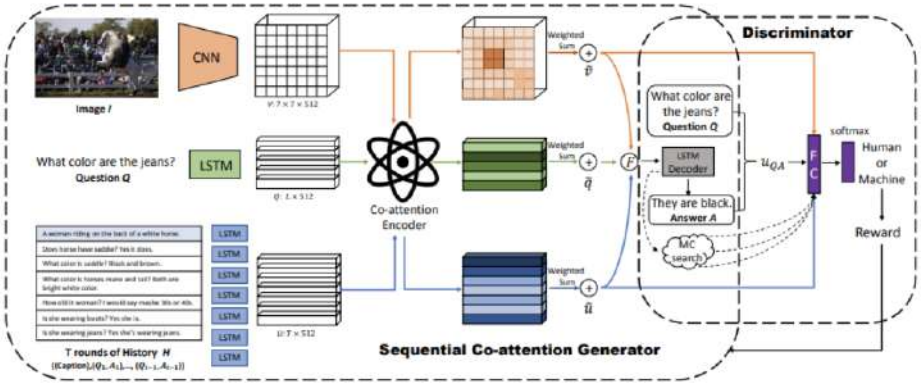
[#322] Yue Qiu

Are You Talking to Me? Reasoned Visual Dialog Generation through Adversarial Learning

Qi Wu, Peng Wang, Chunhua Shen, Ian Reid, Anton Van den Hengel CVPR 2018 741

概要

・強化学習とGANを用いたVisual Dialog回答文を自動生成する手法の提案．・従来のVisual Dialogシステムは画像とDialog履歴に基づきMLEにより回答文の予測を行う．こういった手法では回答文が短い，バリエーションが少ないなどの問題点がある．そこで， co-attentionを利用したジョイントで画像， Dialog履歴をreasonできる回答文生成器を提案した．提案モデルはsequential co-attention生成器と回答文が“human”からか“生成された”かを弁別できる弁別で構成される．



新規性・結果

・GANを用いた提案手法はVisual Dialogタスク従来の学習データの不足，簡潔な回答しか生成できないなどの問題点を改善した．・attentionをGANと組み合わせ， 生成回答文のinterpretabilityを向上した ・ VisDial データセットにおいて,従来の手法より高い精度を達成した．

リンク集

・interactive環境でVisual Dialog回答文の生成ができたら更に様々な場面で応用できる

- 論文

[#323]

Yue Qiu

Density Adaptive Point Set Registration

Felix Järemo Lawin, Martin Danelljan, Fahad Khan, Per-Erik Forssen, Michael Felsberg CVPR 2018 464

概要

・三次元センサーにより取得したPoint Set の密度の変動を対応できるPoint Set Registrationの手法を提案した．・従来の三次元センサー(例Lidar)により取得できるPoint Setの密度が均一ではない，一方，従来の確率的Point Set Registrationの手法は高密度の部分を対応させ，低密度の箇所の対応が重視されない問題点がある．提案手法はシーン構造の確率分布をモデリングすることにより，密度の変化にロバストに対応できる． ・提案手法は3次元シーンの構造及びフレーム間のカメラ移動量を同時にモデリングし， EMベースなフレームワークに基づきKL divergenceを最小化によりパラメータの最適化を行う．

(a) CPPSR (b) DARE-color

	Avg. inlier error (°)	Failure rate (%)
CPPSR	2.57±0.837	87.4
CPPSR-*eub*	1.63±0.807	20.9
JRMPC	2.38±1.01	92.1
JRMPC-*eub*	2.13±0.83	38.6
DARE-color	1.26±0.61	14.5
DARE	1.84±0.80	36.0

新規性・結果

・Lidarを用いたregistrationシステムのPoint Setの密度変化をロバストで対応できた． ・ DAR-ideal、 VPS and TLS ETH datasetsなどのLidarデータセットで従来の確率的マルチビューRegistration手法より良い性能を達成した．

リンク集

・deep learningを用いていない手法

- 論文

Ryota Suzuki

pOSE: Pseudo Object Space Error for Initialization-Free Bundle Adjustment

J. Hong and C. Zach CVPR2018

概要

カメラ姿勢推定，3次元復元に使われるバンドル調整では，適した初期値を与える必要があるが，初期値を与える必要を無くす提案をする．

アフィンバンドル調整問題においては，任意の初期化から到達可能な使いやすいminimaがあることが知られているが，その主な要因は，収束のワイドな領域を持つことで知られているVariable Projection（VarPro）法の導入によるものである．本研究では Pseudo Object Space Error（pOSE）を提案する．これは，アフィンと射影のモデルのハイブリッドで表現される複数カメラにおける目的関数である． この定式化で，VarPro法に適したバイリニア問題構造となり，真の射影復元と近い3D復元結果を得られる． 実験では，ランダムな初期化から高い成功率で正しい3D復元を得られることを確認した．

新規性・結果・なぜ通ったか？

ランダム初期値でもメトリックの正しい3D復元が行える．

コメント・リンク集

- 論文

[#325]

Kazuki Inoue

Finding Tiny Faces in the Wild with Generative Adversarial Network

Yancheng Bai, Yongqiang Zhang, Mingli Ding, Bernard Ghanem CVPR 2018

概要

GANを用いて画像中の顔を検出する研究。検出が難しい顔として小さくかつボケている顔が挙げられるが、これらの顔をGANによって高解像度かつはっきりとした顔にすることで検出精度を向上させる手法を提案。 generatorは高解像度にするsuper resolution network(SRN)と顔の詳細な情報を復元するrefinment network(RN)を結合したネットワークである。 discriminatorはVGG19であり、ロスとしてデータセットの顔/generatorによる顔、顔/顔ではないモノを同時に行うロスを導入。 またよりはっきりとした顔を生成するために、generatorのロスとして物体識別のロスを導入。

Figure1. The detection results of tiny faces in the wild. (a) is the original low-resolution blurry face, (b) is the result of re-sizing directly by a bi-linear kernel, (c) is the generated image by the super-resolution method, and our result (d) is learned by the super-resolution (×4 upscaling) and refinement network simultaneously. Best viewed in color and zoomed in.

新規性・結果・なぜCVPRに通ったか？

- GANによって画像中の顔から高解像度かつはっきりとした顔を生成することで高精度な顔検出手法を提案。
- GANの導入による精度の向上、導入したロスの有効性を確認している。
- state-of-the-artと比較して、最も高い検出精度を達成

コメント・リンク集

- 検出精度が非常に高く、データセットではアノテーションし忘れている顔すらも検出してしまい、これによって精度が悪いように見えてしまうと主張している。
- テスト時も学習時と同様に画像全体ではなくROIを与えているため、実行時間はそれなりにかかりそう。
- 論文
- Project page

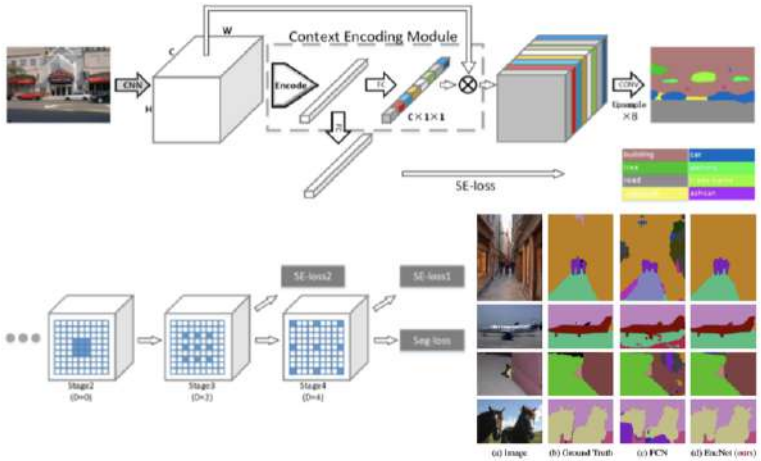
[#326] Yue Qiu

Context Encoding for Semantic Segmentation

Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, Amit Agrawal CVPR 2018 893

概要

・コンテキスト情報の抽出を利用したセマンティックセグメンテーションの効率を上げられるContext Encoding Moduleを提案した．・従来の階層式シーンの高レベルから低レベル特徴の抽出を行うネットワーク(eg. PSPNet)にはシーンのコンテキスト情報の抽出がexplicitではない問題点があり，従来のグローバル特徴抽出ネットワークの知識から，シーンのコンテキスト情報を抽出することにより，セマンティックセグメンテーションの効率を上げられるモジュールを提案した． ・具体的には：Encodingによりシーンのコンテキスト情報をキャプチャーし，クラス依存の特徴マップを選択的に強調表示できるContext Encoding Moduleを提案した； Semantic Encoding Loss (SE-loss)を提案した； Context Encoding Moduleを利用したセマンティックセグメンテーションネットワークEncNetを提案した



新規性・結果・なぜ通ったか？

・ PASCAL VOC 2012において85.9% mIoUを達成した・提案ネットワークをCIFAR-10 datasetに応用し，14層だけのネットワークで100層超えのネットワークと同じレベルの精度を実現した

コメント・リンク集

・シンプルなネットワークでstate-of-the-artな精度を実現したので，将来的に広く用いられそう

- 論文

[#327]

Shusuke Shigenaka

Video Based Reconstruction of 3D People Models

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, Gerard Pons-Moll CVPR 2018

概要

人間が動いている単眼のRGB映像から、正確な3次元物体モデルと任意の人物テクスチャを得る研究。仮想現実や拡張現実、監視やゲームなどの人間の追跡にはアニメーション可能な人間行動の3Dモデルが必要である。この研究では、動的な人間のシルエットに対応するシルエット形状を見つけ出し、テクスチャや骨格を推定して、アニメーション可能なデジタルダブルを作成することができる。

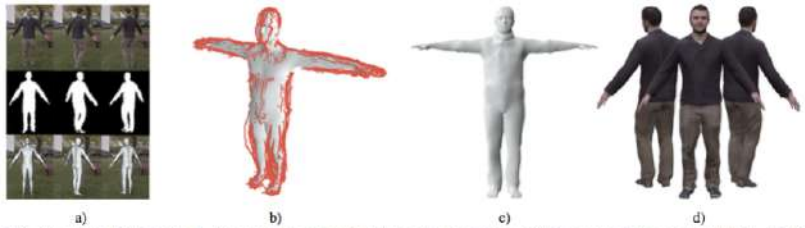


Figure 2. Overview of our method. The input to our method is an image sequence with corresponding segmentations. We first calculate poses using the SMPL model (a). Then we unpose silhouette camera rays (unposed silhouettes depicted in red) (b) and optimize for the subjects shape in the canonical T-pose (c). Finally, we are able to calculate a texture and generate a personalized blend shape model (d).

Figure 6. Our results on image sequences from BUFF and D-FAUST datasets. Left we show D-FAUST: (a) ground truth 3D scan, (b) consensus shape with ground truth poses (consensus-p), (c) consensus-p heatmap, (d) consensus shape (consensus), (e) consensus heat-map (blue means 0mm, red means $\geq$ 2cm). Right we show textured results on BUFF: (a) ground truth scan, (b) consensus-p (c) consensus.

手法・新規性・結果

(a). SMPLモデルを用いてポーズを計算(b). シルエットの赤で描かれていないシルエットを取り除く (c). 正規のTポーズで被写体の形状を最適化 (d). ティクスチャを計算しパーソナライズされた好みの形状を生成 ・単眼のRGBビデオから髪や衣服を含む現実的なアバターを抽出 ・被服を含む4.5mmの精度で人体形状を再構成

リンク集

link

[#328]

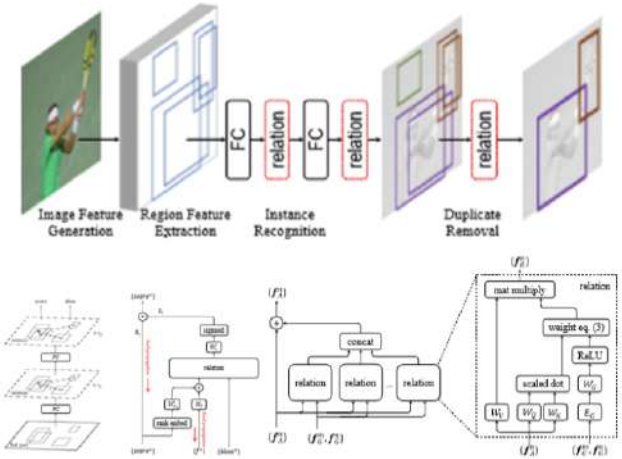
Yue Qiu

Relation Networks for Object Detection

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, Yichen Wei CVPR 2018 439

概要

・マルチオブジェクトのアピアランス特徴及び幾何情報間の関係を取り扱える，様々なタスク（物体検出，VQAなど）に用いられるObject Relation Moduleを提案した．・最近attentionに関する研究が発展し，著者たちがattentionモジュールがelement間の依頼性を学習できる面から，物体検出に応用できるアテンションモジュールを提案した．・提案モジュールを物体検出の2つの段階に応用できる：インスタンス認識段階で提案モジュールによりオブジェクト間の関係を習得でき，精度を上げられる；duplicate removal段階で提案モジュールにより有効的に物体領域を抽出できる．



新規性・結果・なぜ通ったか？

・従来の物体検出手法は物体ごとに推定を行い，物体間の関係を利用しない．提案手法はObject Relation Moduleを提案し，物体間の関係を学習することで，物体検出の精度を更に向上した．

コメント・リンク集

・提案モジュールが付加の監督信号不要，既存なネットワークに追加しやすい特徴があるため，様々なタスクでの応用が期待される

- 論文

[#329]

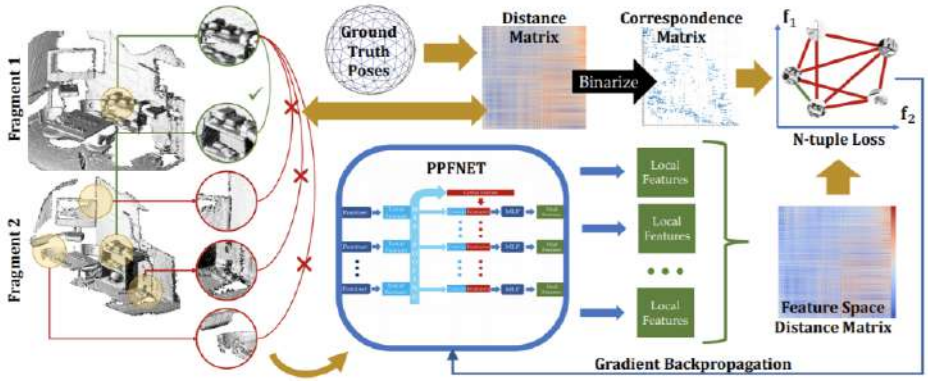
Shuichi Akizuki

PPFNet: Global Context Aware Local Features for Robust 3D Point Matching

Haowen Deng, Tolga Birdal, Slobodan Ilic CVPR2018 45

概要

点群データから直接3Dの局所特徴量を抽出するネットワークを提案．N-Tuple loss(Triplet lossの拡張)によって， 対応点間の特徴量が近く，それ以外の特徴量間の距離が遠くなるような変換を学習する． PPFNetの入力は局所パッチ内の点の座標，法線，Point Pair Featureをまとめたデータ． ネットワークの内部ではPointNetを利用する． 大域的な情報を得るために，各パッチから取得した局所特徴量を Max poolingによって大域特徴量化し，局所特徴と結合する工夫も入れている．



新規性・結果

局所特徴量を生成するネットワークを構築した点，N-Tuple lossによる学習法を提案した点が新しい． キーポイントマッチングのベンチマークでRecall rateが向上． オーバーラップが少ないシーンでのレジストレーションも可能になっている．

コメント・リンク集

- Paper

[#330]

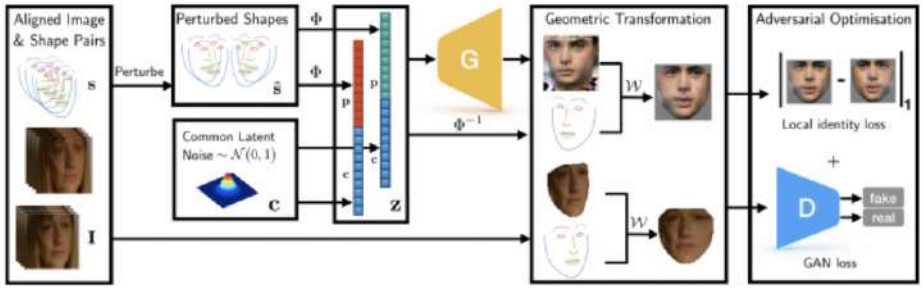
Yuta Matsuzaki

GAGAN: Geometry-Aware Generative Adversarial Networks

Jean Kossaifi, Linh Tran, Yannis Panagakis and Maja Pantic CVPR2018

概要

既存のGANでは考慮されていなかった形状や位置といった幾何学的情報をGANの生成プロセスに組み込んだGeometry-Aware Generative Adversarial Networks (GAGAN) を提案．具体的にGAGANでは，ジェネレータで統計的情報な形状モデルの確率空間から潜在関数をサンプリングする．次にジェネレータの出力値を微分可能な幾何学変換を介して標準座標系にマッピングすることで，物体の形状や位置といった情報を強制し，生成を行う．



新規性・結果・なぜ通ったか？

- GAGANのような幾何学的情報を考慮した生成モデルはなく，GAGANが初
- 入力画像の属性の形状に合わせて，画像を生成することが可能

コメント・リンク集

今後は，(i)より大きな画像の生成，(ii)アフィン変換によって起こりうる変形を緩和するより複雑な幾何学的変換の探索およびそれによるGAGANの拡張，(iii)顔のランドマーク検出のための従来CNNアーキテクチャの拡張に取り組む予定

- 論文

[#331]

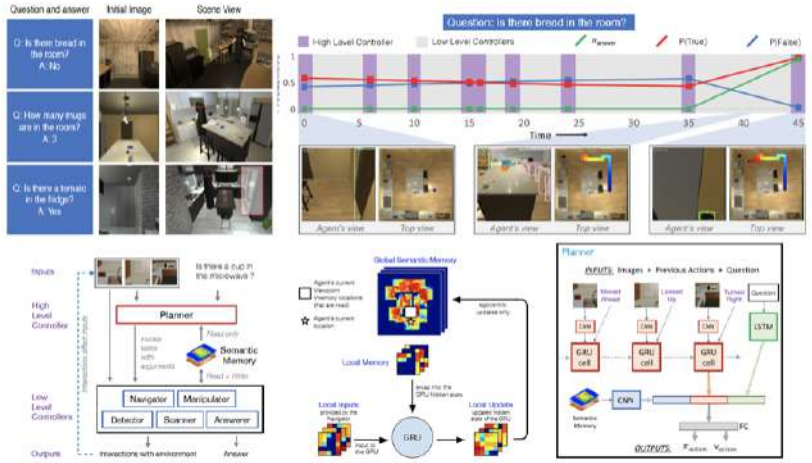
Yue Qiu

IQA: Visual Question Answering in Interactive Environments

Daniel Gordon, Ali Farhadi, Aniruddha Kembhavi, Dieter Fox, Mohammad Rastegari, Joe Redmon CVPR2018 533

概要

・新たな問題設定ー動的環境とインターアクトしながら視覚質問に答える(IQA)を提案した．・具体的には，IQAには4つの設定がある：環境でナビゲートする能力；環境中のオブジェクト，アクション及びアフォーダンスの理解；環境中のオブジェクトとインターアクトする能力；質問文に応じで環境での行動を計画する能力．・提案の問題設定を解決するために，階層的マルチレベルで行動計画及びコントロールするネットワークHIMN及び空間的かつセマンティックなメモリを実現できる新たなrecurrent layer形式Egocentric Spatial GRUを提案した．・更に，75000質問及びCGシーンを含んだデータセットIQUAD V1を提案した．



新規性・結果・なぜ通ったか？

・従来のVQAタスクをCGシーンでの自己ナビゲーションと組み合わせた新たな問題設定を提案した．・IQUAD V1で従来の手法よりstate-of-the-artな精度

コメント・リンク集

・従来のVQAタスクに更に環境での探索および環境中オブジェクトとのインターアクトを取り入れ，従来の問題設定より一層現実に近づいている．・質問文の自動生成にも応用できそう ・特に色々なタスクを取り扱えているので，技術の面では向上する空間がありそう

- 論文

[#332]

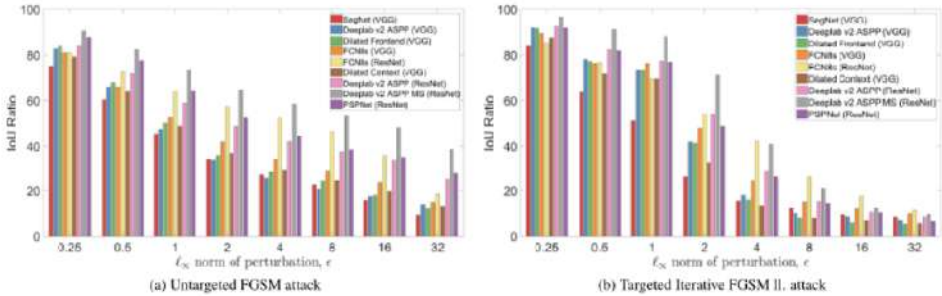
kodai nakashima

On the Robustness of Semantic Segmentation Models to Adversarial Attacks

Anurag Arnab, Ondrej Miksik and Philip H.S. Torr CVPR 2018

概要

adversarial attackに対するロバスト性の評価を, semantic segmentationにおいてstate-of-the-artな性能を持つネットワークを用いて実験した.Pascal VOCとCityscapesのデータセットに対して, FGSM, Interative FGSM, FGSM II, Interative FGSM IIで攻撃したときのIoU Ratioによりロバスト性を評価した.



(a) Untargeted FGSM attack (b) Targeted Iterative FGSM ll. attack

新規性・結果

- ResNetをバックボーンに持つネットワークがロバストであることがわかった. 中でもDeeplab v2が最もロバスト.
- multi-scale processingやmean field CRFによりロバストになる.
- 画像分類の分野で一般的なロバスト性やモデルサイズについての知識がsemantic segmentationでも有用とは限らない.

リンク集

- 論文URL

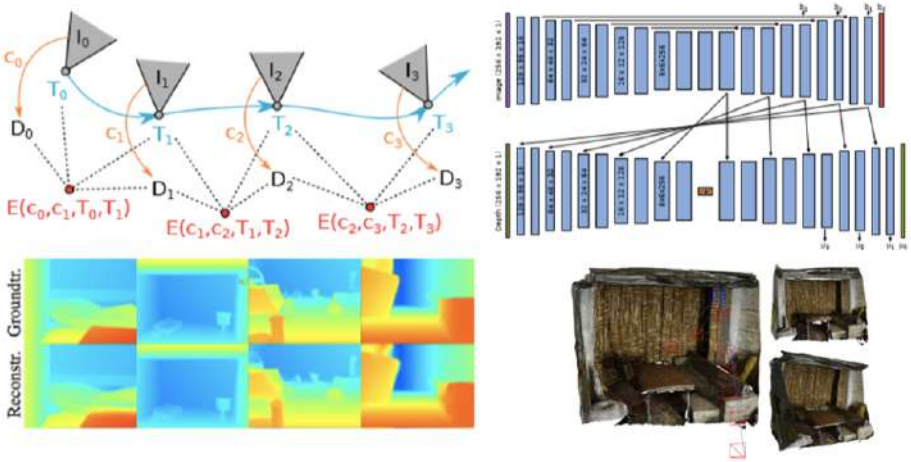
[#333] Yue Qiu

CodeSLAM --- Learning a Compact, Optimisable Representation for Dense Visual SLAM

Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, Andrew Davison CVPR 2018 288

概要

・RGB画像の強度データと少数のパラメータを条件に，ほぼリアルタイムで行えるデンスなシーン幾何を推定手法を提案した．・提案手法UNet構造により強度画像の特徴抽出を行い，更に抽出特徴をauto-encoder構造を用いたデプス情報推定ネットワークに入力することで階層的にデプス情報推定を行う．また，カメラ移動中得られるマルチフレームに対し，フレームごとのデプス推定及びフレーム間のカメラモーションをジョイントで最適化を行う．



新規性・結果・なぜ通ったか？

・デンスなデプス情報推定を行うことでSLAMシステムの更なる精度向上できると宣言した．・初めてのほぼリアルタイムで行えるカメラモーションとシーンのデンス幾何をジョイントで推定する研究である．

コメント・リンク集

・著者たちは将来のワークとして，提案手法をリアルタイムでデンスなSLAMシステムの構築に拡張すると指摘し,将来的な研究を期待している．

- 論文

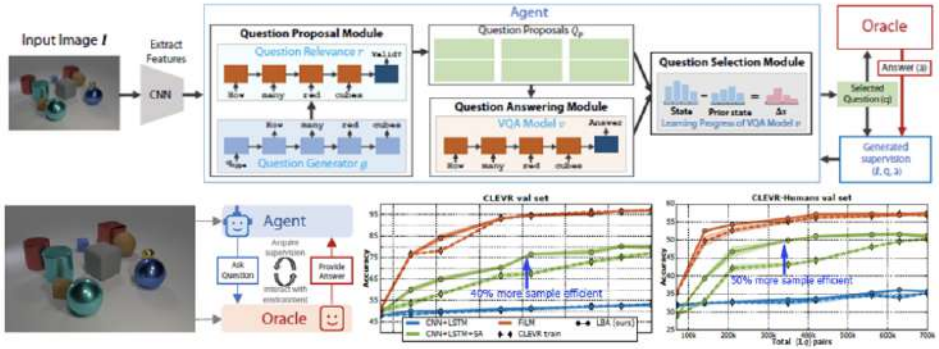
[#334] Yue Qiu

Learning by asking questions

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, Laurens van der Maaten CVPR 2018 3

概要

・VQAタスクに用いられる新たなインターアクティブ学習フレームワークを提案した．・提案フレームワークは入力画像から，question proposal moduleにより問題集を生成し，画像との相関性を基準に問題集をフィルタリングし，残った問題をVQAにより解く．予測した答え，自己の知識及び過去の知識から質問を1つ選び，oracleにより答える．・提案フレームワークにより，効率高い学習サンプルを得られる．また，従来のVQAネットワークで用いられるstate-of-the-artな問題集を生成できる．



新規性・結果・なぜ通ったか？

・従来のあらゆるフレームワークは学習データから学習を行う．この論文で，質問文の自動生成できる及び質問を選択する構造を導入し，自動的でインターアクティブで環境から情報を獲得することを可能にした．・実験を通し，提案手法により質問を選択する規制がsampleの効率を高められる．（従来と同じ精度の場合，学習データ量を40%減らせる）

コメント・リンク集

real-worldバージョンのLBAシステムが実現されたら，機械で学習することは更に人の学習システムに近づく．

- 論文

Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking

Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, Ming-Hsuan Yang CVPR 2018 1353

概要

Spatially Regularized Discriminative Correlation Filters (SRDCF)に空間正則化を導入した一般物体追跡手法Spatial-Temporal Regularized Correlation Filters (STRCF)を提案. SRDCFは複数学習画像を利用するため, 計算量が大きくなってしまうことに着目し, 単一学習画像に対するSRDCFにonline Passive-Aggresive learningの考えに基づいて時間正則化を導入. STRCFはADMMで直接解くことができるため, DCFの高速性を保持したまま高い精度で追跡が可能となっている.

新規性・結果

- 単一学習画像に対するSRDCFに時間正則化を導入することで, 複数学習画像に対するSRDCFを近似したSRTCFを定式化
- online Passive-Aggresive learningを拡張することで, STRCFは大きな見た目の変化に対して頑健である
- SRTCFはADMMを用いて, 3つの部分問題に帰着させ, Eckstein-Bertsekas条件を満たし, 大域的最適解への収束性を保証している
- OTB-2015, Temple-Color, VOT-2016データセットにおいてSRDCFより精度も計算速度も向上させた

コメント・リンク集

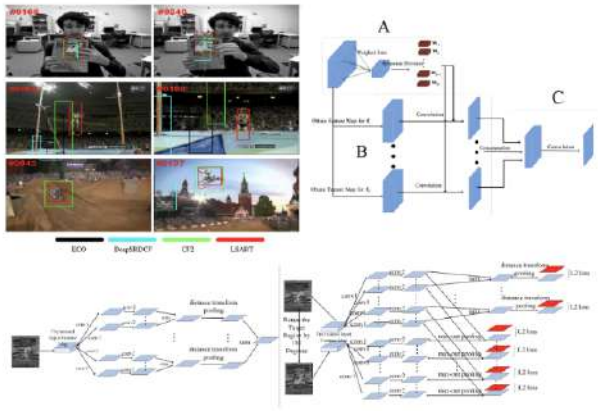
- 論文
- コード

Learning Spatial-Aware Regressions for Visual Tracking

Chong Sun, Huchuan Lu, Ming-Hsuan Yang CVPR 2018 1676

概要

一般物体追跡手法の二大手法であるカーネルリッジ回帰（相関フィルタを含む）とCNNのハイブリッドな手法を提案した.カーネルリッジ回帰は全体的な情報に,CNNは局所的な情報に注目するように設計している.それぞれの導入がどの精度向上に結びついているかも検討している.



新規性・結果

- cross-patch similarityを用いたカーネルリッジ回帰モデルを提案し,それをニューラルネットに再定式化.
- spatially reguralized kernelとdistance transform pool layerを用いて,出力の各チャンネルが特定の領域に反応するようなCNN提案.
- 提案したカーネルリッジ回帰とCNNを相補的に用いることで,OTB-2013,OTB-2015,VOT-2016データセットでstate-of-the-artな精度を達成.

リンク集

- 論文1
- 論文2

[#337] Yue Qiu

Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering

Nguyen Duy Kien, Takayuki Okatani CVPR 2018 739

概要

VQAタスクに用いられるattentionメカニズム“Dense Co-attention Network”(DCN)を提案した．DCNはfully対称的で，階層的にスタックできるため，マルチステップで視覚及び言語特徴のインターアクションを可能にする．具体的には，まず言語から画像の注目マップ及び画像から言語の注目マップを生成し，そして連結によりマルチモデルの特徴を融合する（dense co-attention layer)．そして階層的にdense co-attention layerをスタックにより，さらにマルチモデル特徴を深く探る．

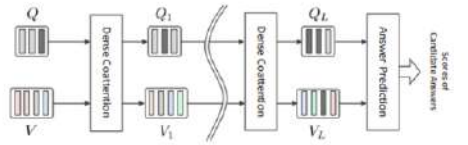


Figure 1: The global structure of the dense co-attention network (DCN).

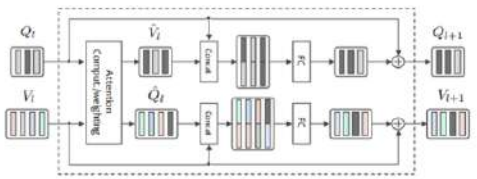


Figure 2: The internal structure of a single dense co-attention layer of layer index $l+1$.

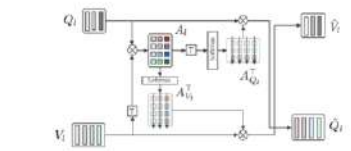


Figure 3: Computation of dense co-attention maps and attended representations of the image and question.

新規性・結果

・従来のattention for VQAタスクより，有効的でデンスな視覚と言語モデルの特徴の融合メカニズムDCN（構造的にも簡潔で拡張しやすい）を提案し，将来の様々なVQAタスクに用いられる．・VQA, VQA2.0データセットで2017 VQA優勝したモデルより良い精度を達成した． ・定性的な実験により，提案モデルが有効的にattentionを抽出できることを証明した

リンク集

- 論文

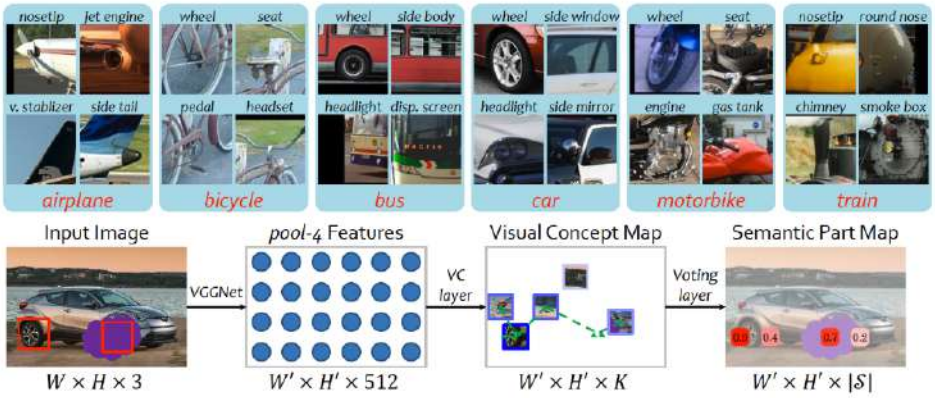
Kensho Hara

DeepVoting: A Robust and Explainable Deep Network for Semantic Part Detection under Partial Occlusion

Z. Zhang et al. CVPR 2018

概要

画像中から物体のパーツ（車のタイヤなど）を検出するための新しい手法を提案．投票ベースの手法でオクルージョンへの頑健性を持つ．Visual ConceptというMid-levelな特徴をベースにして，個々のMid-level特徴から推定されるパーツの位置推定結果を積み重ねていくことでパーツを検出する．Visual Conceptの検出とそれに基づく投票処理はConvolutionによって実装されており，End-to-Endでの学習が可能になっているところがポイント．Faster-RCNNといった物体検出アプローチよりもオクルージョンに頑健なことが実験的に確認できている．



新規性・結果

- CNNベースのVotingによるオクルージョンに頑健なパーツ検出手法を提案
- Visual Conceptの検出から投票までConvolutionで実装
- 人工的なオクルージョン環境下での有効性を確認

コメント・リンク

- 投票処理までConvolutionで表現されているのが面白い
- 論文
- Supplementary Material

Tomoyuki Suzuki

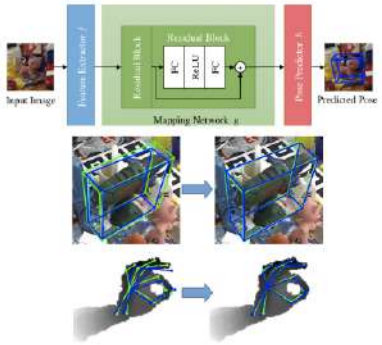
Feature Mapping for Learning Fast and Accurate 3D Pose Inference from Synthetic Images

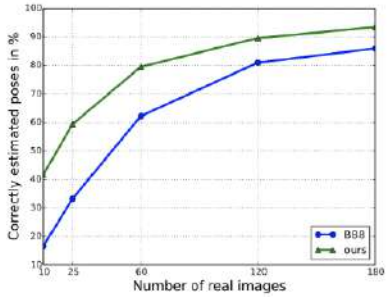
Mahdi Rad et al. CVPR 2018

概要

合成データを利用した、6D pose estimationとdepth based 3D hand pose estimationの研究。

埋め込み空間内で、合成データから実データへのマッピング関数を学習する。その関数の学習のためには実データに対応する(grand truthが同じ)合成データが必要であるので、教師あり実データがある程度あることが前提としてある。





手法

残差構造を持つmapping netを対応するペアを用いて学習する。従来のドメイン適応手法と比較しても提案手法の精度が良く、適応の有無による性能の差も非常に大きい。

メモ・リンク

手法としてはかなりstraight forwardな印象。実データの量を変化させた時の精度変化の結果はあったが、合成データの量を変化させた時の精度変化が気になる。

- 論文

[#340]

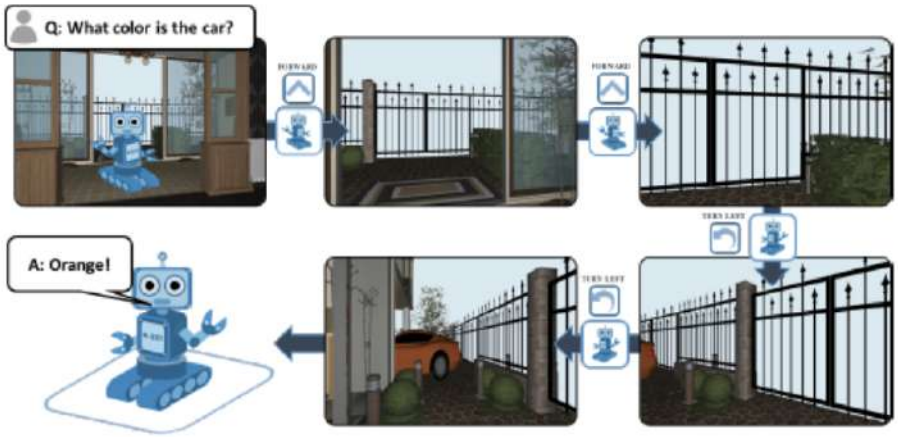
ShintaroYamamoto

Embodied Question Answering

Abhishek Das et al. CVPR 2018

概要

3次元空間において、エージェントに質問の答え（例：車の色は？）を探させる研究。初期位置における視覚情報だけでは答えに行きつかないためにエージェントは移動しながら答えを探していく。 エージェントの移動には、どの方向（forward, rightなど)に進むかを決定するplannerとどこまで進むかを決定するcontrolerによって行う。 目的地(正解が分かる場所)にたどり着いた時点で、最後の5フレームを用いて172の選択肢から正解を出力する。



新規性・結果

LSTMを使った場合の方が目的地により近付けるという結果が得られた。強化学習なしのものは目的地により近づいている一方、ファインチューニング＋強化学習の方が正解率は高いという結果となった。 また、最短経路を与えてVQAによって答えさせる場合でも精度が悪く、答えを導くにあたってどの方向から目的地に近づくかも重要であるということが分かった。

リンク集

- プロジェクトページ

Tomoyuki Suzuki

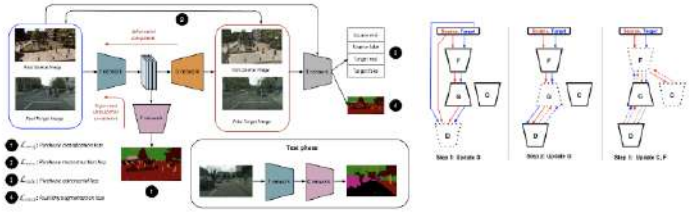
Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation

Sankaranarayanan et al. CVPR 2018

概要

GANによる画像生成の枠組みを中間的に取り入れることでSemantic segmentationにおけるドメイン適応を行う研究。

従来の特徴ベクトルに対する敵対的学習によって埋め込み空間におけるdomain gapを縮める手法に対して、この研究では特徴ベクトルから画像を復元し、その画像が識別器によってどのドメインからの復元か識別できないように埋め込み関数を学習させる。 合成データからのドメイン適応で最も良い精度を達成。



Type	Variants	Description
Within-domain	$\mathcal{L}^s_{adv\_D}$	Classify real source input as *src-real*; fake source input as *src-fake*
	$\mathcal{L}^s_{adv\_G}$	Classify fake source input as *src-real*
	$\mathcal{L}^t_{adv\_D}$	Classify real target input as *tgt-real*; fake target input as *tgt-fake*
	$\mathcal{L}^t_{adv\_G}$	Classify fake target input as *tgt-real*
Cross-domain	$\mathcal{L}^s_{adv\_F}$	Classify fake source input as real target (*tgt-real*)
	$\mathcal{L}^t_{adv\_F}$	Classify fake target input as real source (*src-real*)

手法

Source(S)は教師ありデータ、Target(T)は教師なしデータ。学習のフローは以下である: (1)識別器(D)は入力画像に対してpixel-wiseにsource real(SR), source fake(SF), target real(TR), target fake(TF)の4値分類を学習。(2)生成器(G)は入力特徴ベクトルからDによってSからの特徴はSRに、 Sからの特徴はTRに分類されるよう学習。 (+入力との担保を取るL2Loss)(3)埋め込み関数(F)はSからの入力はTRに、Tからの入力はSRに分類されるように学習。さらにSからのサンプルに対してはFからの特徴マップを入力としてsegmentation taskを解くCNNを学習。

メモ・リンク

論文内にこの手法がうまくいく理由の裏付け的実験や考察が詳細にはなかったが、特徴量から画像再生成を行うことによる入力情報の保存とS/T間の敵対的学習による分布の混合が一つのフローで行えていることが効いているように思えた。実際特徴量に対するS/T間の敵対的学習のみの場合よりも大きく精度が向上している。

- 論文

[#342]

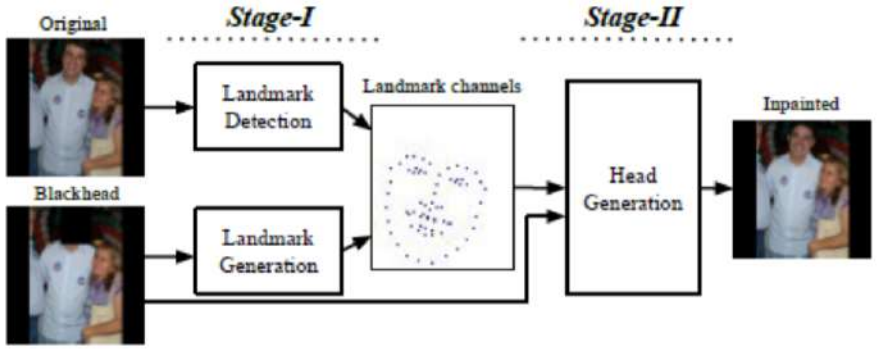
ShintaroYamamoto

Natural and Effective Obfuscation by Head Inpainting

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, Mario Fritz CVPR 2018

概要

SNSなどで共有された画像には、プライバシー保護の問題が生じる。プライバシー保護のために顔領域にぼかしや黒塗りなどの処理がされることが多いが、画像としては不自然さが残ってしまう。 そこで、塗りつぶされた領域に顔を挿入することで自然な画像ではあるが別人のためプライバシーを保護できる画像を生成する。 提案手法は、特徴点検出（生成）と顔の挿入の2つのステップに分かれる。 特徴点検出（生成）では、オリジナルの顔画像が存在する場合は既存の特徴点検出によって特徴点を検出する。 対称の画像が既に黒塗りされているなどで特徴点検出ができない場合は、GANによって特徴点を生成する。 次のステップでは、黒塗りされている顔画像と特徴点を入力し、黒塗りされた領域に顔の挿入を行う。



新規性・結果

特徴点生成器は、GANによって生成することで正解値とのノルム最小化よりも高い精度で生成することを可能にした。画像に対する処理としてぼかしと黒塗りを比較したところ、ぼかしは顔の情報が一部残るため高い精度での生成が可能である一方、元の人物の情報は黒塗りよりも多く残ることが分かった。 また、顔の形状にも個人性が含まれるためオリジナル画像から検出した特徴点よりもGANによって生成した特徴点を使用した方が個人性は損なわれることが分かった。

リンク集

- 論文

[#343] Ryota Suzuki

Augmenting Crowd-Sourced 3D Reconstructions using Semantic Detections

T. Price, J. L. Schonberger, Z. Wei, M. Pollefeys and J.M. Frahm CVPR2018

概要

SfMにおいて，一つの撮影にしか映らないような移動物体を考慮することで，そのシーンの絶対スケールが推定可能になるし，人混みだと見えにくい地平面の復元も成しうる．個々の撮影画像において検出された人を3次元空間に投影し，さらに物体の意味情報（本稿では背の高さの分布）から絶対スケールを推定する． また，人検出結果を用いて地平面推定も行う． ランダムなインターネット画像で手法をデモンストレーションし，量的評価を行う．

人検出はトルソモデルのフィッティングに基づく．画像における肩，腰の位置が推定でき，おおよその立ち位置も分かるということ．

評価点

若干SIGGRAPH的な気風のある，面白い視点を提供する論文．過去の知見に基づく高品質な人検出などを用いて成し得た，正統なアプリケーションに感じる． 動画のインパクトも大きいので，一度視聴を勧める．

リンク集

- 論文
- 動画

[#344]

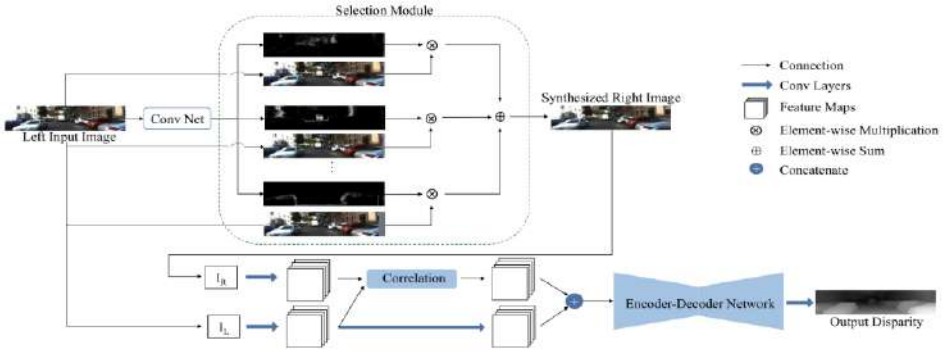
kodai nakashima

Single View Stereo Matching

Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li and Liang Lin CVPR 2018

概要

従来の単眼奥行き推定法では, 推論の際に幾何的な制約を明示的に課していないことや多くのground truth labeled dataが必要といった問題があった.この研究では単眼奥行き推定問題をview synthesis問題とstereo matching問題に分けて考えることにより, 従来法の問題を解決する. view synthesis問題では, 入力を左画像として捉え, view synthesis networkにより右画像を生成する. stereo matching問題では, 左画像を右画像を用いstereo matching networkにより奥行きを推定する.



新規性・結果

- 単眼奥行き推定問題をview synthesis問題とstereo matching問題に分けて考えた.
- 従来法の問題を解決.
- 従来のどの方法よりも精度が高い.

リンク集

- 論文URL

[#345]

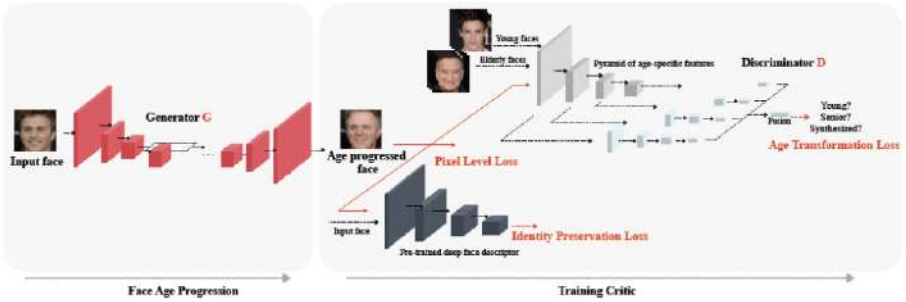
ShintaroYamamoto

Learning Face Age Progression: A Pyramid Architecture of GANs

Hongyu Yang, Di Huang, Yunhong Wang and Anil K. Jain CVPR 2018

概要

入力画像中の人物の老化顔をGANによって生成する手法の提案。Discriminatorには生成した画像が合成画像であるか及び目標年代の特徴を保持しているかを判定させ、それに加え元の画像とのL2ノルム及び元の顔画像と同一人物であるかをロスに加えることで、同一人物性を保持している。 その際、Discriminatorの中間層の各出力を途中で取り出すことにより（ピラミッド型ネットワーク），様々な解像度からの年齢特徴の抽出を行う。



新規性・結果

年齢推定及び個人認証タスクによって有効性を確認した。従来手法では髪や額領域は変化できなかったが、提案手法によってこれらの要素を変化させることを可能とした。 Discriminatorをピラミッド型にすることにより、従来手法に比べてより詳細な老化特徴を取り出すことに成功。

リンク集

- 論文

[#346]

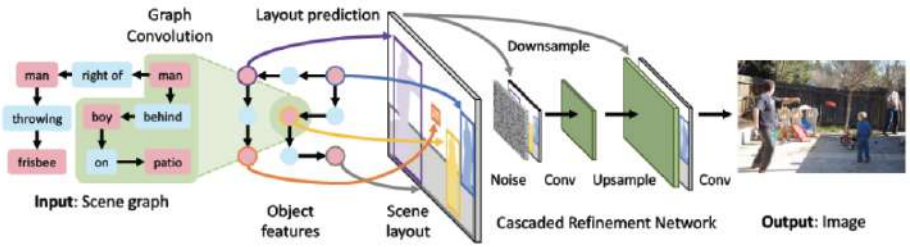
ShintaroYamamoto

Image Generation from Scene Graphs

Justin Johnson et al. CVPR 2018

概要

物体同士の関係を表すScene Graphsから画像を生成する手法の提案。従来のテキストから画像を生成する手法よりも物体の数が多く複雑なシーンの画像を生成することができる。 初めに、Scene Graphsを処理するネットワークによってScene Graphsを表現するベクトルを取得し、そこから画像のレイアウトを作成する。 次にレイアウトからCRN(参考文献)を用いて画像を作成する。 作成された画像は、画像全体のリアルさと各物体のリアルさを評価するDiscriminatorによってリアルな画像であるかを評価する。



新規性・結果

ユーザースタディの結果、StackGANと比較して合成結果が良いと答えた人が68%、認識可能な物体を生成できてると答えた人が59%という結果が得られた。

リンク集

- 論文
- CRN
- StackGAN

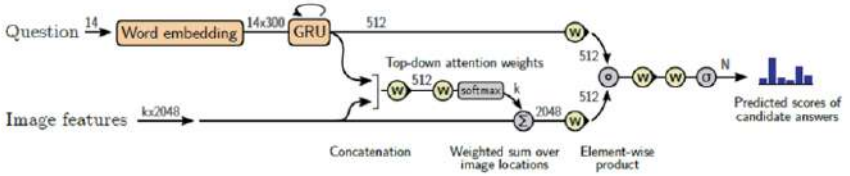
[#347] Yue Qiu

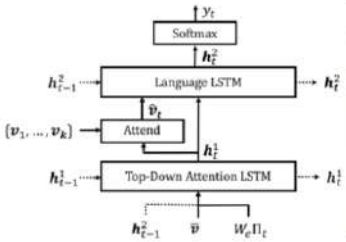
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson, Xiaodong He, Chris Buehler,Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang CVPR 2018 738

概要

Image captioningとVQAタスクに用いられるBottom-upとtop-down attentionをコンバインするメカニズムを提案した．従来のオブジェクトレベルの領域の抽出のほか，salient 領域の抽出も行う．Faster R-CNNを利用したbottom-up的にsalient 領域を特徴ベクトルを抽出し， top-downにより特徴のウェットを決めることをベースに， Image captioningとVQAのアーキテクチャを提案し（右図），両方ともstate-of-artな性能を得られた．





	Yes/No	Number	Other	Overall
Prior [12]	61.20	0.36	1.17	25.98
Language-only [12]	67.01	31.55	27.37	44.26
d-LSTM+n-I [26, 12]	73.46	35.18	41.83	54.22
MCB [11, 12]	78.82	38.28	53.36	62.27
UPMC-LIP6	82.07	41.06	57.12	65.71
Athena	82.50	44.19	59.97	67.59
HDU-USYD-UNCC	84.50	45.39	59.01	68.09
Ours: Up-Down	**86.60**	**48.64**	**61.15**	**70.34**

新規性・結果

・従来のVQAとImage captioningは主にタスクスペシフィックなtop-downタイプのattentionを用いる．この論文で，人の視覚attentionメカニズムから，タスクスペシフィックなtop-downタイプのattentionを及びsalient 領域に注目するBottom-upのattentionを用いることと主張した．・2017 VQA Challengeにおいて優勝した．VQA v2.0 test-standardにおいて70.3%の精度を達成した．また， Image captioning タスクに対しMSCOCO Karpathy testで従来の手法より良い性能を達成した．

リンク集

- 論文

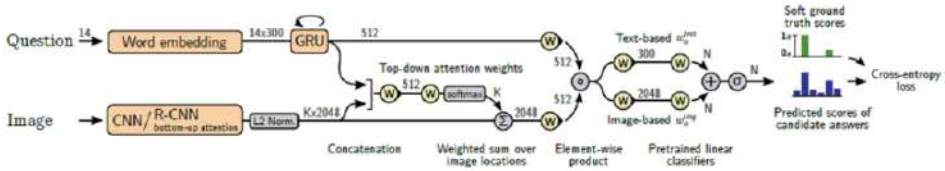
[#348] Yue Qiu

Tips and Tricks for Visual Question Answering: Learning from the 2017 Challenge

Damien Teney, Peter Anderson, Xiaodong He, Anton Van den Hengel CVPR 2018 547

概要

2017 VQA Challengeに優勝したモデルのモデル詳細を紹介し，さらにいかにVQAモデルの精度を上げられるかのコツとテクニックを紹介した．モデルのコアなところは視覚と質問文の意味特徴をジョイントでエンベディングし，さらにマルチ-ラベル予測を行う．



新規性・結果

論文により，VQAの性能上げるために，以下のテクニックがある：1.sigmoid outputsを用いて，マルチアンサーをできるようにする．2．Soft scoresを用いて，分類ではなく回帰を行う．3．Bottom-up attentionから注目領域の画像特徴を用いる．4．Gated tanhを活性化関数に用いる．5．Pre-trainedウェットで初期化する．6．ミニバッチサイズを大きく設定し，training-dataにシャッフリングを用いる

リンク集

- 論文

[#349]

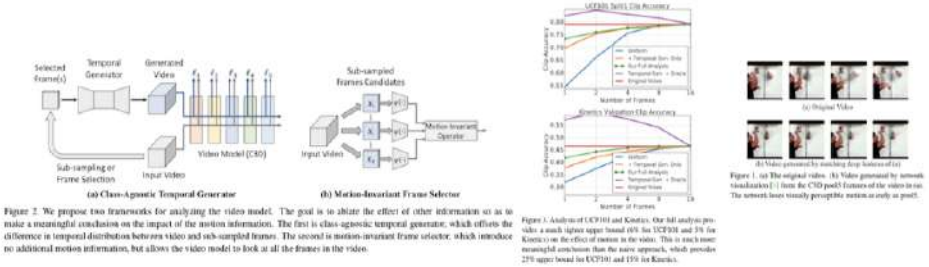
Tomoyuki Suzuki

What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets

Xiaolong Wang et al. CVPR 2018

概要

「3DCNNが実は動き特徴を捉えられていないのではないか」という考えのもと、3DCNNにおける動き特徴の影響の上界を実験的に求める。提案する工夫により、この影響のかなり低い上界を得ることができ、動き特徴を捉えているのではない(例えば実は複数フレーム入力から「重要なフレーム選択」を行っているなど)ことを示唆した。



検証方法

通常の16frames入力で学習したC3Dにおいてtest時にsub-samplingした(動き情報を無くした)設定下でできるだけ精度を上げることで結果的に動き特徴の上界を得る。Naïveにsub-samplingを行うと入力のデータ分布の明らかな違いから動き以外の精度低下への影響をもたらすと考えられるため、 sub-samplingされたclipから元clipを生成するgeneratorを構築。学習はC3Dの中間層の値をMSEで近づける。 またsampling方法によっても精度は変わるという考えから、識別confidenceが最大となるframesをsamplingする。注意として、この際動きに関しては全く考慮せずにsamplingしてきている。

コメント・リンク

結果として、かなりきつい上界を求められ、論文内では3DCNNが2Dよりも精度が良いのは動き特徴ではなく、複数フレーム入力の中で最も識別しやすいフレームを選択可能になるからではと述べられている。

フレーム選択をしているという仮説は面白いし、select frameによって精度が上昇したり、動きが大きい動画はフレーム単位での推定結果の分散が大きいなどから十分ありえそう。これが本当なら、optical flowを3dCNNに導入して大きく精度が向上することともつじつまが合いそう。

- 論文

[#350]

Yue Qiu

Surface Networks

Ilya Kostrikov, Joan Bruna, Daniele Panozzo, Denis Zorin CVPR 2018

概要

3D triangleメッシュから有用的な三次元幾何情報を抽出するネットワークSurface Networkを提案した．従来のLaplace operatorがintrinsic三次元幾何情報しか抽出できない．しかし，様々な応用場面でextrinsic情報が必要となる．この文章で主要なcurvature方向を抽出できるDirac operator を提案し，従来のLaplace operatorより幅広い場面で応用できる．

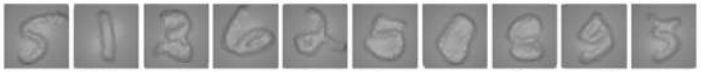


Figure 4: Meshes from the dataset (first five). And meshes generated by our model (last five).

Figure 5: Qualitative comparison of different models. We plot 30th and 40th predicted frames correspondingly.

Model	Receptive field	Number of parameters	Smooth L1-loss (mean per sequence (std))
MLP	1	519672	64.56 (0.62)
AvgPool	-	1018872	23.64 (0.21)
Laplace	16	1018872	17.34 (0.52)
Dirac	8	1018872	16.84 (0.16)

Table 1: Evaluation of different models on the temporal task

新規性・結果

・定性的および定性的な結果によりspatial-temporal predictionsタスクにおいて，従来手法より良い結果を得られている．・variationalエンコーダーを用いたメッシュ合成手法を提案し，有効的に3次元メッシュを生成できる．

リンク集

- 論文

[#351]

Yue Qiu

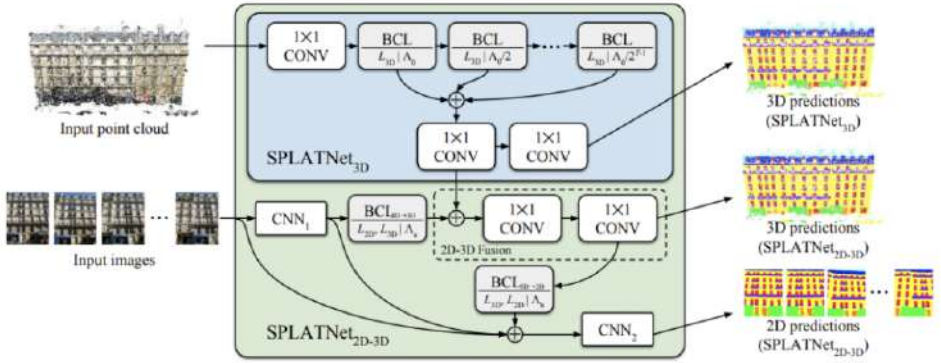
SPLATNet: Sparse Lattice Networks for Point Cloud Processing

Hang Su, University of Massachusetts, Amherst; Varun Jampani, NVIDIA Research; Deqing Sun, NVIDIA; Evangelos Kalogerakis, UMass; Subhransu Maji, ; Ming-Hsuan Yang, UC Merced; Jan Kautz, NVIDIA

CVPR2018

概要

点群情報を直接処理できるSPLATNet（右図）を提案した．SPLATNetは直接点群から階層的な空間情報を抽出可能．また，2D情報と3D情報のマッピングも行えるので，点群とマルチ画像の両方をSPLATNetで処理可能．従来の直接点群情報を処理するネットワークはより局所的な空間情報を損失してしまう問題点がある．提案手法はこの問題を解決するために，BCLs層を用いた． BCLs層は点群をスパースなlatticeにマッピングし，さらにそのスパースなlatticeを畳み込みできる．それにより， unordered点群情報を処理できる上に点群のより局所的な情報も抽出可能にした．



新規性・結果

Façade segmentationタスクにおいて，点群とマルチ画像のラベリングに良い処理スピードと従来手法手法より優れた精度を得られた．ShapeNet part segmentationにおいて従来手法より優れた精度（クラスmIoU：83.7%）を得られた．

リンク集

- 論文

[#352]

Yoshihiro Fukuhara

From Lifestyle Vlogs to Everyday Interactions

Fouhey et al. CVPR 2018. arXiv ID: 1712.02310

概要

従来のデータ取集手法（collection-by-acting）では難しいかった, バイアスの少ない, 多様で大規模な日常生活におけるインタラクションのデータベース Lifestyle VLOG dataset を公開した.

新規性・結果

- 従来のデータセットが想定している陽的なデータ収集とは対照的に隠的なデータ収集方法を行うことで, バイアスを小さくすることに成功した.
- ビデオに対してインタラクションのラベル, フレームに対してインタラクション時の手の状態のラベル付けられている.
- 従来のデータセットのBiasを分析するために, 従来のデータセットで訓練した手法が Lifestyle VLOG データセットに対しても上手く動作するか検証した.

リンク集

- [論文] From Lifestyle Vlogs to Everyday Interactions
- Project Page
- GitHub

[#353]

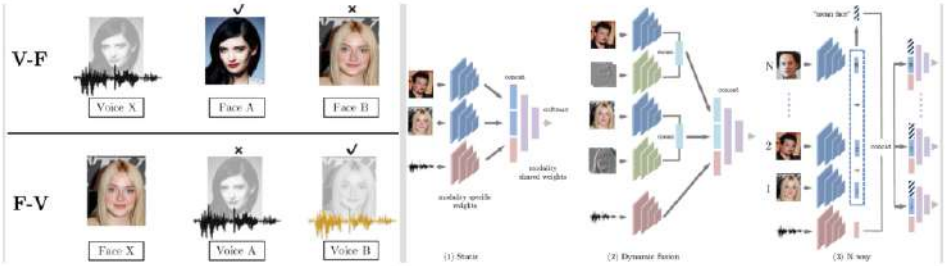
Kensho Hara

Seeing Voices and Hearing Faces: Cross-modal biometric matching

A. Nagrani et al. CVPR 2018

概要

ある音声と2人分の顔画像から，どちらの人物の声かを推定する課題と，ある顔画像と2人分の音声から，どちらの音声がその人物の声かを推定する課題の2つを解くという問題設定の研究． 異なるモダリティ間でのマッチングという課題ということ． ある入力に対応するのがどちらの人物かという2クラス識別の問題設定として定式化． この問題を解くために，3入力を扱う3-streamのネットワーク構造を持つモデルを提案． 音声もスペクトログラムの形式で画像のように扱い，顔画像，音声ともにConvolutionしていくモデル． 実験では80%程度の識別率を達成し，人と同等の結果が出ている． 二人分の選択肢の性別，国籍，年齢などが同じという設定にすると，60%程度の正答率になるが，こちらでは人 (57%) を上回る結果となっている．



新規性・結果

- 人物の顔画像と音声の対応付けという新しい問題設定
- 人間レベルの高い精度を実現

リンク集

- 論文 (arXiv)

[#354]

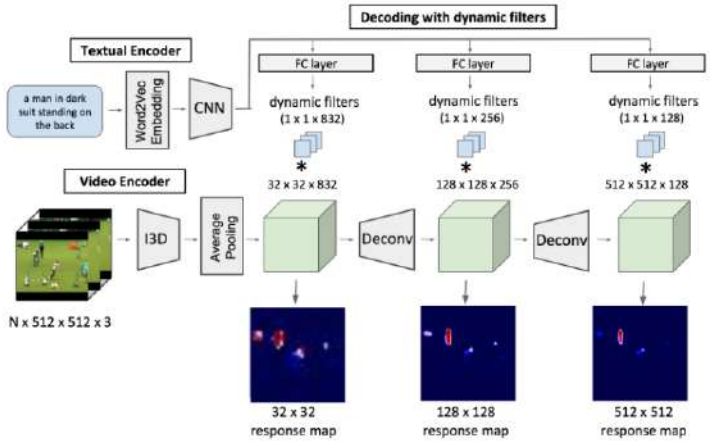
Hirokatsu Kataoka

Actor and Action Video Segmentation from a Sentence

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, Cees G.M. Snoek CVPR 2018 (oral)

概要

センテンスの入力から、行動者と行動（Actor and Action）を同時に特定する研究である。複数の同様の物体から特定の人物など、詳細な分類が必要になる。ここではFully-Convolutional（構造の全てが畳み込みで構成される）モデルを適用してセグメンテーションベースで出力を行うモデルを提案。図は提案モデルを示す。I3Dにより動画像のエンコーディング、自然言語側はWord2Vecの特徴をさらにCNNによりエンコーディング。その後、動画像・言語特徴を統合してDeconvを繰り返しセグメントを獲得していく。



新規性・結果

文章（と動画像）の入力から行動者と行動の位置を特定すべくセグメンテーションを実行するという問題を提起した。また、二つの有名なデータセット（A2D/J-HMDB）を拡張して7,500を超える自然言語表現を含むデータとした。同問題に対してはSoTA。

コメント・リンク集

CVxNLPの問題はここにも進出して来た。画像キャプションに限らず、この手の統合は進められるはず。

- 論文

[#355]

Munetaka Minoguchi

Alive Caricature from 2D to 3D

Qianyi Wu, et al. CVPR 2018

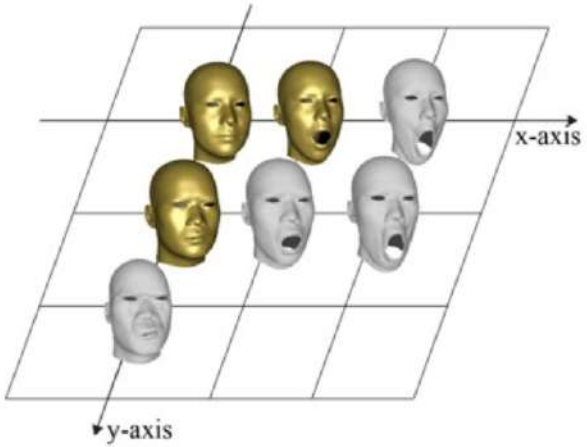
概要

2Dの似顔絵画像から3Dの似顔絵を作成するためのアルゴリズムの提案。似顔絵画像のテストデータとしてはカリカチュアを使用し、カリカチュア画像の3Dモデルとテクスチャ化された画像を生成する。データは、標準の3D顔の変形を座標系に配置(下図、 xは口の開き具合)し、金のオリジナルデータから線形結合によって白い顔を生成する。

新規性・結果・リンク集

カリカチュアを集めたデータセットを作って学習するのではなく、標準の3D顔のデータセットから実装でき、アプリケーションの柔軟さを推している。

3DMMやFaceWareHouseなどの従来手法と比較して、形の歪みが少なく、従来のものよりも綺麗な3D顔の出力が可能。顔以外にも、概形の予測が可能なオブジェクトなら応用できる？



- 論文

[#356]

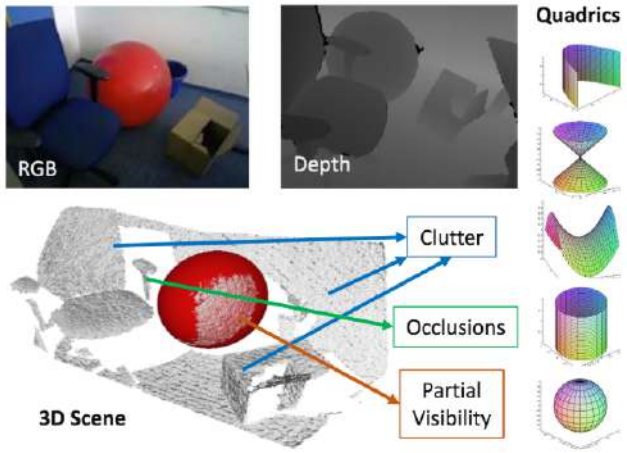
Hirokatsu Kataoka

A Minimalist Approach to Type-Agnostic Detection of Quadrics in Point Clouds

Tolga Birdal, Benjamin Busam, Nassir Navab, Slobodan Ilic, Peter Sturm CVPR 2018

概要

オクルージョンが発生している場合/複雑な環境下でも簡単な形状がポイントクラウドから検出できる枠組みを提案する。手法は3D楕円形状のフィッティング、3次元空間操作、4点取得により構成。



新規性・結果

タイプに依存しない3次元の二次曲面（楕円球形状）検出を点群の入力から行う手法を考案した。さらに、4点探索問題を3点探索にしてRANSACベースの手法で解を求めた。モデルベースのアプローチよりはフィッティングの性能がよいが、キーポイントベースの手法よりは劣る。

コメント・リンク集

曖昧な教示のみで3次元形状探索問題が解決できるようになる？

- 論文

[#357]

Hirokatsu Kataoka

COCO-Stuff: Thing and Stuff Classes in Context

Holger Caesar, Jasper Uijlings, Vittorio Ferrari CVPR 2018

概要

MSCOCOデータセットに対してThing（もの）やStuff（材質）に関する追加アノテーションを行い、さらにコンテキスト情報も追加したCOCO-Stuffを提案した。このデータセットには主にシーンタイプ、そのものがどこに現れそうかという場所、物理的/材質的な属性などをアノテーションとして付与する。COCO2017をベースにして164Kに対して91カテゴリを付与し、スーパーピクセルを用いた効率的なアノテーションについてもトライした。

新規性・結果

材質的なアノテーションは画像キャプションに対して重要であることを確認、相対的な位置関係などデータセットのリッチなアノテーションが重要であること、セマンティックセグメンテーションベースの方法により今回のアノテーションを簡易的に行えたこと、などを示した。

コメント・リンク集

さらにリッチなアノテーションは今後重要になる。この論文ではスーパーピクセルという弱い知識を用い、人間のアノテーションと組み合わせることでボトムアップ・トップダウンを効果的かつ効率的に組み合わせてアノテーションを行っている点が素晴らしい。ラストオーサのVittorio Ferrariは機械と人の協調によるアノテーションが得意（なので、既存データセットへのよりリッチなアノテーションを早いペースで提案できる）。

- 論文
- GitHub

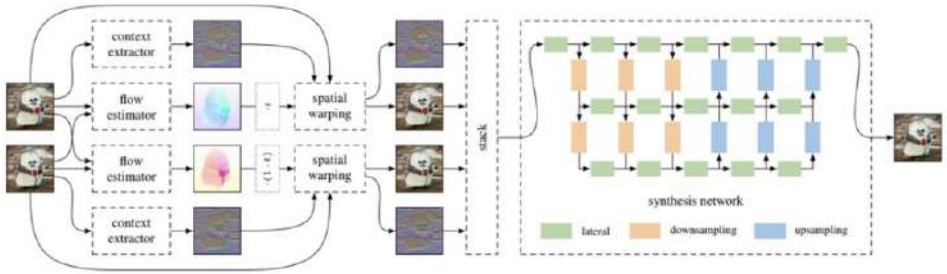
Munetaka Minoguchi

Context-aware Synthesis for Video Frame Interpolation

Simon Niklaus, Feng Liu CVPR 2018

概要

入力フレームだけでなく、ピクセル単位の文脈情報を用いて、高品質の中間フレームを補間するためのコンテキスト認識手法の提案。まず、プレトレインモデルを使用して、入力フレームのピクセルごとのコンテキスト情報を抽出。オプティカルフローを使用して、双方向フローを推定し、入力フレームとそのコンテキストマップの両方をワープする。最後にコンテキストマップをsynthesis networkに入力し、補間フレームを生成。



新規性

従来のビデオフレーム補間アルゴリズムは、オプティカルフローまたはその変動を推定し、それを用いて2つのフレーム間の中間フレームを生成する。本手法では、2つの入力フレーム間の双方向フローを推定し、コンテキスト認識という方式をとることで精度向上を図る。

結果・リンク集

高品質のビデオフレーム補間実験において、従来を上回る性能。

- 論文

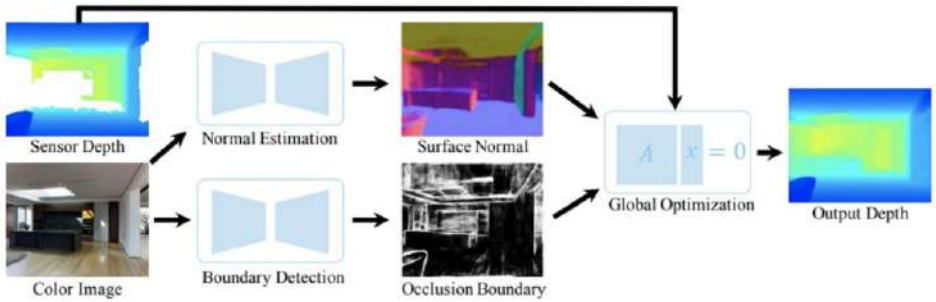
Munetaka Minoguchi

Deep Depth Completion of a Single RGB-D Image

Yinda Zhang, Thomas Funkhouser CVPR 2018

概要

RGB画像から表面の法線とオクルージョン境界を予測し、 RGB-D画像と組み合わせて、欠けている奥行き情報を補完するDeep Depth Completionの提案。また、奥行き画像と対になったRGB-D画像のデータセットであるcompletion benchmark datasetを作成し、性能を評価。これは、低コストのRGB-Dカメラでキャプチャした画像と、高コストの深度センサで同時にキャプチャした画像で構成されている。



新規性

深度カメラは、光沢があり、明るく、透明で、遠い表面の深さを感知しないことが多い。 このような問題を解決するために、本手法ではRGB画像から得た情報と組み合わせて、 RGB-D画像の深度チャネルを完全なものにする。

結果・リンク集

深さ修復および推定において従来よりも優れた性能。

- 論文
- Project webpage

[#360]

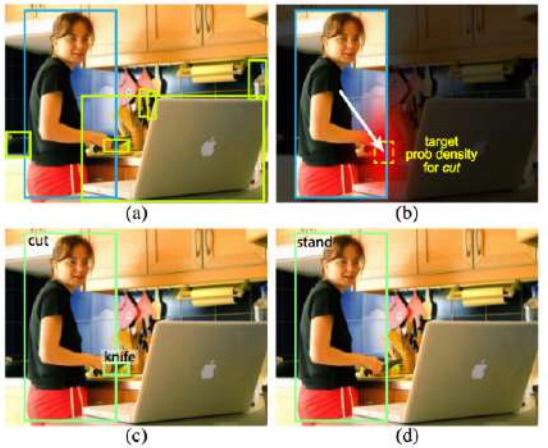
Hirokatsu Kataoka

Detecting and Recognizing Human-Object Interactions

Georgia Gkioxari, Ross Girshick, Piotr Dollár, Kaiming He CVPR 2018 (spotlight)

概要

人物検出と同時に人物行動やその物体とのインタラクションも含めて学習を行うモデルを提案する。本論文では物体候補の中でも特にインタラクションに関係ありそうな物体に特化して認識ができるようにする。さらに、検出されたのペアを用いて学習する（図の場合には）。さらに、その他の行動（図の場合にはstand）を同時に推定することもできる。モデルはFaster R-CNNをベースとするが、物体検出（box, class）、行動推定（action, target）、インタラクション（action）を推定して誤差を計算する。さらに、推定した人物位置に対する対象物体の方向も確率的に計算することが可能。



新規性・結果

人間に特化した検出と行動推定の枠組みを提案した。V-COCO（Verbs in COCO）にて、相対的に26%精度が向上（31.8=>40.0）、HICO-DETデータセットにて27%相対的な精度向上が見られた。計算速度は135ms/imageであり、高速に計算が可能である。

コメント・リンク集

単純な多タスク学習ではなく、人物に特化して対象物体の位置も確率的に推定しているところがGood。

- 論文
- Project
- Verbs in COCO DB

[#361]

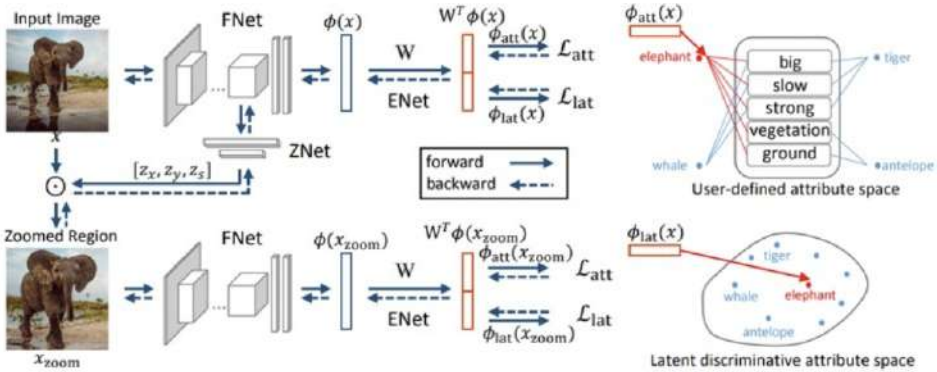
Munetaka Minoguchi

Discriminative Learning of Latent Features for Zero-Shot Recognition

Minghui Yan Li, et al CVPR 2018

概要

Zero-shot learning(ZSL)における、視覚的および意味的インスタンスを別々に表現し学習するLatent Discriminative Features Learning(LDF)の提案。 (1)ズームネットワークにより差別的な領域を自動的に発見することができるネットワークの提案。(2)ユーザによって定義された属性と潜在属性の両方について、拡張空間における弁別的意味表現の学習。



新規性

ZSLは、画像表現と意味表現の間の空間を学習することによって、見えない画像カテゴリを認識する。 既存の手法では、視覚と意味空間を合わせたマッピングマトリックスを学習することが中心的課題。提案手法では、差別的に学習するとうアプローチで識別精度向上を図る。

結果・リンク集

2つのコンポーネントによって、互いに支援しながら学習することで最先端の精度に。

- 論文

[#362]

Hirokatsu Kataoka

Domain Adaptive Faster R-CNN for Object Detection in the Wild

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, Luc Van Gool CVPR 2018

概要

ドメイン変換について、ゲームなどのCG映像から実際の交通シーンに対応して物体検出を行うための学習方法を提案する。本論文では (i) 画像レベルのドメイン変換、(ii) インスタンス（ある物体）に対してのドメイン変換、の二種類の方法を提案し、整合性をとるように正規化する（図のConsistency Regularization; Global/Localな特徴変換を考慮）。ここで、物体検出はFaster R-CNNをベースとしてドメイン変換の手法も二種類（H-divergence、敵対的学習）用意する。

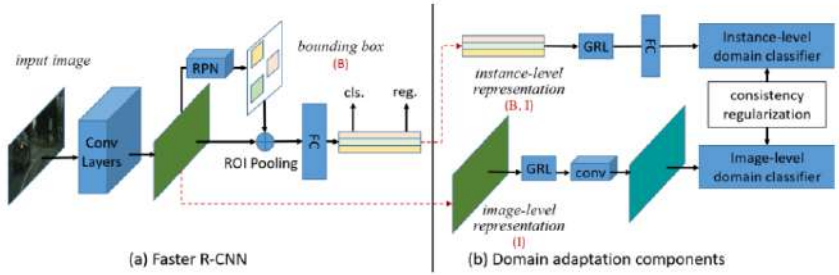


Figure 2. **An overview of our Domain Adaptive Faster R-CNN model:** we tackle the domain shift on two levels, the image level and the instance level. A domain classifier is built on each level, trained in an adversarial training manner. A consistency regularizer is incorporated within these two classifiers to learn a domain-invariant RPN for the Faster R-CNN model.

新規性・結果

CGで学習し実環境における自動運転などで使えるドメイン変換の手法を提案した。実験はCityscapes, KITTI, SIM10Kなどで行い、ロバストな物体検出を実行することができた。例えばCityscapesとKITTIの相互ドメイン変換でベースラインのFaster R-CNNが30.2 (K->C)、53.5 (C->K)のところ、Domain Adaptive Faster R-CNNでは38.5 (K->C)、64.1 (C->K)であった。

コメント・リンク集

データ収集は手動から自動の時代になって来た？データを手作業で集める時代からアルゴリズムを駆使して収集する時代へ移行。

- 論文
- 著者

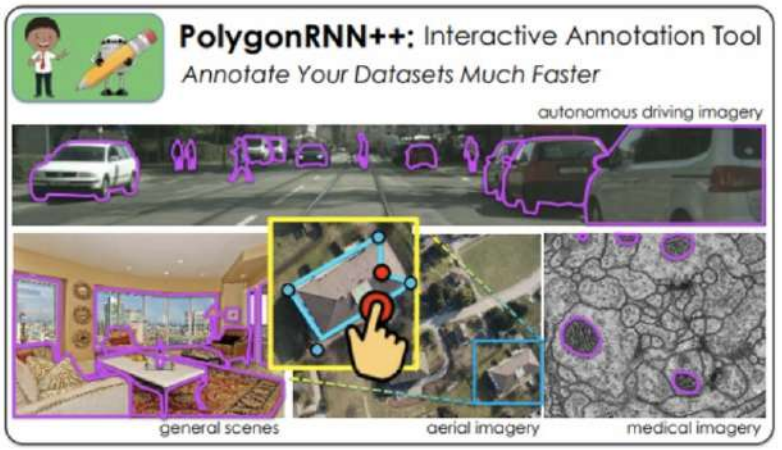
Munetaka Minoguchi

Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++

David Acuna, Huan Ling, Amlan Kar, Sanja Fidler CVPR 2018

概要

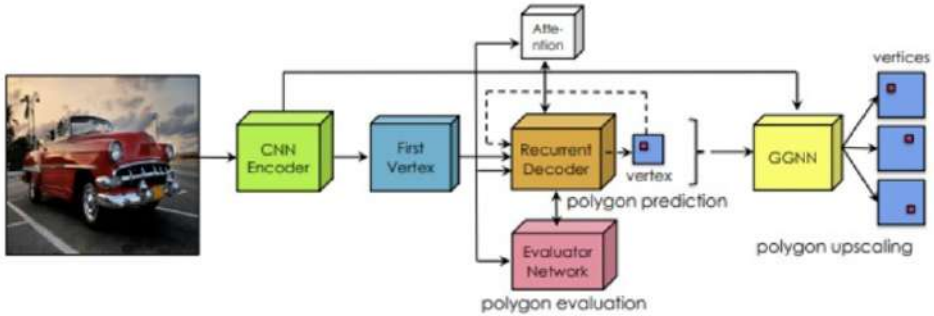
Polygon-RNNのアイデアを踏襲し、ヒューマン・イン・ザ・ループを使って対話的にオブジェクトのポリゴンアノテーションの生成。また、新しいCNNエンコーダアーキテクチャの設計、強化学習によるモデルの効果的な学習、 Graph Neural Networkを使用した出力解像度の向上を行う。これらのアーキテクチャをPolygon-RNN ++と呼ぶ。



新規性・結果・リンク集

アノテーション作成時の負担を軽減。より正確にアノテーションを付加できるため、雑音の多いアノテーターに対しても頑健である。

高い汎化能力となり、既存のピクセルワイズメソッドよりも大幅に改善。ドメイン外のデータセットにも適応可能。



- 論文

[#364]

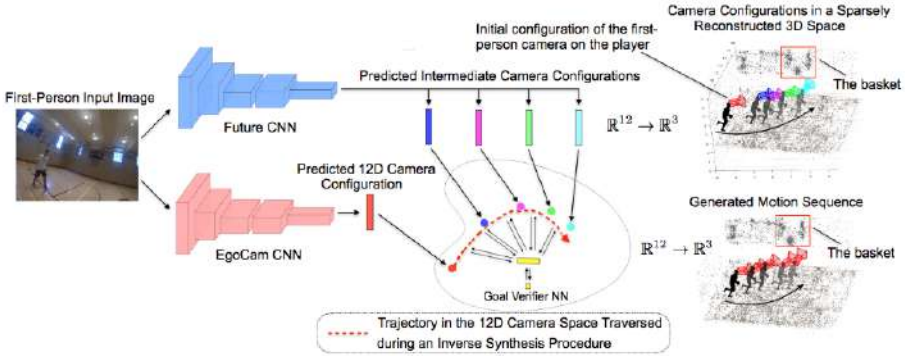
Hirokatsu Kataoka

Egocentric Basketball Motion Planning from a Single First-Person Image

Gedas Bertasius, Aaron Chan, Jianbo Shi CVPR 2018

概要

一人称視点の画像からゴールリングに到達するまでのバスケットボール選手の動線を生成する。本論文では3D位置や頭部方向も記録する。同タスクを実行するため、まずは画像空間から12Dのカメラ空間に投影を行うEgoCam CNNを学習。次に予測を行うCNN（Future CNN）を構築、さらに予測位置やゴールまでの位置が正確かどうかを検証するGoal Verifier CNNを用いることでより正確な推定を行うことができる。



新規性・結果

複数のネットワークの出力（ここではEgoCamCNNとFutureCNN）を検証するVerification Networkという考え方は面白い。他のネットワークの出力を、検証用のネットワークにより正すというのはあらゆる場面で用いることができる。RNN/LSTM/GANsなどよりも高度な推定ができることが判明した。

コメント・リンク集

結果例は動画像を参照。未来予測・３次元投影などコンポーネントがDNNにより高度にできるようになってきたからできた研究。さらに検証用のネットワークを構築することで出力自体を操作している。

- 論文
- YouTube

[#365]

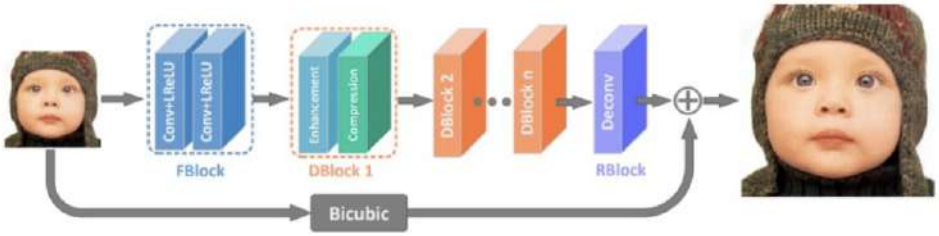
Munetaka Minoguchi

Fast and Accurate Single Image Super-Resolution via Information Distillation Network

Zheng Hui, Xiumei Wang, Xinbo Gao CVPR 2018

概要

元の低解像度画像から高解像度画像を再構築するための、深くてコンパクトなCNNを提案。提案モデルは、特徴抽出ブロック、積み重ね情報蒸留ブロック、再構成ブロックの3部構成。これにより、情報量が豊富かつ効率的に特徴を徐々に抽出できる。



新規性

CNNが超解像殿画像を扱うようになってきたが、ネットワークが増大するにつれて、計算上の複雑さとメモリ消費という問題が生じる。これらの問題を解決するためのコンパクトなCNN。

結果・リンク集

PSNR、SSIM、IFCの4つのデータセットで検証し、精度向上を確認。デシジョンおよび圧縮アーチファクト低減などの他の画像修復問題にも応用可能？

- 論文

[#366]

Hirokatsu Kataoka

Future Frame Prediction for Anomaly Detection -- A New Baseline

Wen Liu, Weixin Luo, Dongze Lian, Shenghua Gao CVPR 2018

概要

先の（未来の）フレーム予測と異常検知を同時に行う手法を提案する論文。予測したフレームと異常検知の正解値により誤差を計算して最適化を行う。図に本論文で提案するネットワークアーキテクチャの図を示す。U-Netにより画像予測やさらにオプティカルフロー推定を行い、RGB空間、オプティカルフロー空間にて誤差を計算しGANの枠組みでそれらがリアルかフェイクかを判定する。同フレームを用いて異常検知を実施する。

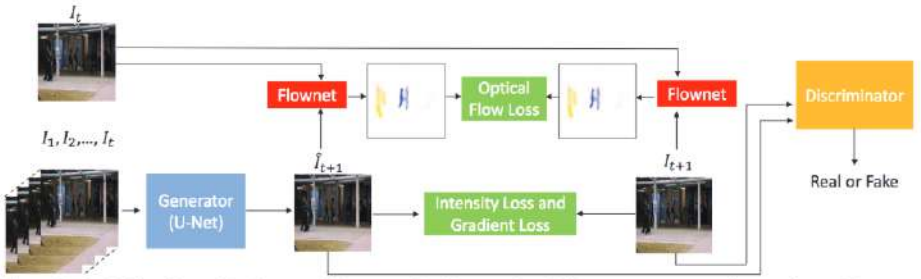


Figure 2. The pipeline of our video frame prediction network. Here we adopt U-Net as generator to predict next frame. To generate high quality image, we adopt the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Here Flownet is a pretrained network used to calculate optical flow. We also leverage the adversarial training to discriminate whether the prediction is real or fake.

新規性・結果

従来は現在フレームを入力として異常検知を行う手法は存在したが、未来フレームを予測して異常検知を行う枠組みは本論文による初めての試みである。異常値の正解値を与えることで画像予測にもフィードバックされるため、画像予測と異常検知の相互学習に良い影響を与える。オープンデータベースにてベンチマークした結果、何れもState-of-the-artな精度を達成。

コメント・リンク集

生成ベースで画像予測+X（Xは任意タスク）というものはSoTAが出せるくらいにはなってきた。

- 論文
- Project

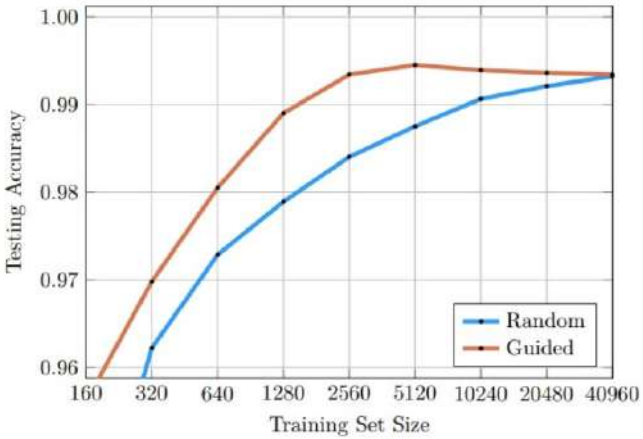
Munetaka Minoguchi

Guided Labeling using Convolutional Neural Networks

Sebastian Stabinger, et al. CVPR 2018

概要

ラベルの付いていないデータに対して、どの画像にラベルを付けてデータセットを構成すればよいかを判断するguided labelingの提案。ラベル付けを行う必要があるサンプルを見定めることで、データセットの量を大幅に減らすことができる。



新規性

大規模データセットにおいて、手動でのラベル付けは大変。選別してラベル付けを行えば、作業を最小限に抑えられる。また、ある意味良いデータを選別できるため、場合によっては精度も向上。

MNISTは、データセットのサイズを1/16に、CIFAR10は1/2に減らすことが可能に。また、MNISTの場合は、全部使った時よりも識別精度が向上した。普遍性を妨げる不必要なデータを取り除けたことが精度向上につながった？

- 論文

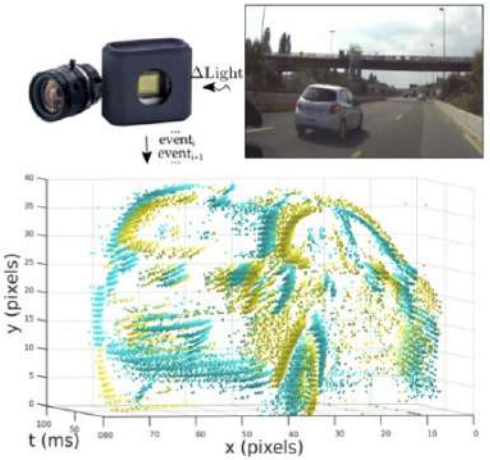
Munetaka Minoguchi

HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification

Amos Sironi, et al. CVPR 2018

概要

イベントベースカメラにおける、識別アルゴリズムの提案。本研究では、(1)イベントベースのオブジェクト分類のための低レベル表現とアーキテクチャの欠如、(2)実世界における大きなイベントベースのデータセットの欠如、の2つの問題に取り組む。新しい機械学習アーキテクチャ、イベントベースの特徴表現(Histograms of Averaged Time Surfaces)、データセット(N-CARS)を提案。



新規性

イベントベースのカメラは、従来のフレームベースのカメラと比較して、高時間分解能、低消費電力、高ダイナミックレンジという点で優れており、様々なシーンで応用が利く。しかし、イベントベースのオブジェクト分類アルゴリズムの精度は未だ低い。特徴表現には過去時間の情報を使用。

結果・リンク集

過去の情報を使うことで、既存のイベントベースカメラによる認識手法よりも優れた結果となった。

- 論文
- データセット

Munetaka Minoguchi

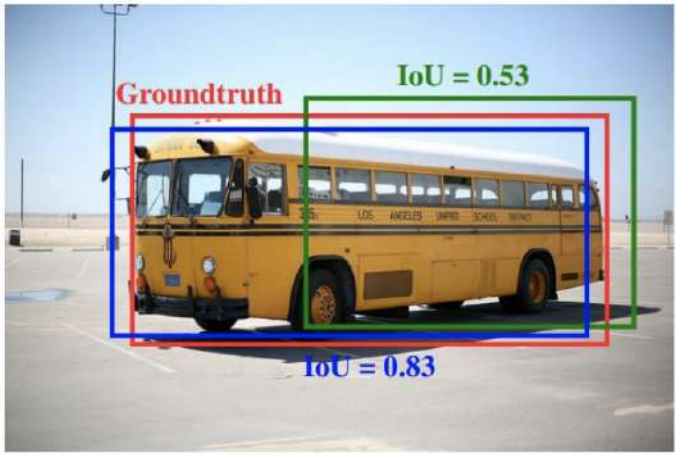
Improving Object Localization with Fitness NMS and Bounded IoU Loss

Lachlan Tychsen-Smith, et al. CVPR 2018

概要

既存のNon-Max Supressionを改良したFitness NMSの提案。Soft NMSも同時に使用するとより効果的。

勾配降下法の収束特性(滑らかさ、堅牢性など)を維持しつつ、IoUを最大化するという目標により適した損失関数であるBounded IoU Loss の提案。これをRoIクラスタリングと組み合わせることで精度が向上する。



新規性

バウンディングボックスのスコアを算出する関数を拡張する。具体的には、グランドトゥルースとのIoUと、クラスの期待値を追加する。これにより、IoUの重なり推定値と、クラス確率の両方が高いバウンディングボックスを優先して学習することができる。

結果・リンク集

MSCOCO、Titan X(Maxwell)使用時では、精度33.6%-79Hzまたは41.8%-5Hz。本論文ではDeNetでテストしたが、別の手法でも精度向上が望めるよう。

- 論文
- ソースコード

Munetaka Minoguchi

Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN

Shuai Li, et al. CVPR 2018

概要

新しいRNN手法であるindependently recurrent neural network (IndRNN)の提案。一枚のレイヤ内のニューロンが独立しており、レイヤ間で接続されている。これにより、勾配消失問題や爆発問題を防ぎ、より長期的なデータを学習することができる。また、IndRNNは複数積み重ねることができるため、既存のRNNよりも深いネットワークを構築できる。

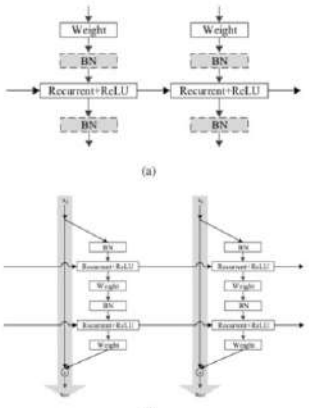


Figure 1. Illustration of (a) the basic IndRNN architecture and (b) the residual IndRNN architecture.

新規性

本手法によって下記の従来手法の問題を解決。

RNNは、勾配の消失や爆発の問題、長期パターンの学習が困難である。LSTMやGRUは、上記のRNNの問題を解決すべく開発されたが、層の勾配が減衰してしまう問題がある。また、RNNは全てのニューロンが接続されているため、挙動の解釈が困難。

結果・リンク集

かなり長いシーケンス(5000回以上の時間ステップ)を処理でき、かなり深いネットワーク（実験では21レイヤー）を構築できる。

- 論文

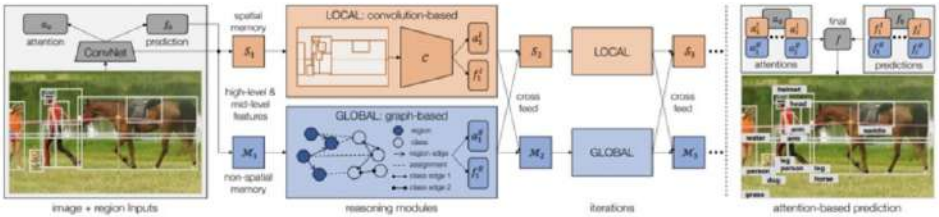
[#371]

Munetaka Minoguchi

Iterative Visual Reasoning Beyond Convolutions

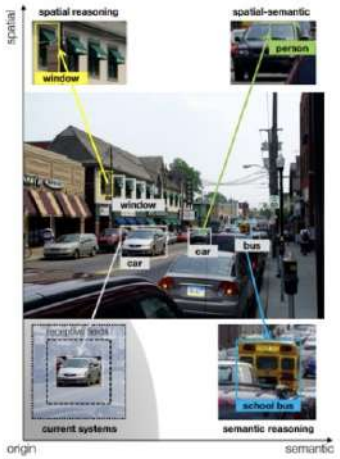
Xinlei Chen, Li-Jia Li, Li Fei-Fei, Abhinav Gupta CVPR 2018

概要



CNNのような理由を突き止める能力がない認識システムを超えた、反復的なvisual reasoningのための新しいフレームワークの提案。畳み込みベースのローカルモジュールとグラフベースのグローバルモジュールの2コアで構成。2つのモジュールのを繰返し展開し、予測結果を相互にクロスフィードして絞り込む。最後に、両方のモジュールの最高値をアテンションベースのモジュールと組み合わせてプレディクト。

新規性・結果・リンク集



ただ畳み込むだけでなく、Spatial(空間的)およびSemanticの空間を探索することができる。下図のように、「人」は「車」を運転するというSpatialとSemanticの双方を兼ね備えた認識を行うことで精度向上を図る。

通常のCNNと比較して、ADEで8.4%、COCOで3.7%の精度向上。

- 論文

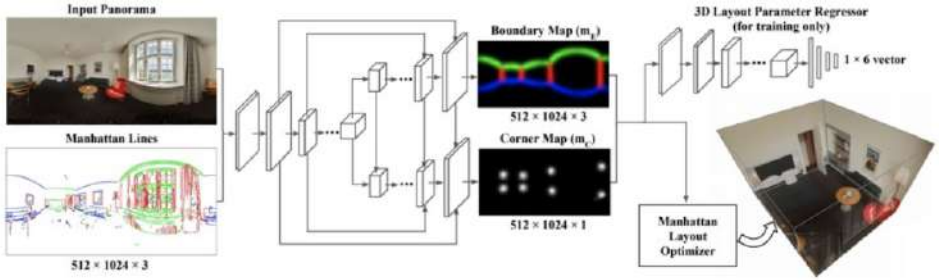
Munetaka Minoguchi

LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image

Chuhang Zou, Alex Colburn, Qi Shan, Derek Hoiem CVPR 2018

概要

単一のパースペクティブまたはパノラマ画像から屋内3Dルームレイアウトを推定するLayoutNetの提案。最初に、消失点を分析し、水平になるように画像を整列。これにより、壁と壁の境界が垂直になり、ノイズ低減。画像からコーナー(レイアウト接合点)と境界を、エンコーダ/デコーダ構造のCNNで出力。最後に、3D Layoutパラメータを、予測したコーナーと境界に適合するように最適化する。



新規性

アーキテクチャはRoomNetと似ているが、消失点に基づいて画像を整列させ、複数のレイアウト要素（コーナー、境界線、サイズ、平行移動）を予測し、“L”形の部屋のような非直方体のマンハッタンレイアウトに対しても適応できる。

結果・リンク集

従来手法と比較して、処理速度と正確さにおいて性能の向上。

- 論文
- ソースコード

[#373]

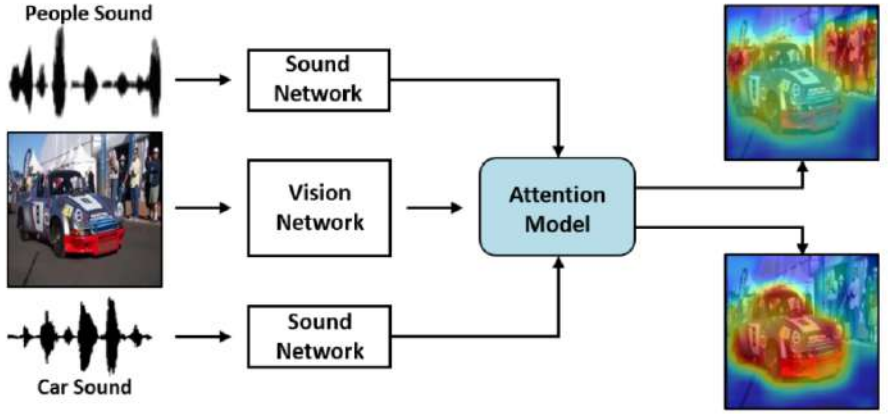
Hirokatsu Kataoka

Learning to Localize Sound Source in Visual Scenes

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, In So Kweon CVPR 2018

概要

画像と音声の入力から、音が画像のどこで鳴っているか（鳴りそうか？）を推定した研究。さらに、人の声なら人の領域、車の音なら車の領域にアテンションがあたるなど物体と音声の対応関係も学習することができる。学習には音源とその対応する物体の位置を対応づけたデータセット（144Kのペアが含まれるSound Source Localization Dataset）を準備した。さらに既存の物体認識と音声を対応づけて（？）Unsupervised/Semi-supervisedに学習することにも成功した。



新規性・結果

教師あり、教師なし、半教師あり、いずれの枠組みでも音声ー物体の対応関係を学習することができるようにした。音源とそれに対応する物体領域の尤度がヒートマップにて高く表示されている。結果はビデオを参照されたい。教師なし学習はTriplet-lossにより構成され、ビデオと近い/遠い音声の誤差により計算。

コメント・リンク集

非常に面白い問題設定、プラス誤差関数を柔軟に抽出可能というところが上手。精読しても良いと感じた論文。CVにおいてビデオの音声は今まで使用しないことも多かった（もしくは精度向上のために活用していた）が、これからは使用方法を見直してもよいと感じた。

- 論文
- YouTube

[#374]

Hirokatsu Kataoka

Learning to Segment Every Thing

Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, Ross Girshick CVPR 2018

概要

ラベルが完全に手に入らない際にでも転移学習が可能なセグメンテーション手法（論文中ではPartially Supervised Training Paradigm, weight transfer functionを紹介）を提案する。条件として、bboxが手に入っている物体に対してセグメンテーション領域を学習可能。Mask R-CNNをベースとしているが、Weight Transfer Functionを追加、セグメントの重みを学習・推定して誤差計算と学習繰り返し。

新規性・結果

Visual Genome Datasetから3,000の視覚的概念を獲得、MSCOCOから80のマスクアノテーションを獲得した。

コメント・リンク集

弱教師付き学習が現実的な精度で動作するようになってきた？アノテーションはお金や知識があっても非常に大変なタスクであり、いかに減らすかという方向に研究が進められている。（What's next? ー弱教師/教師なしの先とは？）

- 論文
- 著者
- Kaiming He

[#375]

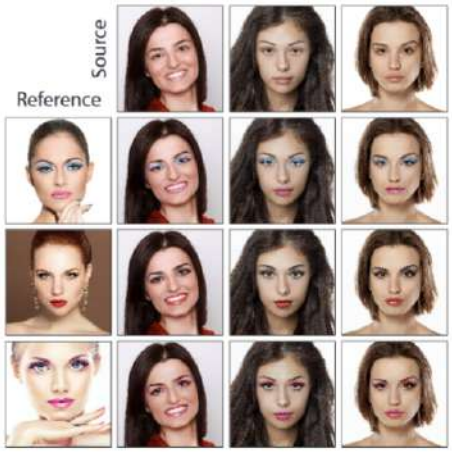
ShintaroYamamoto

MakeupGAN: Makeup Transfer via Cycle-Consistent Adversarial Networks

Huiwen Chang et al. CVPR 2018

概要

ソース画像のメイクをターゲット画像へ転写やメイクの除去をする研究。ターゲット画像とメイク済み画像の2枚を入力としメイクを転写するネットワークGとメイク済み画像らメイクを取り除くネットワークFを考え、2つのネットワークによって元の画像に戻るように学習していく。 その際、Fによってxに付与されたメイクがyのメイクと同じものであるかを評価するロスを加えることでメイクの特徴を捉える。 従来手法ではメイク転写・除去を独立した問題として考えていたが、この研究ではセットとして考えている。



新規性・結果

Youtubeのメイクチュートリアルの動画から、1148枚のメイクなし画像と1044枚のメイクあり画像を収集。ユーザースタディによって2つの既存手法と比較し、提案手法が一番いいと答えた人が65.7%（2番目と答えた人が31.4%） 従来手法では肌の色や表情の違いがあると上手くいかないのに対し、ソースとターゲット間でこれらが違ってもうまく転写できる。

リンク集

- プロジェクトページ

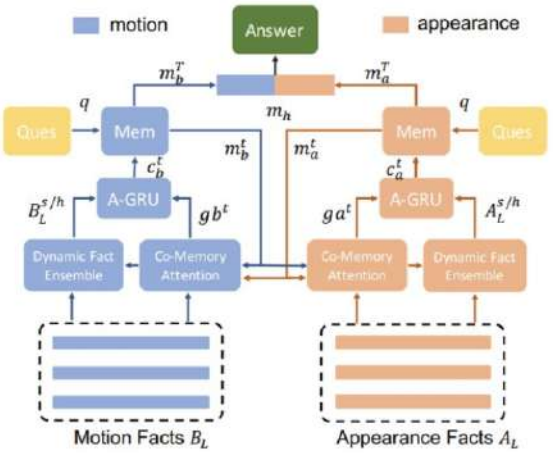
Munetaka Minoguchi

Motion-Appearance Co-Memory Networks for Video Question Answering

Jiyang Gao, Runzhou Ge, Kan Chen, Ram Nevatia CVPR 2018

概要

ビデオQAのための、 Dynamic Memory Network(DMN) のコンセプトに基づいたmotion-appearance comemory networkの提案。本研究の特徴は次の3つである。(1)アテンションを生成するために動きと外観情報の両方を手がかりとして利用する共メモリアテンションメカニズム。(2) multi-level contextual factを生成するための時間的conv-deconv network。(3)異なる質問に対して動的な時間表現を構成するdynamic fact ensemble method。



新規性

本手法は、次のようなvideo QA特有の属性に基づいている。(1)豊富な情報を含む長い画像シーケンスを扱う。(2)動き情報と出現情報を相互に関連付け、アテンションキューを他の情報に応用できる。(3)答えを推論するために必要なフレーム数は質問によって異なる。

結果・リンク集

TGIF-QAの4つのタスクすべてにおいて、最先端技術よりも優れている。

- 論文

[#377]

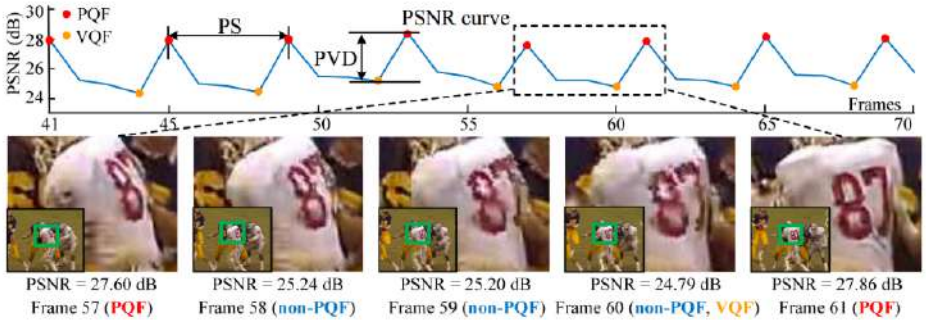
Hirokatsu Kataoka

Multi-Frame Quality Enhancement for Compressed Video

Ren Yang, Mai Xu, Zulin Wang, Tianyi Li CVPR 2018

概要

圧縮した動画像に対して画質を向上させる研究。Peak Quality Frames (PQFs)を用いたSVMベースの手法やMulti-Frame CNN (MF-CNN)を提案。提案法により、圧縮動画における連続フレームからアーティファクトを補正するような働きが見られた。



新規性・結果

動画の画質改善手法においてState-of-the-art。動画に対する画質改善の結果は図を参照。

リンク集

- 論文
- GitHub

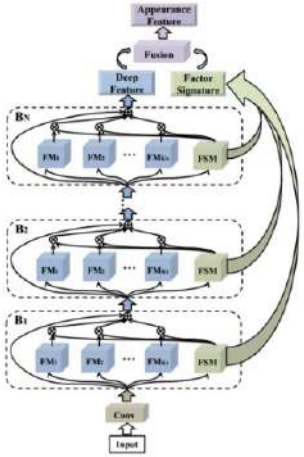
Munetaka Minoguchi

Multi-Level Factorisation Net for Person Re-Identification

Xiaobin Chang, Timothy M. Hospedales, Tao Xiang CVPR 2018

概要

人間の視覚的外観を、人の手によるアノテーションなしかつ、複数のセマンティックレベルで識別因子に分解する Multi-Level Factorisation Net(MLFN)の提案。 MLFNは、複数のブロックで構成されており、各ブロックには、複数の因子モジュールと、各入力画像の内容を解釈するための因子選択モジュールが含まれている。



新規性

効果的なRe-IDを目指すには、高低のセマンティックレベルでの人の差別化かつ視界不変性をモデル化することである。 近年(2018)のdeep Re-IDモデルは、セマンティックレベルの特徴表現を学習するか、アノテーション付きデータが必要となる。MLFNではこれらを改善する。

結果・リンク集

3つのRe-IDと、CIFAR-100の結果で最先端。

- 論文

[#379]

Tomoyuki Suzuki

Non-local Neural Networks

Xiaolong Wang et al. CVPR 2018

概要

NLPなどで効果を発揮しているself-attentionを多次元に一般化し、2D/3DCNNに導入することで新たな「non-local block」を形成し、画像や動画での実験を行った。 行動認識@Kineticsでは非常に高い精度を達成。Instance segmentationやkey point detectionなどのタスクでも汎用的に効果を発揮。

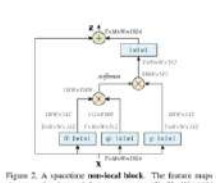


Figure 2. A spacetime **non-local block**. The feature maps are shown as the shape of their tensors, e.g., $T\times H\times W\times 1024$ for 1024 channels (proper reshaping is performed when noted). "⊗" denotes matrix multiplication, and "⊕" denotes element-wise sum. The softmax operation is performed on each row. The blue boxes denote $1\times1\times1$ convolutions. Here we show the embedded-Gaussian version, with a bottleneck of 512 channels. The vanilla Gaussian version can be done by removing θ and ϕ, and the dot-product version can be done by replacing softmax with scaling by $1/N$.

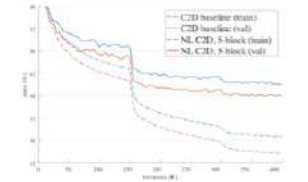


Figure 4. Curves of the training procedure on Kinetics for the ResNet-50 C2D baseline (blue) vs. non-local C2D with 5 blocks (red). We show the top-1 training error (dash) and validation error (solid). The validation error is computed in the same way as the training error (so it is 1-clip testing with the same random jittering at training time); the final results are in Table 2c (R50, 5-block).

$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})}\sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j)g(\mathbf{x}_j).$$

Figure 1. A spacetime **non-local** operation in our network trained for video classification. A position $\mathbf{x}_i$'s response is computed by the weighted average of the features of *all* positions $\mathbf{x}_j$ (only the highest weighted ones are shown here). In this example computed by our model, note how it relates the ball in the first frame to the ball in the last two frames. More examples are in Figure 3.

手法

位置jと位置iに依存してアテンションを出力する関数f(.)とjのみに依存する関数g(.)の積を入力位置jに関して和をとることによって位置iの出力値を決定する。 位置情報の保存、可変入力サイズ、などの性質を持ち、全結合、畳み込みを特殊な形として含む。またf(.)の定義の仕方によってはself-attentionと一致する。 f(.)は様々な形が提案されているが、種類によらず効果を発揮している。実際に使用する場合は図のような残差構造を使用している。

コメント・リンク

効果のインパクトがすごい。学習曲線からもうまくいっていることが明らか。C2Dに対してspace-timeにnon-local blockを適用すると3Dconvよりも時系列方向への拡大として効果があったのが興味深い。 結局残差を用いたnon-local blockを使用していたので、単純にnon-local layerのみでの性能もきになる。 位置情報の保存は重要でも、局所性はあまり重要ではなかったのかと感じられる。

- 論文

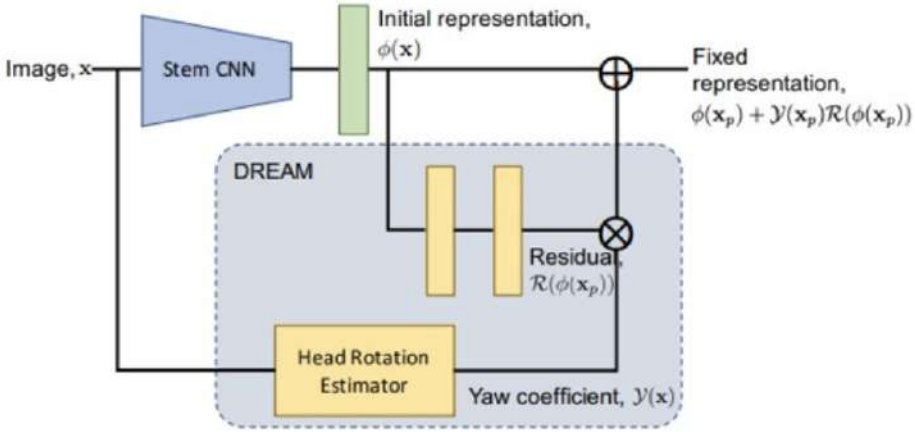
Munetaka Minoguchi

Pose-Robust Face Recognition via Deep Residual Equivariant Mapping

Kaidi Cao, Yu Rong, Cheng Li, Xiaoou Tang, Chen Change Loy CVPR 2018

概要

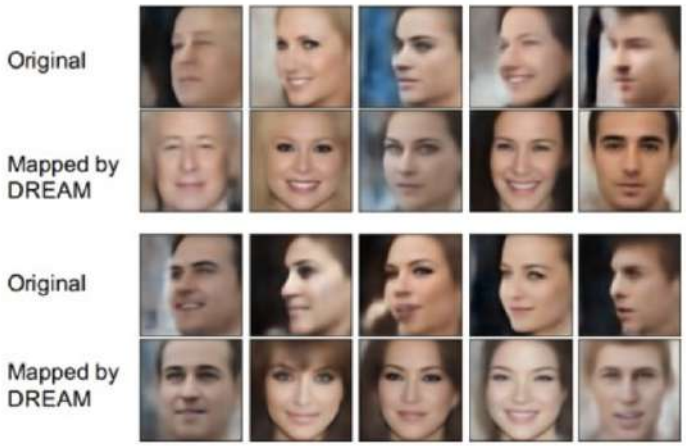
横顔の認識精度を高めるためにDeep Residual EquivAriant Mapping (DREAM)の提案。正面と側面の顔間のマッピングを行うことで特徴空間を対応付ける。これにより、横顔を正面の姿勢に変換して認識を単純化。



新規性・手法・リンク集

正面と側面のトレーニング数の不均衡から、現代の顔認識モデルの多くは、正面と比べて横顔を処理するのが比較的貧弱。本手法は姿勢変動を伴う顔認識に限定されない顔認識が可能で、横顔のデータを増やさなくても精度向上。

上図より、DREAMをCNNに追加し、入力に残差を動的に追加。下図はマッピングによる姿勢変換の例。



- 論文
- ソースコード

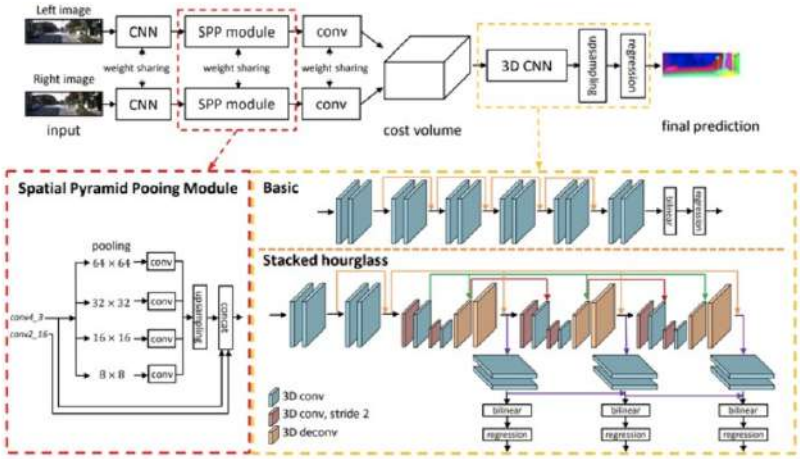
Munetaka Minoguchi

Pyramid Stereo Matching Network

Jia-Ren Chang, Yong-Sheng Chen CVPR 2018

概要

空間ピラミッドプーリングと3D CNNの2つのモジュールから構成された、ステレオ画像対からの奥行き推定を行うPyramid Stereo Matching Network(PSMNet)の提案。空間ピラミッドプーリングは、異なるスケールおよび位置でコンテキストを集約し、コストボリュームを形成する。 3D CNNは、複数のhourglass networksを重ねて、コストボリュームを規則化することを学習。



新規性

現在(2018)ではステレオ画像からの奥行き推定を、CNNの教師あり学習で解決されてきている。 コンテキスト情報を利用することで精度向上を図る。

結果・リンク集

最先端の手法よりも優れている結果。

- 論文
- ソースコード

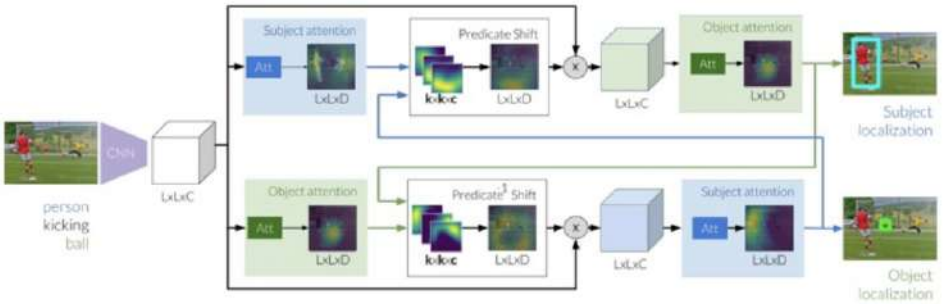
[#382]

Munetaka Minoguchi

Referring Relationships

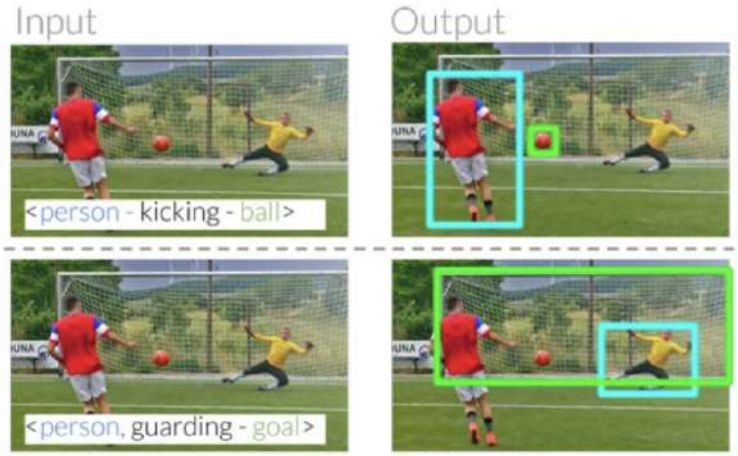
Ranjay Krishna, Ines Chami, Michael Bernstein, Li Fei-Fei CVPR 2018

概要



referring relationshipsを利用して同カテゴリのエンティティ間の曖昧さを解消するタスクの提案。特徴抽出後、アテンションを生成。述語を使用することで、アテンションをシフトさせる。この述語シフトモジュールを介して、subjectとobjectの間でメッセージを反復的に渡すことで、2つのエンティティをローカライズ。

新規性



画像中のエンティティ間の関係にはそれぞれ意味があり、画像の理解に役立つ。例えば、図のサッカーの試合の画像では、複数の人写っているが、それぞれは異なる関係を持っている。一人はボールを蹴っており、もう一人はゴールを守っている。 に着目すると、述語の”kick”を理解することにより、画像内のどの人物が”ball”を蹴っているのかを正しく識別する。

- 論文
- ソースコード

[#383]

Hirokatsu Kataoka

Rethinking Feature Distribution for Loss Functions in Image Classification

Weitao Wan, Yuanyi Zhong, Tianpeng Li, Jiansheng Chen CVPR 2018 (spotlight)

概要

本論文ではLarge-margin Gaussian Mixture (L-GM) Lossを提案して画像識別タスクに応用する。Softmax Lossとの違いは、学習セットにおけるディープ特徴の混合ガウス分布をフォローしつつ仮説を設定するところである。識別境界や尤度正則化においてL-GM Lossは非常に高いパフォーマンスを実現している。

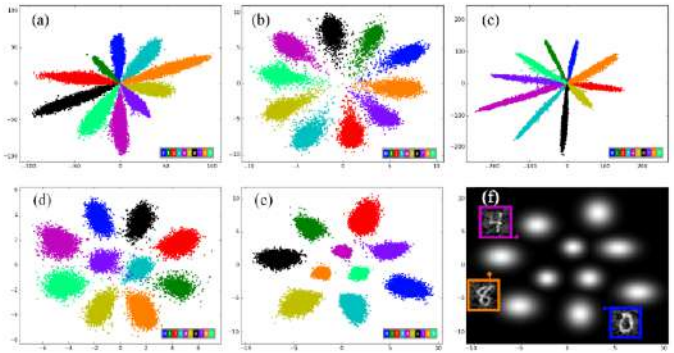


Figure 1. Two-dimensional feature embeddings on MNIST training set. (a) Softmax loss. (b) Softmax loss + center loss [32]. (c) Large-margin softmax loss [22]. (d) GM Loss without margin ($\alpha = 0$). (e) Large-margin GM loss ($\alpha = 1$). (f) Heatmap of the learned likelihood corresponding to (e). Higher values are brighter. Several adversarial examples generated by the Fast Gradient Sign Method [8] have extremely low likelihood according to the learned GM distribution and thus can be easily distinguished. This figure is best viewed in color.

新規性・結果

L-GM Lossは画像識別においてSoftmax Lossよりも精度が高いことはもちろん、特徴分布を考慮するため例えばAdversarial Examples（摂動ノイズ）などにおいても対応できる。MNIST, CIFAR, ImageNet, LFWにおける識別や摂動ノイズを加えた実験においても良好な性能を確かめた。

コメント・リンク集

Softmax Lossよりも有意に精度向上が見られている。導入が簡単なら取り入れて精度向上したい。

- 論文

[#384]

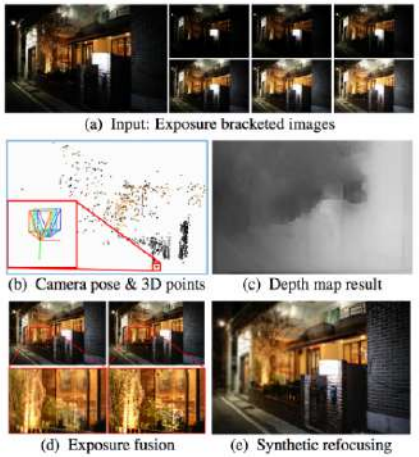
Hirokatsu Kataoka

Robust Depth Estimation from Auto Bracketed Images

Sunghoon Im, Hae-Gon Jeon, In So Kweon CVPR 2018

概要

HDRの画像の明るさを補正するためのブラケット撮影からの距離画像やカメラ姿勢を同時推定する手法を提案する論文。ブラケット撮影とは通常の露出撮影以外に意図的に「少し明るめの写真」と「少し暗めの写真」を同時に撮影。距離画像推定は幾何変換をResidual-flow Networkに統合したモデルにより行う。ここでは学習ベースのMulti-view stereo手法（Deep Multi-View Stereo; DMVS）を幾何推定（Structure-from-Small-Motion; SfSM）と組み合わせる。



新規性・結果

距離画像推定において、スマートフォンやDSLRカメラなど種々のデータセットにてSoTAな精度を達成。モバイル環境でも動作するような小さなネットワークと処理速度についても同時に実現した。

リンク集

- 論文
- 著者

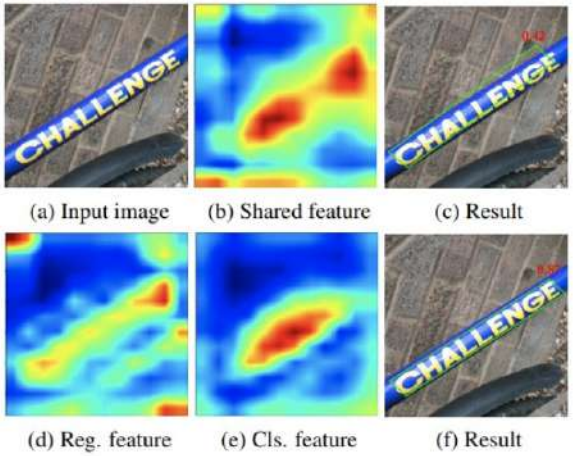
Munetaka Minoguchi

Rotation-Sensitive Regression for Oriented Scene Text Detection

Minghui Liao, et al. CVPR 2018

概要

自然画像から文字を検出する。単なる検出ではなく、文字の方向を考慮したバウンディングボックスによる検出手法であるRotation-sensitive Regression Detector (RRD)の提案。回帰ブランチによって、畳み込みフィルタを回転させて回転感知特徴を抽出。分類ブランチによって、回転感性特徴をプーリングすることによって回転不変特徴を抽出。



新規性

文字をテーマにした研究では(1)テキストの向きを無視した分類方法と，(2)向きを考慮したバウンディングボックスによる回帰がある。従来研究では、両方のタスクの共有の特徴を使用していたが、互換性がなかったためにパフォーマンスが低下(図b)。そこで、異なる2つのネットワークから抽出した、異なる特性の特徴を分類および回帰することを提案(図d,e)。

結果・リンク集

ICDAR 2015、MSRA-TD500、RCTW-17およびCOCO-Textを含む3つのシーンテキストのデータセットで最先端のパフォーマンスを達成。向きがある一般物体検出にも応用可能？

- 論文

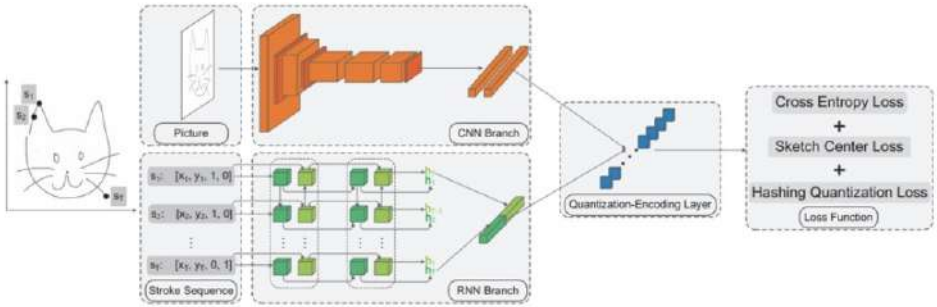
Munetaka Minoguchi

SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval

Peng Xu, et al. CVPR 2018

概要

スケッチ検索のためのディープハッシングフレームワークの提案。3.8mの大規模スケッチデータセットを構築。CNNでスケッチの特徴抽出。RNNでペンストロークの時間情報をモデル化。CNN-RNNでエンコードすることで、スケッチ性質に対応した新しいhashing lossを導入。



新規性・差分

従来のスケッチ認識タスクに従う代わりに、より困難な問題のスケッチハッシュ検索を行う。ネットワークをスケッチ認識のために再利用することもでき、どちらも高パフォーマンス。大規模なデータセットを利用することで、従来の文献ではあまり研究されていなかった、スケッチのユニークな特性を見出す。

リンク集

- 論文
- ソースコード/データセット

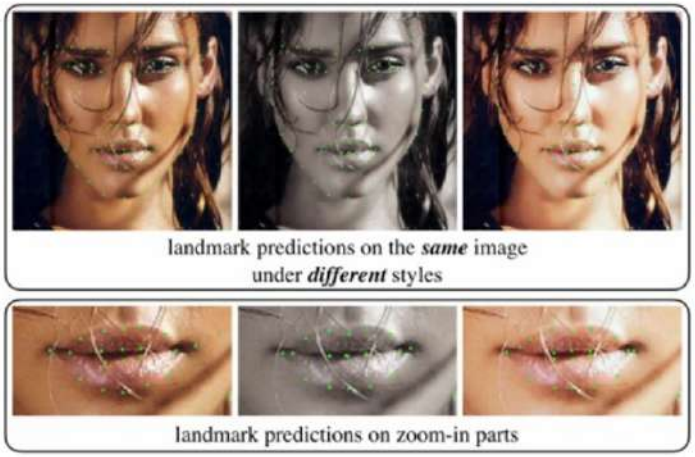
Munetaka Minoguchi

Style Aggregated Network for Facial Landmark Detection

Xuanyi Dong, Yan Yan, Wanli Ouyang, Yi Yang, University of Technology Sydney, The University of Sydney CVPR 2018

概要

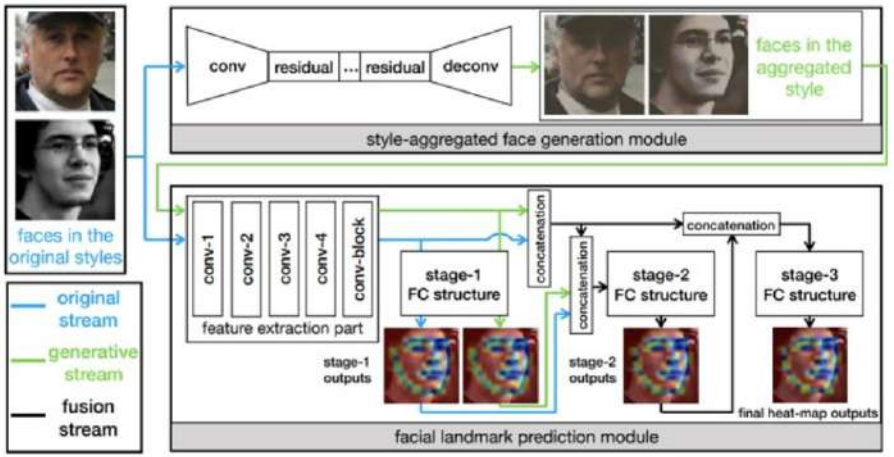
顔のランドマーク検出。顔そのもののばらつきの他に、グレースケールやカラー画像、明暗などの画像スタイルが変わっても同様に検出できるStyle Aggregated Network(SAN)の提案。まず、(1)入力画像をさまざまなスタイルに変換し、スタイルを集約し、(2)顔のランドマーク予測する。(2)は、元画像とスタイルを集約した特徴の両方を入力し、融合してカスケード式のヒートマップ予測を生成する。



結果・リンク集

Flickr8kとFlickr30kを使った実験において、最先端モデルと同等かそれ以上の結果。より正確で、より多様なキャプション生成。

- 論文
- ソースコード



[#388]

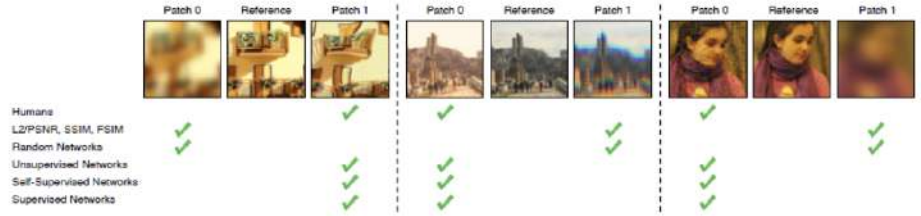
ShintaroYamamoto

The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

Richard Zhang et al. CVPR 2018

概要

2枚の画像の類似度を表す指標は数多く提案されているが、その類似度は必ずしも人間の知覚と一致していない。近年はDNNにより高次の特徴を得ることが可能となっており、人間の知覚に近づいている。 そこで、既存の類似度の評価尺度とDNNベースの類似度判定を比較することでDNNベースの手法がより人間の知覚に近い類似度を表現できることを確認した。 具体的には、ある画像を異なる方法で加工したもの2つを用意し、どちらが元の画像に近いかを人間とコンピュータ両方に判定させることで検証を行った。



新規性・結果

データセットとして、画像に様々な加工を施したデータを人間に類似度を評価してもらったものを作成。加工の例としては、ノイズの付与やオートエンコーダによる画像の復元などが挙げられる。 検証の結果、ＤＮＮベースの類似度の方が既存の尺度より人間の知覚に乗っ取ってることを示した。 また、DNNのネットワーク構造そのものは重要ではないことが分かった。

リンク集

- プロジェクトページ

[#389]

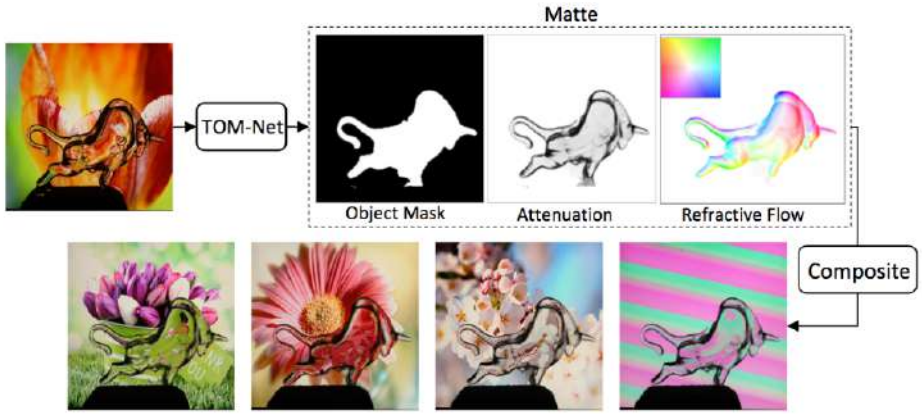
Hirokatsu Kataoka

TOM-Net: Learning Transparent Object Matting from a Single Image

Guanying Chen, Kai Han, Kwan-Yee K. Wong CVPR 2018 (spotlight)

概要

透明物体の切り抜き（Transparent Object Matting; TOM）と反射特性を推定することが可能なネットワークTOM-Netを提案する。TOM-Netにより、物体の反射特性を保存しながら他の画像にレンダリングして、同画像のテクスチャを反映させることができる。同問題を反射フローの推定問題と捉えてDNNのモデルを構築することで解決した。荒い部分は多階層のEncoder-Decorderで推定し、詳細な部分はResidualNetで調整する。この問題を解決するために、データセットを構築した。



新規性・結果

178Kの画像を含むデータセットを構築した。同DBには876サンプル、14の透明物体、60種の背景を含む。透明物体の推定と反射特性のレンダリングはGitHubページを参照。

コメント・リンク集

- 論文
- Project
- GitHub

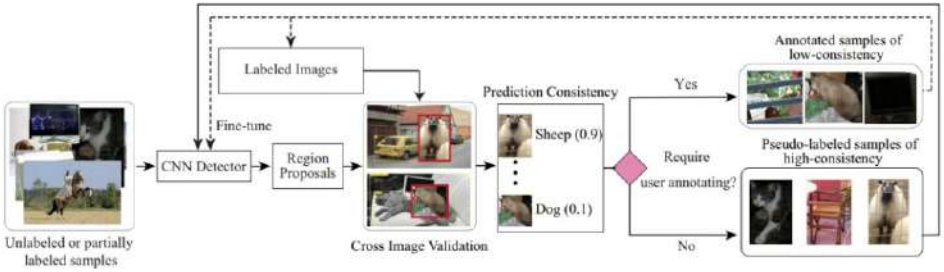
Munetaka Minoguchi

Towards Human-Machine Cooperation:Self-supervised Sample Mining for Object Detection

Keze Wang, et al. CVPR 2018

概要

物体検出の課題を考慮し、既存のActive Learning(AL)の欠点を改善することを目的とした、Self-Supervised Sample Mining(SSM)の提案。ラベルなし、もしくは一部ラベルのないデータを使って学習することができる。交差検証後のスコアによってサンプルを選別。低い場合にはユーザによってアノテーション、高い場合にはそのままラベルとして採用。



新規性

既存のAL法では主に、単一の画像コンテクスト内でサンプル選択基準を定義し、大規模な物体検出において最適ではなく、頑強性および非実用的である。SSMによって、ユーザが必要な部分にだけ介入し、アノテーションの作業を軽減。

結果・リンク集

アノテーションが少ないデータセットにおいても最先端の精度。

- 論文

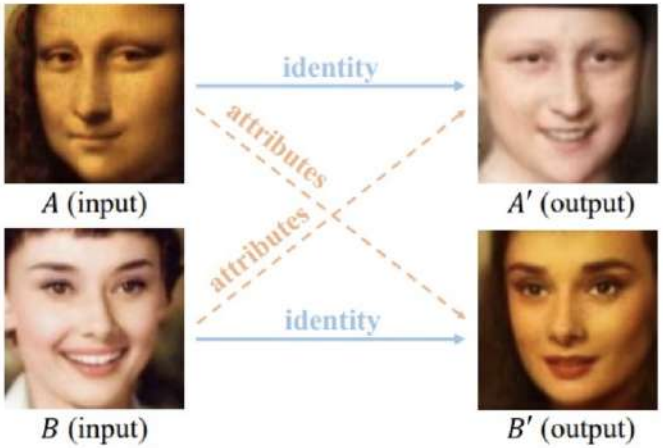
Munetaka Minoguchi

Towards Open-Set Identity Preserving Face Synthesis

Jianmin Bao, et al. CVPR 2018

概要

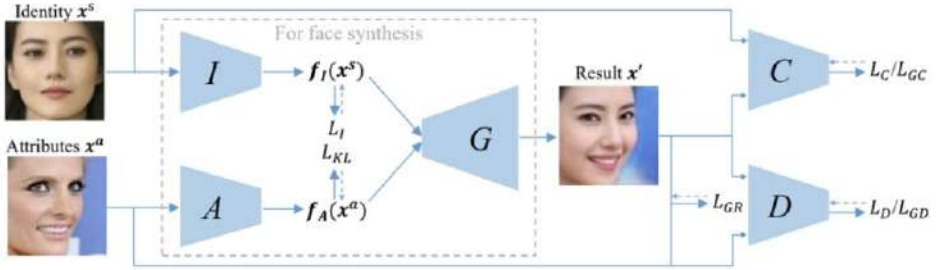
顔画像からidentityとattributesを別々に再構成する、GANに基づいたOpen-Set Identity Generating Adversarial Networkの提案。 face synthesis networkは、ポーズや感情、照明、背景などをキャプチャする属性ベクトルを抽出することができる。図中の2つの入力画像AおよびBから抽出された識別を再結合することによって、A0およびB0を生成することができる。



新規性・結果・リンク集

顔の正面化、顔属性モーフィング、 face adversarial example detectionなど、より広範なアプリケーションに応用可能。

- 論文



[#392]

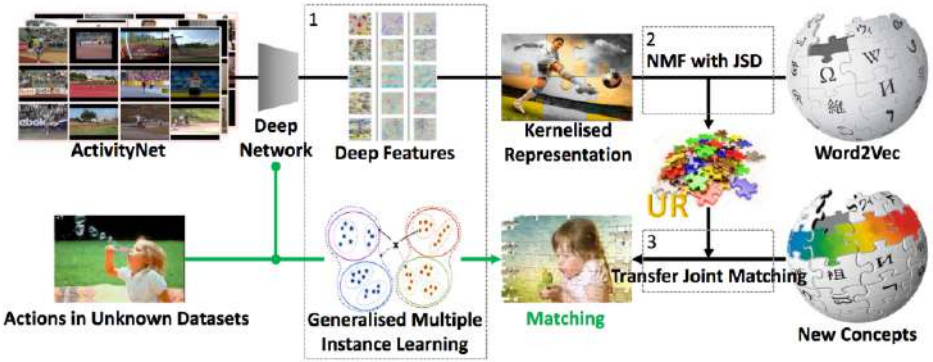
Hirokatsu Kataoka

Towards Universal Representation for Unseen Action Recognition

Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, Ling Shao CVPR 2018

概要

学習画像がなくても行動認識を実現する「Unseen Action Recognition (UAR)」についての研究。UARの問題をMIL（Multiple Instance Learning）の一般化（GMIL）として扱い、ActivityNetなど大規模動画データから分布推定して表現を獲得。図は提案手法であるCross-Domain UAR (CD-UAR)である。ビデオから抽出したDeep特徴はGMILによりカーネル化される。Word2Vecとの投稿によりURを獲得し、ドメイン変換により新しい概念を獲得する。



新規性・結果

従来法では見た/見てないの対応関係をデータセット中に含ませていたが、本論文での提案はUniversal Representation（ユニバーサル表現）を獲得して同タスクを解決する。

リンク集

- 論文

Munetaka Minoguchi

Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatial-Temporal Patterns

Jianming Lv, et al. CVPR 2018

概要

歩行者の時空間パターンを用いた、教師なし学習の人物再同定アルゴリズムであるTFusionを提案。既存の人物再同定アルゴリズムのほとんどは、小サイズのラベル付きデータセットを用いた教師付き学習手法である。そのため、大規模な実世界のカメラネットワークに適応することは困難である。また、そこで、ラベルなしデータセットも用いたクロスデータセット手法によって精度向上を図る。

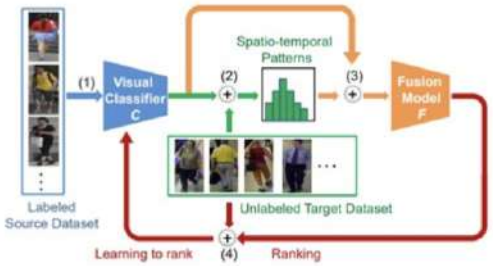


Figure 1: The *TFusion* model consists of 4 steps: (1) Train the visual classifier $\mathcal{C}$ in the labeled source dataset (Section 4.2); (2) Using $\mathcal{C}$ to learn the pedestrians' spatio-temporal patterns in the unlabeled target dataset (Section 4.3); (3) Construct the fusion model $\mathcal{F}$ (Section 4.4); (4) Incrementally optimize $\mathcal{C}$ by using the ranking results of $\mathcal{F}$ in the unlabeled target dataset (Section 4.6).

手法

まず、歩行者の空間的-時間的パターンを学習するために、ラベル付きデータセットを用いて学習した視覚的分類器を、ラベルなしデータセットに転送。次に、Bayesian fusion modelによって、学習された時空間パターンを視覚的特徴と組み合わせて、分類器を改善。最後に、ラベルのないデータを用いて分類器を段階的に最適化。

結果・リンク集

人物再同定のための、教師なしクロスデータセット学習手法の中では最先端。

- 論文

[#394]

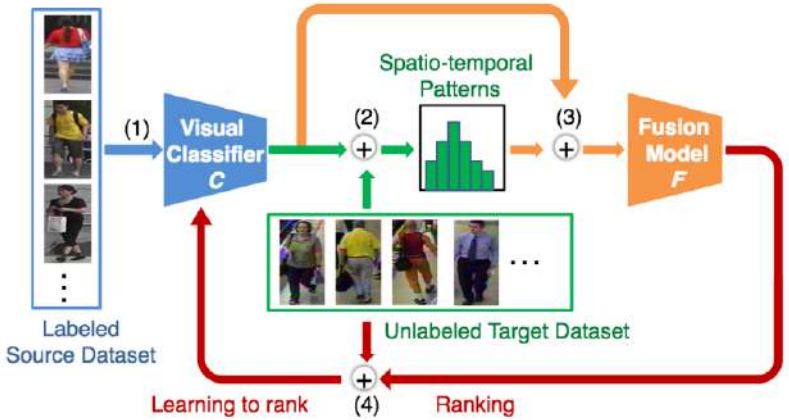
Hirokatsu Kataoka

Unsupervised Cross-dataset Person Re-identification by Transfer Learning of Spatio-temporal Patterns

Jianming, Lv and Weihang, Chen and Qing, Li and Can, Yang CVPR 2018

概要

ラベルなし、ドメインが異なる環境に対して人物再同定を行う手法を提案する。モデルであるTFusionは４ステップにより構築（１）教師あり学習により識別器を構築（２）ターゲットであるラベルなしデータにより時空間特徴パターン（Spatio-temporal Pattern）を学習（３）統合モデルFを学習（４）ラベルなしのターゲットデータにて徐々に識別器を学習する（１～４は図に示されている）。Bayesian Fusionを提案して、時空間特徴パターンと人物のアピアランス特徴を統合してドメイン変換を行う。



新規性・結果

従来の人物再同定の設定では比較的小さいデータセットであり、完全に教師ありの環境を想定していたが、本論文ではラベルなし、ドメインが異なる環境に対して人物再同定を実行するため、非常に難しい問題となる。

コメント・リンク集

- 論文
- GitHub

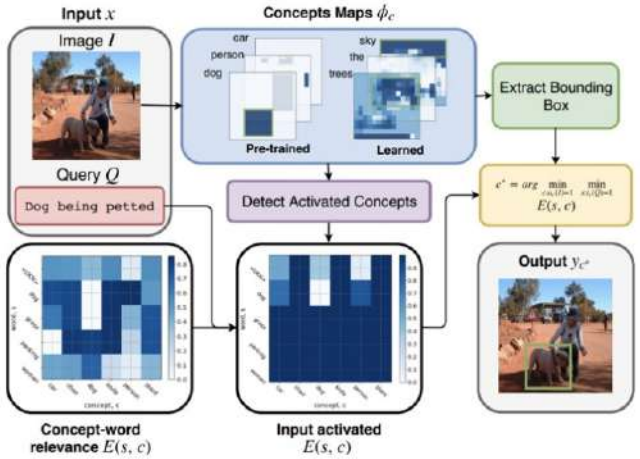
Munetaka Minoguchi

Unsupervised Textual Grounding: Linking Words to Image Concepts

Raymond A. Yeh, Minh N. Do, Alexander G. Schwing CVPR 2018

概要

単語を検出された画像の概念に関連付けるための、仮説検定を用いた教師なしTextual grounding手法の提案。ネットワークにはVGG-16を採用し、画像内のオブジェクト/単語の空間情報やクラス情報、およびクラス外の新しい概念を学習できる。



新規性

Textual grounding、すなわち画像内のオブジェクトと単語をリンクさせる既存の技法は、教師付きのディープラーニングとして定式化されており、大規模なデータセットを用いてバウンディングボックスを推定する。しかし、データセットの構築には時間やコストがかかるので教師なしの手法を提案。

結果・リンク集

ReferIt GameとFlickr30kを用いたベンチマークでそれぞれ7.98%と6.96%以上の精度。

- 論文

[#396]

Hirokatsu Kataoka

Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments

Peter Anderson, et al. CVPR 2018

概要

自然言語のナビゲーションを入力として、実空間の中をロボットが動き目的地に到達できるかどうかを競うベンチマーク（Visually-grounded natural language navigation in real buildings）を提案。データセットは3Dのシミュレータによりキャプチャされ、22Kのナビゲーション、文章の平均単語数は29で構成される。

新規性・結果

(1) Matterport3Dデータセットを強化学習を行えるように拡張。(2) 同タスクが行えるようなベンチマークであるRoom-to-Room (R2R) を提案して言語と視覚情報から実空間にてナビができるようにした。(3) seq-to-seqをベースとしたニューラルネットによりベンチマークを構築。VQAをベースにしていて、ナビゲーション（VQAでいう質問文）と移動アクション（VQAでいう回答）という組み合わせで同問題を解決する。

コメント・リンク集

自然言語の問題はキャプションや質問回答の枠を超えて実空間、さらにいうとロボットタスクに導入されつつある。この研究はビジョン側からのアプローチだが、ロボット側のアプローチが現在どこまでできているか気になる。すでに屋内環境をある程度自由に移動するロボットが実現しているとこの実現可能性が高くなる。SLAMとの組み合わせももう実行できるレベルにある？

- 論文
- Project
- GitHub
- Matterport3D dataset

[#397]

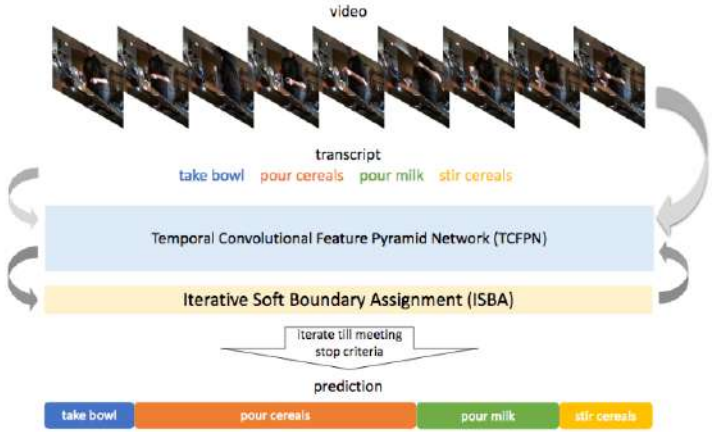
Hirokatsu Kataoka

Weakly-Supervised Action Segmentation with Iterative Soft Boundary Assignment

Li Ding, Chenliang Xu CVPR 2018

概要

時系列の行動検出/セグメンテーション（Action Segmentation）に関する問題をWeakly-Supervised（WS学習）に解いた。ここではTemporal Convolutional Feature Pyramid Network (TCFPN)とIterative Soft Boundary Assignment (ISBA)を繰り返すことで行動に関する条件学習ができてくるという仕組み。TCFPNではフレームの行動を予測し、ISBAではそれを検証、それらを繰り返して行動間の境界線を定めながらWS学習の教師としていく。さらに、WS学習を促進するためにより弱い境界として行動間の繋がりを定義することでWS学習の精度を向上させる。学習はビデオ単位の誤差を最適化することで境界についても徐々に定まる（ここがWS学習の所以）ように学習する。



新規性・結果

Breakfast dataset, Hollywood extended datasetにて弱教師付き学習とテストを行いState-of-the-artな精度を達成した。

コメント・リンク集

弱い教師データを大量に集めると、そろそろ（ある程度の）教師ありデータによる精度を超えそう？もっと汎用的に学習できる枠組みが必要か。

- 論文

[#398]

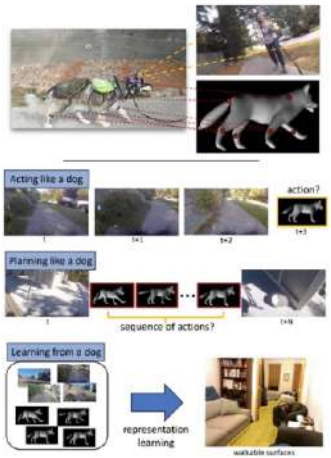
Munetaka Minoguchi

Who Let The Dogs Out? Modeling Dog Behavior From Visual Data

Kiana Ehsani, et al. CVPR 2018

概要

犬視点の大規模ビデオデータセットを作成し、このデータを使用した、犬の行動や行動計画のモデル化。次の3つの問題に焦点を当てる。(1)犬の行動予測。(2)入力された画像対から犬のような行動計画を見出す。(3)例えば、歩行可能な表面推定などのタスクについて、学習された表現を利用。



新規性

視覚情報からintelligent agent(知的エージェント)を直接的にモデリングするタスク。犬の視覚情報を使うことで、行動をモデル化する斬新な取り組み。得られたモデルをAIなどに応用する。特に、歩行可能な表面推定のタスクで良い結果となる。

結果・リンク集

様々なエージェントやシナリオで使用でき、ラベルがないにもかかわらず有用な情報を学習することが可能。今後は、モデルやデーセットの拡張に挑む。

- 論文

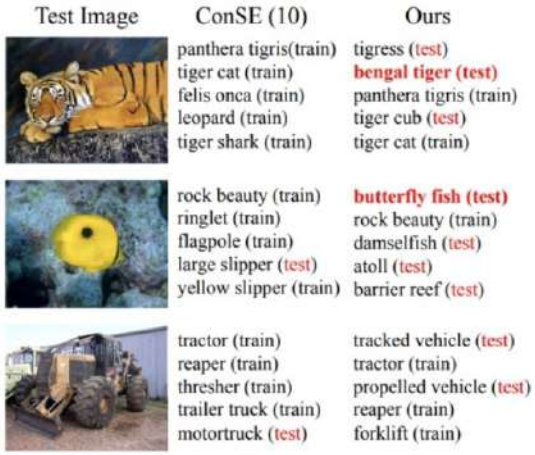
Munetaka Minoguchi

Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs

Xiaolong Wang, Yufei Ye, Abhinav Gupta, The Robotics Institute, Carnegie Mellon University CVPR 2018

概要

カテゴリの単語の埋め込みと他のカテゴリとの関係(視覚データが提供される)を使用するだけで、学習例がないカテゴリの分類器を学習するゼロショット認識モデルを提案。 knowledge graph (KG) を入力とし、Graph Convolutional Network(GCN)を基に、セマンティック埋め込みとカテゴリの関係の両方を使用して分類器を予測する。



手法

学習済のKGが与えられると、各ノードに対する意味的埋め込みとして入力を得る。一連のグラフ畳み込みの後、各カテゴリの視覚的分類器を予測する。トレーニング中に、カテゴリの視覚的分類器が与えられ、GCNパラメータを学習。テスト時に、これらのフィルタを使用して、見えないカテゴリの視覚的分類器を予測する。

結果・リンク集

KGのノイズに対してロバストであり、最先端の精度。

- 論文

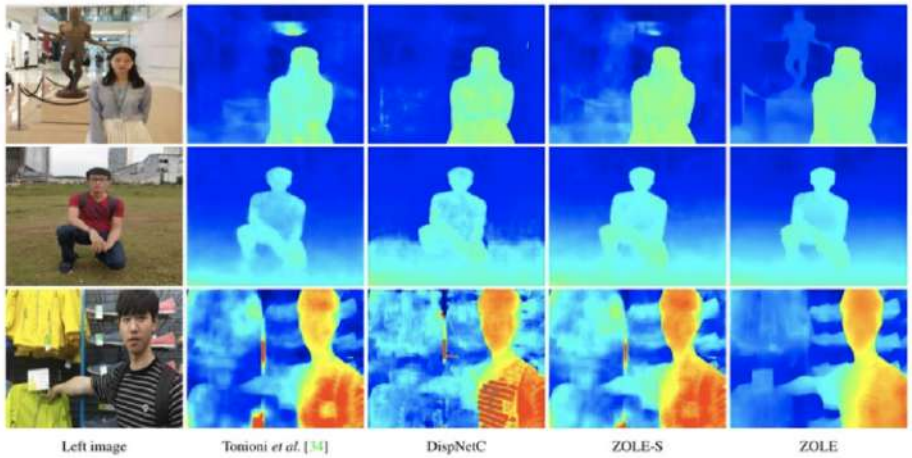
Munetaka Minoguchi

Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains

Jiahao Pang, et al. CVPR 2018

概要

学習済みデータと新しいドメイン(ground-truthなし)の両方を用いて、ディープステレオマッチングを行うZoom and Lean(ZOLE)の提案。これにより，他のドメインに一般化できるプレトレインモデルを作成することができる。一般化に際する不具合を抑制しながらアップサンプリングを行う、反復最適化問題を定式化する。



新規性

ground-truthデータが不足しているため、CNNを用いたステレオマッチングでは学習済みステレオモデルを新規ドメインに一般化することが困難とされていた。CNN学習時のイテレーションごとに最適化していくイメージ。

結果・リンク集

スマートフォンで収集したデータを従来の手法に入力すると、物体のエッジがぼやけてしまうが、提案手法のZOLEではこれらを改善できる。

- 論文

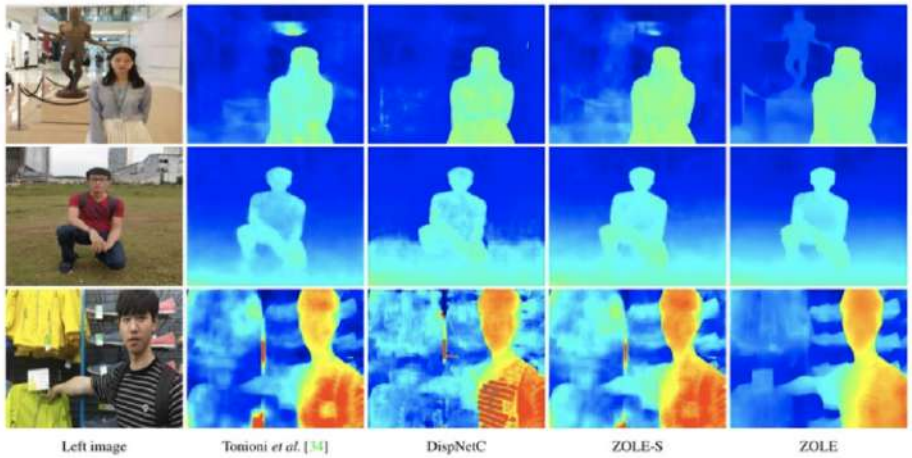
Munetaka Minoguchi

Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains

Jiahao Pang, et al. CVPR 2018

概要

学習済みデータと新しいドメイン(ground-truthなし)の両方を用いて、ディープステレオマッチングを行うZoom and Lean(ZOLE)の提案。これにより，他のドメインに一般化できるプレトレインモデルを作成することができる。一般化に際する不具合を抑制しながらアップサンプリングを行う、反復最適化問題を定式化する。



新規性

ground-truthデータが不足しているため、CNNを用いたステレオマッチングでは学習済みステレオモデルを新規ドメインに一般化することが困難とされていた。CNN学習時のイテレーションごとに最適化していくイメージ。

結果・リンク集

スマートフォンで収集したデータを従来の手法に入力すると、物体のエッジがぼやけてしまうが、提案手法のZOLEではこれらを改善できる。

- 論文

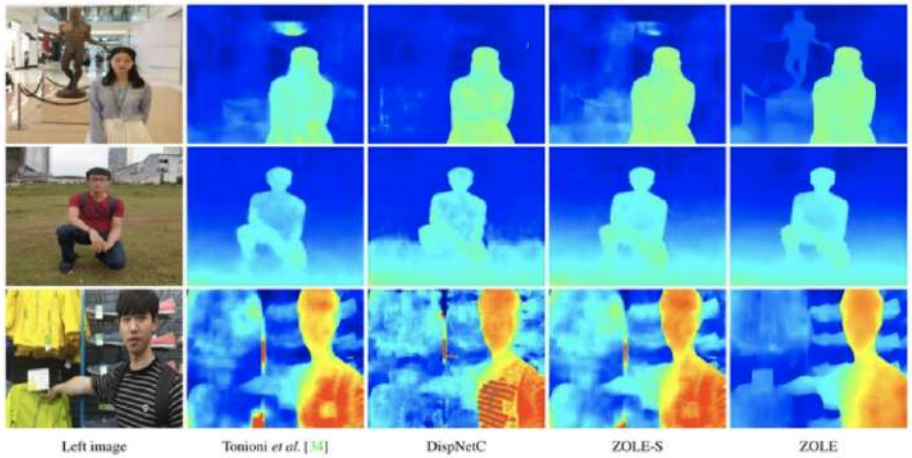
Munetaka Minoguchi

Zoom and Learn: Generalizing Deep Stereo Matching to Novel Domains

Jiahao Pang, et al. CVPR 2018

概要

学習済みデータと新しいドメイン(ground-truthなし)の両方を用いて、ディープステレオマッチングを行うZoom and Lean(ZOLE)の提案。これにより，他のドメインに一般化できるプレトレインモデルを作成することができる。一般化に際する不具合を抑制しながらアップサンプリングを行う、反復最適化問題を定式化する。



新規性

ground-truthデータが不足しているため、CNNを用いたステレオマッチングでは学習済みステレオモデルを新規ドメインに一般化することが困難とされていた。CNN学習時のイテレーションごとに最適化していくイメージ。

結果・リンク集

スマートフォンで収集したデータを従来の手法に入力すると、物体のエッジがぼやけてしまうが、提案手法のZOLEではこれらを改善できる。

- 論文

ご質問・コメント等ありましたら，cvpaper.challenge@gmail.com / Twitter@CVPaperChallengまでお願いします．