cvpaper.challenge

CVPR 2023 速報

片岡裕雄，八木拓真，Qiu Yue，中村凌，山田亮佑，
篠田理沙，Yanjun Sun，高島空良，速水亮，
松澤郁哉，田所龍，福原吉博，和田唯我，千葉直也，
舘野将寿，柴田優斗，西山翔大，上田樹，大谷豪

DNNの動向・CVのトレンド

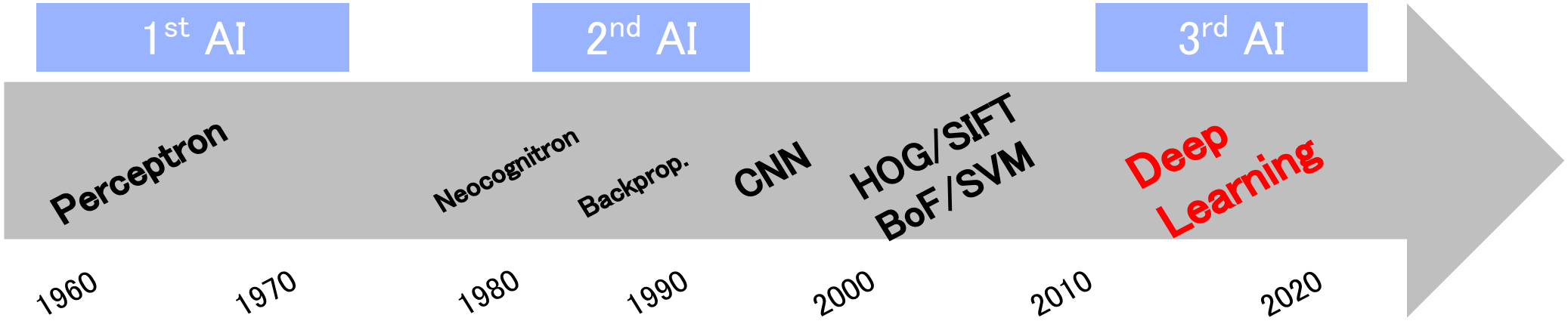
- 現在までの流れをおさらい
- CV分野の最近のトレンドも記載
- 下記資料をベースに新規の内容も追加
- トレンドご存知の方はp.54からご覧ください

DNNの動向・CVのトレンド(1/51)

DNN時代以前の動向

- Perceptron, MLP, Neocognitron, BackProp, CNN
- DNNが流行る直前の画像認識では局所特徴が使用
- 深層学習(Deep Learning)の隆盛期にある現在は第3次AIブーム

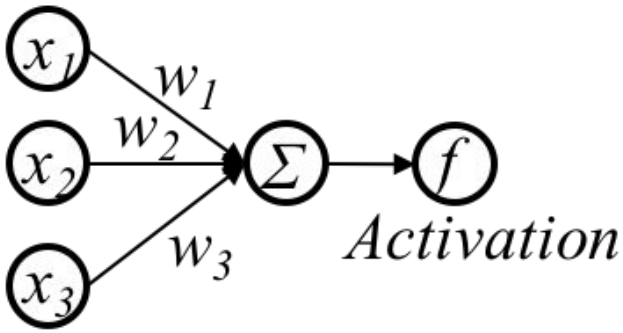
1st – 3rd AIまでの流れ



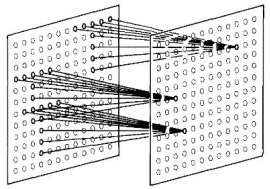
DNNの動向・CVのトレンド(2/51)

Perceptron, MLP, Neocognitron/ConvNet

- Perceptron
  - 入力とコネクション(重み)の線形和，活性化関数により構成
- MLP: Multi-layer Perceptron
  - Perceptronの多層化
- Neocognitron/ConvNet
  - 畳込みの概念を導入，特に隣接ピクセルに類似関係のある画像処理に有効

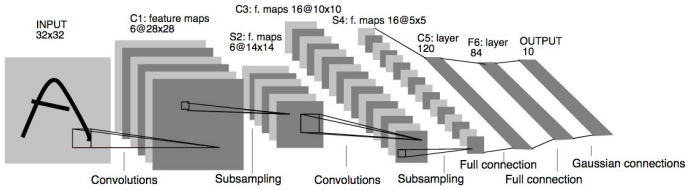


Perceptron(パーセプトロン)



Neocognitron

K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," Biol. Cybenetics 36, pp.193-202, 1980.
https://www.rctn.org/bruno/public/papers/Fukushima1980.pdf



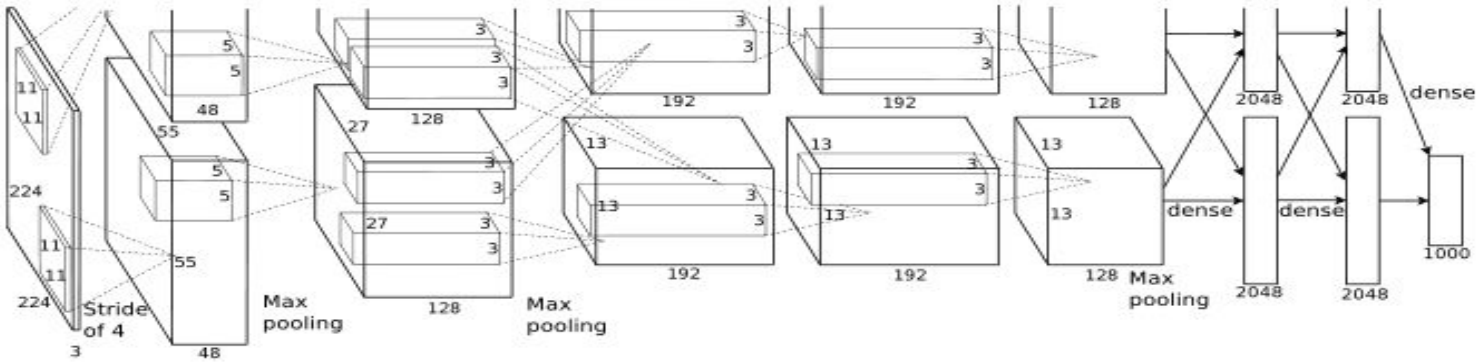
Convolutional Neural Net

Y. LeCun et al. "Gradient-Based Learning Applied to Document Recognition," IEEE, 1998.
http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

DNNの動向・CVのトレンド(3/51)

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
  - もはや専門家には説明不要だが，2位に10%以上の大差で圧勝
  - 10年弱で100,000+回も引用される論文となる
- 背景には構造をDeepにする技術が揃っていた
  - 最適化，活性化関数，正則化など



AlexNet

A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," in NIPS 2014.
https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

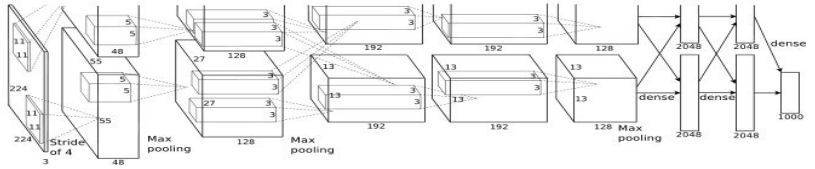
DNNの動向・CVのトレンド(4/51)

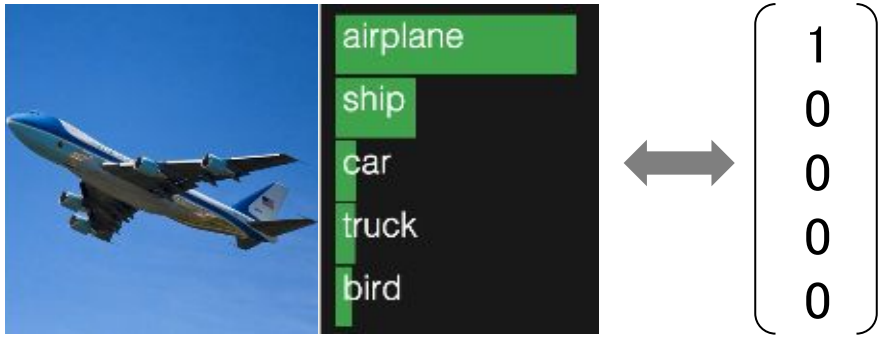
ネットワークの学習について

- 基本は推論(順伝播)と誤差逆伝播の繰り返し
- まずは推論(左図)により確率分布を計算
- 次いで誤差計算と逆伝播(右図)を行いネットワークを調整

画像入力 → 確率分布出力

誤差を順にネットワークに反映して調整することで学習





推論により得た確率分布(左)と正解値(右; 基本は正解ラベルのみ1のベクトル)を比較して誤差を計算

DNNの動向・CVのトレンド(5/51)

DNNが勝てた背景

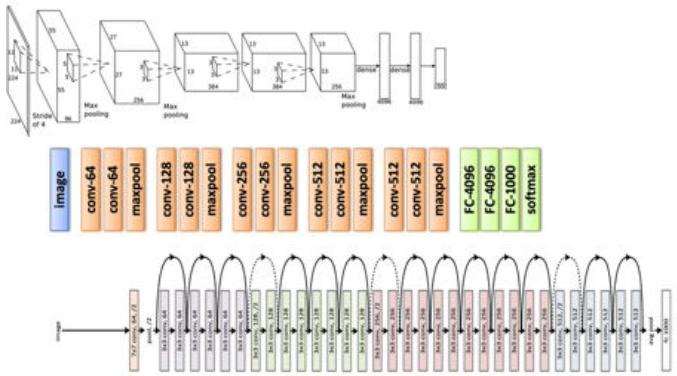
- DNN Architecture!(モデル自体や学習戦略)
- Large-scale Image Datasets!(データは最重要)
- Machine Power!(圧倒的な計算力; NVIDIA!)

Why 3rd AI?

Architecture(Model)
ConvNet, Transformer, GAN…

Image Data
ImageNet, Places, COCO…

Machine Power
GPU, TPU, Super Computer…



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

NVIDIA

https://nvidianews.nvidia.com/multimedia/corporate/corporate-nvidia-logos-legal-filings

DNNの動向・CVのトレンド(6/51)

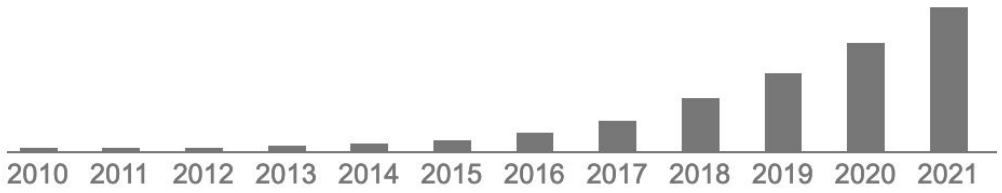
大規模画像データセット(ImageNet)の収集について

- 14,197,122 画像 / 21,841 カテゴリ https://www.image-net.org/
- 2007年からデータを収集，2009年CVPR発表
- その後，画像認識コンペを開催し2012年大会でAlexNetが提案される

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一世界平和一を示す(らしい)

総被引用数　引用元 35595



https://scholar.google.com/citations?view_op=view_citation&hl=ja&user=rDfyQnIAAAAJ&citation_for_view=rDfyQnIAAAAJ:qjMakFHDy7sC

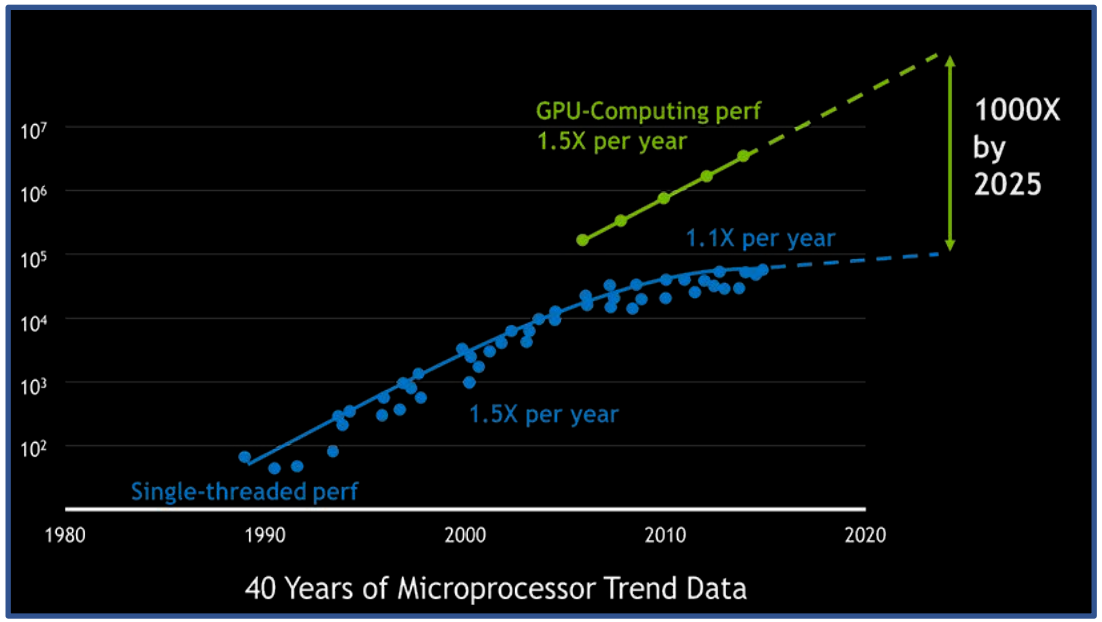
Fei-Fei氏のTED動画では資金繰りの苦労について言及，ワークショップ等では2000年代当時はアルゴリズム至上主義でデータを収集することが理解されなかったとも
https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures

ImageNet論文(CVPR 2009)の被引用数：コンピュータビジョン分野における深層学習の隆盛を象徴する存在

DNNの動向・CVのトレンド(7/51)

計算機環境(主にGPU)の発展

- 特に第3次AIブームからはNVIDIAの隆盛ぶりがすごい
- 当初ゲーム用ボードを売っていたが深層学習始めAI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を進めるに至る



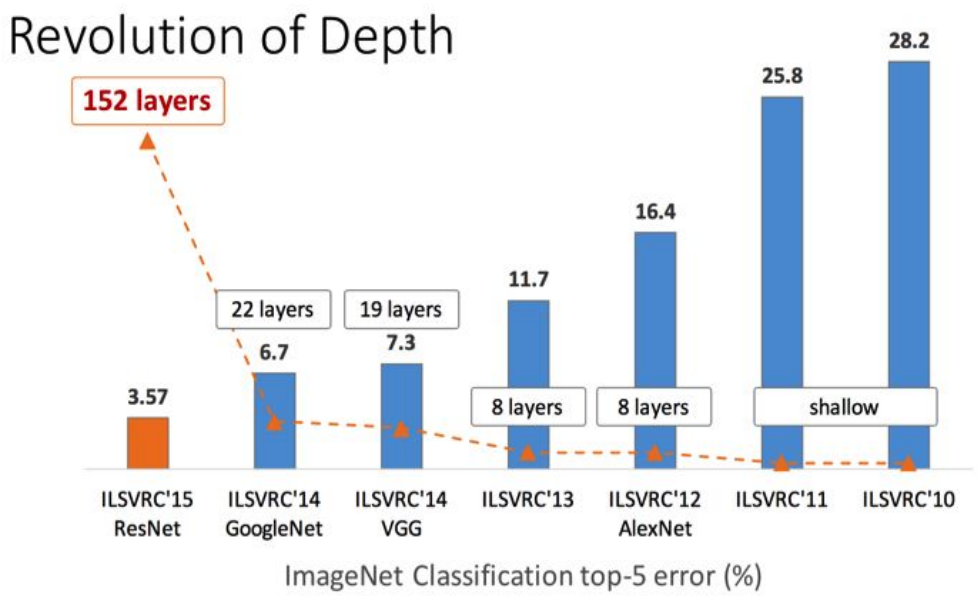
https://www.nvidia.com/ja-jp/about-nvidia/ai-computing/

年々1.5倍のペースでコンピューティングが進化，2025年までには1,000倍にまで到達すると予想
(NVIDIA公式ページより)

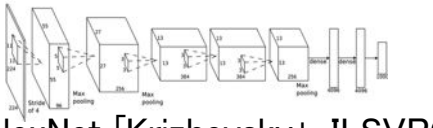
DNNの動向・CVのトレンド(8/51)

構造の深化(2014～2016)

- 2014年頃から「構造をより深くする」ための知見が整う
  - AlexNet, VGGNet, GoogleNetは，ほぼ単純に層を追加
  - ResNetは残差(Residual)の考え方を導入，Skip Connectionにより大幅に深化



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.
http://kaiminghe.com/ilsvrc15/ilsvrc2015_deep_residual_learning_kaiminghe.pdf



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

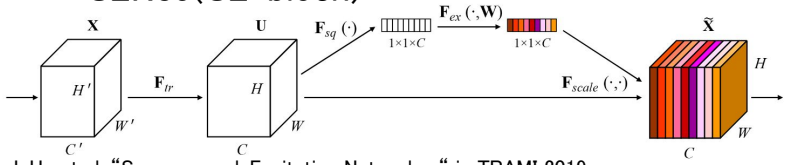
ResNet [He+, ILSVRC2015/CVPR2016]
ILSVRC2015 winner， 152層！(実験では10^3+層も)

DNNの動向・CVのトレンド(9/51)

構造の複雑化・自動化(2016～2019)

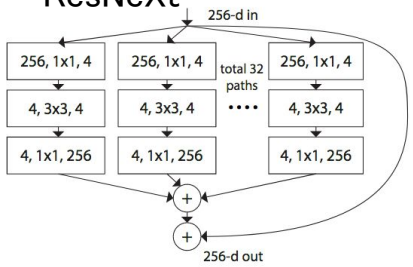
- ResNet以降のアーキテクチャ
  - ResNeXt, DenseNet, SENet, ...
- 自動化・効率化(Neural Architecture Search, EfficientNet)
  - (P)NAS, EfficientNetV1/V2...

SENet(SE-block)



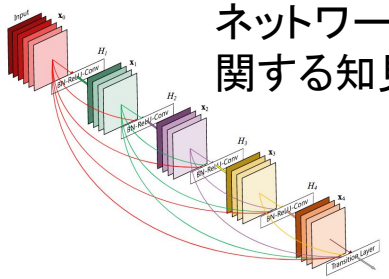
J. Hu et al. "Squeeze-and-Excitation Networks, " in TPAMI 2019.
https://arxiv.org/abs/1709.01507

ResNeXt



S. Xie et al. "Aggregated Residual Transformations for Deep Neural Networks, " in CVPR 2017.
https://arxiv.org/abs/1709.01507

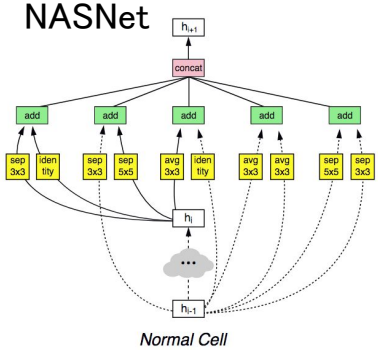
DenseNet



ネットワークの接続に関する知見が強化

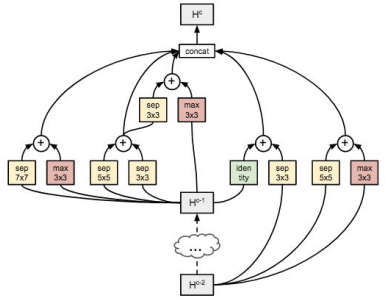
G. Huang et al. "Densely Connected Convolutional Networks, " in CVPR 2017.
https://arxiv.org/abs/1608.06993

NASNet



B. Zoph et al. "Learning Transferable Architectures for Scalable Image Recognition," in CVPR 2018.
https://arxiv.org/abs/1707.07012

PNASNet

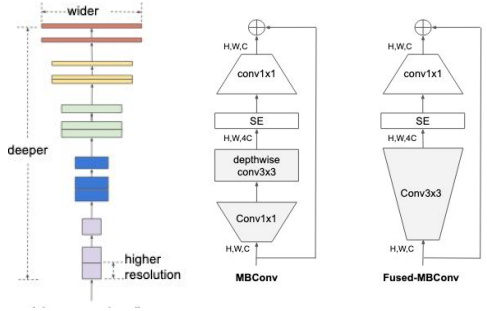


C. Liu et al. "Progressive Neural Architecture Search," in CVPR 2018.
https://arxiv.org/abs/1712.00559

NASはデータセットに合わせてその場で構造を探索するという方針(但し計算リソースを大量に要する)

EfficientNetは深さ・幅・入力に関するサイズのバランスが重要と解析

EfficientNetV1/V2

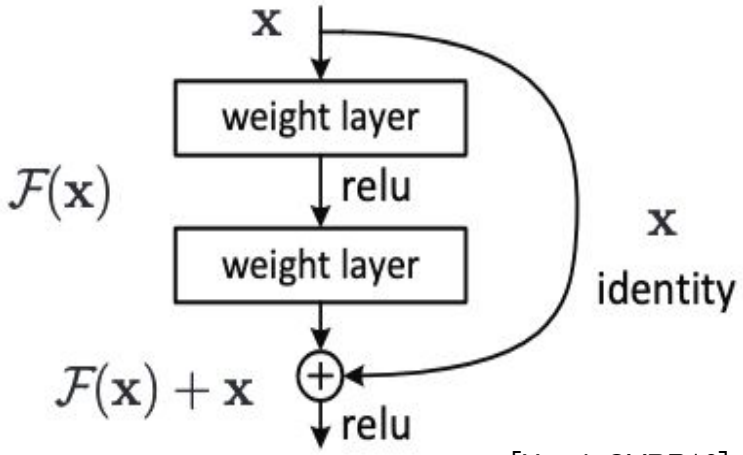


M. Tan et al. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in ICML 2019.
https://arxiv.org/pdf/1905.11946.pdf

M. Tan et al. "EfficientNetV2: Smaller Models and Faster Training," in ICML 2021.
https://arxiv.org/pdf/2104.00298.pdf

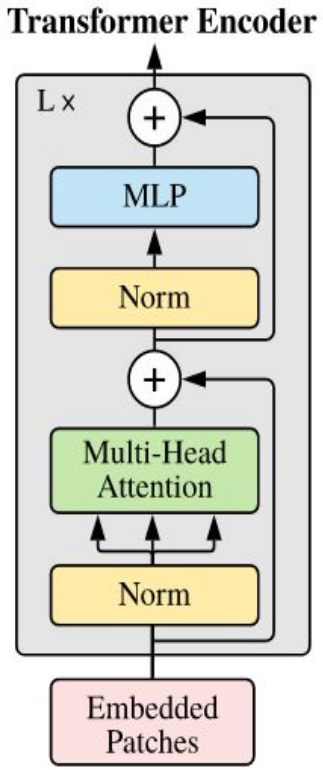
DNNの動向・CVのトレンド（10/51）

畳み込み（Convolution）と自己注視（Self-attention）



[He al. CVPR16]

周辺の特徴を集約して良いものを取り出す
（より局所的な特徴に着目）



[Vaswani al. NIPS17]
Figure from [Dosovitskiy al. ICLR21]

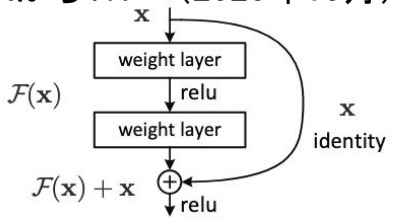
全体の特徴を把握して良いものを取り出す
（より大局的な特徴に着目）

DNNの動向・CVのトレンド(11/51)

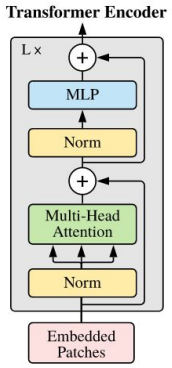
Vision Transformer(ViT)導入(2020～Present)

- 2017にTransformer提案，2020にViT提案
- 2021年，CV分野はTransformerの一年だった
- ViT派生モデル/MLP等で同様のモデル提案に至る

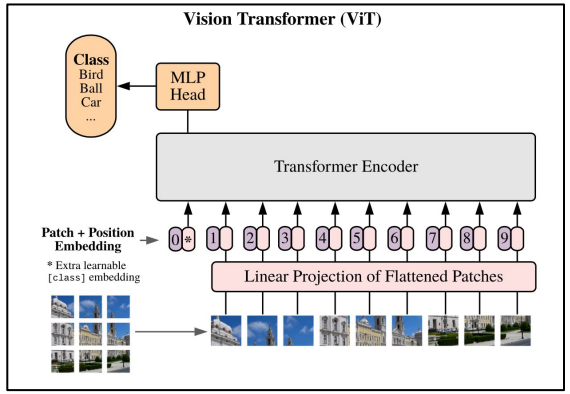
Computer Visionのパラダイムシフト
CNNからViTへ(2020年10月)



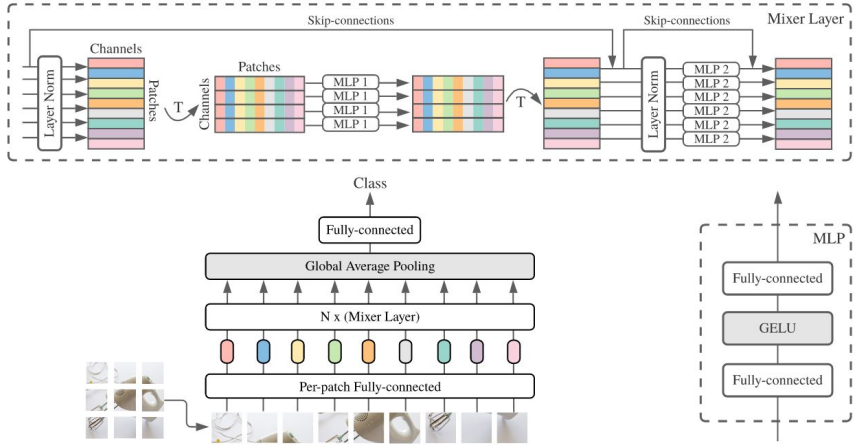
[He al. CVPR16]



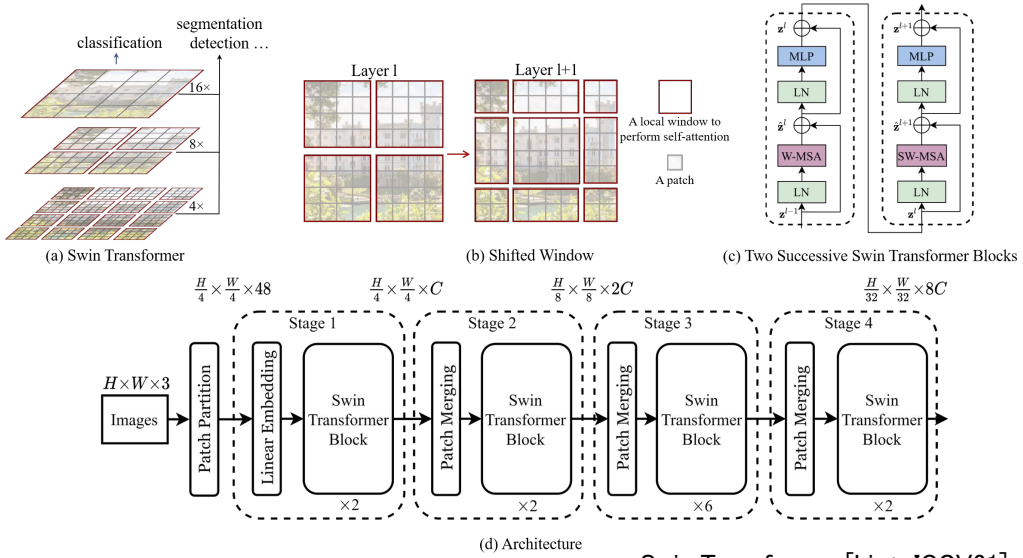
[Vaswani al. NIPS17]
Figure from [Dosovitskiy al. ICLR21]



ViT [Dosovitskiy+, ICLR21]



MLP-Mixer [Tolstinhin+, NeurIPS21]



(a) Swin Transformer
(b) Shifted Window
(c) Two Successive Swin Transformer Blocks
(d) Architecture

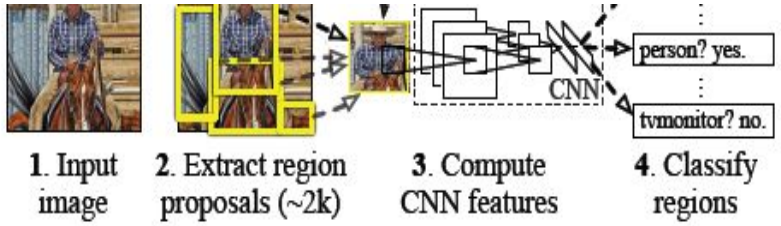
Swin Transformer [Liu+, ICCV21]

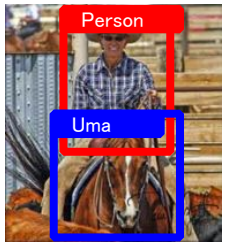
ViTはエンコーダのみのモデル，同様の機能をMLPで実現(MLP-Mixer)，近年のベースラインはSwinTransformerか(V2も2021年末に提案)

DNNの動向・CVのトレンド(12/51)

他タスクへの応用(画像認識・動画認識・マルチモーダル)

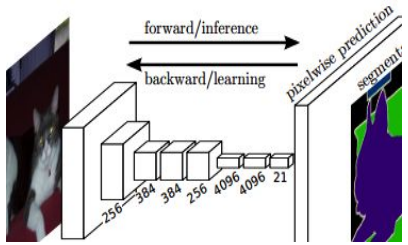
- 物体検出: R-CNN, Fast/Faster R-CNN, YOLO, SSD,,,
- 領域分割: FCN, SegNet, U-Net,,,
- Vision & Language: 画像説明文, VQA, Visual Dialog,,,
- 動画認識: Two-stream ConvNets, 3D Conv., (2+1)D Conv. ,,,





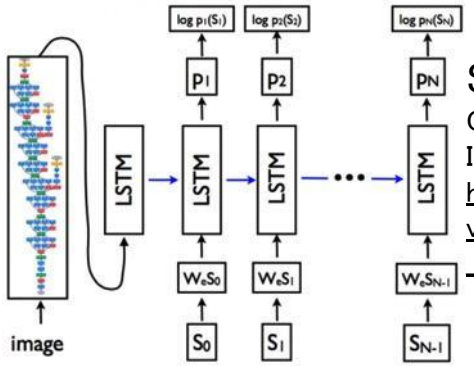
R-CNN

R. Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in CVPR 2014.
https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html



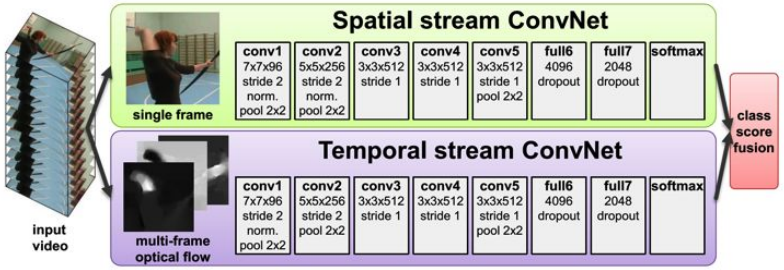
FCN

J. Long et al., "Fully Convolutional Networks for Semantic Segmentation," in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html



Show and Tell

O. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html



Two-stream ConvNets

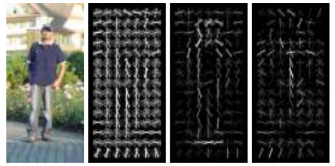
K. Simonyan et al., "Two-Stream Convolutional Networks for Action Recognition in Videos," in NIPS 2014.
https://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14b/simonyan14b.pdf

DNNの動向・CVのトレンド(13/51)

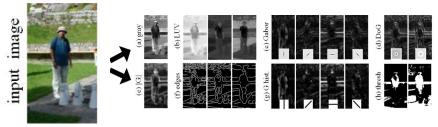
物体検出の流れ(2001～2017)

Hand-crafted feature時代　基礎/枠組みの構築

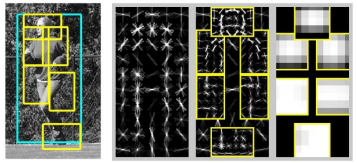
Haar-like [Viola+, CVPR01]
+ AdaBoost



HOG [Dalal+, CVPR05]
+ SVM

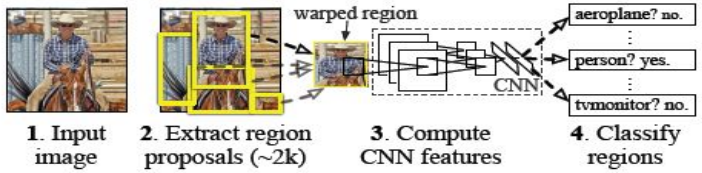


ICF [Dollár+, BMVC09]
+ Soft-cascade

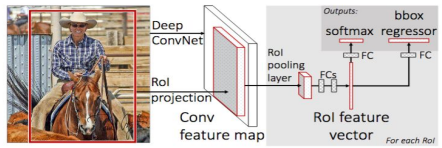


DPM [Felzenszwalb+, TPAMI12]
+ Latent SVM

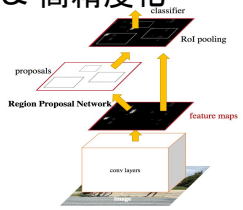
R-CNN時代(それ以前は"Hand-crafted" ObjectNess)高速化 & 高精度化



R-CNN [Girshick, CVPR14]
Selective Search + CNN

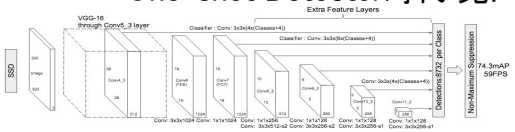


Fast R-CNN [Girshick, ICCV15]
ROI Pooling, Multi-task Loss

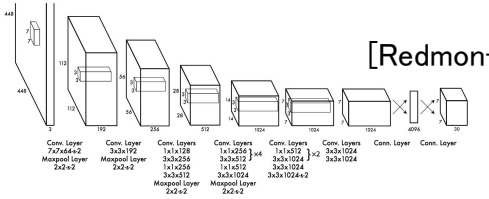


Faster R-CNN [Ren+, NIPS15]
RPN

One-shot Detector時代 兎にも角にも(精度を保ちつつ)高速化



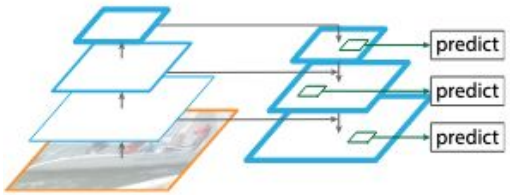
SSD [Liu+, ECCV16]
One-shot detector



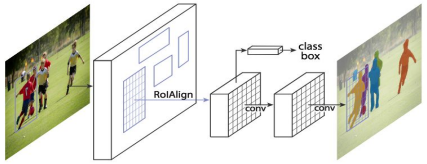
YOLO(v1)/v2/v3
[Redmon+, CVPR16/CVPR17/arXiv18]
One-shot detector

…

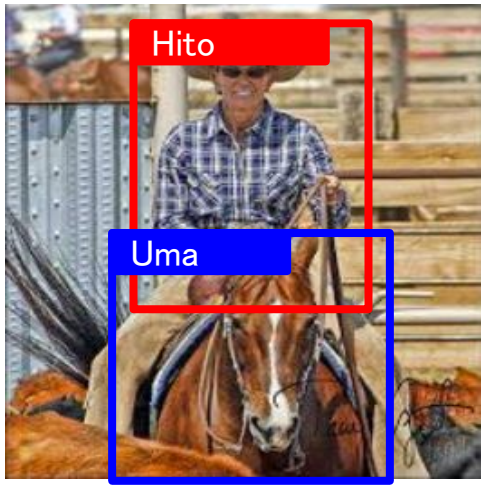
Latest Algorithm 精度重視, 高速



Feature Pyramid Networks (FPN)
[Lin+, CVPR17]



Mask R-CNN [He+, ICCV17]
RoI Align, Det+Seg



【物体1】
カテゴリ:Hito
位置:(x_1, y_1, w_1, h_1)

【物体2】
カテゴリ:Uma
位置:(x_2, y_2, w_2, h_2)

物体検出は物体カテゴリ識別と画像上の位置座標を回帰する問題

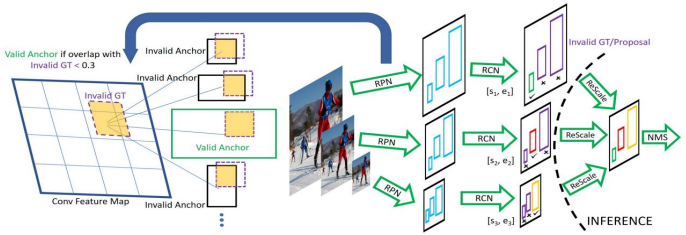
DNNの動向・CVのトレンド(14/51)

物体検出の流れ(2018～2020)

CNNによる工夫:
解像度による調整(上段), 座標回帰による調整(下段)

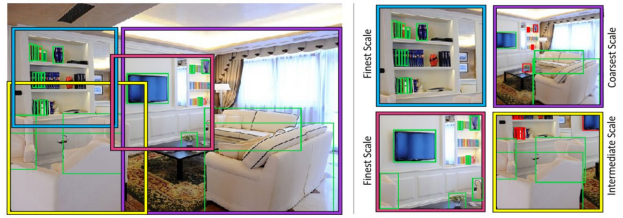
SNIP

B. Singh et al. "An Analysis of Scale Invariance in Object Detection," in CVPR 2018.
https://arxiv.org/abs/1711.08189



SNIPER

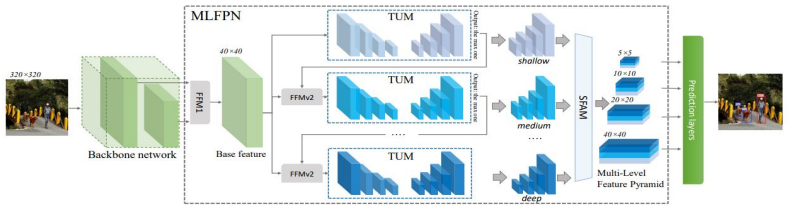
B. Singh et al. "SNIPER: Efficient Multi-Scale Training," in NeurIPS 2018.
https://papers.nips.cc/paper/8143-sniper-efficient-multi-scale-training.pdf



物体候補が多く位置する領域を拡大

M2Det

Q. Zhao et al. "M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network," in AAAI 2019.
https://arxiv.org/abs/1811.04533



多重解像度の特徴マップ使用

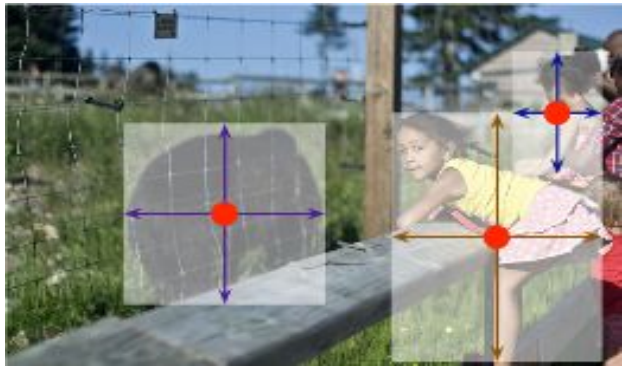
CornerNet

H. Law et al. "CornerNet: Detecting Objects as Paired Keypoints," in ECCV 2018.
https://arxiv.org/abs/1808.01244

2点(左上, 右下)を回帰

CenterNet

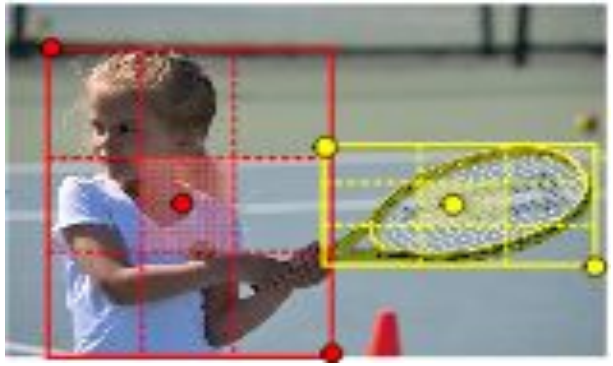
Z. Xingyi et al. "Objects as Points," in arXiv 2019.
https://arxiv.org/abs/1904.07850



中央からxyサイズを回帰

CenterNet

K. Duan et al. "CenterNet: Keypoint Triplets for Object Detection," in ICCV 2019.
https://arxiv.org/abs/1904.08189



中央+2点, xyサイズを回帰

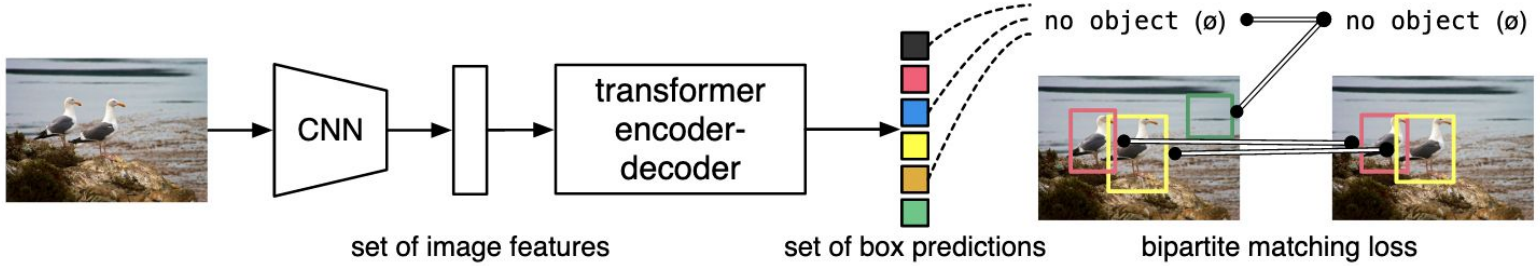
MSCOCO 基準で Mask R-CNN 41.7 -> CenterNet 47.0まで検出精度向上！

DNNの動向・CVのトレンド(15/51)

物体検出の流れ(2020～Present)

DETR

N. Carion et al. "End-to-End Object Detection with Transformers," in ECCV 2020.
https://arxiv.org/abs/2005.12872

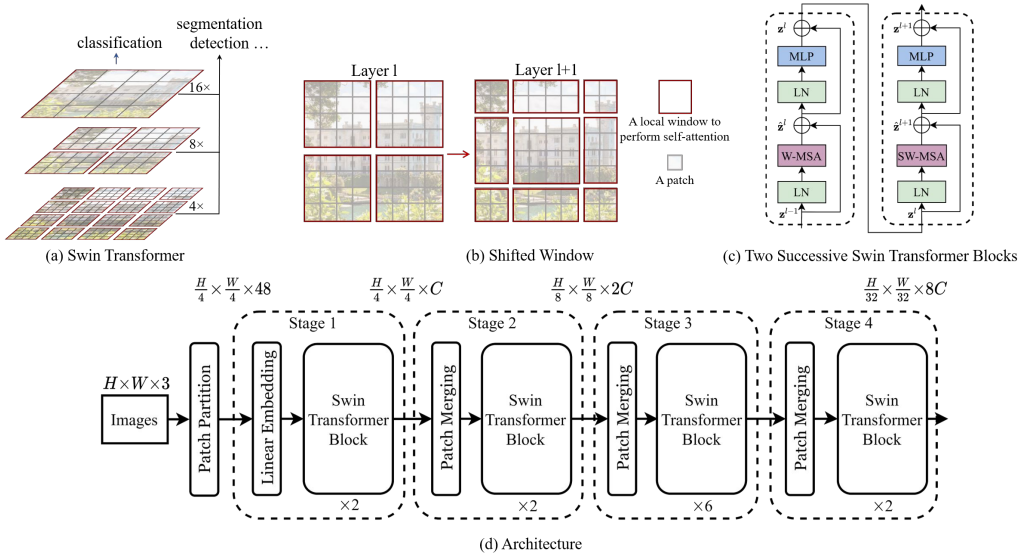


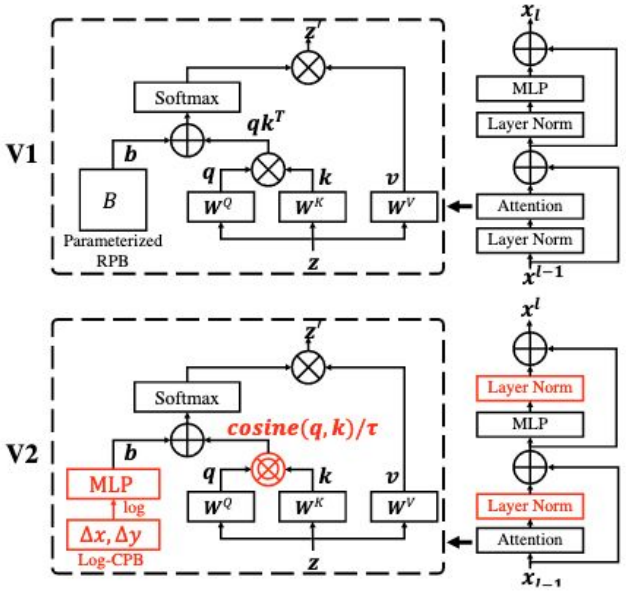
Transformerによる物体検出のパイオニア：Backboneには比較的うまくいくことが明らかなCNNも組み合わせている

SwinTransformer V1/V2

Z. Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in ICCV 2021.
https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html

Z. Liu et al. "Swin Transformer V2: Scaling Up Capacity and Resolution," in ICCV 2021.
https://arxiv.org/abs/2111.09883





ViT特有のパッチ入力の問題点を，階層化&シフトにより改善，画像識別のみならず物体検出にも有効

DNNの動向・CVのトレンド（16/51）

セマンティック/インスタンスセグメンテーション（2006～2017）

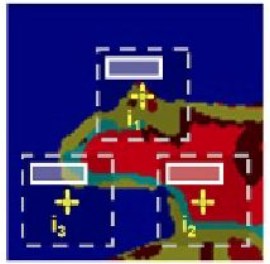
- BackBone（ベースとなるネットワーク）
- Head（ピクセルごとの回帰を出力）
  - 両者の組み合わせ
  - 文脈把握とスケール変動を考慮する傾向

【セマンティック/インスタンスセグメンテーション】

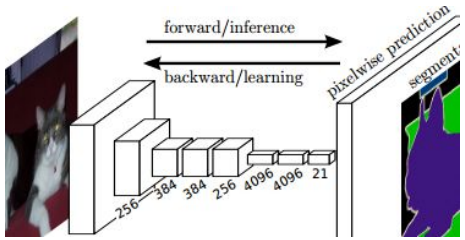
ピクセルごとに画像カテゴリを割り当てる問題設定

※下はセマンティック/インスタンスセグメンテーションを両方含む

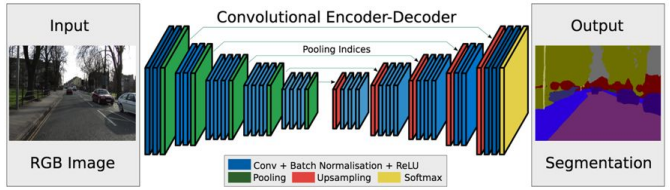
ベースアルゴリズム



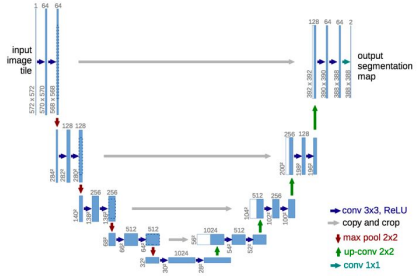
TextonBoost
[Shotton, ECCV06]



FCN [Long, CVPR15]
全層畳み込み，チャネル和

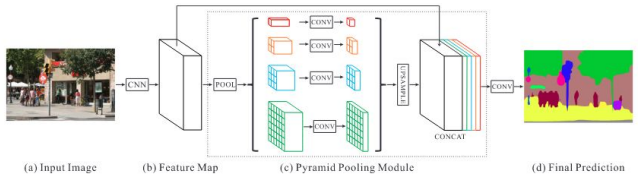


SegNet [Kendall, arXiv15]

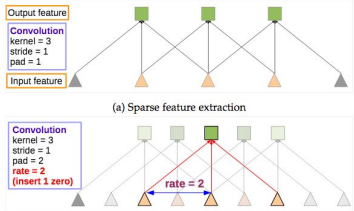


U-Net [Ronneberger, MICCAI15]
位置情報保持，チャネル連結

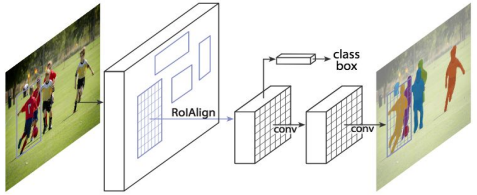
精度重視



PSPNet [Zhao, CVPR17]
特徴マップの階層化，コンテクスト情報



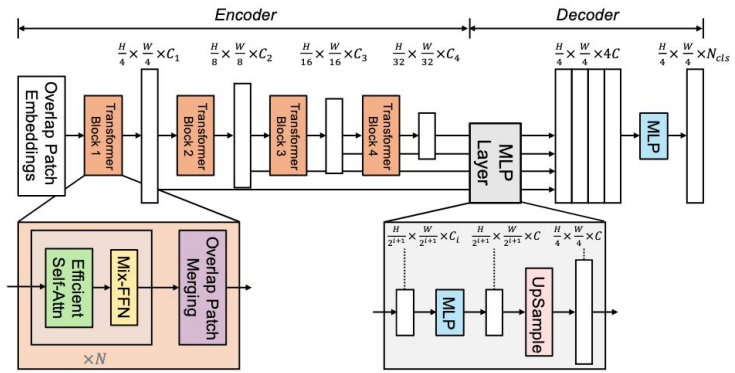
DeepLab(v1,v2,v3) [Chen, TPAMI17]
Dilated Conv，特徴マップの並列化



Mask R-CNN [He, ICCV17]
RoI Align, Det+Seg

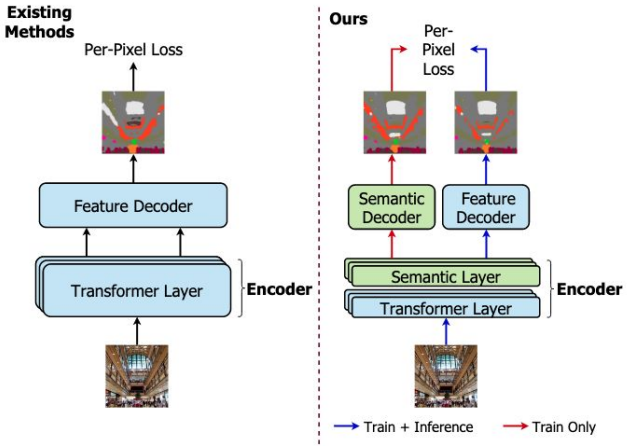
DNNの動向・CVのトレンド(17/51)

セマンティック/インスタンスセグメンテーションの流れ(2021～Present)



SegFormer

E. Xie et al. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in arXiv 2105.15203, 2021.
https://arxiv.org/abs/2105.15203



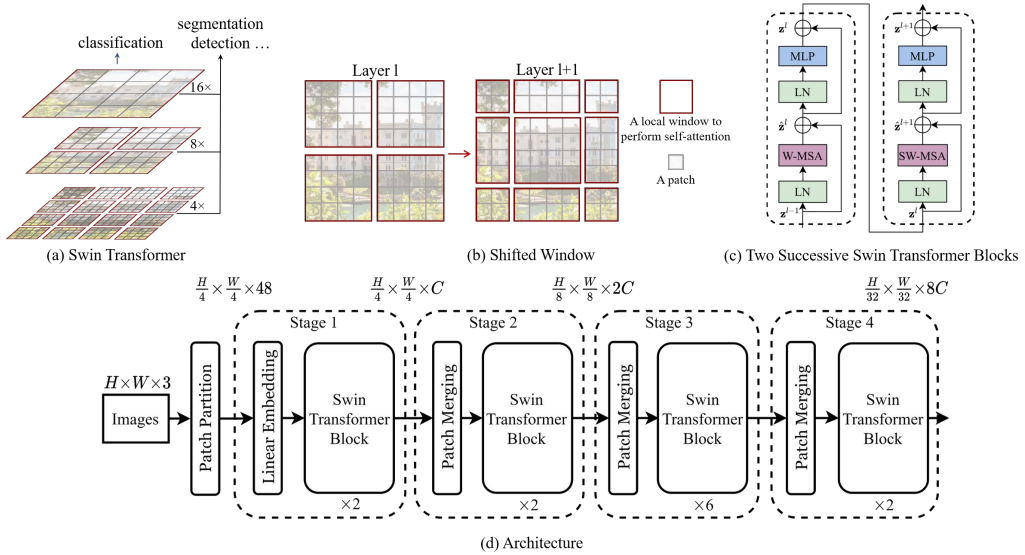
SeMask

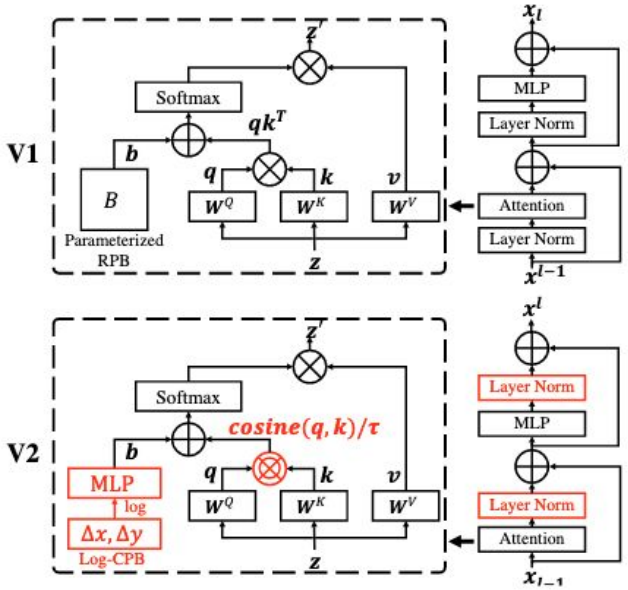
J. Jain et al. "SeMask: Semantically Masked Transformers for Semantic Segmentation," in arXiv 2112.12782 2021.
https://arxiv.org/abs/2112.12782

SwinTransformer V1/V2

Z. Liu et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in ICCV 2021.
https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html

Z. Liu et al. "Swin Transformer V2: Scaling Up Capacity and Resolution," in ICCV 2021.
https://arxiv.org/abs/2111.09883



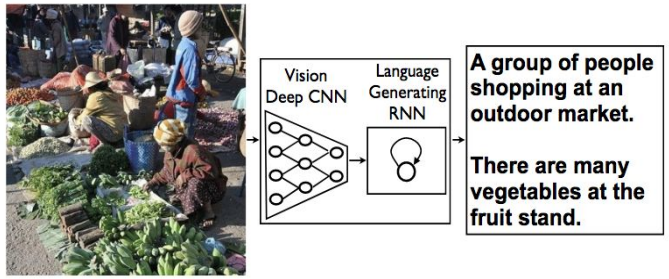


DNNの動向・CVのトレンド(18/51)

CVとNLP(自然言語処理)の融合分野

- 画像説明文(Image Captioning)
- 視覚的質問回答(Visual Question Answering; VQA)
- Visual Dialog

【Image Captioning】

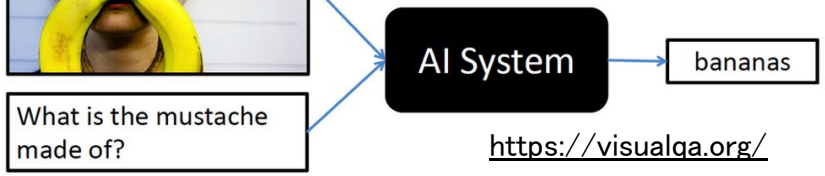


画像を入力として文章を出力

Show and Tell
O. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html

【VQA】

画像 (Visual)と質問文 (Q)を入力,
回答 (A)を返却

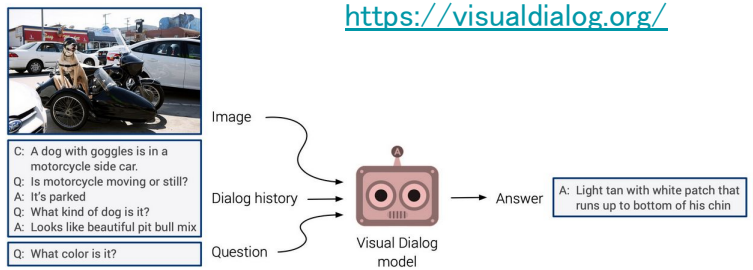


https://visualqa.org/

VQA
A. Agrawal et al., "VQA: Visual Queation Answering," in ICCV 2015.
https://arxiv.org/pdf/1505.00468.pdf

【Visual Dialog】
画像とそれに対する対話を繰り返しながら回答を行う

https://visualdialog.org/



Visual Dialog
A. Das et al., "Visual Dialog," in CVPR 2017.
https://arxiv.org/abs/1611.08669

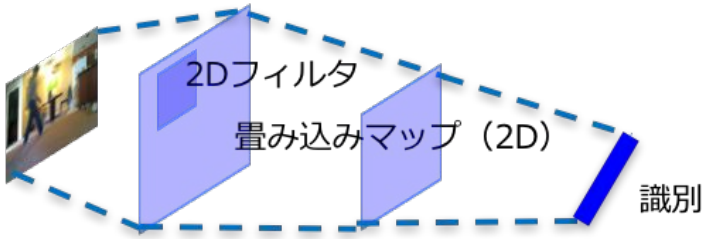
【その他のキーワード】
Text-to-Image(テキストから画像生成), Visual Navigation(画像からの道案内),
Cross-modal Summarization(要約タスク)など多数

DNNの動向・CVのトレンド(19/51)

動画像認識のモデル(CNN; N-Dimension Convolution)

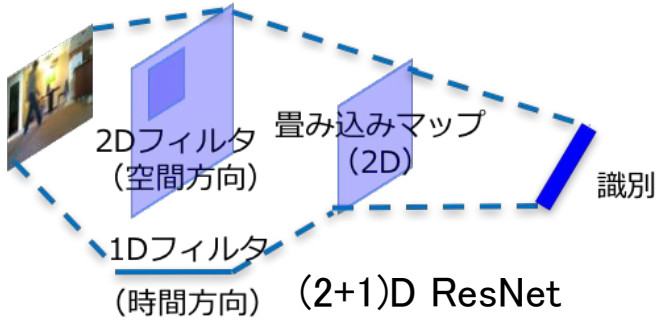
- 2D: Two-Stream ConvNets(フロー画像を同時に準備)
- 2D+再帰モデル: CNN+LSTM(最近ベンチマークとしてしか使われなくなった)
- 3D: 3D CNN(データが用意できればこれが本命)
- (2+1)D(少量データの場合のオプション)

2D畳み込み(TS ConvNet, CNN+LSTM)



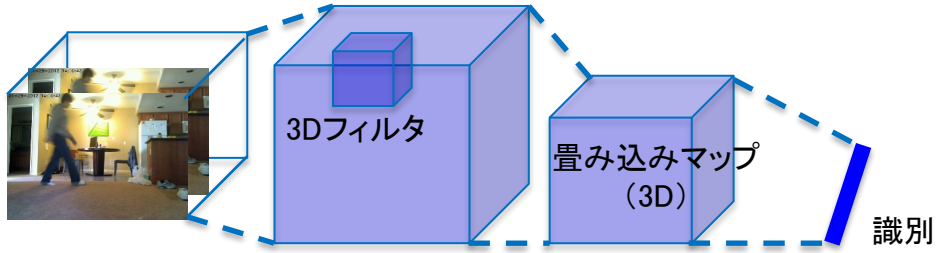
Two-stream ConvNets
K. Simonyan et al., "Two-Stream Convolutional Networks for Action Recognition in Videos," in NIPS 2014.
https://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14b/simonyan14b.pdf

(2+1)D畳み込み(Separable Conv.)



(2+1)D ResNet
D. Tran et al. "A Closer Look at Spatiotemporal Convolutions for Action Recognition," in CVPR 2018.
https://arxiv.org/abs/1711.11248

3D畳み込み(C3D, 3D ResNet, I3D)

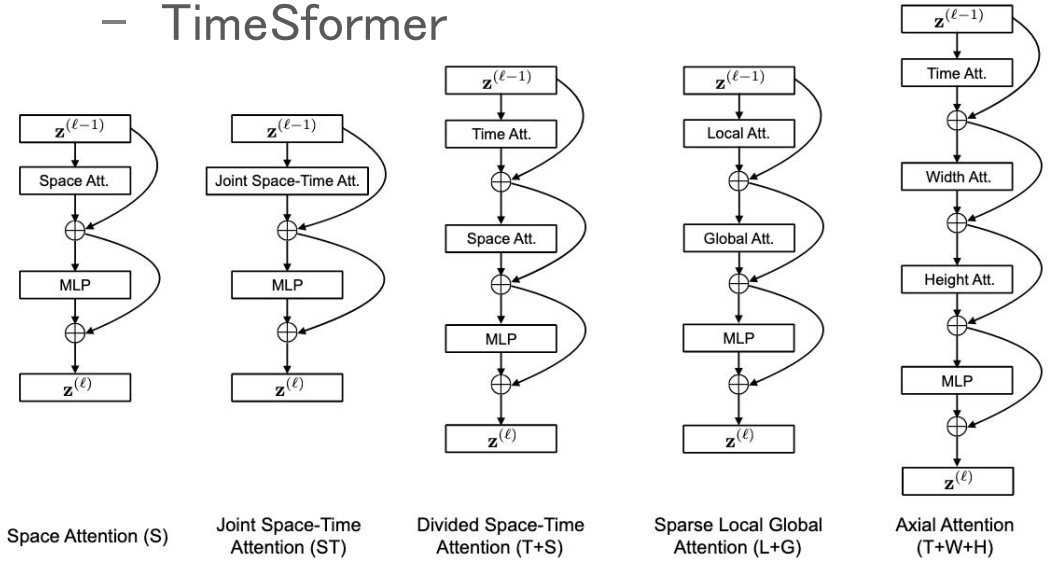


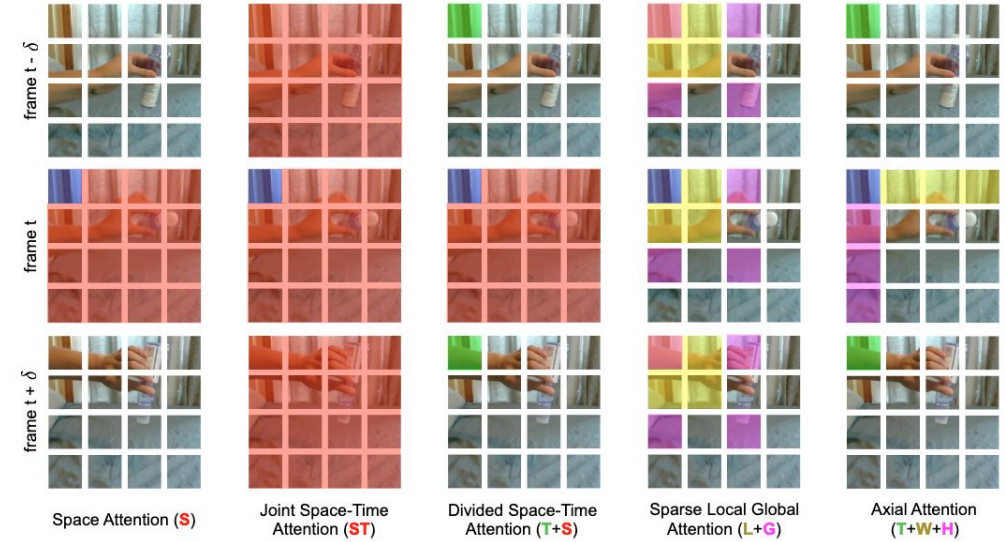
3D ResNet
K. Hara et al. "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in CVPR 2018.
https://arxiv.org/abs/1711.09577

DNNの動向・CVのトレンド(20/51)

動画像認識のモデル(Video Transformer)

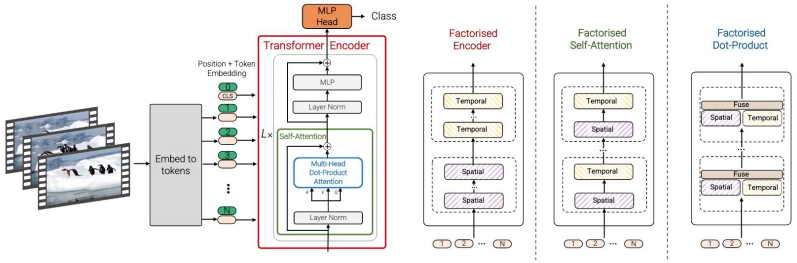
- TimeSformer





TimeSformer
G. Bertasius et al., "Is Space-Time Attention All You Need for Video Understanding?," in ICML 2021.
https://arxiv.org/abs/2102.05095

ViViT
A. Arnab et al., "ViViT: A Video Vision Transformer," in ICCV 2021.
https://arxiv.org/abs/2103.15691v2



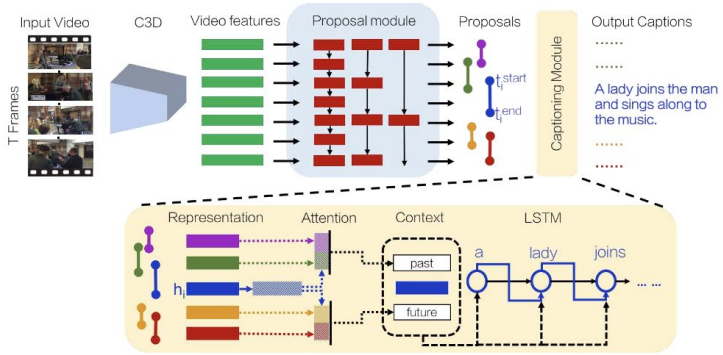
- 時間と空間の処理を実行するモジュールを複数種類比較実験
- 結果から，時間tと空間xyを分割して処理する方式(Divided Space-Time Attention)が有効
- 図中，中央のモジュールとアテンション

DNNの動向・CVのトレンド(21/51)

動画認識タスク

Dense Captioning Events

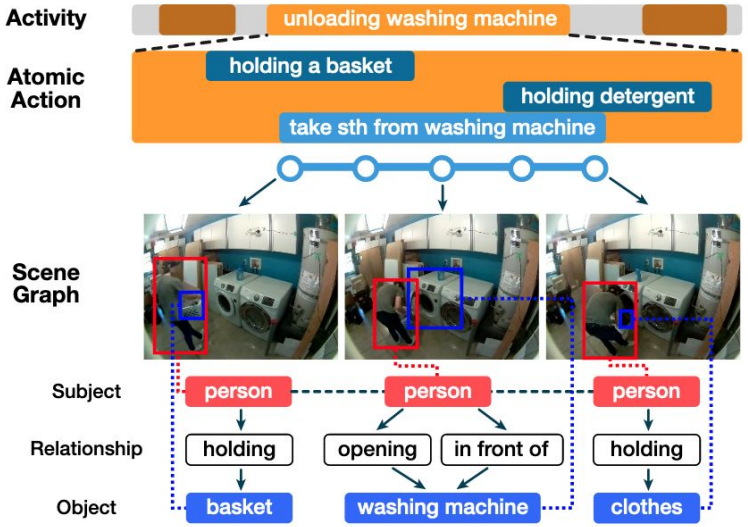
R. Krishna et al. "Dense Captioning Events in Videos," in ICCV 2017.
https://arxiv.org/abs/1705.00754



動画像から説明文出力，動画像の検索にも用いることができる

Home Action Genome

N. Rai et al. "Home Actioin Genome: Contrastive Compositional Action Understanding," in CVPR 2021.
https://homeactiongenome.org/



マルチモーダル・複数視点・シーングラフラベル付与により，包括的な動画の理解をサポート

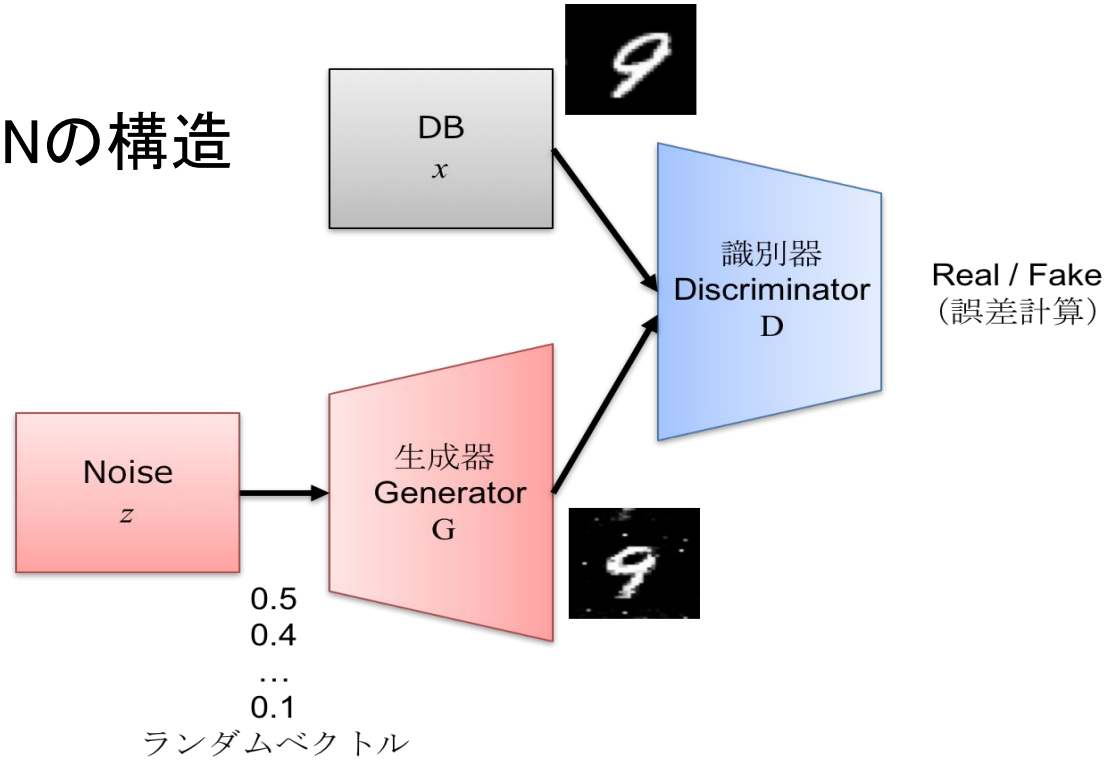
cvpaper.challenge

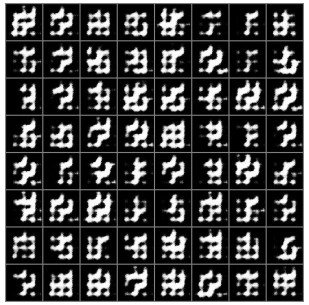
DNNの動向・CVのトレンド(22/51)

GAN: 画像生成を行うための構造として提案

- 現在，生成画像/実画像の分布を近づけることで生成モデルを学習
- 敵対的学習は超解像，異常検知，データ拡張 など多様な場面に応用

GANの構造





徐々に鮮明になるデータ

https://medium.com/@sunnerli/the-missing-piece-of-gan-d091604a615a

(注)下はGANにより生成された画像です

BigGAN https://arxiv.org/pdf/1809.11096.pdf

DNNの動向・CVのトレンド(23/51)

GANの主要な流れ

1. GAN(オリジナルのGAN)
   - [Goodfellow, NIPS2014] https://papers.nips.cc/paper/5553-generative-adversarial-nets.pdf
2. DCGAN(畳み込み層の使用)
   - [Radford, ICLR2016] https://arxiv.org/abs/1511.06434
3. Pix2Pix(pixel同士が対応付くという意味でConditionalなGAN)
   - [Isola, CVPR2017] https://arxiv.org/abs/1611.07004
4. CycleGAN(pix2pixの教師なし版)
   - [Zhu, ICCV2017] https://arxiv.org/pdf/1703.10593.pdf
5. ACGAN(カテゴリ識別も同時に実施してコンディションとした)
   - [Odera, ICML2017] https://arxiv.org/abs/1610.09585
6. WGAN/SNGAN(学習安定化)
   - [Arjovsky, ICML2017] http://proceedings.mlr.press/v70/arjovsky17a.html
   - [Miyato, ICLR2018] https://arxiv.org/abs/1802.05957
7. PGGAN(高精度化)
   - [Karras, ICLR2018] https://arxiv.org/abs/1710.10196
8. Self-Attention GAN(アテンション機構を採用)
   - [Zhang, arXiv 1805.08318] https://arxiv.org/abs/1805.08318
9. BigGAN(超高精細GAN)
   - [Brock, ICLR2019] https://arxiv.org/abs/1809.11096
10. StyleGAN(超高精細GAN)
    - [Karras, CVPR2019] https://arxiv.org/abs/1812.04948

2018年12月時点での調査

DNNの動向・CVのトレンド(24/51)

鮮明な画像生成: VAE, Diffusion Model！？

- キレイな画像生成はGANが優勢だった(～2019/05)
  - BigGAN, StyleGAN etc.
- VQVAEでVAEが，2021年はDiffusion Modelが話題になった
  - VAE: GANのようにMode Collapseを起こさないと主張
  - Diffusion Model: 学習時のサンプルノイズを徐々に取り除くことで生成の多様性と信頼性を向上（BigGAN よりも高精度であることから注目）

左：VQ-VAE-2，右：BigGAN

A. Razavi et al. “Generating Diverse High-Fidelity Imageswith VQ-VAE-2,” NeurIPS, 2019.
https://arxiv.org/pdf/1906.00446.pdf

+Diffusion Model

P. Dhariwal et al. “Diffusion Models Beat GANs on Image Synthesis,” arXiv, 2105.05233, 2021.
https://arxiv.org/abs/2105.05233

DNNの動向・CVのトレンド(25/51)

DALL·Eの衝撃

– テキストから画像を生成するText-to-Imageの問題設定で革新的な成果

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓

- 「アボカド形の腰掛け椅子」のような見たことがない物体も概念を組み合わせて画像を生成
- 1200億パラメータのモデルをインターネットから集めた2.5億の画像・テキストのペアから学習

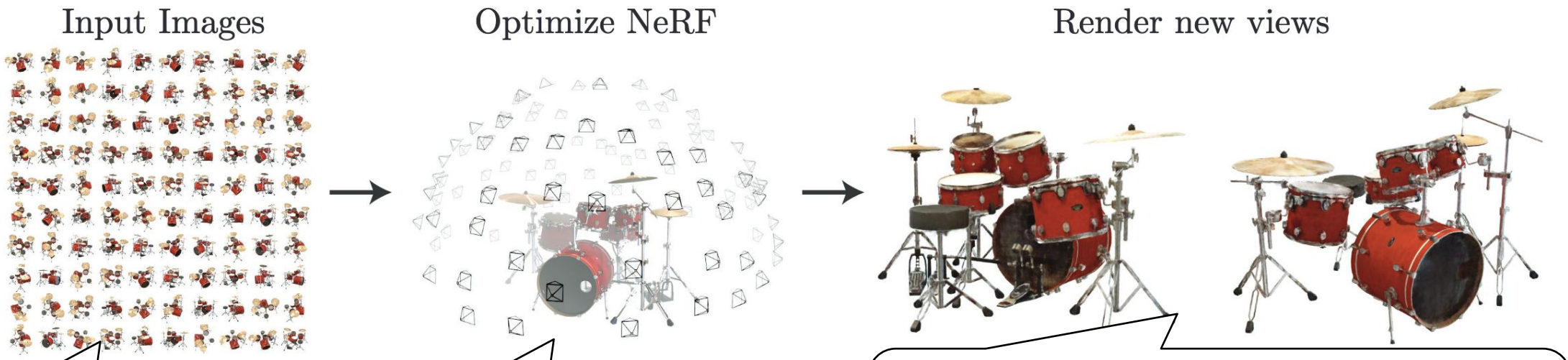
DALL·E: Creating Images from Text
https://openai.com/blog/dall-e/

DNNの動向・CVのトレンド(26/51)

Neural Radiance Fields(NeRF)

B. Mildenhall, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis" European Conference on Computer Vision (ECCV), 2020.
https://arxiv.org/pdf/2003.08934.pdf

- Radiance Field(x,y,z, θ , ϕ ; ある3次元空間xyzの光と透明度 θ , ϕ)をNeural Netにより全て推定する問題設定
- 3次元空間自体を再構成するので，未観測の新規視点も推定可能



入力は多数視点の2次元画像，各画像にはRFを推定するヒントを含む

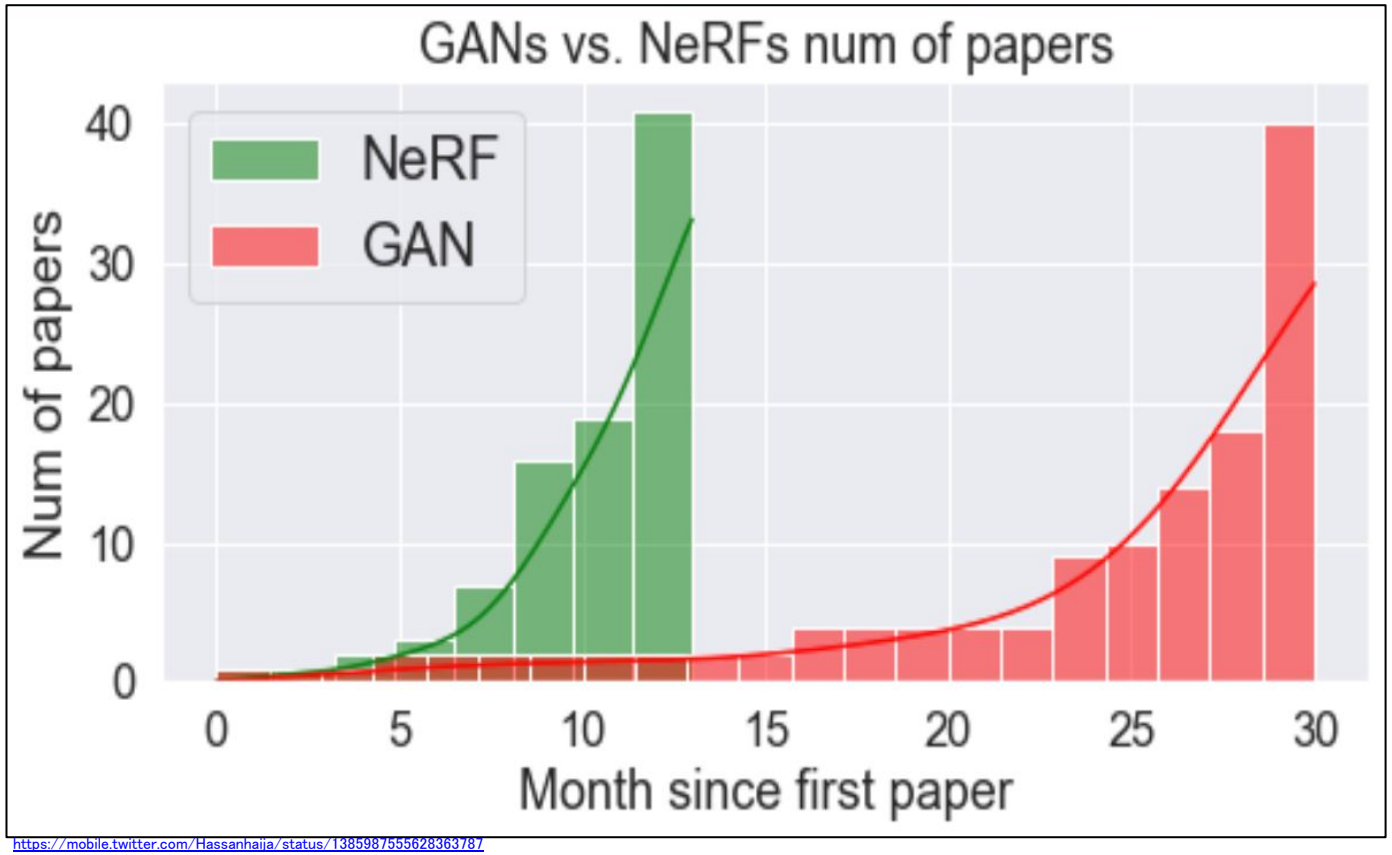
既知視点(またはその場で推定)の画像をNeuralNetに入力，5次元空間推定

推定した3次元空間の光・透明度により新しい視点を再構成; もちろんこの2例以外の視点も再構成可能

DNNの動向・CVのトレンド(27/51)

NeRFの派生技術

■ 2020年3月公開，トップ国際会議ECCV 2020 Awardを受け爆発的に拡がる



https://mobile.twitter.com/Hassanhaija/status/1385987555628363787

B. Mildenhall, et al. “NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images“ CVPR, 2022.
https://arxiv.org/abs/2111.13679

I. Ueda, et al. “Neural Density-Distance Fields“ ECCV, 2022. https://arxiv.org/abs/2207.14455

M. Tancik, et al. “Block-NeRF: Scalable Large Scene Neural View Synthesis”, CVPR, 2022.
https://arxiv.org/abs/2202.05263

DNNの動向・CVのトレンド(28/51)

教師なし/少量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
  - Transfer Learning
  - Domain Adaptation
  - Reinforcement Learning
- 教師がない/間接的に教師を与える，ような仕組みに対する競争も激化
- 巨大IT企業のように大量のラベルを持た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・CVのトレンド(29/51)

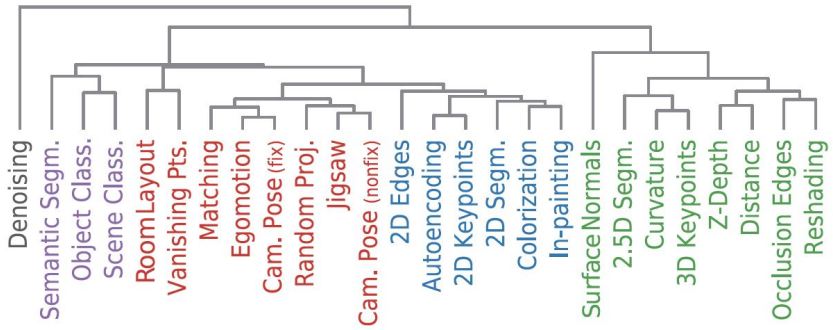
学習法の簡単な整理

- {Un-, Semi-, Weak-, Self-} supervision
  - Un-supervision(教師なし学習)
    アノテーションが一切ないデータで学習
  - Semi-supervision(半教師あり学習)
    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(弱教師付き学習)
    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(自己教師あり学習)
    自ら教師を作り出して特徴表現を学習する「自己教師学習」
    - 特定タスクの前に自ら教師を作り出し特徴表現を学習するため，その後に特定タスクのためのファインチューニングを伴う
    - ex) 領域分割した画像でジグソーパズルを解く，回転を当てる

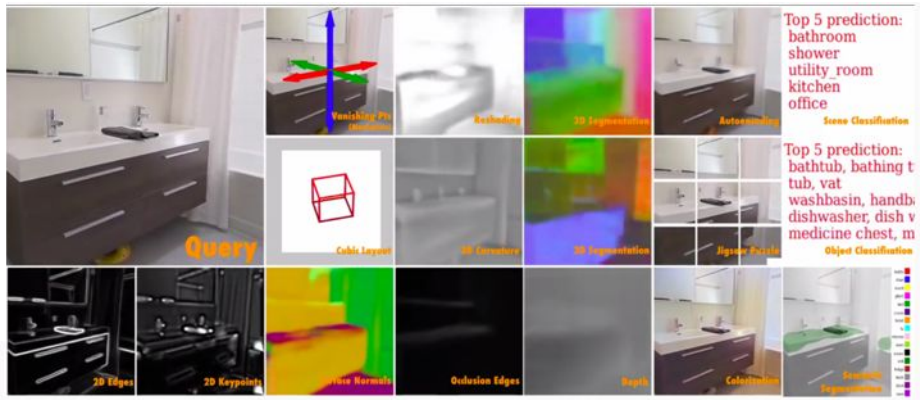
DNNの動向・CVのトレンド(30/51)

転移学習(Transfer Learning)

- 学習した知識を別の領域の学習に適用する技術
- 転移学習の網羅的探索:Taskonomy [Zamir, CVPR2018]
  http://taskonomy.stanford.edu/
  - CVPR 2018 Best Paper Award
  - 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ていたサブタスクを繋げる
    - 効果を最大化する転移学習の関係性を明らかにした



Task Similarity Tree: 類似するタスク間の関係性を可視化



データセットは26タスクに対しラベル付け

DNNの動向・CVのトレンド(31/51)

教師あり学習
Supervised Learning: SL

- 人間が実画像に対して手動で教師ラベル付け
- 入力画像から教師ラベルに基づき教師あり学習，
  それを通して視覚的特徴表現を獲得

教師ありデータセットの例

- ImageNet
- PASCAL VOC

ゴリラ

自己教師あり学習
Self-Supervised Learning: SSL

- 人間によるラベル付けは不要
- 擬似タスク(Pretext Task)を自動的に生成
- 入力画像から擬似タスクに基づき教師あり学習，
  それを通して視覚的特徴表現を獲得

自己教師手法の例

- Rotation，Jigsaw，MAE
- SimCLR，MoCo（対照学習）

数式ドリブン教師あり学習
Formula-driven Supervised Learning: FDSL

- 規則性に基づいた数式から人工画像を自動生成
- パラメータを元に自動で教師ラベル付け
- 入力画像から教師ラベルに基づき教師あり学習，
  それを通して視覚的特徴表現を獲得

FDSLデータセットの例

- FractalDB
- Perlin Noise

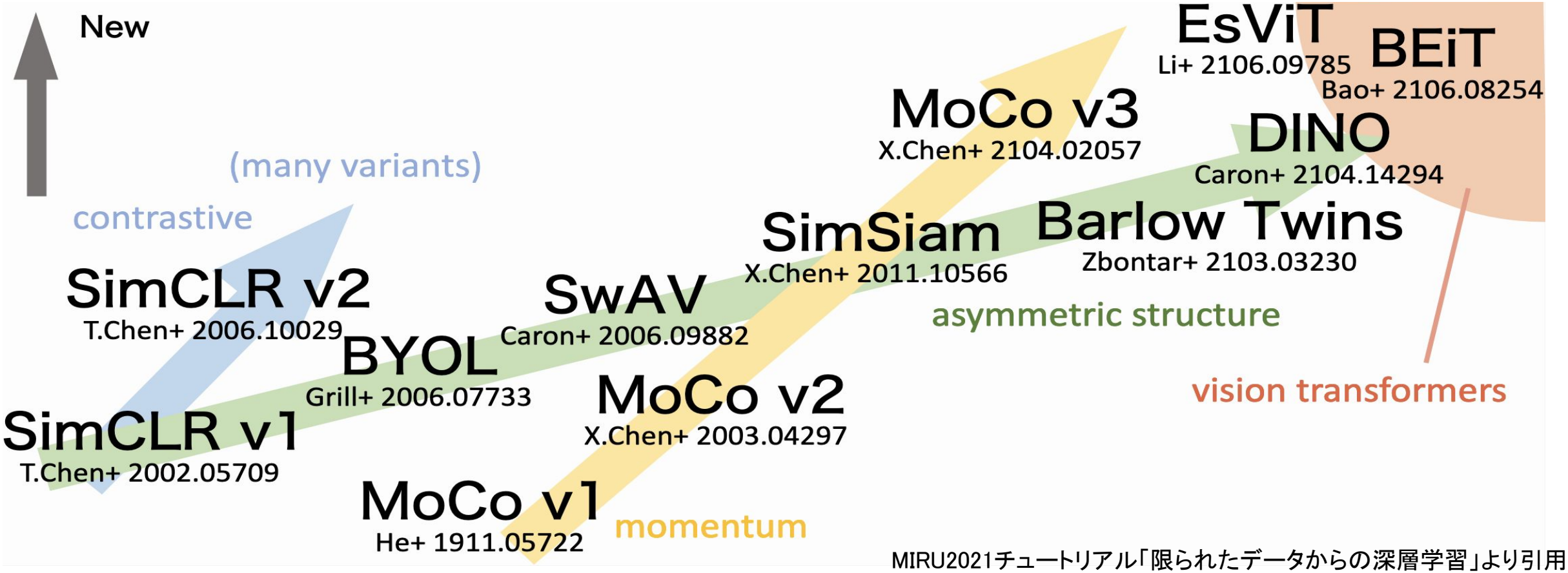
258 ←自動（数式）

←自動（数式）

DNNの動向・CVのトレンド(32/51)

最近の主流は自己教師あり学習

- 一部で人間の教師を置き換える性能まで到達



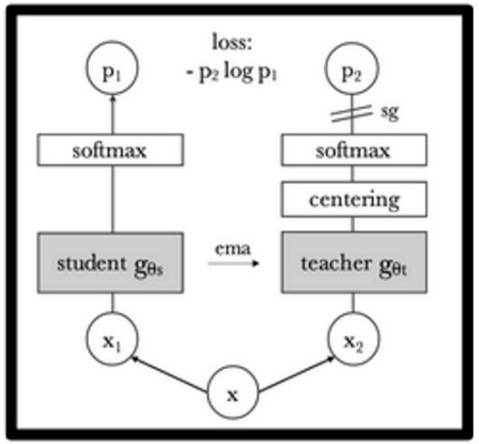
MIRU2021チュートリアル「限られたデータからの深層学習」より引用

https://raw.githubusercontent.com/hirokatsukataoka16/Formula-Driven-DataBase-Group/main/docs/material/MIRU21_Tutorial1.pdf

DNNの動向・CVのトレンド(33/51)

2022年 CV分野の激戦領域:ViTの自己教師あり学習

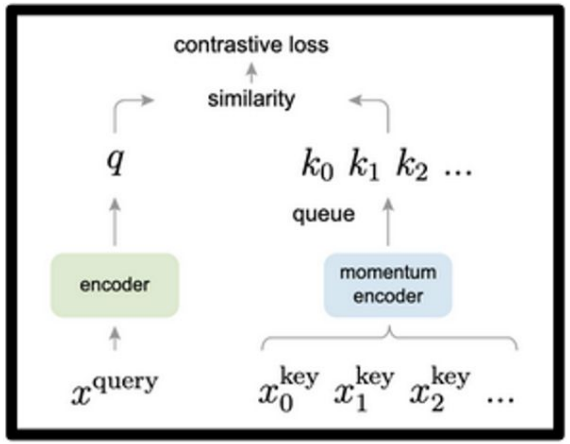
- 実画像に対して自動で一貫した教師を付与
  - DINO:ラベルなしでの自己蒸留 https://arxiv.org/abs/2104.14294
  - MoCoV3:対照学習MoCoと同様だが, ViT向けに改善 https://arxiv.org/abs/2104.02057
  - MAE:パッチの復元(BERTの画像版タスク) https://arxiv.org/abs/2111.06377



PT: DINO + ViT-B/16

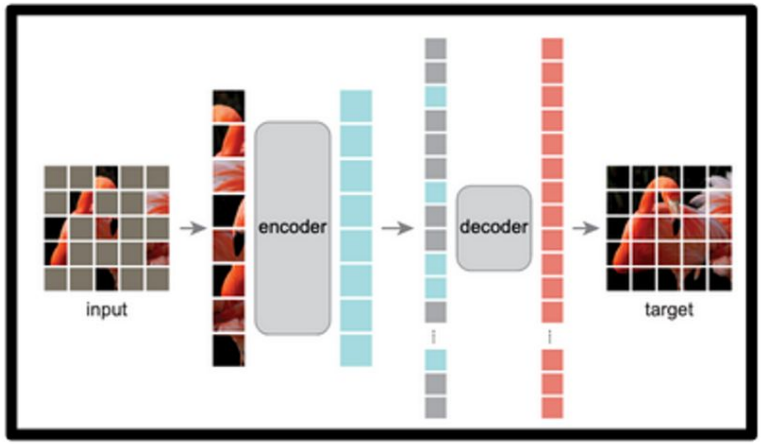
FT: 78.2 @ ImageNet-1k val.

(ViT-B/8では80.1まで向上)



PT: MoCoV3 + ViT-B/16

FT: 83.2 @ ImageNet-1k val.



PT: MAE + ViT-B/16

FT: 83.6 @ ImageNet-1k val.

(ViT-H_{448}では87.8まで向上)

マルチモダリティにて有効なData2Vecも登場(上記のスコアで84.2)

https://arxiv.org/abs/2202.03555

DNNの動向・CVのトレンド(34/51)

数式ドリブン教師あり学習 (FDSL; Formula-Driven Supervised Learning)

- 生成規則から画像パターンと教師ラベルを生成，大規模画像データセット構築
  - 下記はフラクタル幾何を生成規則としたFDSL
  - ベースとなるカテゴリ数/インスタンス数は1k/1kの100万画像

生成規則(式)

ラベルの生成 (ランダムサンプリング)

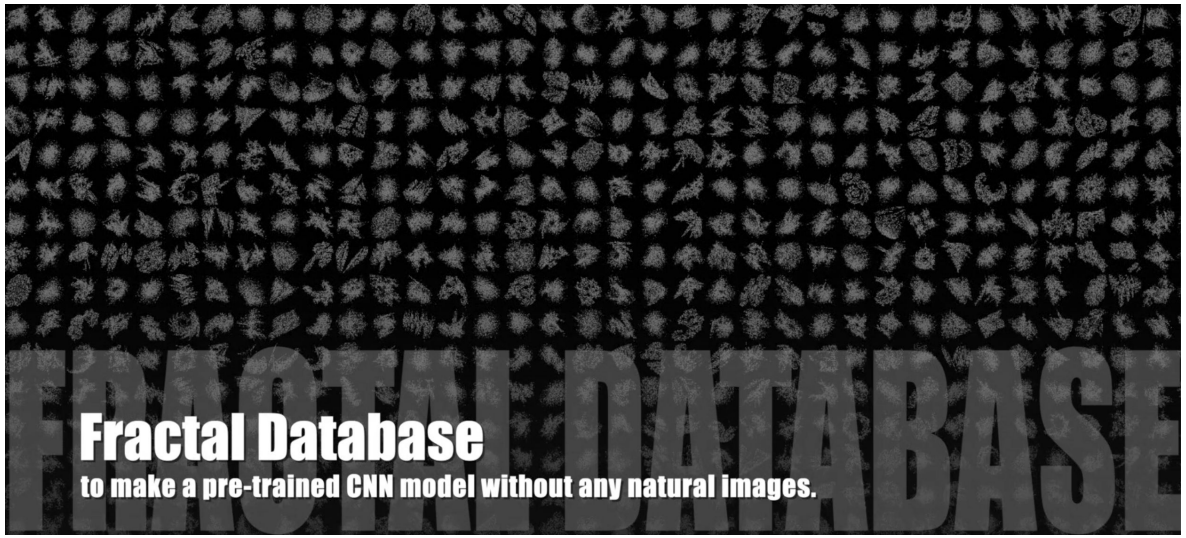
$$\Theta = \{(\theta_i, p_i)\}_{i=1}^{N}$$

データの生成 (IFS)

$$\mathrm{IFS} = \{\mathcal{X}; w_1, w_2, \cdots, w_N; p_1, p_2, \cdots, p_N\}$$

$$w_i(\boldsymbol{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

$$p_i = p(w^* = w_i) \qquad \boldsymbol{x}_{t+1} = w^*(\boldsymbol{x}_t)$$



FDSL / FractalDB

H. Kataoka et al., "Pre-training without Natural Images," in IJCV 2022.
https://hirokatsukataoka16.github.io/Pretraining-without-Natural-Images/

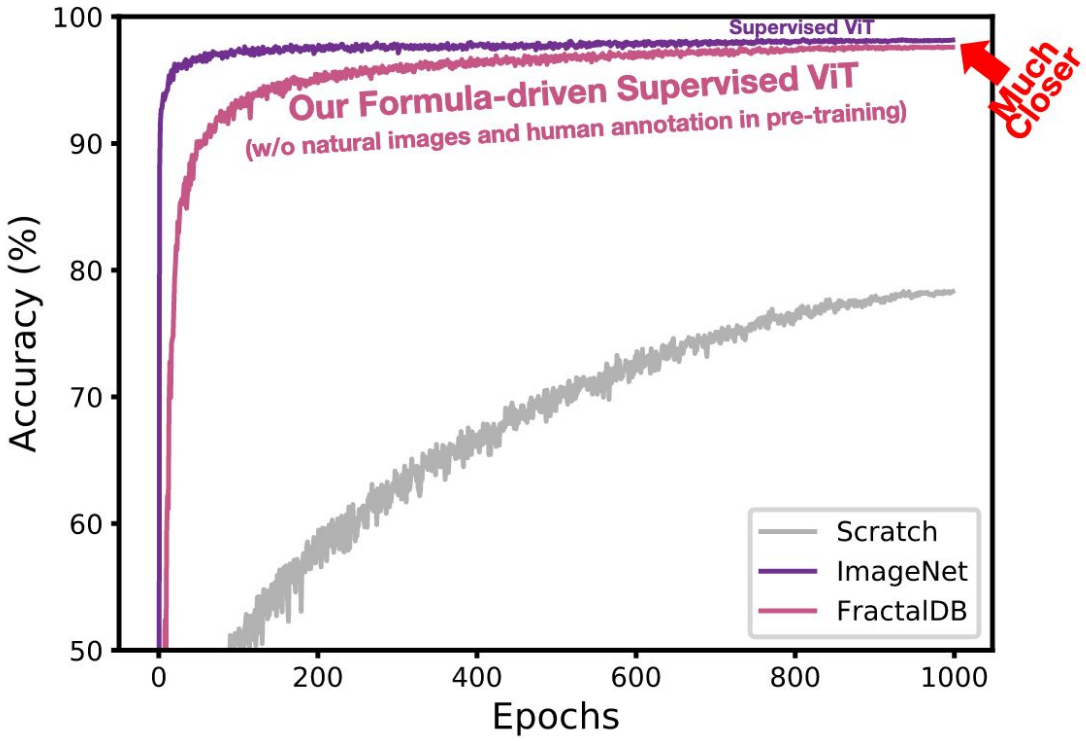
FractalDB Pre-trained ViT

K. Nakashima et al., "Can Vision Transformers Learn without Natural Images," in AAAI 2022.
https://hirokatsukataoka16.github.io/Vision-Transformers-without-Natural-Images/

DNNの動向・CVのトレンド(35/51)

数式ドリブン教師あり学習 (FDSL; Formula-Driven Supervised Learning)

- 実画像+人間教師を用いない事前学習でImageNet事前学習に近接する精度に到達
- 理論上は無限に画像カテゴリ・インスタンスを生成可能
  - 今後，大規模分散学習によりデータセット・モデルサイズを巨大化予定

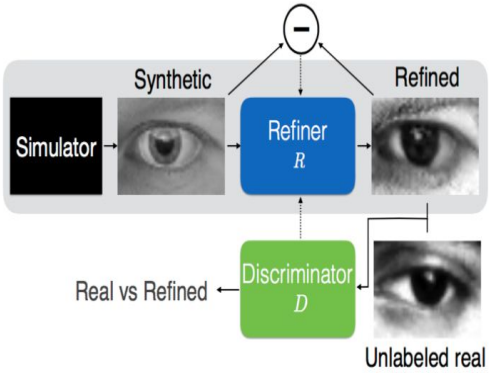


ImageNet vs. FractalDB：実画像も人間による教師ラベルも用いないFractalDBによる学習でScratch学習よりも圧倒的に高い精度，かつImageNetに近い精度まで到達．例えば，CIFAR-10では ImageNet 98.0 vs. FractalDB 97.8.

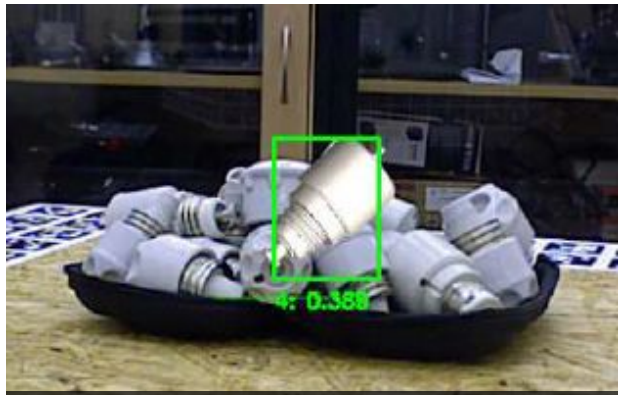
DNNの動向・CVのトレンド(36/51)

生成データ学習への期待

- CGを敵対的学習によりリアルに近づける(SimGAN)
- CGから大量の画像を生成(Domain Randomization)
- 切抜き画像をいかに自然に別画像に挿入するか(Cut-and-Paste学習)
- ノイズ画像と自己教師あり学習で視覚特徴を学習(Visual Representation)

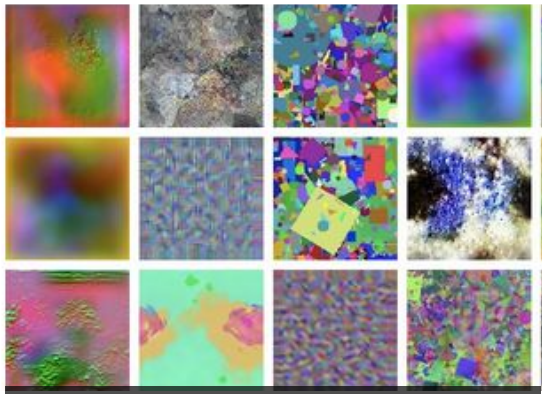


[Shrivastava, CVPR2017] BP
CGをより写実的に表現するRefinerを適用



[Sundermeyer, ECCV2018]Oral,BP
ラベル無しCGデータで実時間6D検出，さらに教師有りを倒した

[Remez, ECCV2018]Oral
Cut/Pasteで既存セグメントラベルを増加，教師有りに接近する精度

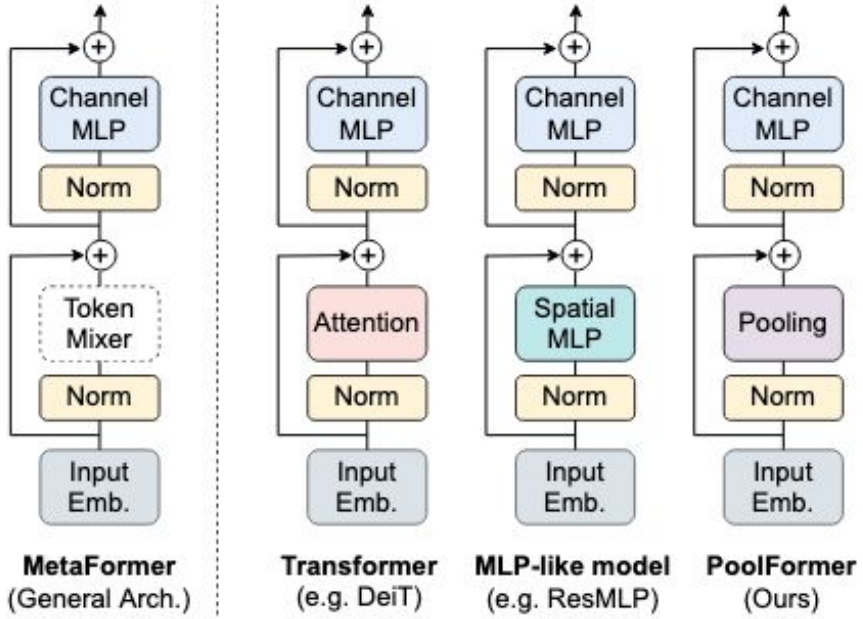


[Baradad, NeurIPS21]Spotlight
ノイズ画像と自己教師で視覚特徴を獲得可能

Keyword: Synthetic Data, Adversarial Learning, Data Augmentation, Domain Randomization, Formula-driven Supervised Learning

DNNの動向・CVのトレンド(37/51)

モデル / データの再考

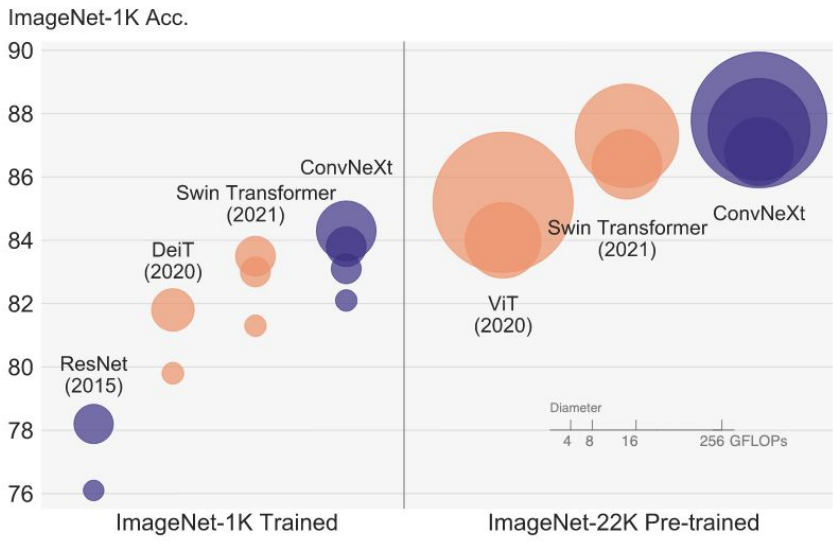


(a) MetaFormer

MLP-Mixer
I. Tolstikhin et al. "MLP-Mixer: An all-MLP Architecture for Vision," in NeurIPS 2021.
https://arxiv.org/abs/2105.01601

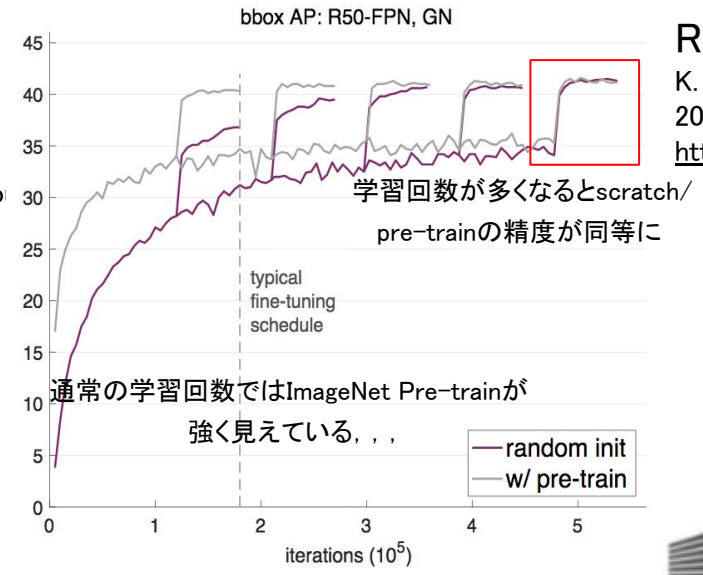
W. Yu et al. "MetaFormer is Actually What You Need for Vision," in arXiv:2111.11418, 2021.
https://arxiv.org/abs/2111.11418

ViTの精度/機能は他のモジュールでも再現できるのではないか？多層パーセプトロン / Poolingで自己注意と同等のことができる.
→本質はどこにあるのか？



ConvNeXt
Z. Liu et al. "A ConvNet for the 2020s," in arXiv:2201.03545, 2022.
https://arxiv.org/abs/2201.03545v1

CNNでも改善すればViTと同等以上にまで改善
- ·グループ化 / チャンネル数増加
- ·ボトルネック改善
- ·カーネルサイズ
- ·など



Rethinking ImageNet Pre-training
K. He et al. "Rethinking ImageNet Pre-training," in ICCV, 2019.
https://arxiv.org/abs/1811.08883

ImageNetは他のタスクの精度向上に貢献する？
- ·検出ではしない(左図参照)
- ·スクラッチで長く学習すれば同等の精度まで到達
  - ≫ ただし，10K以上のラベルは必要
- ·収束の高速化は分野に貢献

DNNの動向・CVのトレンド(38/51)

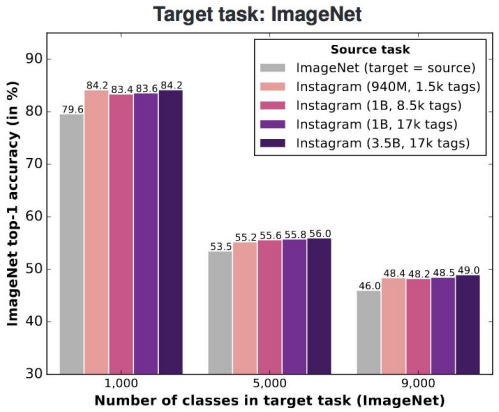
画像/動画DBの大規模化

- 画像識別において，35億画像を含むデータセットが存在
- その他，物体検出や動画認識においても大規模化

【画像識別 / 動画認識】

[Mahajan, ECCV2018]
FBはSNSのHashtagでラベル付けなし，弱教師付きの3.5B枚画像DB構築
https://venturebeat.com/2018/05/02/facebook-is-using-instagram-photos-and-hashtags-to-improve-its-computer-vision/

無/弱/半教師付きの文脈で大量画像とその教師を与えられればモデルを強化できる

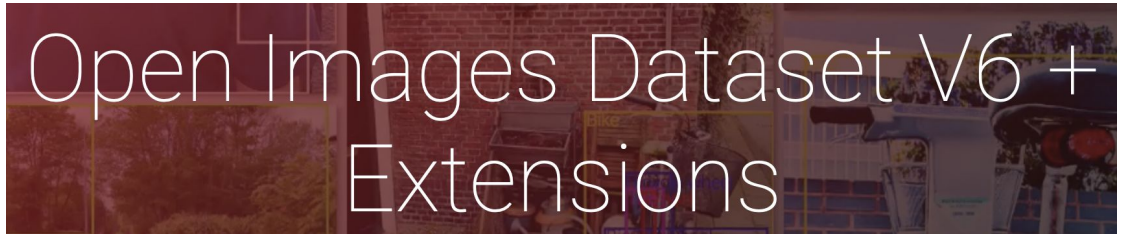


Instagram-3.5B / 65M

D. Mahajan et al. "Exploring the Limits of Weakly-Supervised Pretraining," in ECCV, 2018.
https://arxiv.org/abs/1805.00932

D. Ghadiyaram et al. "Large-scale Weakly-Supervised Pre-Training for Video Action Recognition," in ICCV, 2019.
https://openaccess.thecvf.com/content_CVPR_2019/html/Ghadiyaram_Large-Scale_Weakly-Supervised_Pre-Training_for_Video_Action_Recognition_CVPR_2019_paper.html

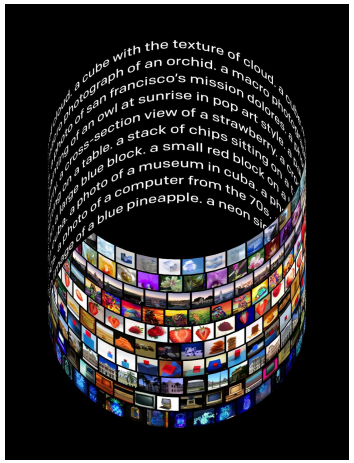
【物体検出】



OpenImages

I. Krasin et al. "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," 2017.
https://storage.googleapis.com/openimages/web/index.html

【画像生成】



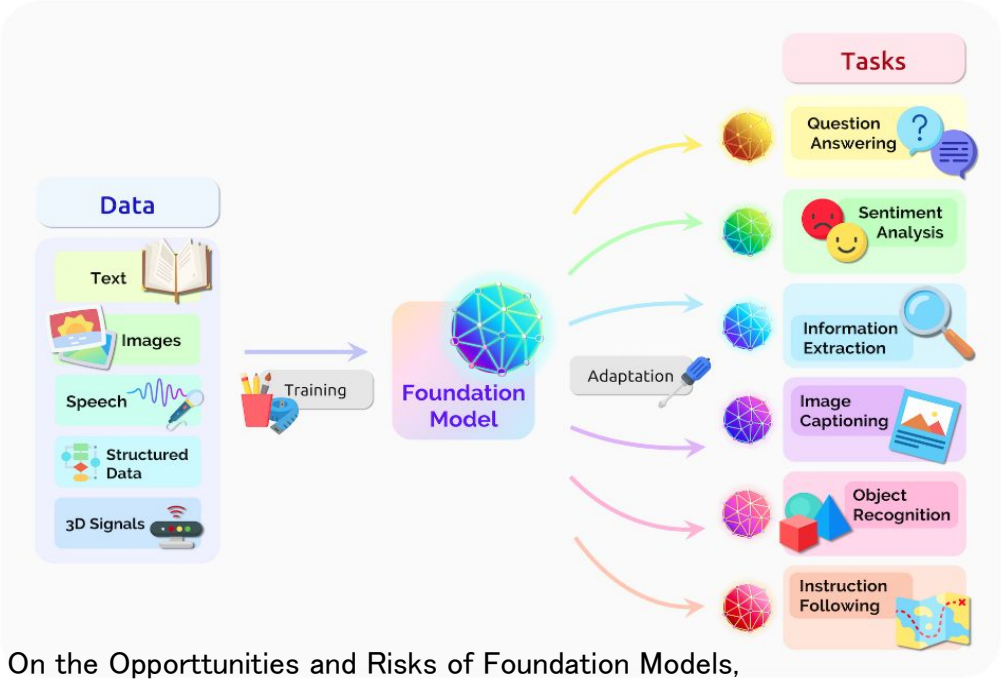
物体検出向けのOpenImageV6は15M画像・座標，DALL·E(前述)は250M画像・テキストを含むデータセットを構築

DALL·E: Creating Images from Text
https://openai.com/blog/dall-e/

DNNの動向・CVのトレンド(39/51)

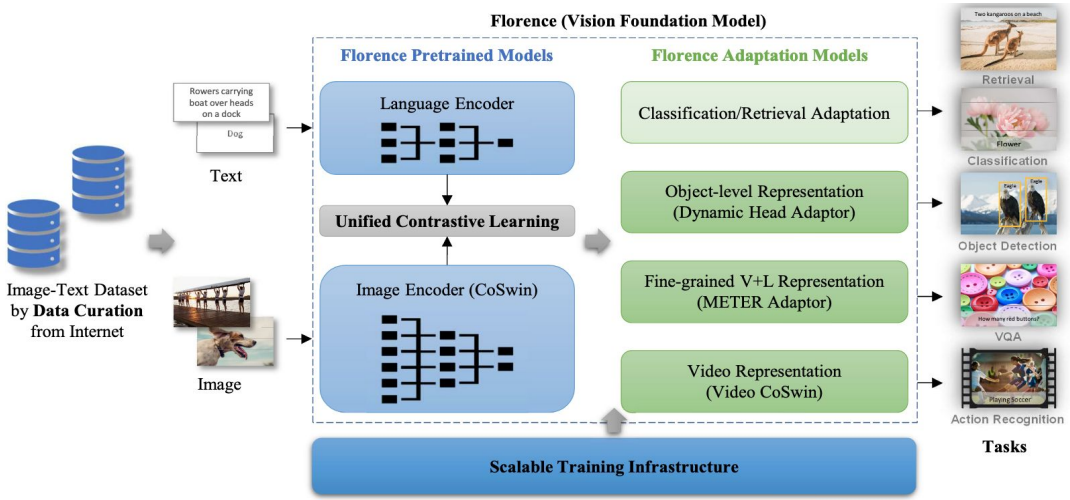
Foundation Model基盤モデル: 汎用適に適用可能なモデル(汎用人工知能への挑戦)

- NLP分野ではBERT, GPT-{1, 2, 3}, DALL-Eなど
- CV分野では最近Florenceなる基盤モデルが提案される



On the Opporttunities and Risks of Foundation Models,
https://arxiv.org/abs/2108.07258

- 基盤モデルのホワイトペーパー的立ち位置
- FMの可能性と危険性を議論

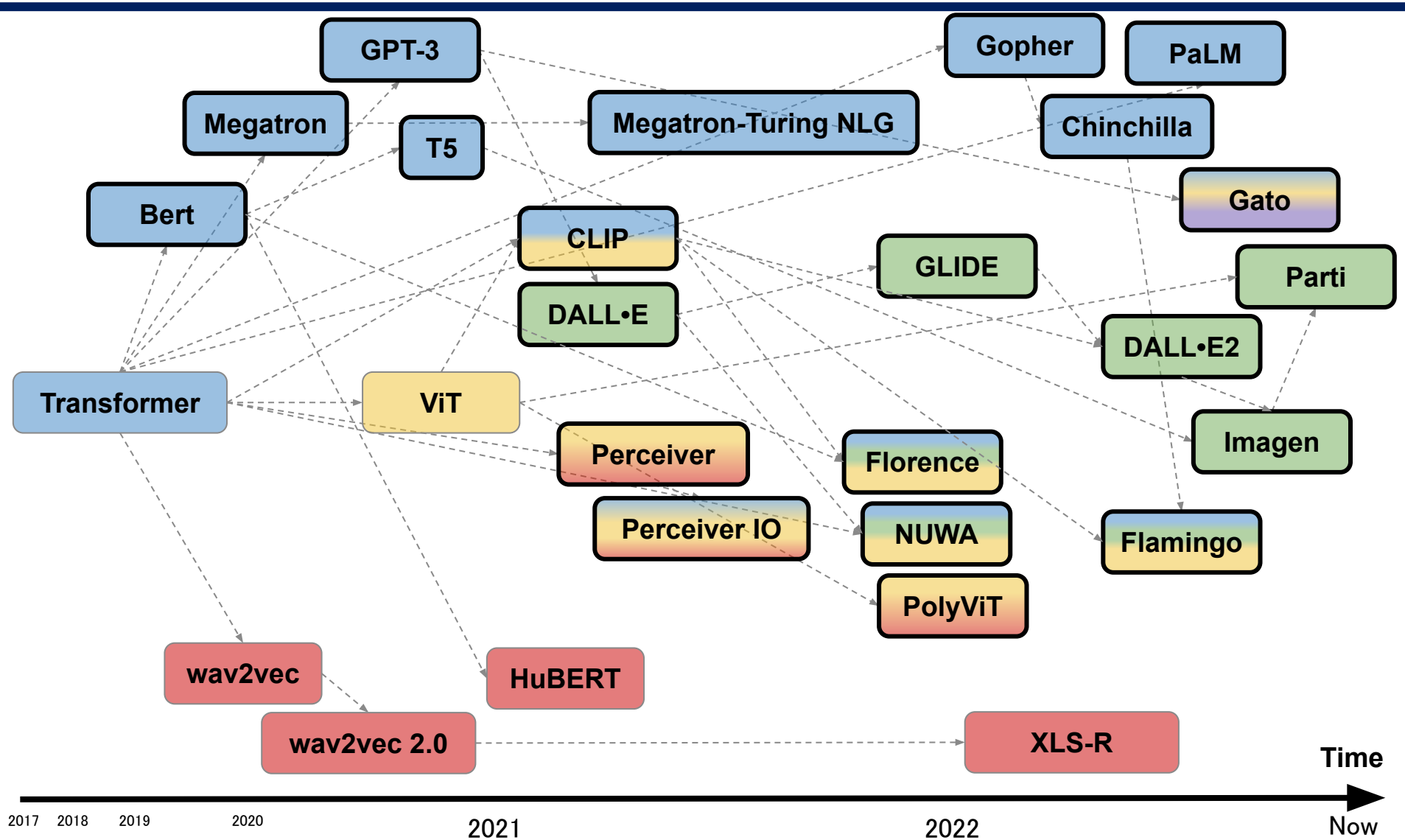


Florence: A New Foundation Model for Computer Vision,
https://arxiv.org/pdf/2111.11432.pdf

- 44タスクを処理可能
- 9億の画像・言語ペアデータを構築

→より最近ではPolyViT(2111.12993) / NuWA(2111.12417) / Omnivore (2201.08377)なるモデルが同時多発的に提案

DNNの動向・CVのトレンド（40/51）



DNNの動向・CVのトレンド(41/51)

自然言語処理(NLP)分野でTransformerが誕生

- **Transformer**
  - Self-Attention(自己注意)機構により

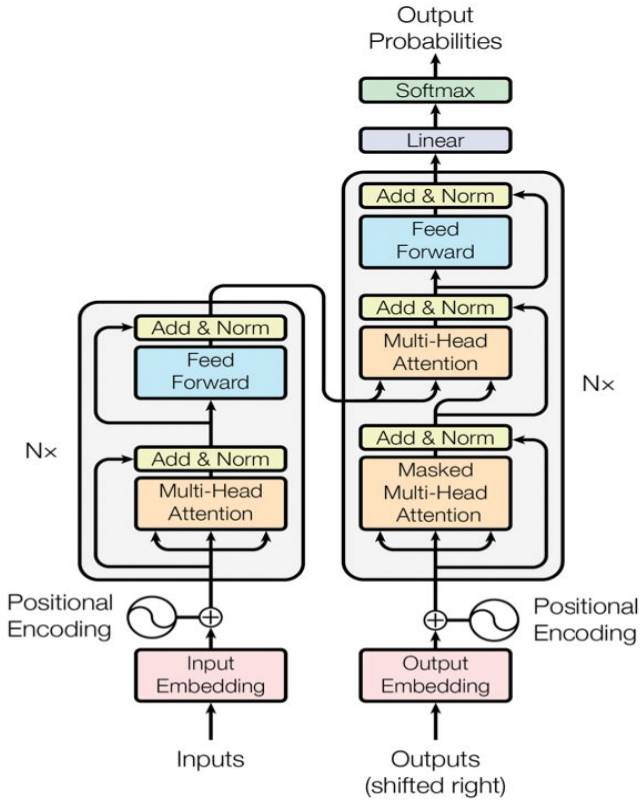
    系列データを一括同時処理可能に

    (従来のRNNでは逐次的に処理

    →並列計算できず処理に時間がかかる,

    離れた単語関係が見えず長文理解は困難)

    - GPUフレンドリーで
      容易に並列計算による学習高速化可能
    - **入力シーケンス全体を考慮可能**
  - 学習時間短縮/性能向上を同時に実現
  - ブレイクスルーを引き起こした
    (当時の翻訳タスクでSoTAを達成)



[Vaswani(Google)+, NeurIPS17]

Transformerについてはこちらも参照
https://www.slideshare.net/cvpaperchallenge/transformer-247407256

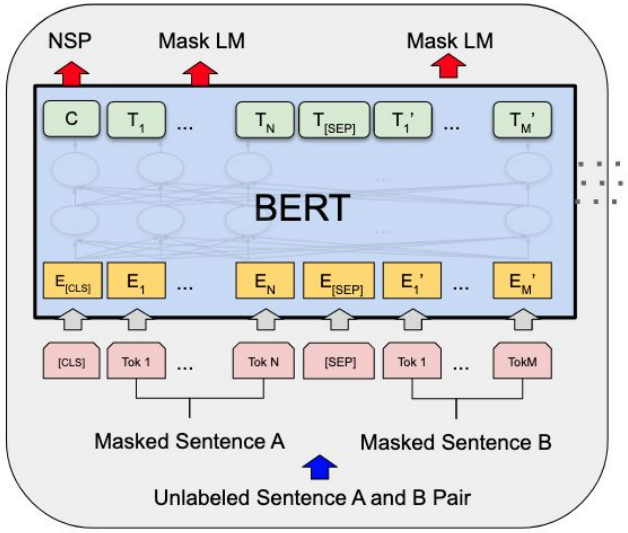
急発達するHPCによるGPU/TPU並列計算技術と相性抜群.
近代のほぼ全ての大規模基盤モデルは, Transformerがベースに

DNNの動向・CVのトレンド(42/51)

NLP分野にてTransformerが拡がる

- **BERT**(Bi-directional Encoder Representations from Transformers) [Devlin(Google)+, NAACL19]
  - 大量の汎用テキストデータで事前学習→タスクに合わせて追加学習
  - 翻訳・予測などNLPのタスクを幅広く解くことができるモデル
  - 文章の「意味を理解」することができるようになったと話題

- **なぜBERTは躍進した？**
  - **自己教師学習によりラベルなし文章を学習に適用可能**
  - 双方向モデルにつき，単語の前後から文脈を把握



Pre-training

BERTでは多くのタスクを単一モデルで解くことができるが，その学習は「文章のマスクと復元」の自己教師あり学習により実施される

Attention is All You Need.(元データ)
↓ 意図的に欠損作成
Attention is All __ Need.(復元前)
↓ BERTにより推定
Attention is All You Need.(復元後)

DNNの動向・CVのトレンド(43/51)

人間レベルの文章生成を可能に

- **GPT**(Generative Pre-trained Transformer) [Brown(OpenAI)+, NeurIPS20]
  - 与えられた文章の先を予測して文章生成
  - 拡張される度にパラメータ数 / 学習テキストサイズが一気に増加
    - GPT : 1.2億パラメータ
    - GPT-2 : 15億パラメータ, 40GBテキスト
    - GPT-3 : 1750億パラメータ, 570GBテキスト(3000億トークン)
    - 爆発的なパラメータ数の増加により大幅な性能改善が見られた
  - 「シンギュラリティが来た」と言われるほどに文章生成能力を獲得

Announcing the NeurIPS 2020 award recipients

Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell and Marc'Aurelio Ranzato

NeurIPS 2020 Program Chairs

In this blog post, we are excited to announce the various awards that are presented at NeurIPS 2020 and to share information about the selection processes for these awards.

NeurIPS 2020 Best Paper Awards

The winners of the NeurIPS 2020 Best Paper Awards are:

NeurIPS 2020 Best Paper Awards

The winners of the NeurIPS 2020 Best Paper Awards are:

- **Language Models are Few-Shot Learners** by Tom B. Brown (OpenAI), Benjamin Mann (OpenAI), Nick Ryder (OpenAI), Melanie Subbiah (OpenAI), Jared D. Kaplan (Johns Hopkins University), Prafulla Dhariwal (OpenAI), Arvind Neelakantan (OpenAI), Pranav Shyam (OpenAI), Girish Sastry (OpenAI), Amanda Askell (OpenAI), Sandhini Agarwal (OpenAI), Ariel Herbert-Voss (OpenAI), Gretchen M. Krueger (OpenAI), Tom Henighan (OpenAI), Rewon Child (OpenAI), Aditya Ramesh (OpenAI), Daniel Ziegler (OpenAI), Jeffrey Wu (OpenAI), Clemens Winter (OpenAI), Chris Hesse (OpenAI), Mark Chen (OpenAI), Eric Sigler (OpenAI), Mateusz Litwin (OpenAI), Scott Gray (OpenAI), Benjamin Chess (OpenAI), Jack Clark (OpenAI), Christopher Berner (OpenAI), Sam McCandlish (OpenAI), Alec Radford (OpenAI), Ilya Sutskever (OpenAI), and Dario Amodei (OpenAI). This paper will be presented on Monday December 7th at 6:00 PM PST in the Language/Audio Applications track.

GPT-3論文はNeurIPS 2020にて
Best Paper Awardを獲得

https://neuripsconf.medium.com/announcing-the-neurips-2020-award-recipients-73e4d3101537

DNNの動向・CVのトレンド(44/51)

CLIPは言語を介して画像の意味を捉えるレンズ

- 従来の画像エンコーダーのように**「そのものに意味がない記号(ラベル)」によって画像を捉えるのではなく,**
  「そのものに意味がある言語表現」によって画像を捉えるため, より解像度の高い自然な画像理解を達成し得る

Class: 3

従来の画像分類で事前学習

CLIPで事前学習

"白と黒のぶち猫が座ってこちらを見ている"

CLIP [Radford(OpenAI)+, CVPR21]

多種多様な画像認識タスク, Vision and Languageタスクへの応用が可能

- 画像分類
- 物体検出
- 画像キャプション妥当性評価
- テキストからの画像生成
- 他モダリティ拡張
- 画像自動キャプショニング
- 視覚言語ナビゲーション
- NeRFによるシーン生成
- VQA

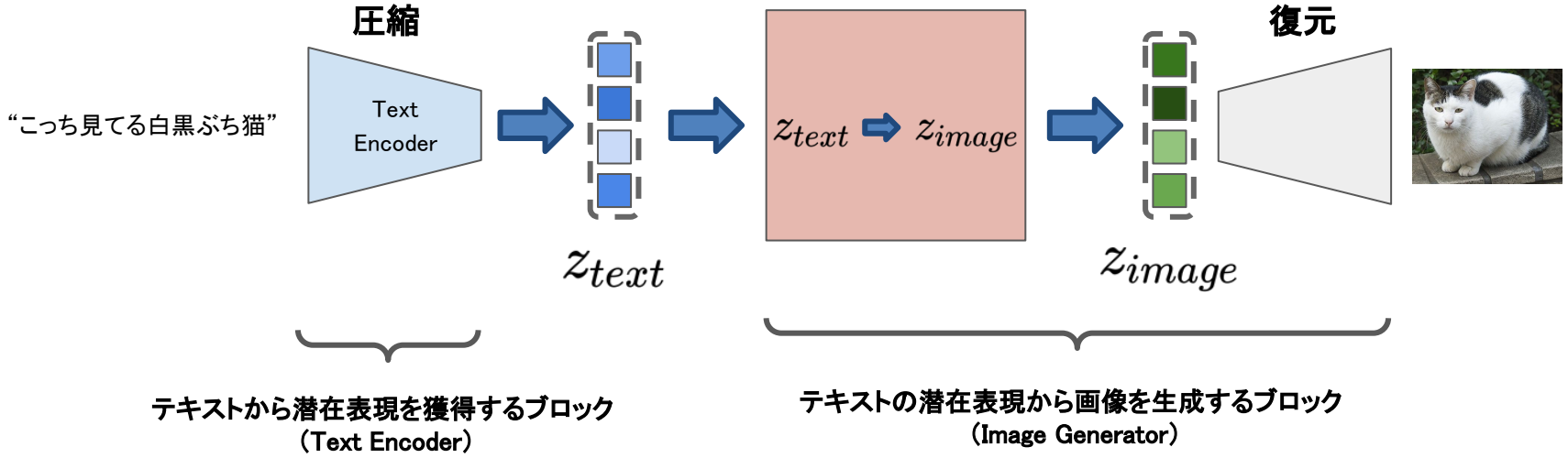
DNNの動向・CVのトレンド(45/51)

次々に更新される文章から画像生成する基盤モデル

- DALL·E [Ramesh(OpenAI)+, ICML21]
- CLIDE [Nichol(OpenAI)+, ACL22]
- VQGAN-CLIP [Crowson(EleutherAI)+, arXiv22]
- DALL·E 2 [Ramesh(OpenAI)+, arXiv22]
- Imagen [Saharia(Google)+, arXiv22]
- Parti [Yu(Google)+, arXiv22]

etc…

大体どのモデルも生成時のお気持ちはこんな感じ.



基盤モデル間で異なる主な要素は,

- **Text Encoder に何を使うか**
- **Image Generator に何を使うか**

DNNの動向・CVのトレンド(46/51)

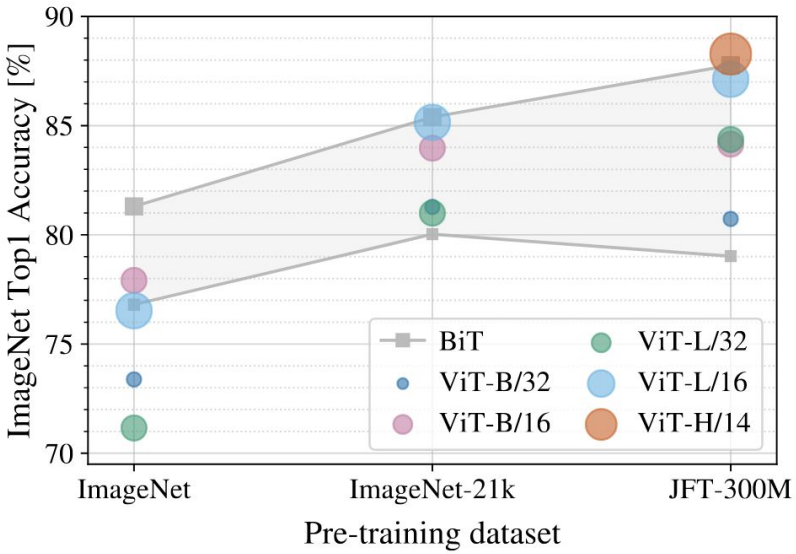
Transformer → Vision FMs 3つのポイント

- **複数モダリティ・単一モデル**
  - 画像・言語・音声を一つのTransformerモデルで扱える
  - 同時に扱えるモダリティは増加傾向
- **モデル・データのサイズ増加**
  - データ:数億～数十億規模(e.g., WebImageText(400M), JFT-4B)
  - モデル:百億パラメータ規模(言語だと数兆とかもっと多い)(e.g.,Flamingo, Parti)
- **ラベルなしデータで事前学習**
  - 自己教師あり学習の台頭
  - マルチモーダルの教師なし学習

DNNの動向・CVのトレンド(47/51)

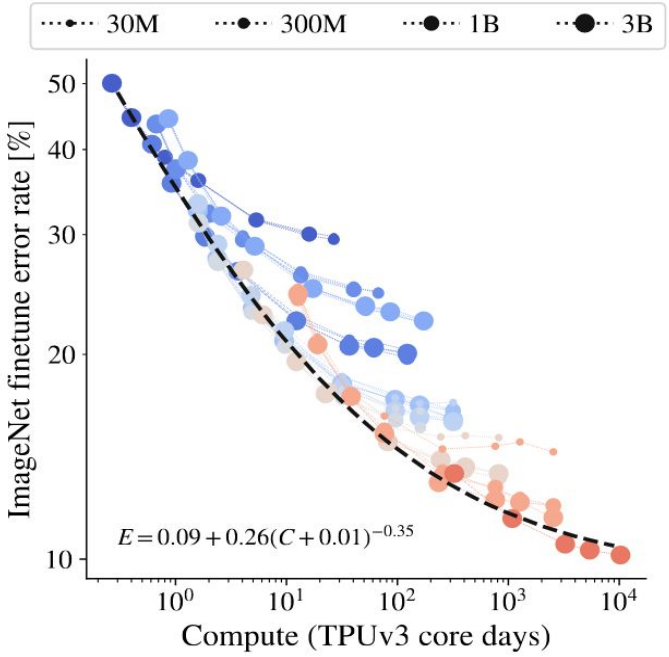
超大規模データセット JFT-300M/3B

■ JFT-300M/3Bはそれぞれ3億/30億画像含む画像データセット



Vision Transformers, https://arxiv.org/pdf/2010.11929.pdf

ImageNet-1k Top-1 Accuracy: 88.55 (ViT-H/14)



Scaling Vision Transformers,
https://arxiv.org/pdf/2106.04560.pdf

ImageNet-1k Top-1 Accuracy: 90.45 (ViT-G/14)

自己教師/数式教師は人間教師+超大規模画像データセットに匹敵するのか？

DNNの動向・CVのトレンド（48/51）

巨大Tech企業による大規模データセット寡占化

○ 独自に保有するプラットフォームで

非公開な超大規模データを収集し利用

Google

JFT-300M/3B/4B （Google, 2017/2021/2022）

→約3億/30億/40億枚の画像データセット

4 Training and Evaluation Datasets

4.1 Training Datasets

Parti [Yu(Google)+, arXiv22]

We train on a combination of image-text datasets for all Parti models. The data includes the publicly available LAION-400M dataset [43]; FIT400M, a filtered subset of the full 1.8 billion examples used to train the ALIGN model [9]; JFT-4B dataset [44], which has images with text annotation labels. For textual descriptions of JFT, we randomly switch between the original labels as text (concatenated if an image has multiple labels) or machine-generated captions from a SimVLM model [45]. We discuss

Meta

IG-3.5B （Facebook, 2018）

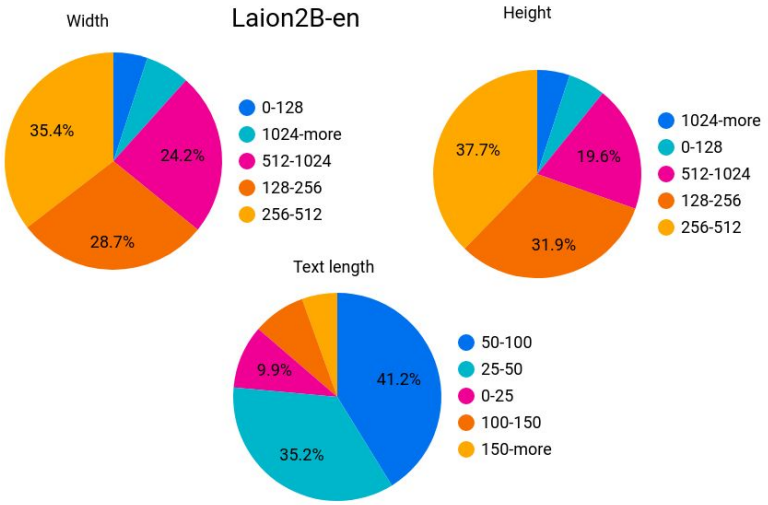
→約35億枚の画像データセット

プライバシーの問題や誤用の懸念がある以上，
容易に公開はできず難しい問題

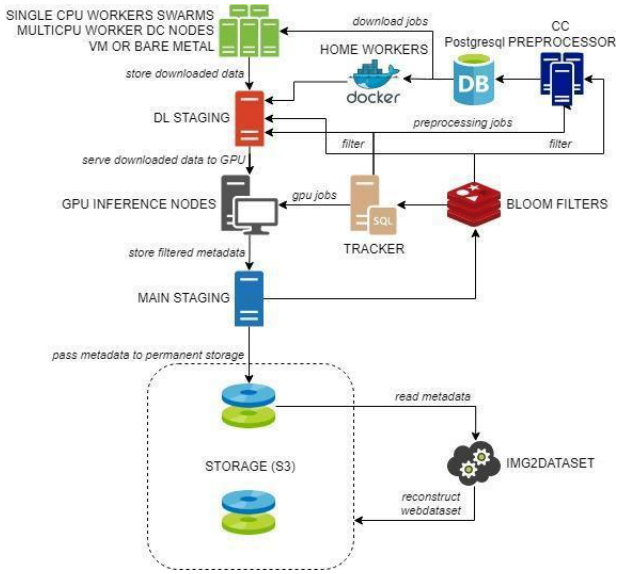
DNNの動向・CVのトレンド(49/51)

非営利組織によって構築された画像・言語データが登場

- LAION(Large-scale Artificial Intelligence Open Network)
  - LAION-400M / 5Bはそれぞれ4億 / 50億(!)の画像・テキストのペアデータ
  - 真のOPEN AIを自称ー100% 非営利, 100%無料
  - 寄付や公的研究費を使用, 志を持つ研究者などが集合してデータセットを構築
  - データはCreative Common CC-BY 4.0 License



LAION2B-enの統計量
(他の統計情報もあり)



- Common Crawl dataset (Peta-byte scale)からURL/テキストのペア構築
- 単一URLから多重に画像取得を避けつつ画像ダウンロード
- データをCLIPにて後処理(複数GPUノードを数日間実行),
  画像とテキストの類似度判定

https://laion.ai/#top

DNNの動向・CVのトレンド(50/51)

High Performance Computing: 研究の加速

- AWS/Azure/Google Cloud, 日本でもTsubame3.0/ABCI/富岳
  - Tsubame3.0: 2,160 GPUs / 540 Nodes
    - NVIDIA Tesla P100 x 2,160
  - ABCI: 5,312 GPUs / 1,208 Nodes
    - NVIDIA V100 x 4,352 + NVIDIA A100 x 960
- ImageNet 世界最速記録の変遷
  - 29h > 1h > 30m > 15m > 6.6m > 1.8m > 2.0m > 1.2m

東工大TSUBAME 3.0

産総研ABCI

	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 → 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia et al. [5]	65,536	Tesla P40 × 2,048	TensorFlow	6.6 mins	75.8 %
Ying et al. [6]	65,536	TPU v3 × 1,024	TensorFlow	1.8 mins	75.2 %
Mikami et al. [7]	55,296	Tesla V100 × 3,456	NNL	2.0 mins	75.29 %
This work	**81,920**	**Tesla V100 × 2,048**	**MXNet**	**1.2 mins**	**75.08%**

M. Yamazaki, et al. "Yet Another Accelerated SGD: ResNet-50 Trainingon ImageNet in 74.7 seconds," arXiv pre-print, 1903.12650, 2019.
https://arxiv.org/pdf/1903.12650.pdf

DNNの動向・CVのトレンド（51/51）

コンピュータビジョンおよび周辺分野で起こっていること一部まとめ

- **深層学習において，人間の教師ラベルを用いずとも特定タスクにおいては人間よりも優れた視覚機能を獲得**
  - 自己教師あり学習による画像認識
- **大規模画像・言語データセットにて学習した画像生成モデルは想像力を持ったと言われる**
  - 画像生成AI
- **画像・言語問わず汎用的にタスクを実行するモデルが登場**
  - 基盤モデル
- **3次元空間を超高精細に再現**
  - Neural Radiance Fields（NeRF）

CVPR 2023 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

CVPR 2023 の動向・気付き(1/259)

(Senior) Area Chairs / Outstanding Reviewers

- Senior Area Chairsは流石に経験豊富な研究者が多いが，Area Chairs / Outstanding Reviewersは若手研究者から何人も選出されている
- Area Chair は(おそらく)紹介制度だがイケてる研究者が選出される
- Outstanding Reviewer は時の運であり，印象で決まる部分もあるが研究を見る「良い目」がある傾向にあると言える
- 日本の研究者も，全項目で選ばれている！
  - 日本にも「経験豊富な」「イケてる」「良い目」を持った研究者がいることを示す

*CVPR 2023 Program Guide (Main Conference)
https://media.icml.cc/Conferences/CVPR2023/CVPR_2023_MainConference_ProgramGuide.pdf

CVPR 2023 の動向・気付き(2/259)

Program Guide*より「投稿側」の詳細

- ”CVPR 2023 received 9155 submissions”
  - 9,155論文が投稿され，前年よりも12%(CVPR 2022時は8,161論文)投稿数が伸びた

- ”2359 papers were accepted (25.8% acceptance rate)”
- “235 (10%) were selected as highlights … 12 (0.5%) papers have been shortlisted as best paper award candidates”
  - １国際会議に2,300本を超える論文が採択
  - うち10%に当たる235件がハイライト発表，12件のみが受賞候補論文としてリスト

*CVPR 2023 Program Guide (Main Conference)
https://media.icml.cc/Conferences/CVPR2023/CVPR_2023_MainConference_ProgramGuide.pdf

CVPR 2023 の動向・気付き(3/259)

Program Guide*より「査読側」の詳細

- ”The review process was managed by 400+ area chairs and, new to the process this year, 30 senior area chairs”
  - 査読者をまとめるエリアチェア(AC)だけで400名超，さらに上位のシニアエリアチェア(Senior AC)を30名割り当て
  - このさらに上位にプログラム委員長(PC)がいる

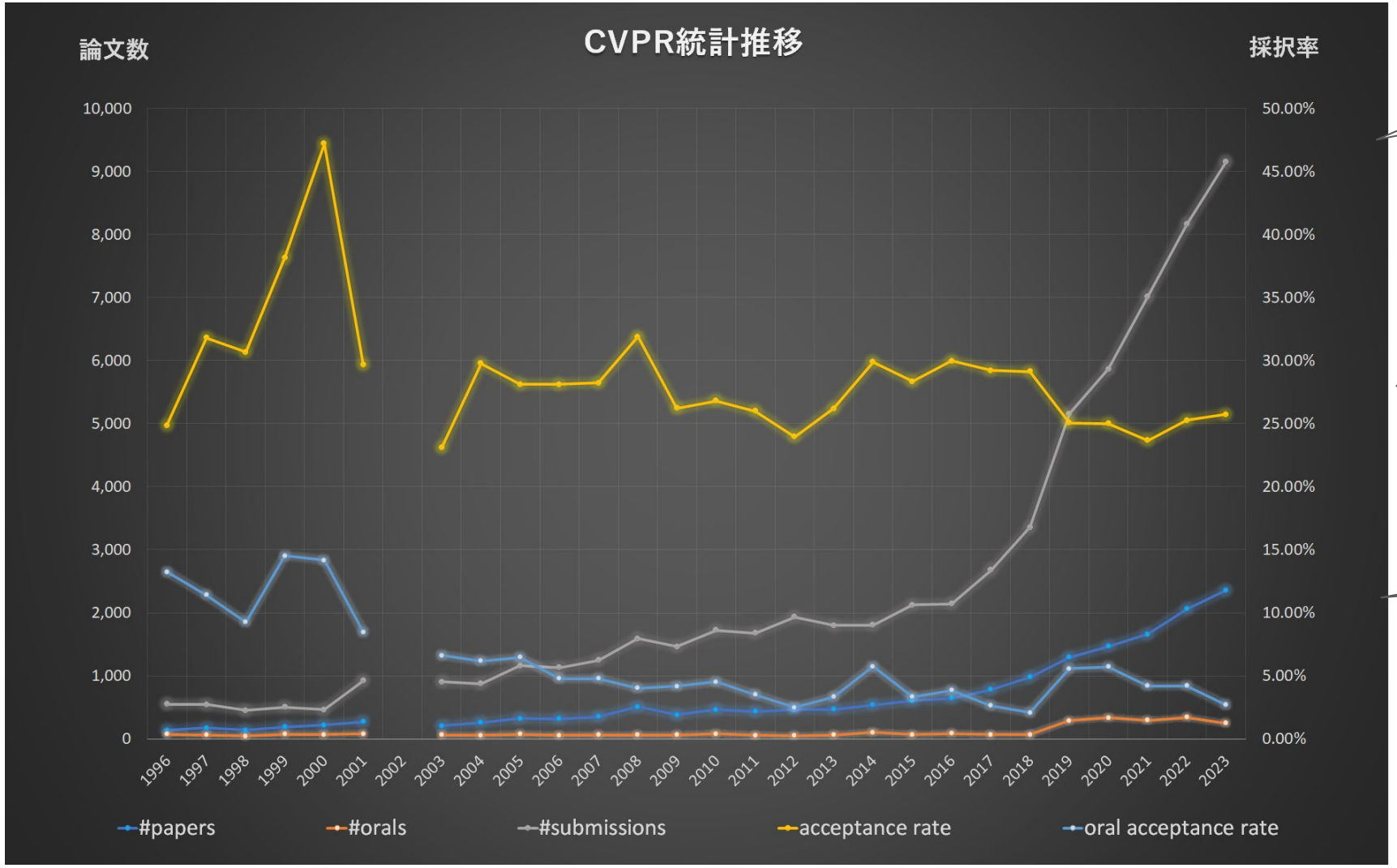
- ”During the review phase, each paper received at least 3 reviews from the pool of 6625 reviewers”
  - 実に6,625名もの査読者が査読に取り組んだ
  - 1論文3査読以上とすると，有効論文数9,155に対して27,465以上の査読が集まった(実際にはより多くの査読が集まったと考えられる)

→ 査読者の多大なる尽力により学会が成り立ち，研究や論文が進展している！

*CVPR 2023 Program Guide (Main Conference)
https://media.icml.cc/Conferences/CVPR2023/CVPR_2023_MainConference_ProgramGuide.pdf

CVPR 2023 の動向・気付き(4/259)

CVPR投稿数・採択率の変遷



「投稿数」は深層学習時代に激増！2024に10,000投稿に届こうか，という勢い

「採択率」は25%前後に調整

「採択数」は2018-2019で1,000本超，2022, 2023は2,000本超まで到達
（2014でも500本程度であった）

ResearchPort -「CVPR2023-速報-」ResearchPortトップカンファレンス定点観測 vol.9より
https://research-p.com/column/1237

CVPR 2023 の動向・気付き(5/259)

Program Overview*より

- まずOralがない…？
  - Award candidates のみ口頭発表，oralはよく見るとhighlightという名前
  - CVPRのYouTubeチャンネルから highlight として紹介:
    https://www.youtube.com/playlist?list=PL_bDvITUYucBlYzTTt3DEsLBBr-wAH6MS

	TUESDAY	WEDNESDAY	THURSDAY
8:00 AM	Breakfast	Breakfast	Breakfast
8:30 AM	Opening	Awards	
9:00 AM	Plenary keynote: Rodney Brooks	Plenary keynote: Yejin Choi	Plenary keynote: Larry Zitnick
9:30 AM			
10:00 AM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
10:30 AM	Poster session TUE-AM	Poster session WED-AM	Poster session THU-AM
11:00 AM			
11:30 AM			
12:00 PM			
12:30 PM	Lunch	Lunch	Lunch
1:00 PM			
1:30 PM			

⋮ 見やすさのため分割してます

2:00 PM	Plenary panel: History and Future of Artificial Intelligence and Computer Vision	Plenary panel: Vision, Language, and Creativity	Plenary panel: Scientific Discovery and the Environment
2:30 PM			
3:00 PM	Award candidates TUE	PAMI-TC meeting	Award candidates THU
3:30 PM			
4:00 PM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
4:30 PM	Poster session TUE-PM	Poster session WED-PM	Poster session THU-PM
5:00 PM			
5:30 PM			
6:00 PM			
6:30 PM	Socials		
7:00 PM		Reception and Musical Performances	
7:30 PM			
8:00 PM			
8:30 PM			

*CVPR 2023 Program Overview (Main Conference)
https://cvpr2023.thecvf.com/Conferences/2023/ProgramOverview

CVPR 2023 の動向・気付き(6/259)

Program Overview*より

- Keynote(招待講演)が復活！
  - 投票を反映して一時期 keynote はプログラムから除外され(確か)よりoralに着目
  - 今回逆にoralを外し，その結果本会議を4日間から3日間に短縮？

	TUESDAY	WEDNESDAY	THURSDAY
8:00 AM	Breakfast	Breakfast	Breakfast
8:30 AM	Opening	Awards	
9:00 AM	Plenary keynote: Rodney Brooks	Plenary keynote: Yejin Choi	Plenary keynote: Larry Zitnick
9:30 AM			
10:00 AM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
10:30 AM	Poster session TUE-AM	Poster session WED-AM	Poster session THU-AM
11:00 AM			
11:30 AM			
12:00 PM			
12:30 PM	Lunch	Lunch	Lunch
1:00 PM			
1:30 PM			

⋮ 見やすさのため分割してます

⋮

2:00 PM	Plenary panel: History and Future of Artificial Intelligence and Computer Vision	Plenary panel: Vision, Language, and Creativity	Plenary panel: Scientific Discovery and the Environment
2:30 PM			
3:00 PM	Award candidates TUE	PAMI-TC meeting	Award candidates THU
3:30 PM			
4:00 PM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
4:30 PM	Poster session TUE-PM	Poster session WED-PM	Poster session THU-PM
5:00 PM			
5:30 PM			
6:00 PM			
6:30 PM	Socials		
7:00 PM		Reception and Musical Performances	
7:30 PM			
8:00 PM			
8:30 PM			

*CVPR 2023 Program Overview (Main Conference)
https://cvpr2023.thecvf.com/Conferences/2023/ProgramOverview

CVPR 2023 の動向・気付き(7/259)

Program Overview*より

- Posterは2時間
  - よく考えると，本会議で2,359論文が採択されている
  - 2時間あっても，セッションあたり400本弱の論文が発表される計算(oh･･･)

「まず読み/見切れない(諦め)」「論文読もうね(お任せ)」のメッセージか

	TUESDAY	WEDNESDAY	THURSDAY
8:00 AM	Breakfast	Breakfast	Breakfast
8:30 AM	Opening	Awards	
9:00 AM	Plenary keynote: Rodney Brooks	Plenary keynote: Yejin Choi	Plenary keynote: Larry Zitnick
9:30 AM			
10:00 AM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
10:30 AM	Poster session TUE-AM	Poster session WED-AM	Poster session THU-AM
11:00 AM			
11:30 AM			
12:00 PM			
12:30 PM	Lunch	Lunch	Lunch
1:00 PM			
1:30 PM			

⋮ 見やすさのため分割してます

2:00 PM	Plenary panel: History and Future of Artificial Intelligence and Computer Vision	Plenary panel: Vision, Language, and Creativity	Plenary panel: Scientific Discovery and the Environment
2:30 PM			
3:00 PM	Award candidates TUE	PAMI-TC meeting	Award candidates THU
3:30 PM			
4:00 PM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
4:30 PM	Poster session TUE-PM	Poster session WED-PM	Poster session THU-PM
5:00 PM			
5:30 PM			
6:00 PM			
6:30 PM	Socials		
7:00 PM		Reception and Musical Performances	
7:30 PM			
8:00 PM			
8:30 PM			

*CVPR 2023 Program Overview (Main Conference)
https://cvpr2023.thecvf.com/Conferences/2023/ProgramOverview

CVPR 2023 の動向・気付き(8/259)

Program Overview*より

- Plenary panelの登場！
  - そのトピックにおけるプロ中のプロを招待して，トレンドを把握するセッション
  - “History and Future of AI and CV”, “Vision, Language, and Creativity”, and “Scientific Discovery and the Environment”の3トピック

パネリスト一覧

Panelists:

- **History and Future of Artificial Intelligence and Computer Vision**
  Chelsea Finn (Stanford), Dan Huttenlocher (MIT), Linda Shapiro (University of Washington), Jamie Shotton (Wayve)
- **Vision, Language, and Creativity**
  Aaron Hertzmann (Adobe), Michal Irani (Weizmann Institute of Science), Devi Parikh (Georgia Tech), Jason Salavon (University of Chicago)
- **Scientific Discovery and the Environment**
  Elizabeth A. Barnes (Colorado State University), Sara Beery (MIT), Josh Bloom (UC Berkeley), Kyle Cranmer (UW Madison)

	TUESDAY	WEDNESDAY	THURSDAY
12:00 PM			
12:30 PM	Lunch	Lunch	Lunch
1:00 PM			
1:30 PM			

⋮ 見やすさのため分割してます

⋮

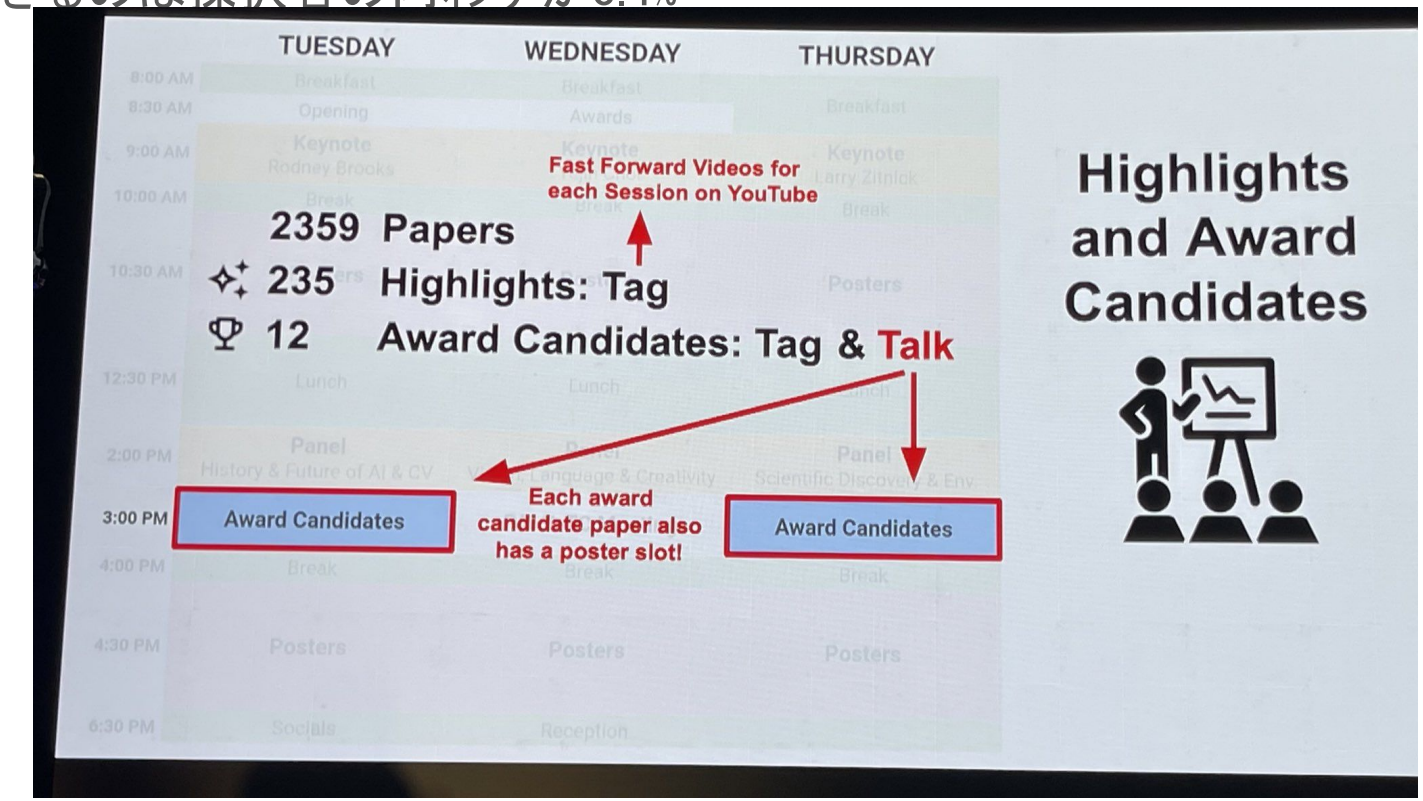
2:00 PM	Plenary panel: History and Future of Artificial Intelligence and Computer Vision	Plenary panel: Vision, Language, and Creativity	Plenary panel: Scientific Discovery and the Environment
2:30 PM			
3:00 PM	Award candidates TUE	PAMI-TC meeting	Award candidates THU
3:30 PM			
4:00 PM	Coffee break / Poster setup	Coffee break / Poster setup	Coffee break / Poster setup
4:30 PM	Poster session TUE-PM	Poster session WED-PM	Poster session THU-PM
5:00 PM			
5:30 PM			
6:00 PM			
6:30 PM	Socials		
7:00 PM		Reception and Musical Performances	
7:30 PM			
8:00 PM			
8:30 PM			

*CVPR 2023 Program Overview (Main Conference)
https://cvpr2023.thecvf.com/Conferences/2023/ProgramOverview

CVPR 2023 の動向・気付き(9/259)

Program Overview*より

- Award Candidates
  - 採択された論文が持ち時間10分で(実質)Oral発表
  - 登壇できるのは採択者の内わずか0.4%

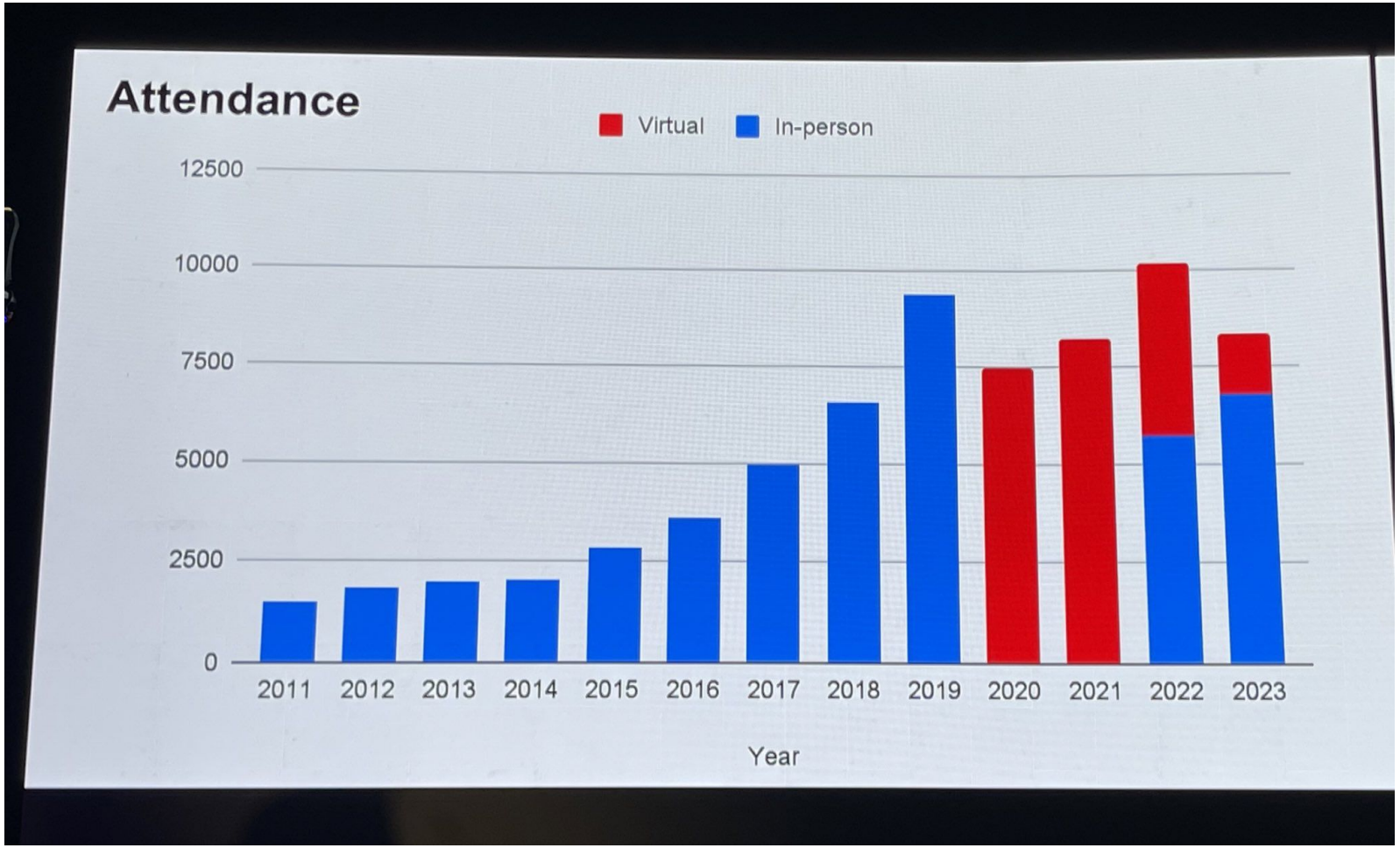


⋮

CVPR公式Twitterより引用
https://twitter.com/CVPR/status/1671181305596485632?s=20

CVPR 2023 の動向・気付き(10/259)

CVPR参加者の推移



参加者は昨年に比べて減少

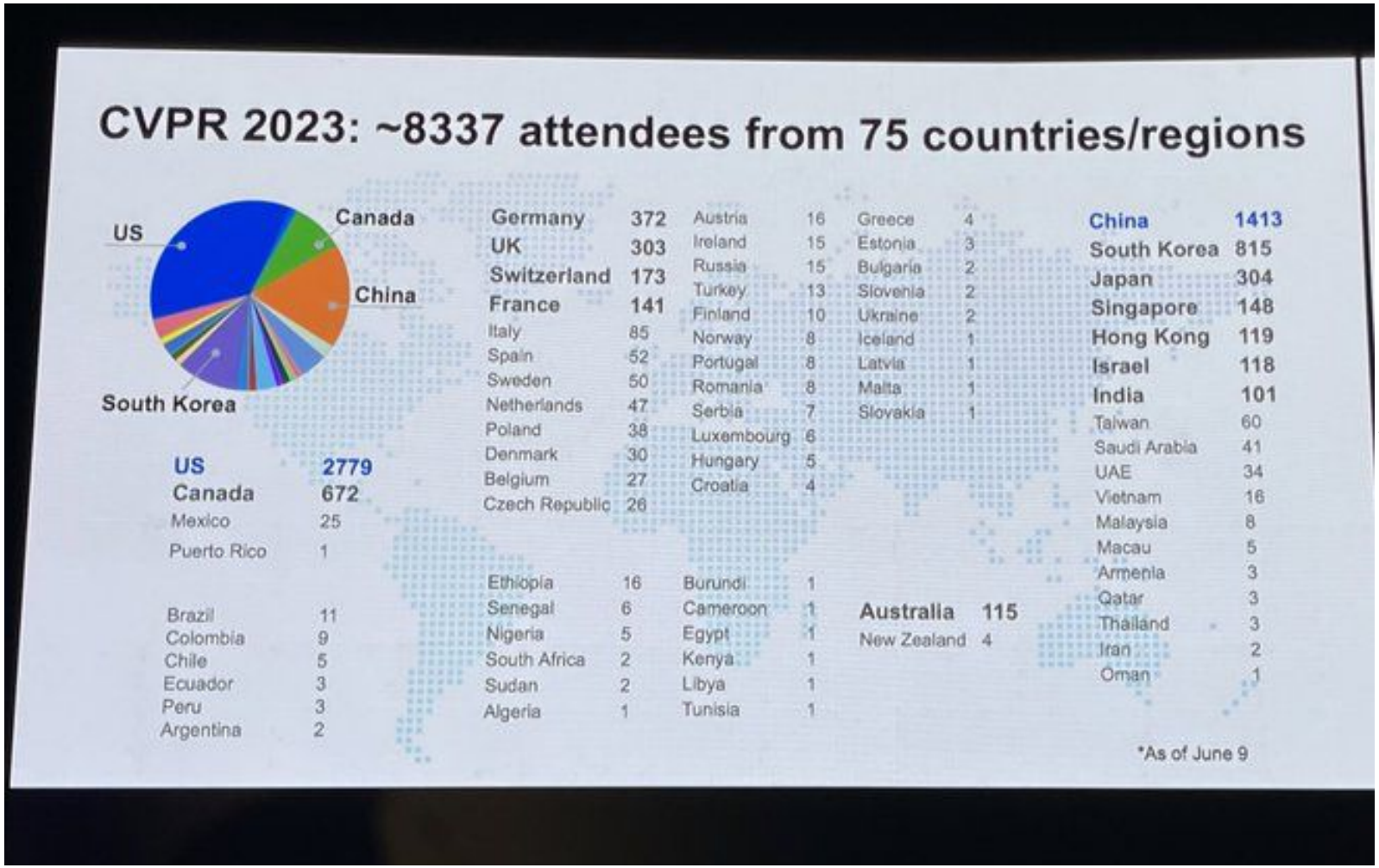
一部Virtualで参加される方も

バンクーバー開催は中国の方がvisaの取得が困難で参加者が減った？？

CVPR公式Twitterより引用
https://twitter.com/CVPR/status/1671181305596485632?s=20

CVPR 2023 の動向・気付き(11/259)

CVPR参加者の内訳



CVPR参加登録者は8337人

内訳は，1位US 2279，2位中国 1413，3位 韓国 815，4位 カナダ 672

中国の参加者は多いが実際に会場に来ている人は少なかった印象

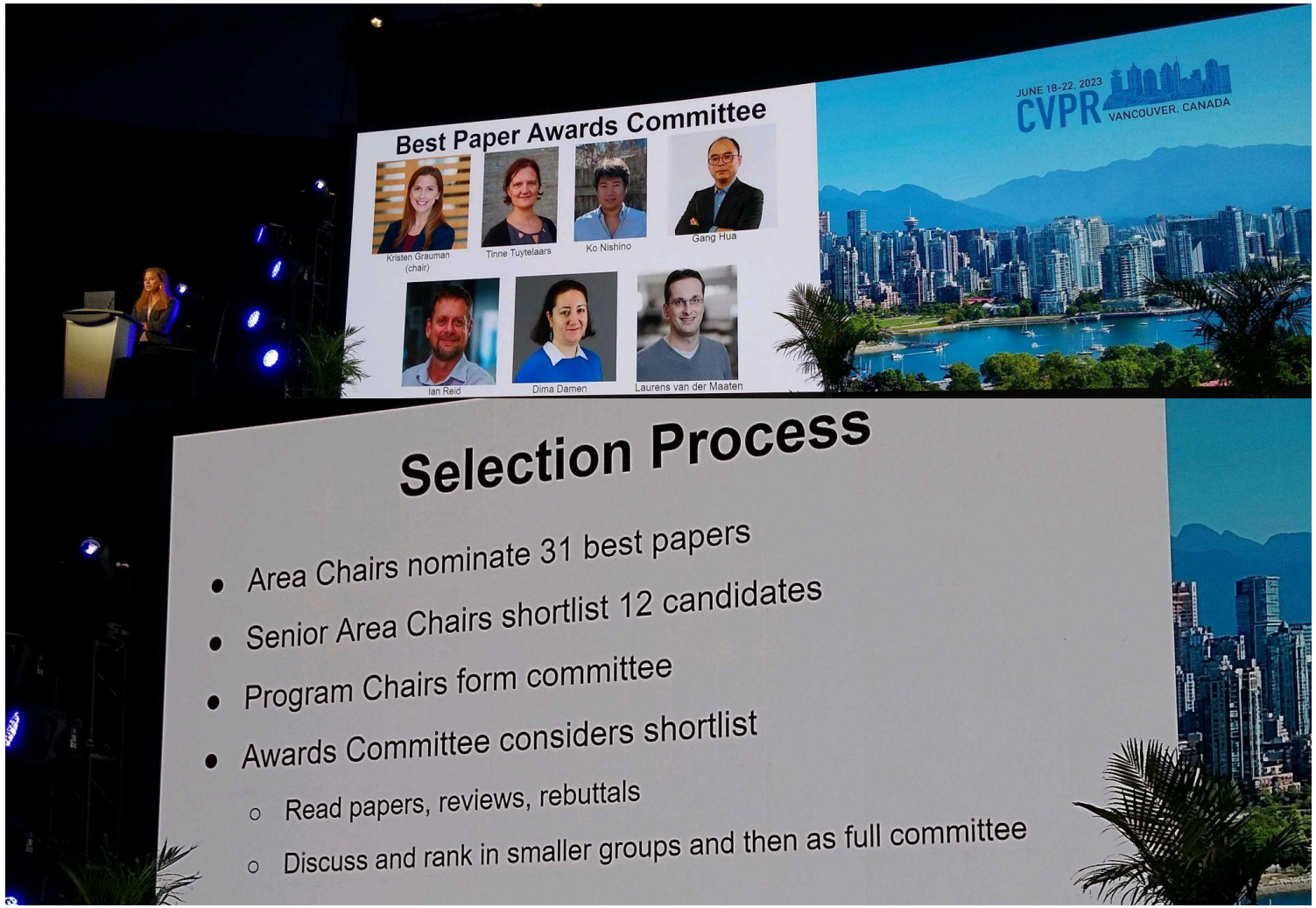
日本は304名参加，順位としては全体の6位

CVPR公式Twitterより引用
https://twitter.com/CVPR/status/1671181305596485632?s=20

CVPR 2023 の動向・気付き(12/259)

Best Papers

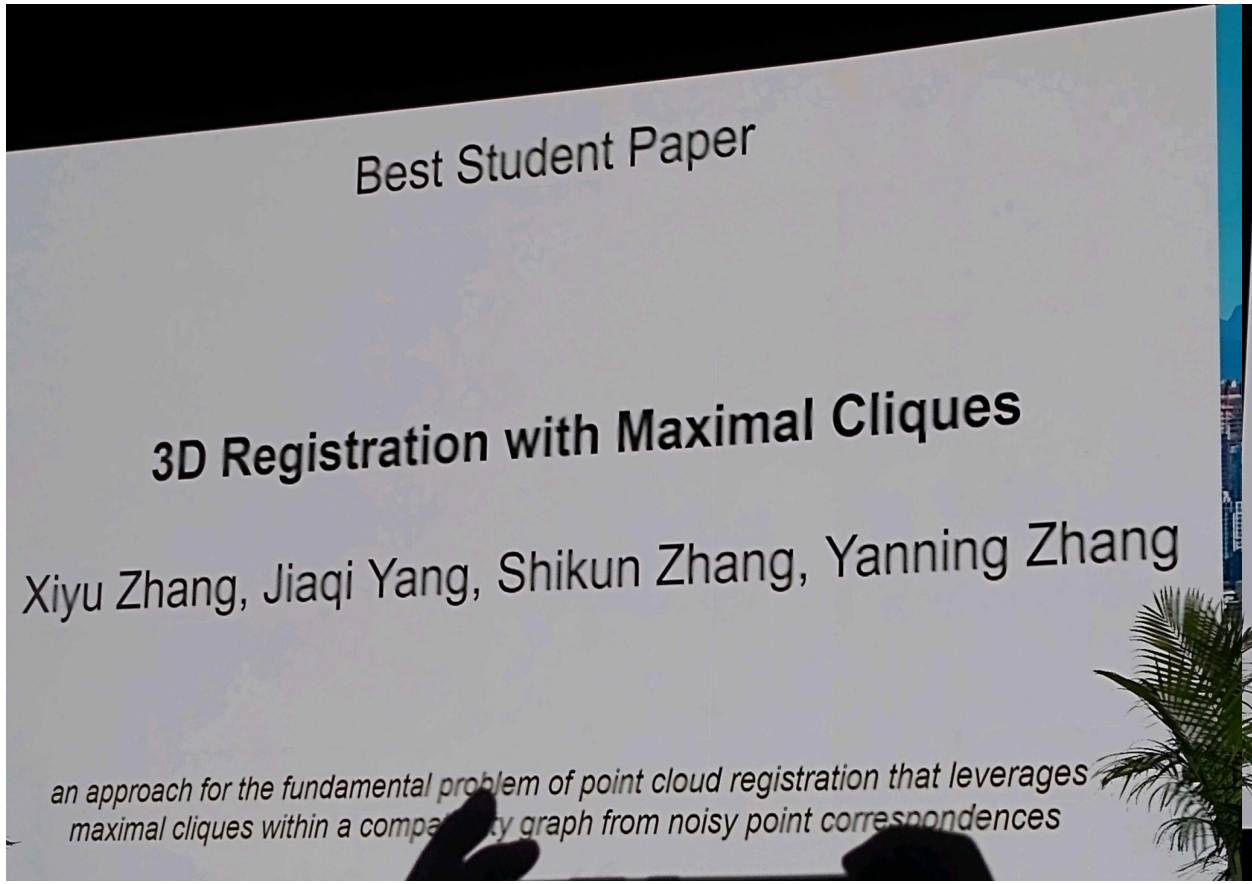
- まずはAwards Committee(賞候補委員会)・選定プロセスから

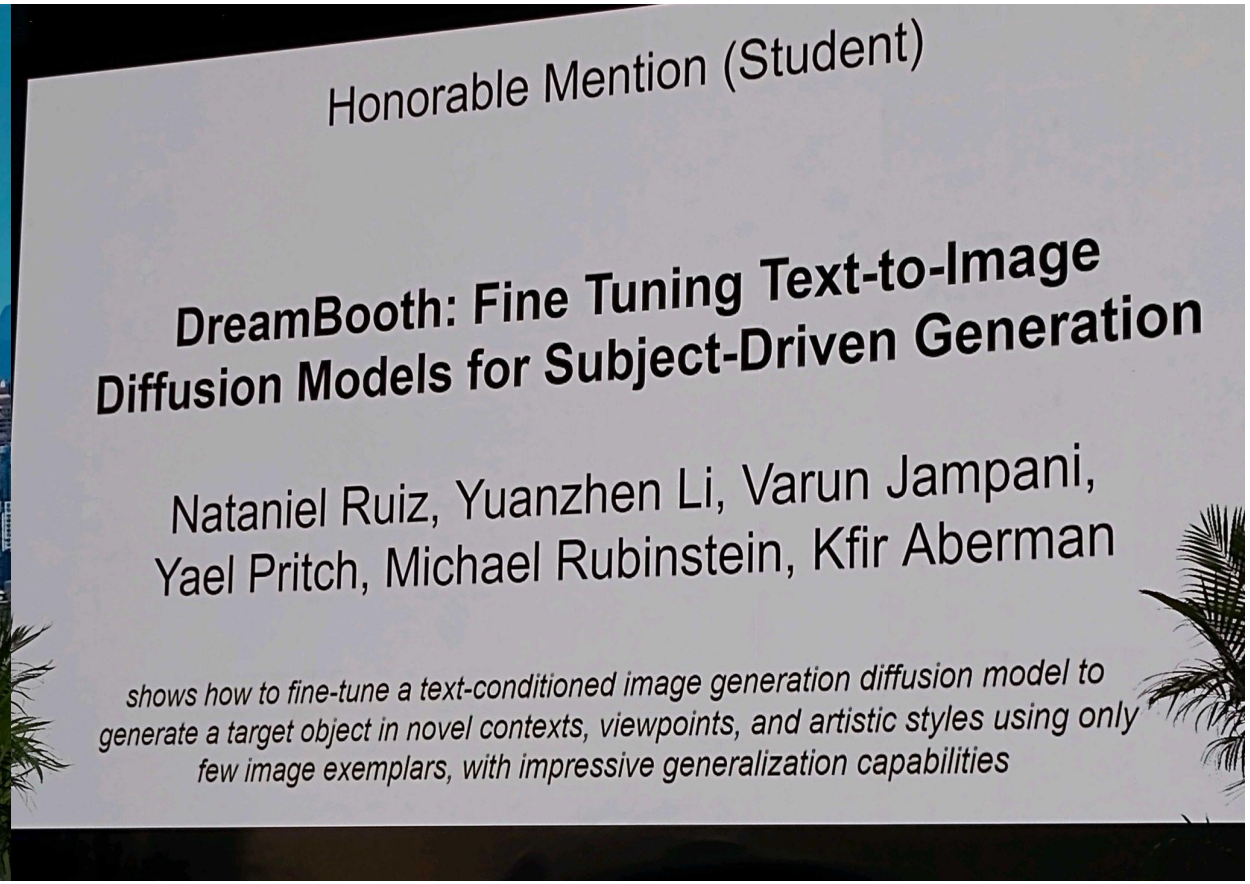


CVPR 2023 の動向・気付き(13/259)

Best Paper Awards

- Best Student Paper / Honorable Mention

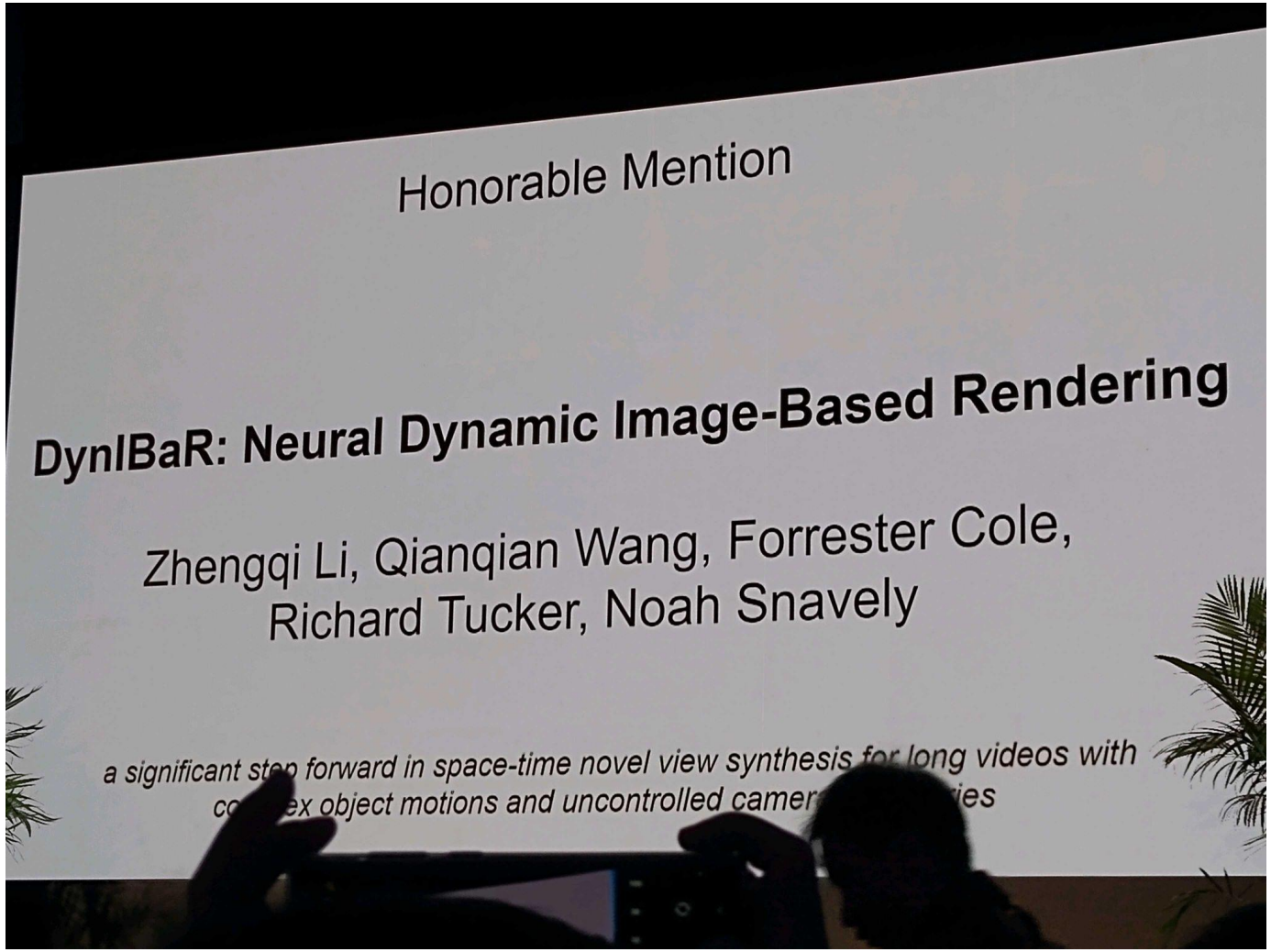




CVPR 2023 の動向・気付き(14/259)

Best Paper Awards

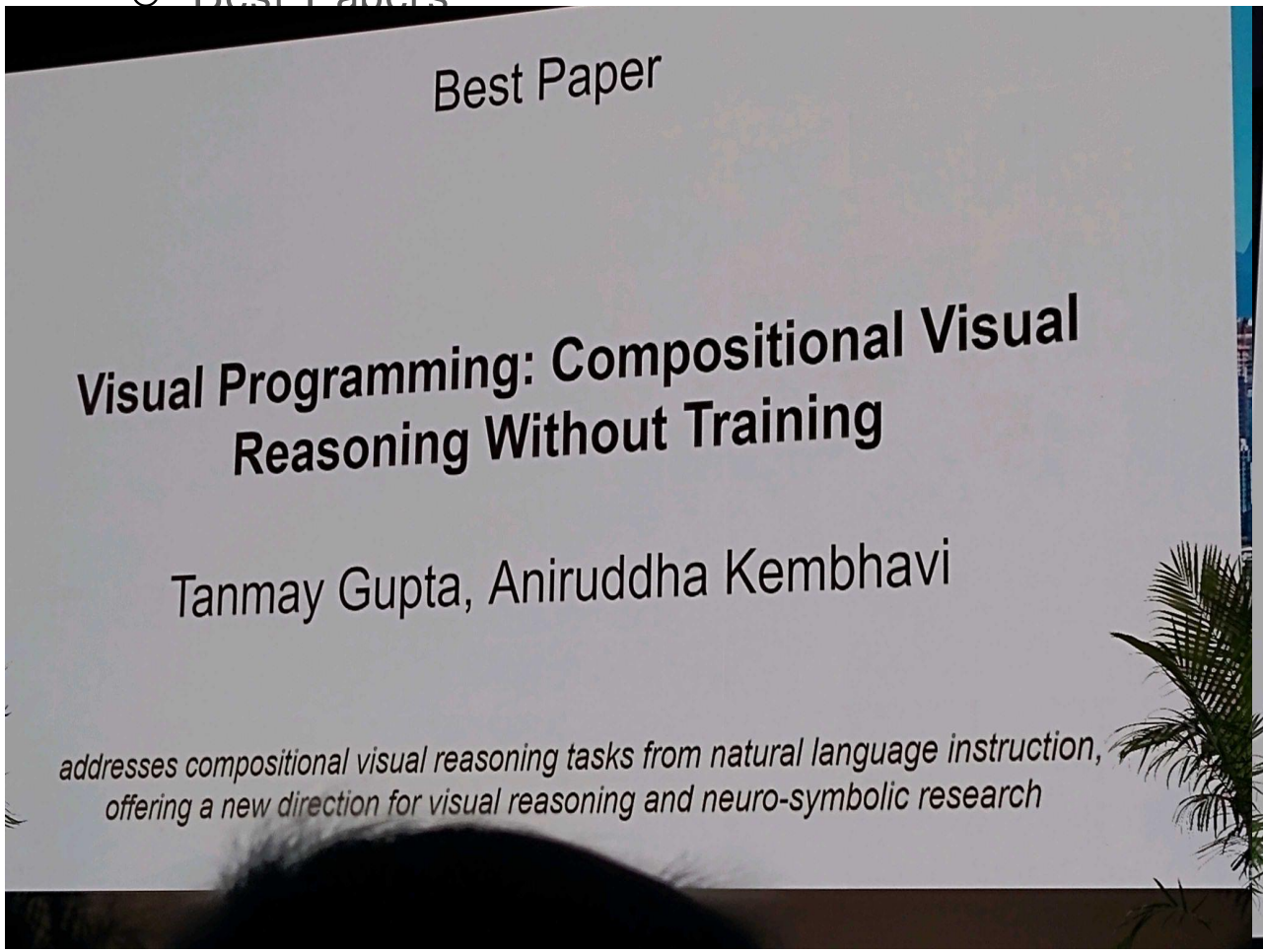
- Best Paper Honorable Mention

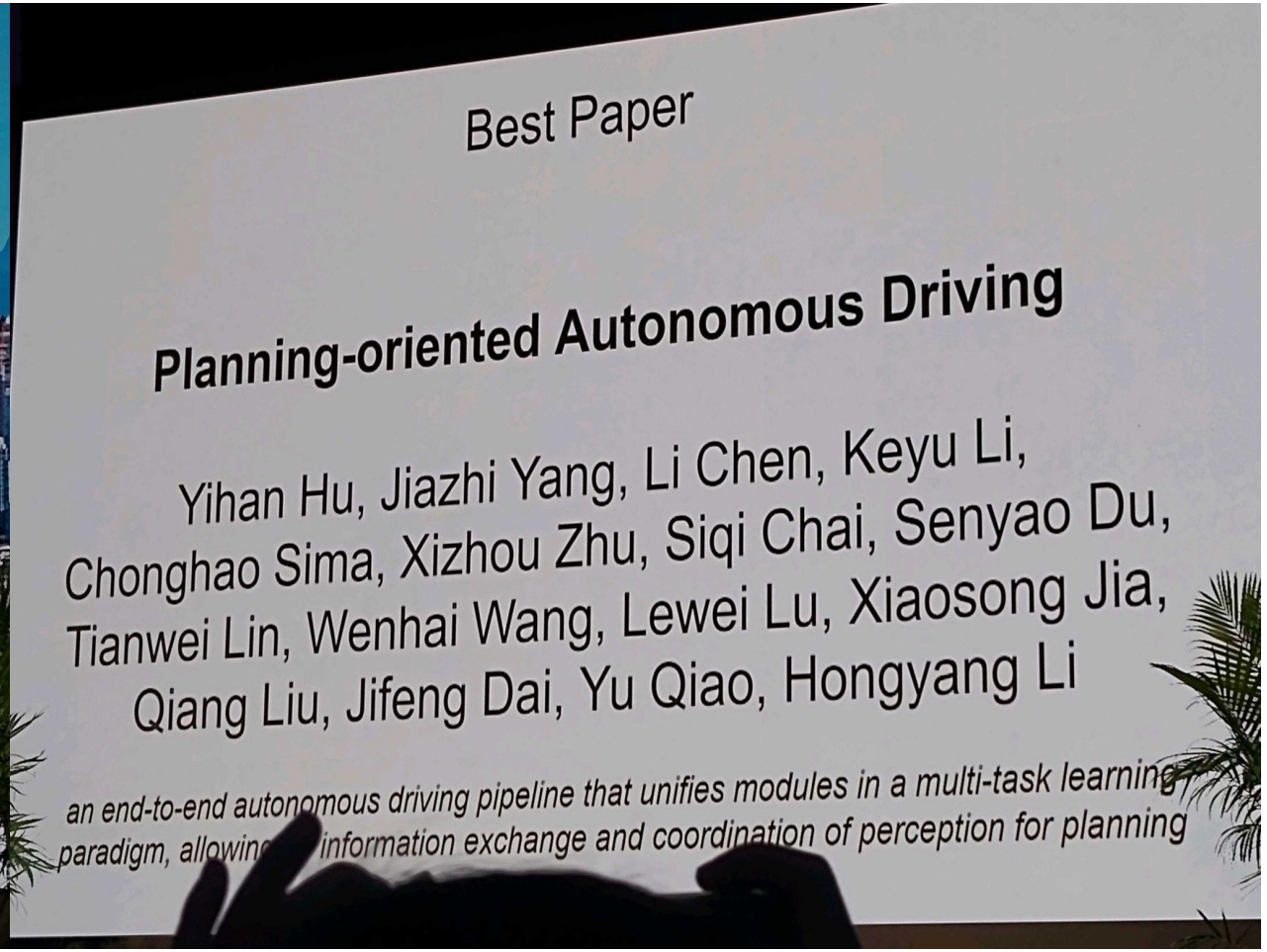


CVPR 2023 の動向・気付き(15/259)

Best Paper Awards

- Best Papers

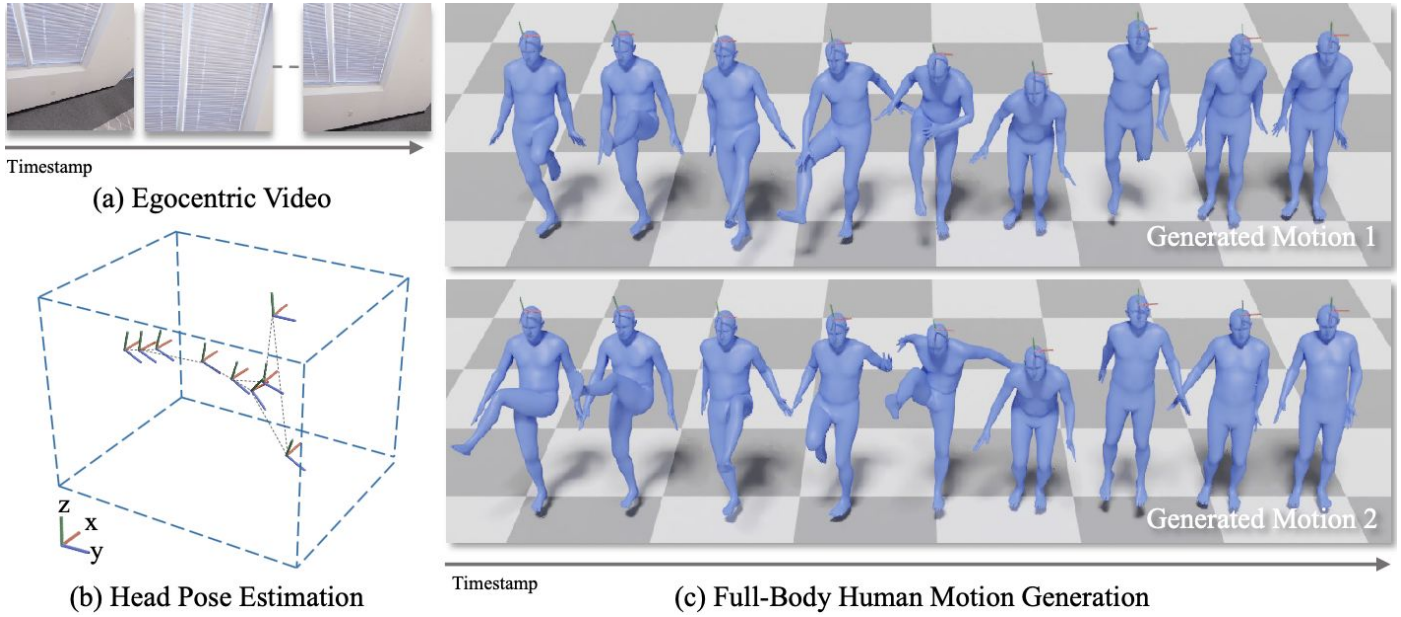




CVPR 2023 の動向・気付き(16/259)

Best Paper Finalist 1: Ego-Body Pose Estimation via Ego-Head Pose Estimation, Jiaman Li · Karen Liu · Jiajun Wu

- 一人称視点から(の他人ではなくカメラ装着者自身！)の姿勢推定
- Stage 1: 頭部カメラから頭部の角度・姿勢推定
- Stage 2: 頭部モーションから人物姿勢を回帰
- Cross-domainのデータセット構築(AMASS-Replica-Ego-Syn)

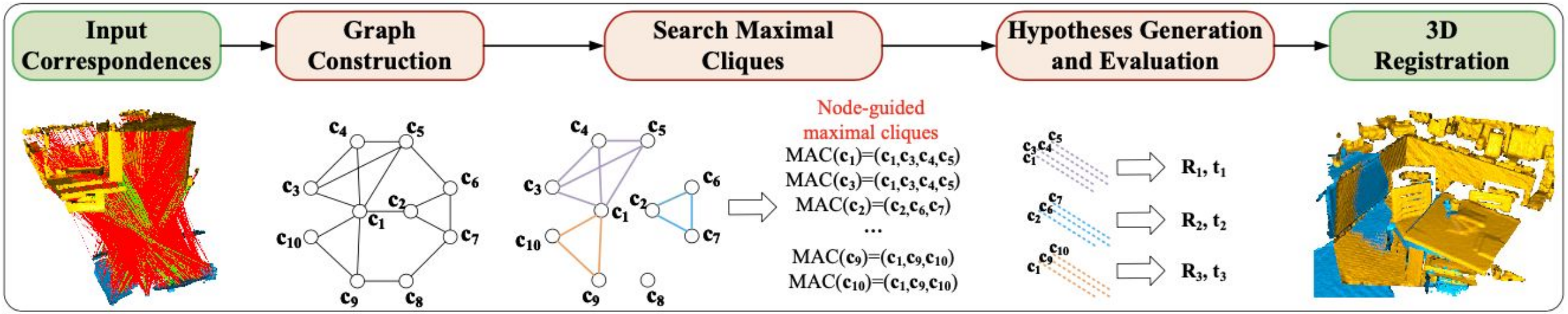


装着者のプライバシを保護しつつ，装着者の3D姿勢を推定する研究

CVPR 2023 の動向・気付き(17/259)

Best Paper Finalist 2(Best Student Paper Award): 3D Registration with Maximal Cliques, Xiyu Zhang · Jiaqi Yang · Shikun Zhang · Yanning Zhang

- 特徴/画像空間ではなく，グラフ空間で(RANSAC的な)仮説検証を実施
  - 従来，6DoF姿勢推定はノイズに脆弱・計算量膨大という問題
  - グラフのclique(環状形状)を緩和することで6DoF姿勢推定表現
  - グラフ空間の方が高精度であることを確認，まだまだ局所特徴も用いることができる，と実証した意味では価値のある論文



CVPR 2023 の動向・気付き(18/259)

Best Paper Finalist 3: OmniObject3D: Large Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation, Tong Wu · Jiarui Zhang · Xiao Fu · Yuxin WANG · Jiawei Ren · Liang Pan · Wenyan Wu · Lei Yang · Jiaqi Wang · Chen Qian · Dahua Lin · Ziwei Liu

- 汎用3DデータセットOmniObject3Dを提案
  - 3D認識・新規視点画像生成・表面形状推定・3Dモデル生成に対応
  - 6,000モデル，200クラス前後の物体を収録

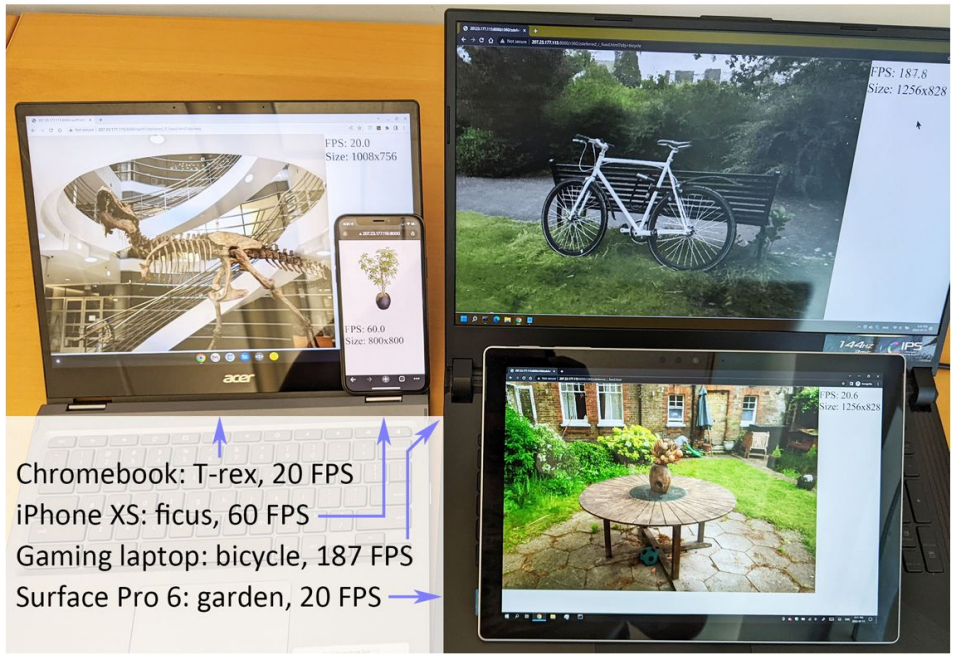
【従来データセットとの比較】

- 合成データ: 低品質，写実的では無い
- CO3D: 3Dの正解が無い
- 3Dスキャン: ほぼ日常物体
- 提案手法: 画像・動画・3Dモデルとそのラベルがある

CVPR 2023 の動向・気付き(19/259)

Best Paper Finalist 4: MobileNeRF: Exploiting the Polygon Rasterization Pipeline for Efficient Neural Field Rendering on Mobile Architectures, Zhiqin Chen・Thomas Funkhouser・Peter Hedman・Andrea Tagliasacchi

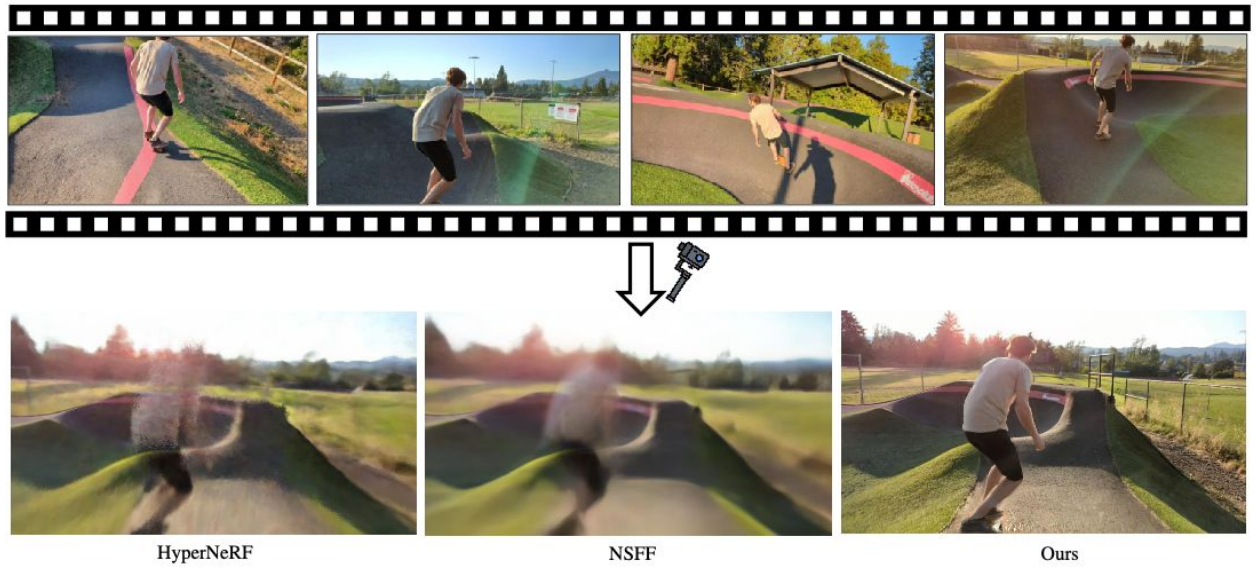
- NeRF をスマートフォンで動かす
  - NeRF表現としてのメッシュ(あとはテクスチャを割り振るのみ)
  - MLP fragment shader optimizations
  - NeRFシーンも編集可能である



CVPR 2023 の動向・気付き(20/259)

Best Paper Finalist 5(Best Paper Honorable Mention Award): DynIBaR: Neural Dynamic Image-Based Rendering, Zhengqi Li · Qianqian Wang · Forrester Cole · Richard Tucker · Noah Snavely

- NeRFの中でも難しいとされているDynamic View Synthesis
  - 物体・背景が常時動いている動画像においてNeRFを実施
  - 従来手法は顔のみが動くなど比較的動作が少ない状況のRadiance Fieldsを復元
  - 提案手法はImage-based Rendering(IBR)をNeRFに導入(IBRNetの提案)

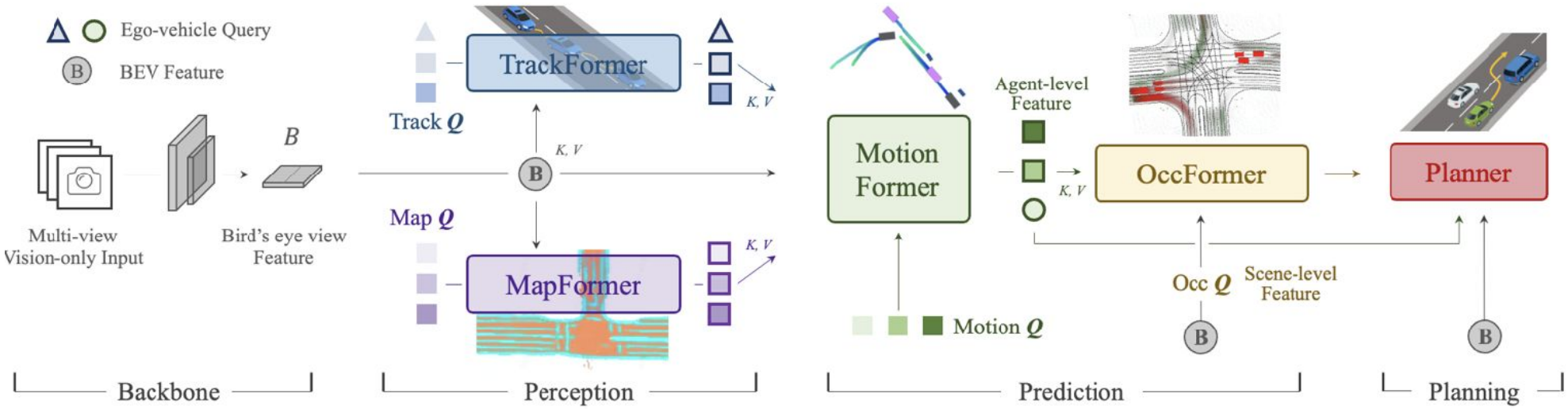


成功

CVPR 2023 の動向・気付き(21/259)

Best Paper Finalist 6(Best Paper Award): Planning-oriented Autonomous Driving, Yihan Hu・Jiazhi Yang・Li Chen・Keyu Li・Chonghao Sima・Xizhou Zhu・Siqi Chai・Senyao Du・Tianwei Lin・Wenhai Wang・Lewei Lu・Xiaosong Jia・Qiang Liu・Jifeng Dai・Yu Qiao・Hongyang Li

- 環境認識と走行計画を統合的に実施
  - Track, Map, Motion, Occupancy, Plannerをひとつのネットワークに収めたUnified Autonomous Driving (UniAD)



CVPR 2023 の動向・気付き(22/259)

Best Paper Finalist 7(Best Student Paper Honorable Mention Award): DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Nataniel Ruiz・Yuanzhen Li・Varun Jampani・Yael Pritch・Michael Rubinstein・Kfir Aberman

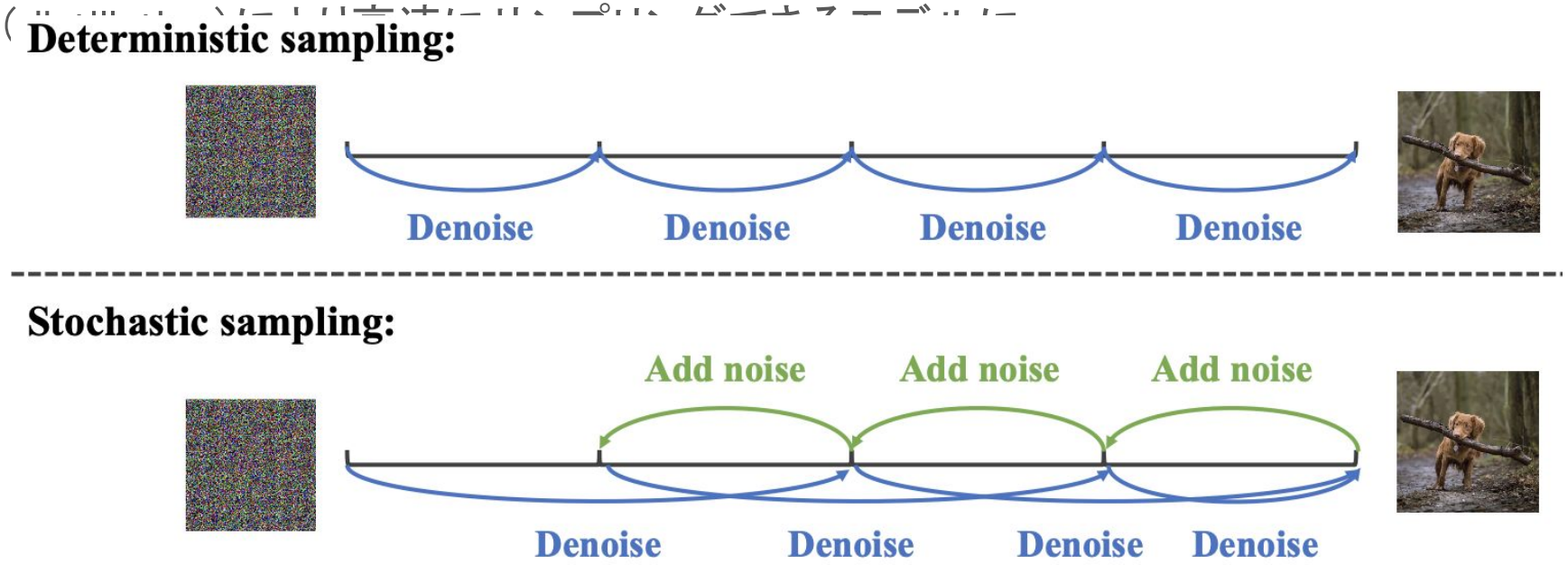
- 数枚の画像追加で生成モデルを個別化
  - Few-shotのfine-tuningで生成モデルを調整
  - 手持ち数枚の写真でモデルをカスタマイズ
  - Reconstruction Loss, Class-specific Prior Preservation Loss

CVPR 2023 の動向・気付き(23/259)

Best Paper Finalist 8: On Distillation of Guided Diffusion Models

Chenlin Meng · Robin Rombach · Ruiqi Gao · Diederik Kingma · Stefano Ermon · Jonathan Ho · Tim Salimans

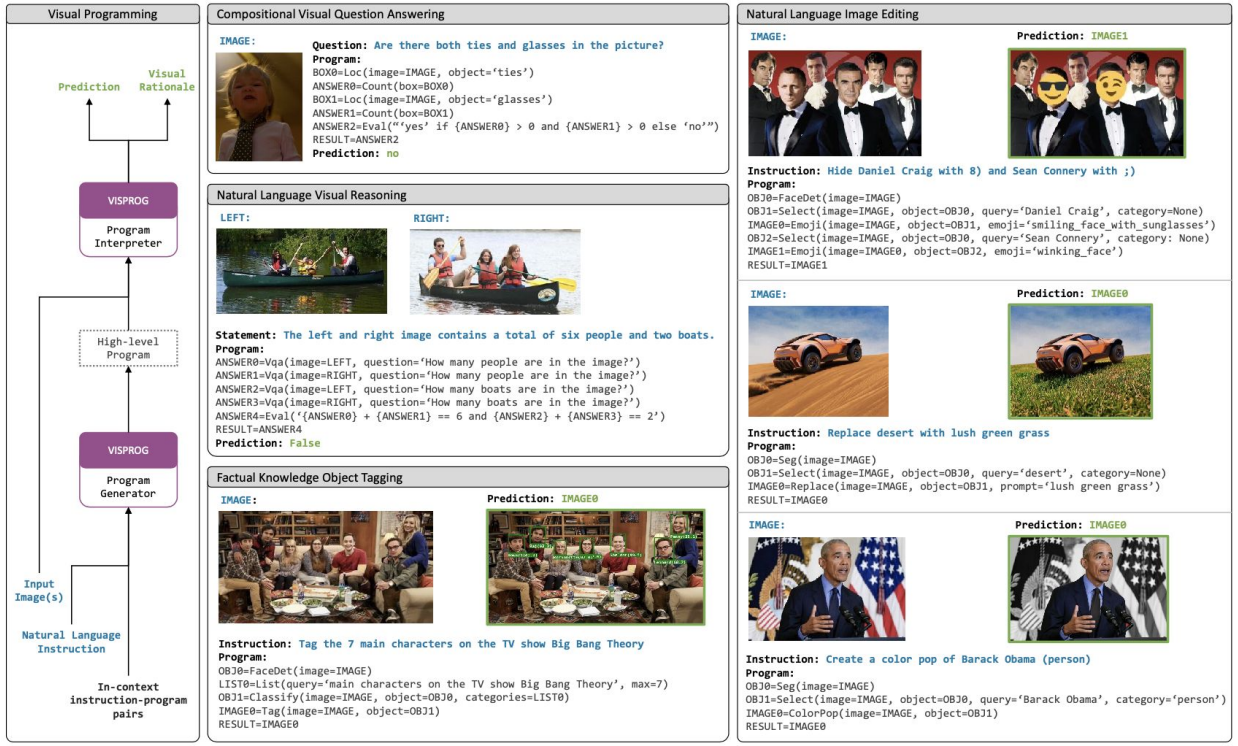
- Diffusion Modelの高速なサンプリング手法を提案
  - Diffusionでは何回もサンプリングを行う必要がある
  - 蒸留(



CVPR 2023 の動向・気付き(24/259)

Best Paper Finalist 9(Best Paper Award): Visual Programming: Compositional visual reasoning without training, Tanmay Gupta・Aniruddha Kembhavi

- CVの新規タスクを言語モデルからのプログラミングで説明・実行・解決
  - 人間はPromptingによりCVタスクを指示
  - GPT-3がプログラミング，その場で実行して結果を出力

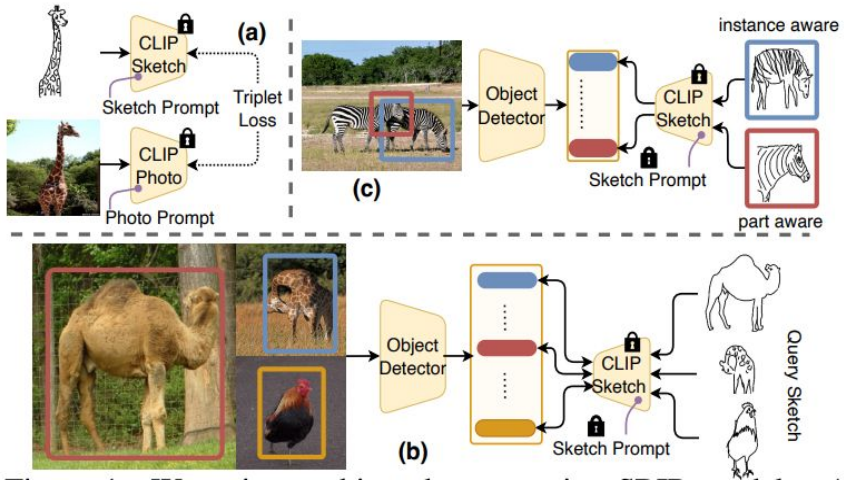


Create a color pop of the woman in blue and the woman in red and blur the background

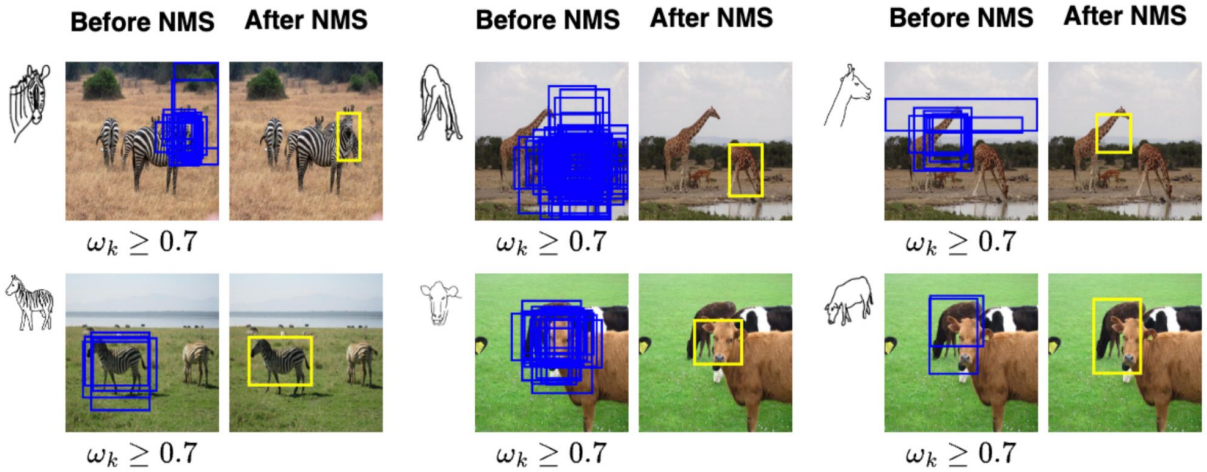
CVPR 2023 の動向・気付き(25/259)

Best Paper Finalist 10: What Can Human Sketches Do for Object Detection?, Pinaki Nath Chowdhury · Ayan Kumar Bhunia · Aneeshan Sain · Subhadeep Koley · Tao Xiang · Yi-Zhe Song

- 手書きスケッチをプロンプトとして物体検出を実現
  - 2D Bboxやカテゴリラベルは必要とせず，教師ラベルはスケッチベースの画像検索をのみ
    - fine-grained object detection: スケッチによる詳細な表現から注目領域を指定
    - category-level object detection: インスタンスレベルの物体検出可能



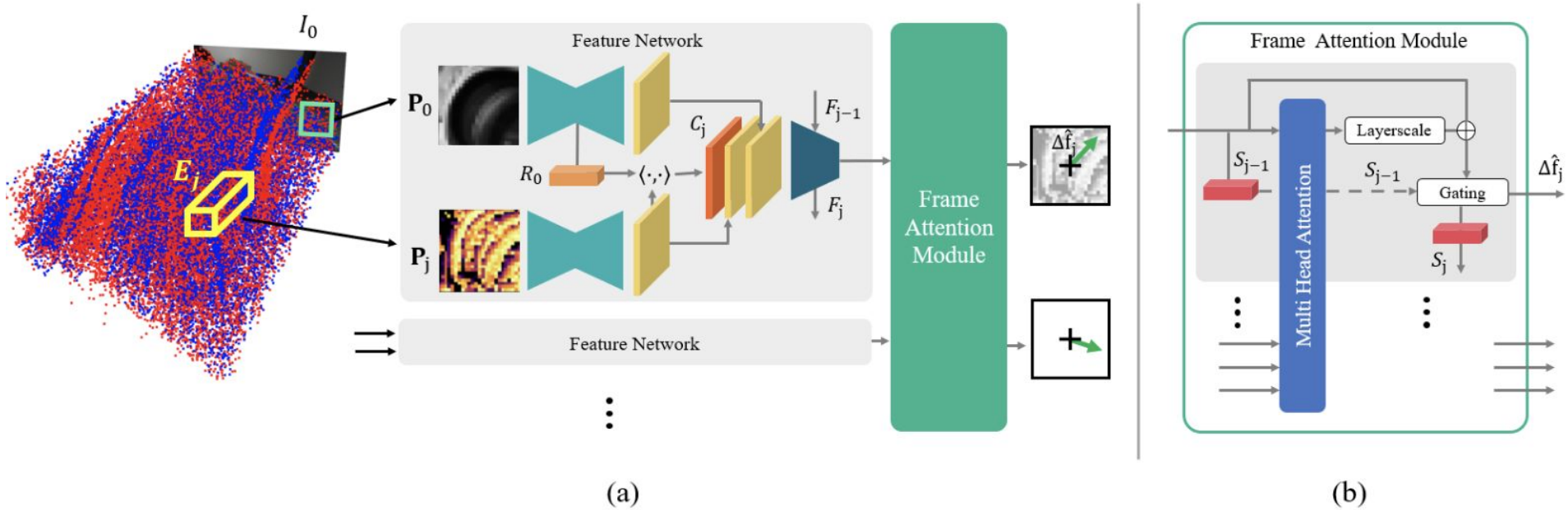
ッチ



CVPR 2023 の動向・気付き(26/259)

Best Paper Finalist 11: Data-driven Feature Tracking for Event Cameras, Nico Messikommer · Carter Fang · Mathias Gehrig · Davide Scaramuzza

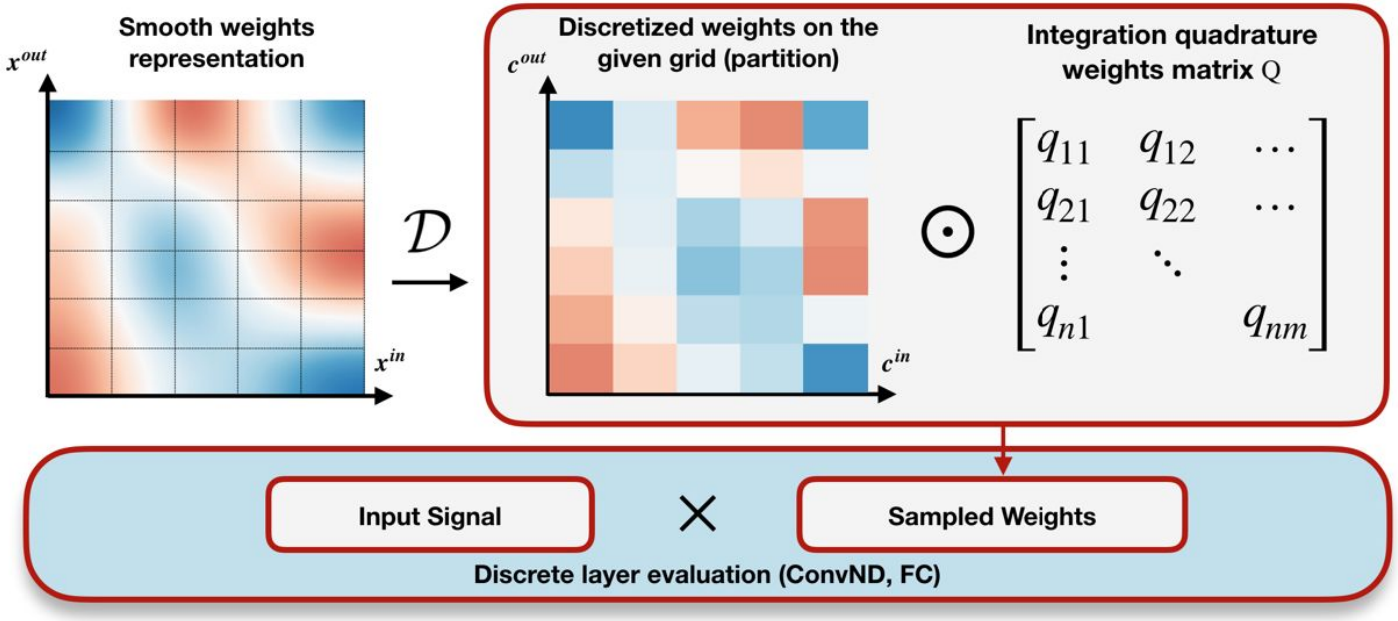
- イベントカメラを用いた物体追跡の研究
  - イベントカメラもデータから学習して追跡する時代が来た
  - ブラーや手ブレなど激しく難しい複合的な動きも追跡する



CVPR 2023 の動向・気付き(27/259)

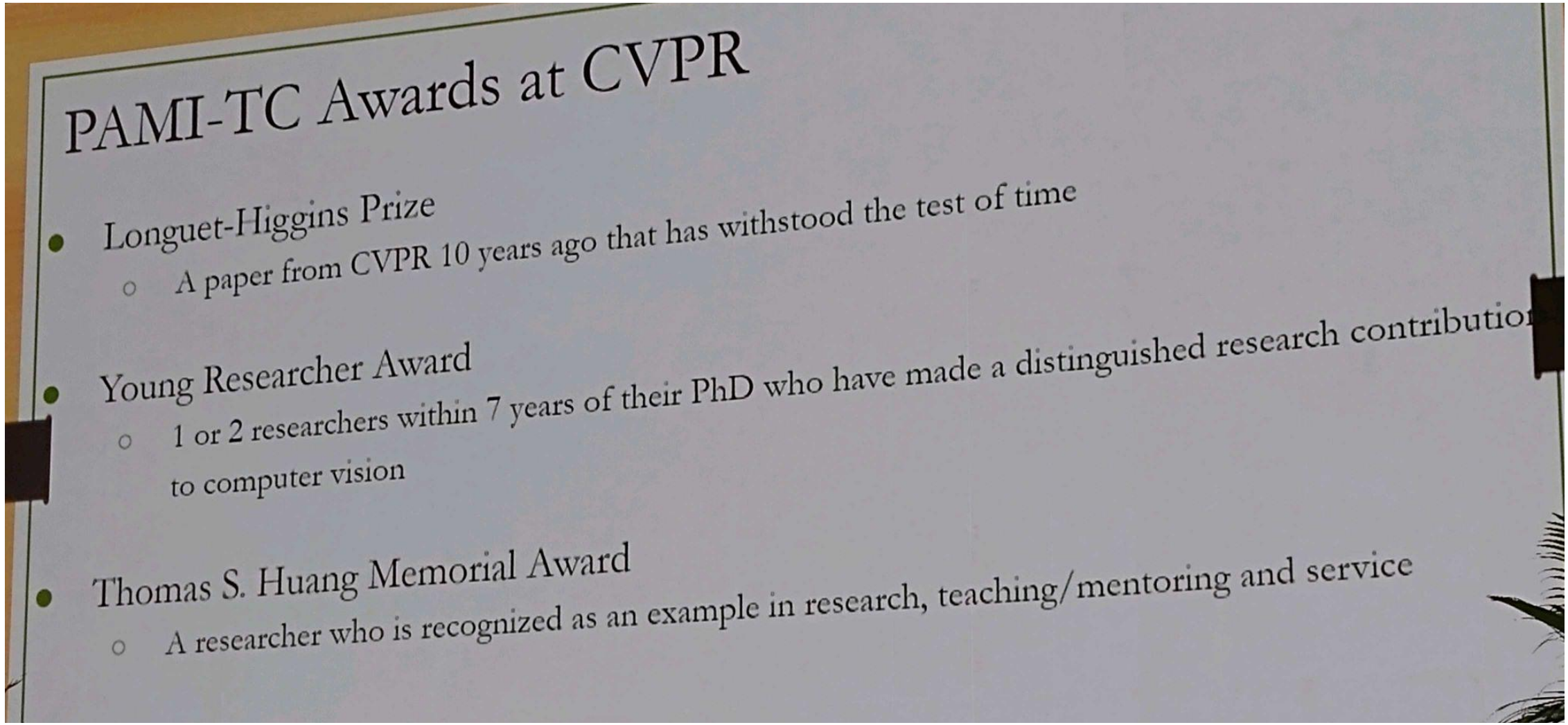
Best Paper Finalist 12: Integral Neural Networks, Kirill Solodskikh・Azim Kurbanov・Ruslan Aydarkhanov・Irina Zhelavskaya・Yury Parfenov・Dehua Song・Stamatios Lefkimmiatis

- Integral Neural Networks (INN)の提案論文
  - 層重みをテンソルではなく，フィルターとチャンネルの次元に沿った連続的な層表現を使用
  - INNはN次元の超立方体上に定義された連続関数として表現することで，様々な離散化間隔を持つ任意のサイズのネットワークの離散化を可能に
  - エッ



CVPR 2023 の動向・気付き(28/259)

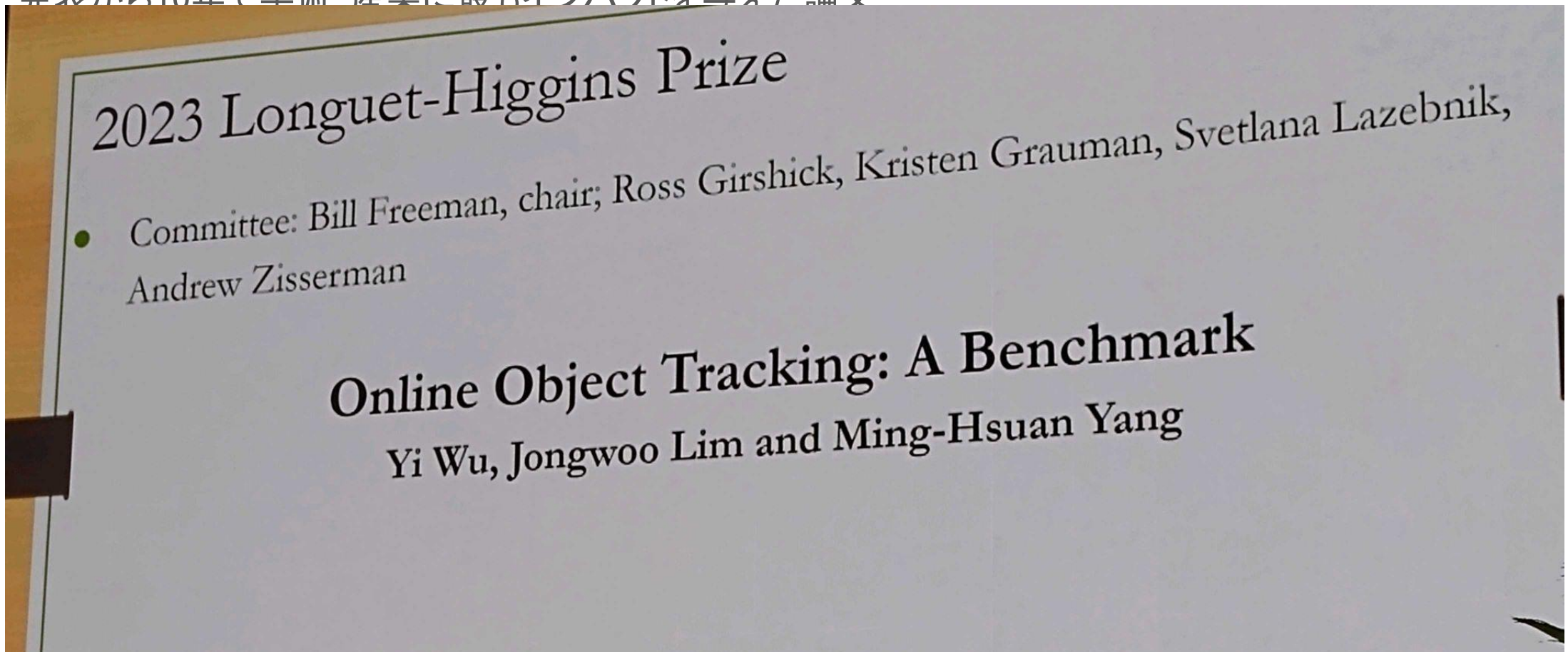
PAMI-TC Awards



CVPR 2023 の動向・気付き(29/259)

PAMI-TC Awards

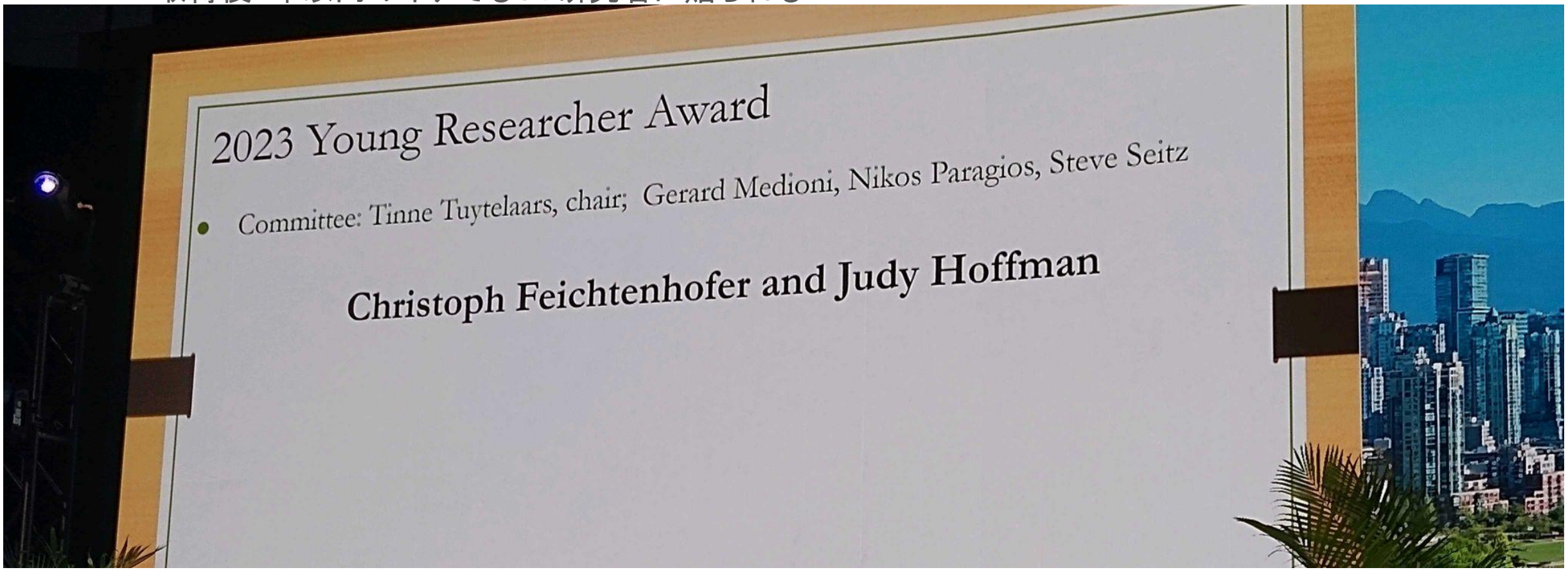
- Longuet-Higgins Prize
- 発表から10年で学術・産業に最もインパクトを与えた論文



CVPR 2023 の動向・気付き(30/259)

PAMI-TC Awards

- Young Researcher Award
- Ph.D.取得後7年以内のイケてるCV研究者に贈られる



CVPR 2023 の動向・気付き(31/259)

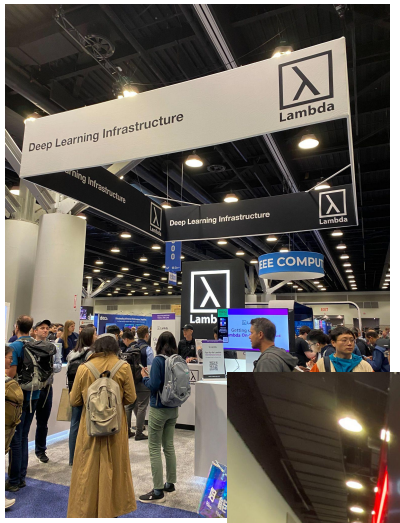
PAMI-TC Awards

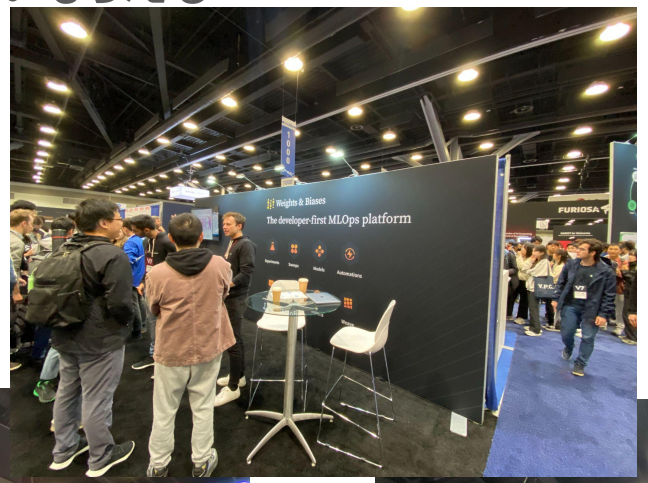
- Thomas S. Huang Memorial Award
- 学生・研究者のメンタリングへの功労賞

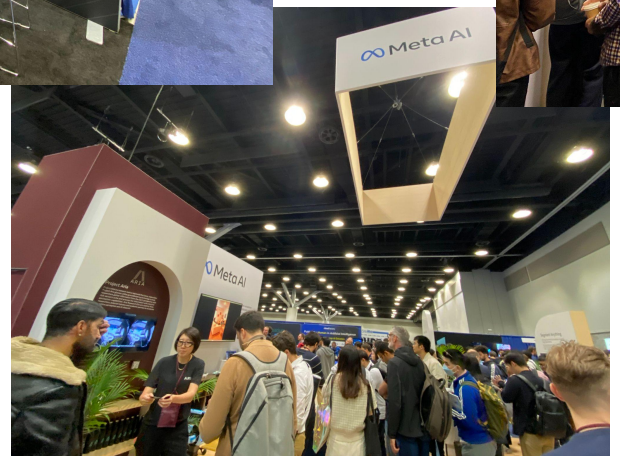
CVPR 2023 の動向・気付き(32/259)

Exhibits

- 企業によるdemoが多く集まっていた
  - Apple, Meta, Tesla, Lambda, Amazon, TikTok, Weghits&Biaesなど
  - 企業によっては景品がもらえる







CVPR 2023 の動向・気付き(33/259)

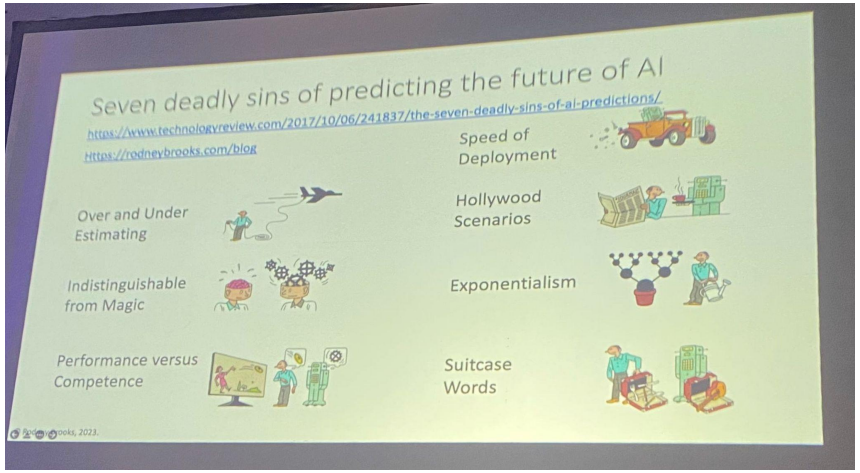
Plenary keynote (本会議初日)

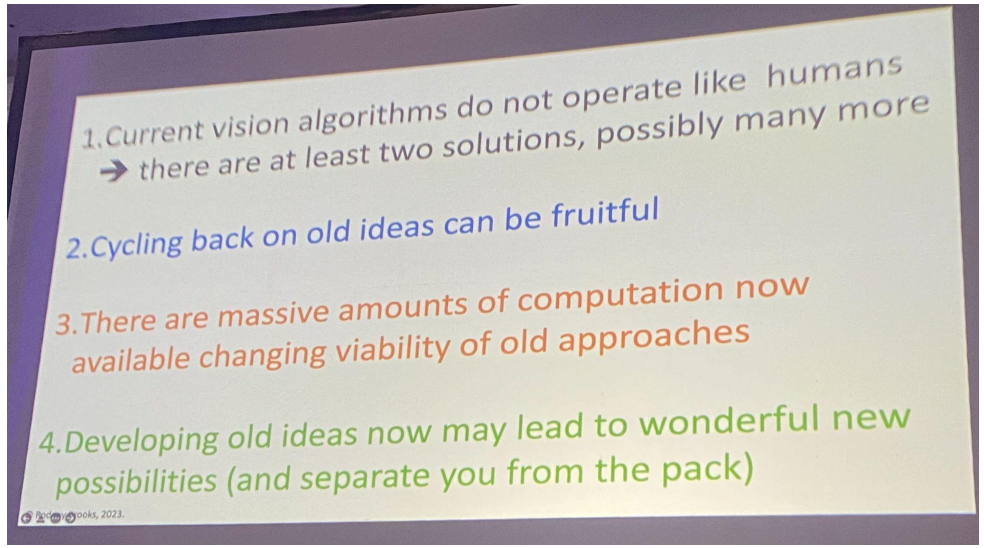
- Rodney Brooks氏(MIT)による講演

タイトル:“Revisiting Old Ideas With Modern Hardware”

概要:ハードウェアの進化によって, 昔は軽視されていたアイデアを現代で実現することができるか？

- ニューラルネットの誕生から始まる画像認識の歴史を1から説明していた
  - ネオコグニトロンの福島先生やIJCVやCVPRの創始者である金出先生など日本人の名前も出てきてすごいと思った
  - CVの進化とハードウェアの進化の両側面について





CVPR 2023 の動向・気付き(34/259)

Plenary keynote (本会議2日目)

- Yejin Choi氏(University of Washington)による講演

タイトル:“2050: An AI Odyssey: Dark Matter of Intelligence”

概要:

- Chapter1: The Possible Impossibilities?
  - ChatGPTやGPT4などのLLMでも5×5ケタなどの簡単な掛け算ができない
  - 掛け算以外の“構成性”タスクについても評価できていないだけでできないのでは？なぜもっと難しそうな“構成性”タスクができていて，簡単な掛け算もできない？
- Chapter2: The Impossible possibilities?
  - 超大規模事前学習+RLHF(人間フィードバックによる強化学習)が最強で勝てないのでは？
    →例えばTask-specific knowledge distillationならLLM(Large Language Model)よりも強くなる
- Chapter3: The Paradox
  - 言語における“ダークマター”とは？
    - 言語に影響があるような，この世界の暗黙のルール
  - Commonsense Paradox
  - Moravec’s Paradox
  - Generative AI Paradox

CVPR 2023 の動向・気付き(35/259)

Plenary talk day 2: Yejin Choi

- タイトル：An AI Odyssey: the Dark Matter of Intelligence
- 内容概要：
  - Possible impossibilities
    - LLM, Transformerはfull mastery 可能？
      → compositional problemsでもstruggling
      → graph collapse
  - Impossible possibilities
    → 部分的なデータセット(commonsense / visual reasoningなど)ではデータと手法で工夫したら，small LMs はLLMsに勝つ場合もある
  - Paradox
    - Commonsense paradox：dark matter in language: non-spoke rules of how world works.
    - Moravec's Paradox：could AGI be possible without robotics and embodiment?
    - Generative AI Paradox: easy is hard and hard is easy (recognition vs. generation, human vs. AI)

CVPR 2023 の動向・気付き(36/259)

Plenary talk day 2: Yejin Choi

- MentionされたLLMsのIssuesをaddress系の関連論文:
  - Faith and Fate: Limits of Transformers on Compositionality
  - Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing
  - The False Promise of Imitating Proprietary LLMs
  - Symbolic Knowledge Distillation: from General Language Models to Commonsense Models
  - PlaSma: Making Small Language Models Better Procedural Knowledge Models for (Counterfactual) Planning
  - Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker
  - Vera: A General-Purpose Plausibility Estimation Model for Commonsense Statements

(現状のLLMsの問題点と分析，今後の可能な方向性について知りたい方はぜひリストの中の論文を読みましょう)

CVPR 2023 の動向・気付き(37/259)

Stable Diffusion系の研究が様々な分野で活用されている

- 画像識別:
  - Fake it Till you make it:ImageNetをStable Diffusionモデルで学習データ生成・評価
- Open-Vocabulary Segmentation:
  - Open-Vocabulary Panoptic Segmentation:
  - Side Adapter Network(上と類似した非diffusion model)

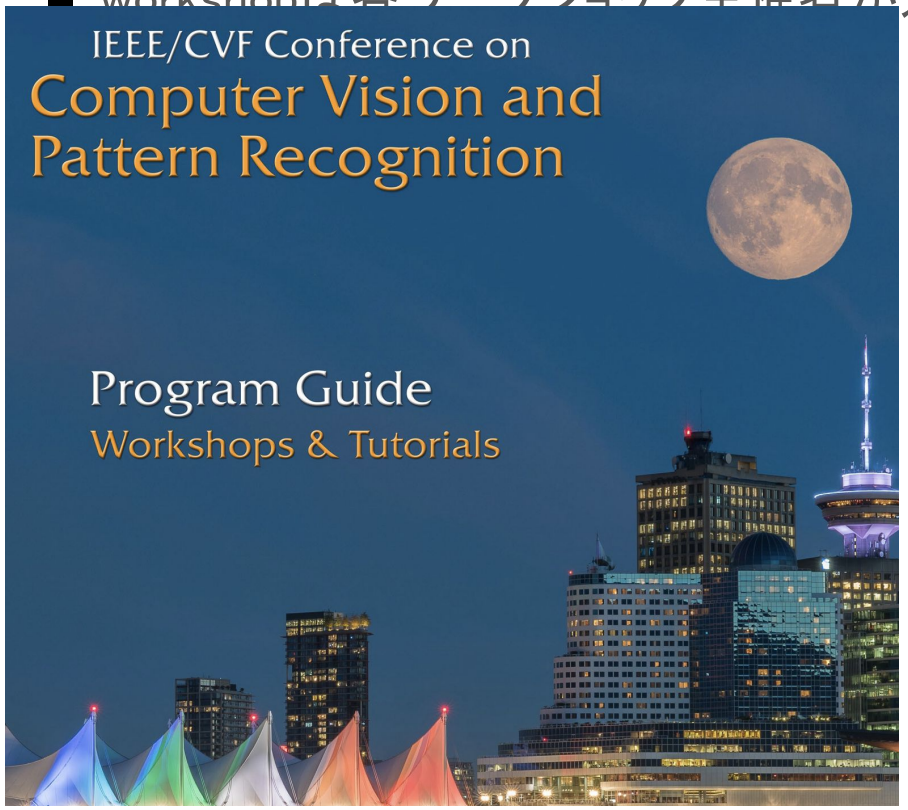
    (他には多数あり, 同時期に類似するモデルが頻出)

- 画像編集:
  - Imagic:テキストベースでハイレベル画像編集(例:立っている犬から座っている犬)
- 多種類の画像生成:
  - Unite and conquer: マルチモーダルを条件とした画像生成
- 学習データ生成:
  - HandsOff:少ないサンプルから学習データ生成

CVPR 2023 の動向・気付き(38/259)

CVPR2023 workshop

- 6/18~19までCVPR workshopが開催
  - workshopの数は100！！
  - workshopは各ワークショップ主催者が運営をしているので，開催時間もバラバラ



Sunday, June 18

NOTE: Workshop rooms are subject to change. Refer to the online site for up-to-date locations. Use the QR code for each workshop to see its schedule. Here is the QR code for the CVPR 2023 Workshops page.

0700–1700 Registration (West Ballroom Foyer)

0700–0900 Breakfast (West Ballrooms A–D)

1000–1045 Morning Break West Ballrooms A–D

1145–1345 Lunch (West Ballrooms A–D)

1500–1545 Afternoon Break (West Ballrooms A–D)

Fair, Data-Efficient, and Trusted Computer Vision

Organizers: Nalini Ratha, Srikrishna Karanam, Ziyan Wu, Mayank Vatsa, Richa Singh, Kuan-Chuan Peng, Michele Merler, Kush R. Varshney, Yiming Ying, Sharath Pankanti

Location: West 217-219

Time: Full Day (0800-1630)

Summary: The CVPR 2023 Workshop on Fair, Data-efficient, and Trusted Computer Vision aims to gather researchers and practitioners from academia and industry to discuss advances in all aspects of fairness, data-efficiency, and trust in computer vision. In addition to invited talks from experts in academia and industry, the workshop will solicit and provide a focused venue for new research ideas that seek to address problems related to topics above in a variety of application areas.

Autonomous Driving

Organizers: Vincent Casser, Alexander Liniger, Henrik Kretzschmar, Jose M. Alvarez, Fisher Yu, Yan Wang, Zhaoqi Leng, Maying Shen, Li Erran Li, Dragomir Anguelov, John Leonard, Luc Van Gool

Location: East Ballroom C

Time: Full Day (0915-1815)

End-to-End Autonomous Driving: Emerging Tasks and Challenges

Organizers: Hongyang Li, Kashyap Chitta, Holger Caesar, Shenlong Wang, Ziwei Liu, Tai Wang, Enze Xie, Huijie Wang, Yang Li

Location: West 110

Time: Full Day (0900-1800)

Summary: The area of autonomous driving has come to a rapid development to handle complicated scenarios and face the challenge of deploying algorithms to feasible massive production. With aid of various machine learning and computer vision techniques, many autonomous driving problems have been resolved. And yet certain key issues, such as safety and explainability for robust L4 solutions, end-to-end autonomous driving framework (and the benefits), bird's-eye-view perception, etc., have not been fully discussed. We have seen the successful holding of recent events at NeurIPS 2022 (incoming), CVPR 2022 (e.g., Embodied AI, Workshop on Autonomous Driving by Waymo/etc), and believe such a workshop is necessary for both the machine learning and computer vision community. This workshop, besides existing editions held at similar venues, serves a brand-new perspective to discuss broad areas of end-to-end framework design for autonomous driving on a system-level consideration. This workshop aims to bring together leading researchers and practitioners to discuss upcoming paradigms for autonomous vehicles. Central to the program is a series of invited talks and four new challenges in the self-driving domain. Each challenge combines new perspectives of multiple components in perception and planning compared to conventional pipelines. Winners of the challenges will present their results and insights as part of the workshop. We invite researchers around the world to build new algorithms to tackle these challenging, real-world autonomous driving tasks!

Generative Models for Computer Vision

Organizers: Adam Kortylewski, Fangneng Zhan, Lingjie Liu, Vincent Sitzmann, Alan Yuille, Christian Theobalt

Location: East Exhibit Hall B

Time: Full Day (0830-1715)

Summary: Recent advances in generative modeling leveraging generative adversarial networks, auto-regressive models, neural fields and diffusion models have enabled the synthesis of near photorealistic images, drastically increasing the visibility and popularity of generative modeling across the computer vision research community. However, these impressive advances in generative modeling have not yet found wide adoption in computer

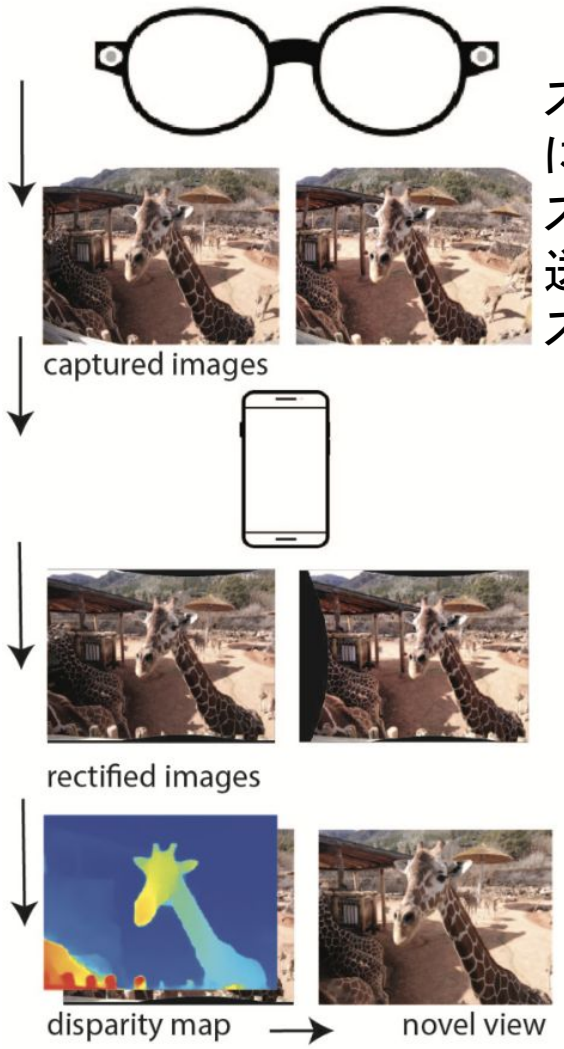
https://media.icml.cc/Conferences/CVPR2023/CVPR_2023_WorkshopsTutorials_ProgramGuide.pdf

Workshopに関するまとめを一部紹介

cvpaper.challenge

CVPR 2023 の動向・気付き(39/259)

Metaverse時代を見据えたデバイス融合型のCV技術も発展



スマートグラスからの画像をスマートフォンに送り視差画像や新規視点画像を生成．スマートグラスからはステレオ視の画像を送信，スマートフォンのCPUにて処理．システム構築も同時進行．

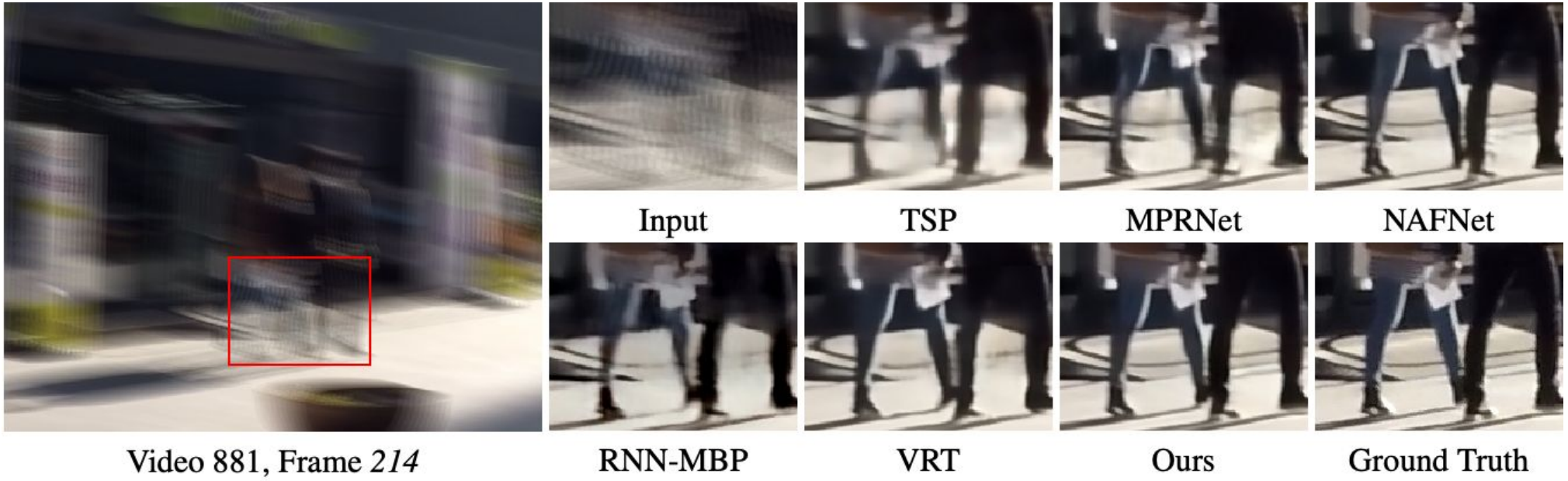
この手の研究だとエンジニアリング力が強く，実験データよりもシステムの完成度で押し切って通しているようにも見える

J. Wang et al. “A Practical Stereo Depth System for Smart Glasses”, in CVPR 2023.

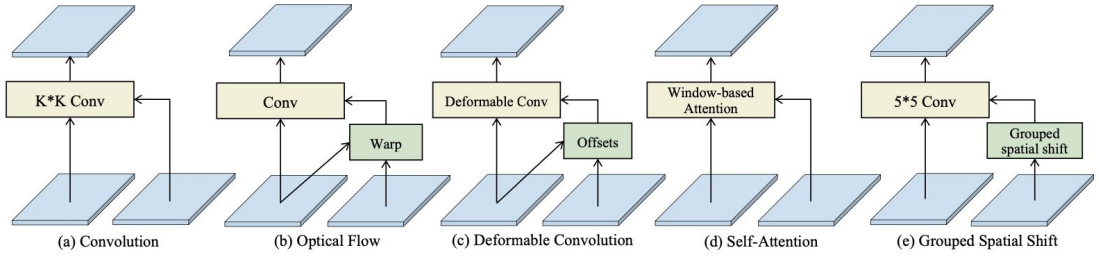
CVPR 2023 の動向・気付き(40/259)

動画の手ブレ補正

- 正確にはGoProにおいて頭部の動きによる激しいブラーから如何に正しく画像を復元する
- 原型を止めていないようにも見えるが，結構綺麗に復元



時空間シフト(spatio-temporal shift)に基づき復元，シンプル・処理が軽いながらも強力なベースラインとなった



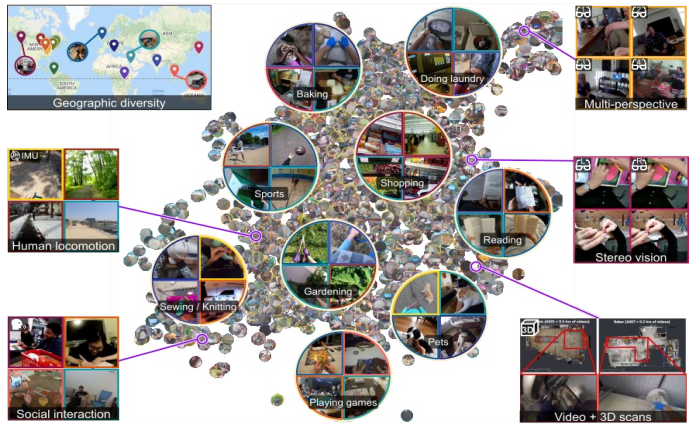
5種類のベースラインの違い：grouped spatial shiftが提案であり，時系列的に連続する2フレームを入力．

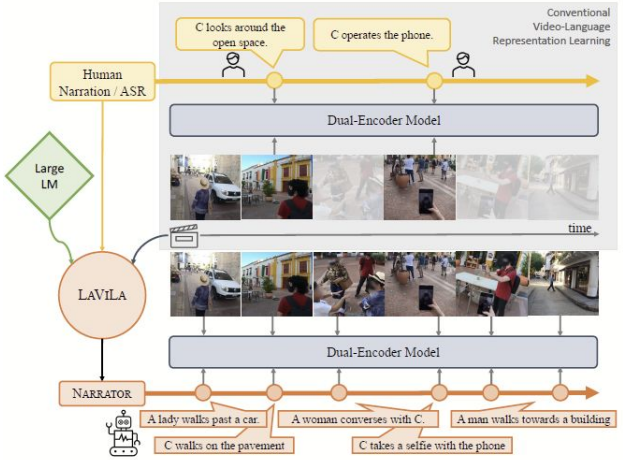
D. Li et al. “A Simple Baseline for Video Restoration with Grouped Spatial-temporal Shift”, in CVPR 2023.

CVPR 2023 の動向・気付き(41/259)

一人称ビジョンの静かな普及

- 3,600時間の一人称視点映像データセットEgo4Dを用いた研究が一人称ビジョン外の分野を含むタスクで急増（公開1年半で200引用超）
  - 使用先：Video QA, ロボット操作の事前学習，手物体インタラクション解析，Video Object Segmentationなど
- 普及の理由：(1) 長時間の未編集映像 (2) 真に自然な行動（実験室環境でない） (3) 映像に付随する密な行動ナレーション (4) 権利関係がクリア といった
  チャレンジングかつ真に生活に入り込んだデータであるため
- 一人称ビジョンにスポットライトがあたる日は来るか？



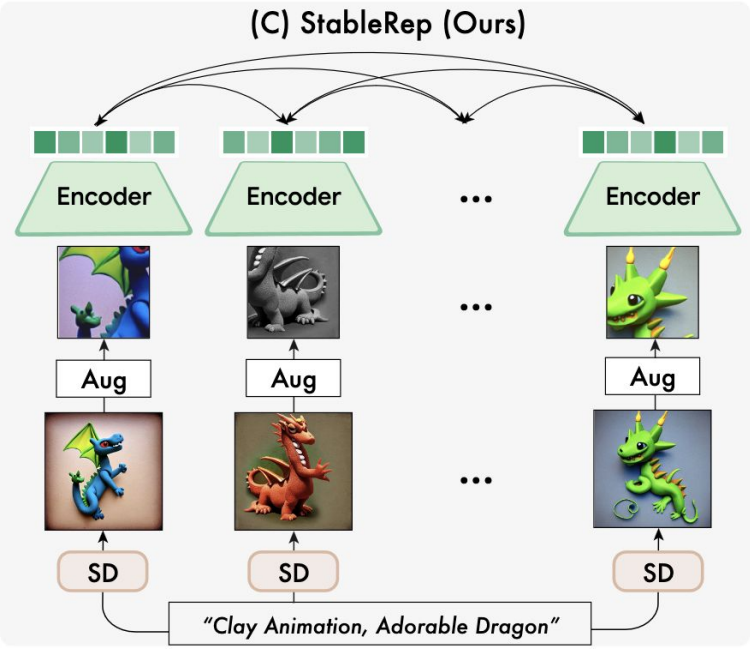


Tokmakov, P., Li, J., & Gaidon, A. (2023). Breaking the" Object" in Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22836-22845).
Yue Zhao, Ishan Misra, Philipp Krähenbühl, Rohit Girdhar, Learning Video Representations from Large Language Models, CVPR2023

CVPR 2023 の動向・気付き(42/259)

Workshop: Generative Models for Computer Vision

- Phillip Isola (MIT): Pix2Pix/CycleGANなどの著者
- Title: “Generative Models as Data++”
  - Generative Models + Model Learningのススメ
  - 生成モデルを用いることで「より多くのデータ利用」と「学習画像の操作性向上」という利点をどちらも取れる
  - StableRep (https://arxiv.org/abs/2306.00984): Stable Diffusion(text-to-image)とContrastive Learning 持徴表現学習.



基本方針: 同一のテキスト(prompt)から生成された画像によりContrastive Learning(対照学習).
SimCLR: 画像のみを用いて対照学習. 同一テキストから生成された異なる画像がいわばデータ拡張的に作用.
CLIP: テキスト・画像共に用いてVisual-Language Pre-trainingを実施.

Y. Tian et al. “StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners”, in arXiv:2306.00984, 2023.

CVPR 2023 の動向・気付き(43/259)

Workshop: Generative Models for Computer Vision

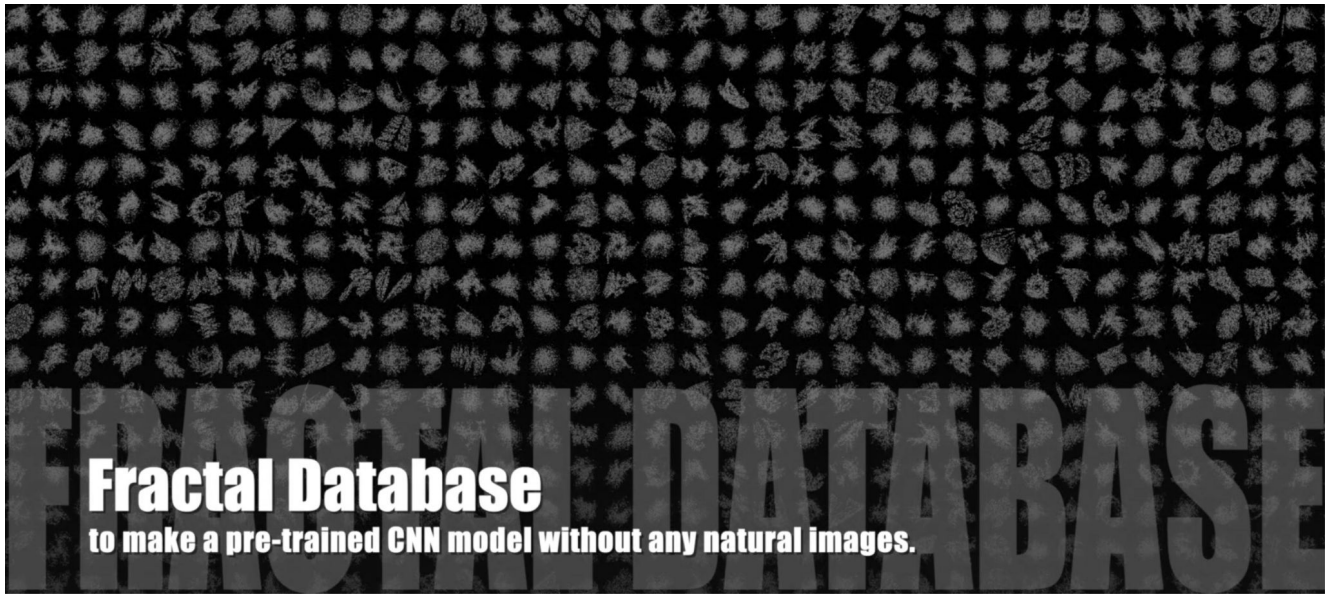
- Phillip Isola (MIT): Pix2Pix/CycleGANなどの著者
- Title: “Generative Models as Data++”
- 講演の最後のTake Home Message
  - Generative AI(含:Diffusion Model, NeRF, LLM)を用いて学習モデルをよくするという方向性が良いのではないか，と主張
  - 面白い研究をしている人は特定トピックにおいても，幅広い実験からの洞察により研究テーマを立てている．短絡的に前の論文の改善(受け売り)ではなくて自らの実験結果と分析からテーマを始める．
  - ↑を上手いこと人間とLLMのコラボレーションでスムーズに進めるフレームワークを考えたい(！？)
  - Foundation Models via Synthetic Pre-trainingができると良さそう．StalbeRepの発展がそうなっていくのか．ただし，まだ生成モデルの学習において権利面の問題は残される．

Y. Tian et al. “StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners”, in arXiv:2306.00984, 2023.

CVPR 2023 の動向・気付き(44/259)

Workshop: Neural Architecture Search & Lightweight NAS Challenge

- Exploring the Potential of Neural Dataset Search
  - Neural Dataset Search(NDS)の提案
  - NASが学習中にアーキテクチャを探索する概念なのに対して，NDSは学習中にデータセット(を生成するパラメータ)を探索して柔軟にデータが変化



数式ドリブン教師あり学習は数式からラベル付き画像を生成．パラメータによりカテゴリやインスタンスを生成するので，学習中にデータセット構築のパラメータを探索すると良いということらしい．

計算リソースさえ整えば「モデル」と「データ」両方が変化する学習ができそう

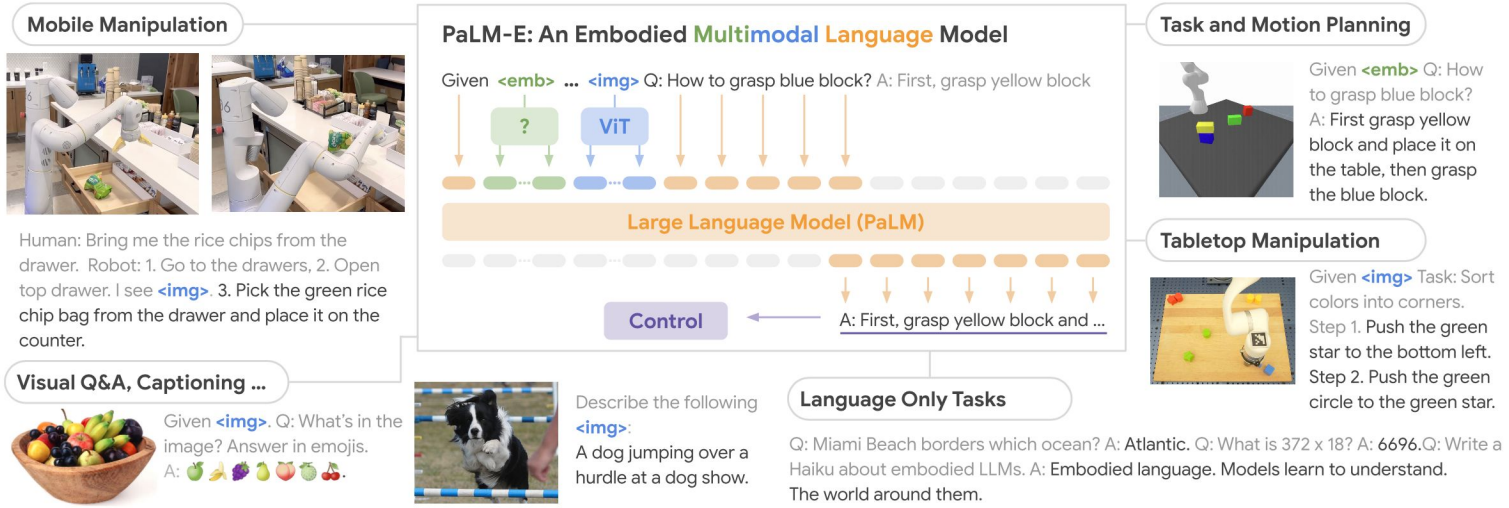
Source: https://hirokatsukataoka.net/temp/presen/230201MIT.pdf

R. Yamada, R. Shinoda, H. Kataoka, “Exploring the Potential of Neural Dataset Search”, in CVPR 2023 Workshop on Neural Architecture Search, 2023.

CVPR 2023 の動向・気付き(45/259)

Workshop: 3D Vision and Robotics

- Pete Florence(Google Deepmind)
  - Googleのアルゴリズム(NeRF / LLM; PaLM-E)をロボットに実装している
  - NeRFでは3D空間(x,y,zとその色, 透明度)の把握がリッチになる
  - LLMでは人間による自然言語の高度な把握と認識結果の拡張などを実現する
  - CVの3D空間把握・認識モデルがNLPの言語モデルにより翻訳, Roboticsのコントロール技術と繋げて実世界で動作する(前者: CV/NLP が賢くなると後者: Roboticsに活きる)



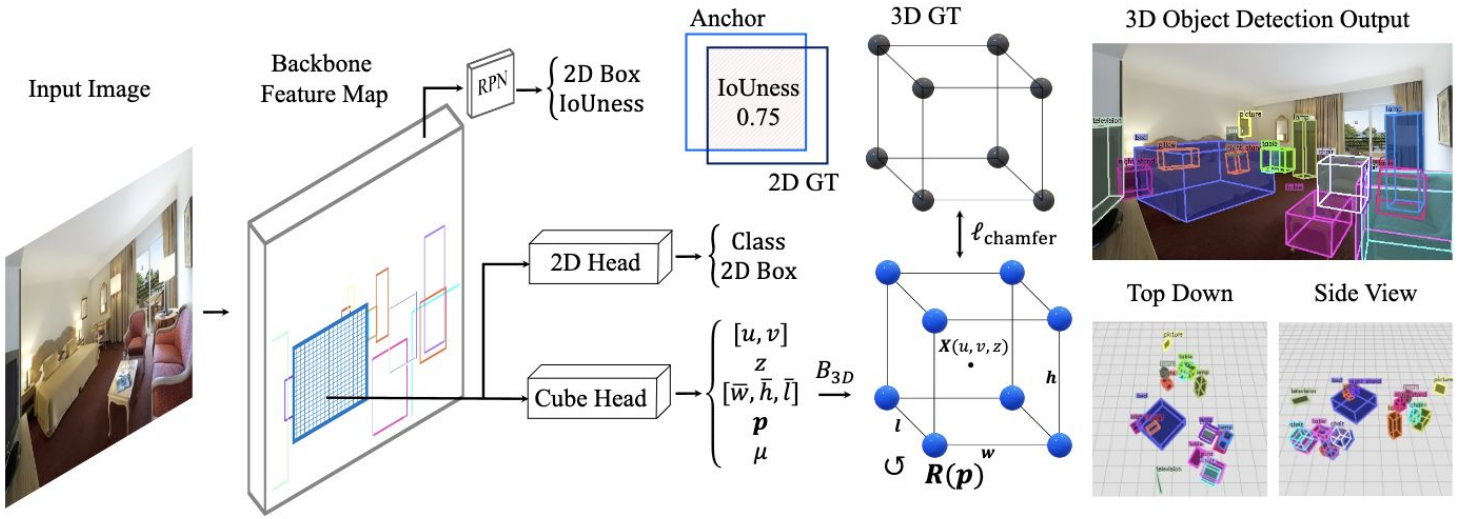
講演ではPaLM-Eの話も登場

Source: https://palm-e.github.io/

CVPR 2023 の動向・気付き(46/259)

Workshop: 3D Vision and Robotics

- Georgia Gkioxari(Caltech)
  - Mask R-CNN (ICCV 2017 Best Paper) / PyTorch3Dの著者として知られる
  - 今回は3D Prior(3次元空間の事前情報)をいかに2DのRGB画像入力と物体認識に使用するか，という講演
  - 主にCVPR 2023採択のOmni3D(データセット構築と各種アルゴリズムのベンチマーク評価)について説明



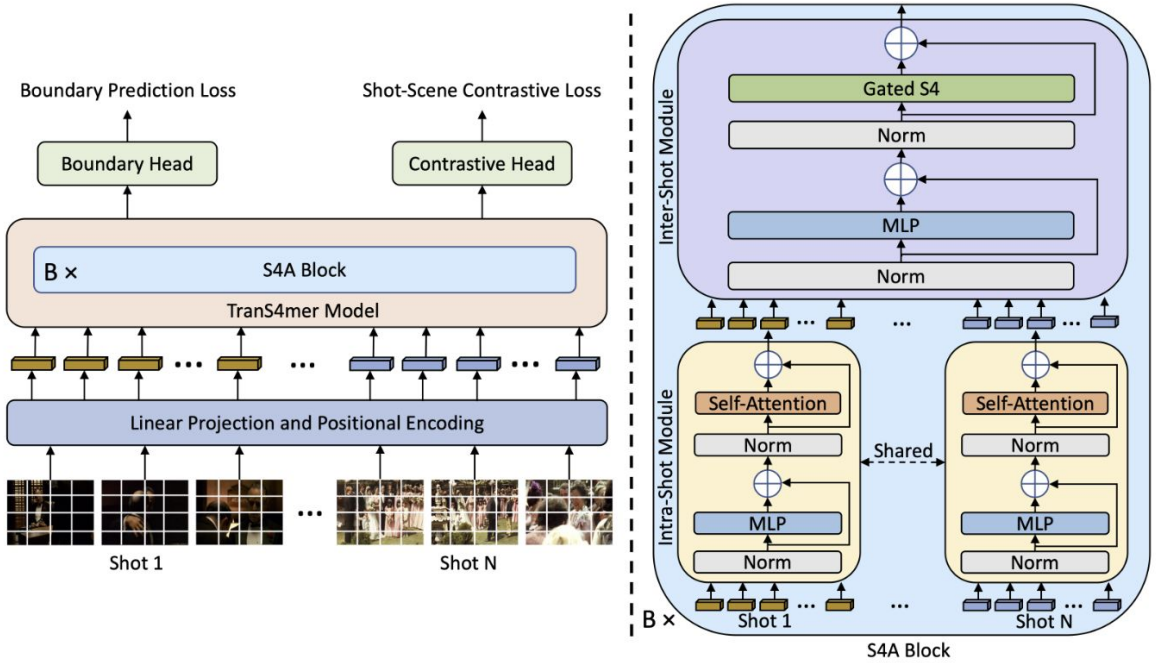
Omni3D Benchmark: 従来，3D向けのベンチマークは比較的小規模であった．今回は23.4万画像・300万3Dモデルを含むベンチマークを準備した．
Cube R-CNN(左図): 2DのRGB画像の入力から2D bbox, 3D cube, class labelを回帰・推定する．

G. Brazil et al. “Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild”, in CVPR 2023.
https://arxiv.org/abs/2207.10660

CVPR 2023 の動向・気付き(47/259)

Workshop: Long-form Video Understanding and Generation

- Gedas Bertasius(University of North Carolina, Chapel Hill)
  - 長時間にわたる動画の認識を提案
  - 従来，通常の動画認識モデルは数秒～10秒くらいだったが，提案法は数分レベルの認識でも効果を発揮 → 映画・ニュースなどの認識を想定し，各シーン3フレーム程度を入力して依存関係を捉えるモデル



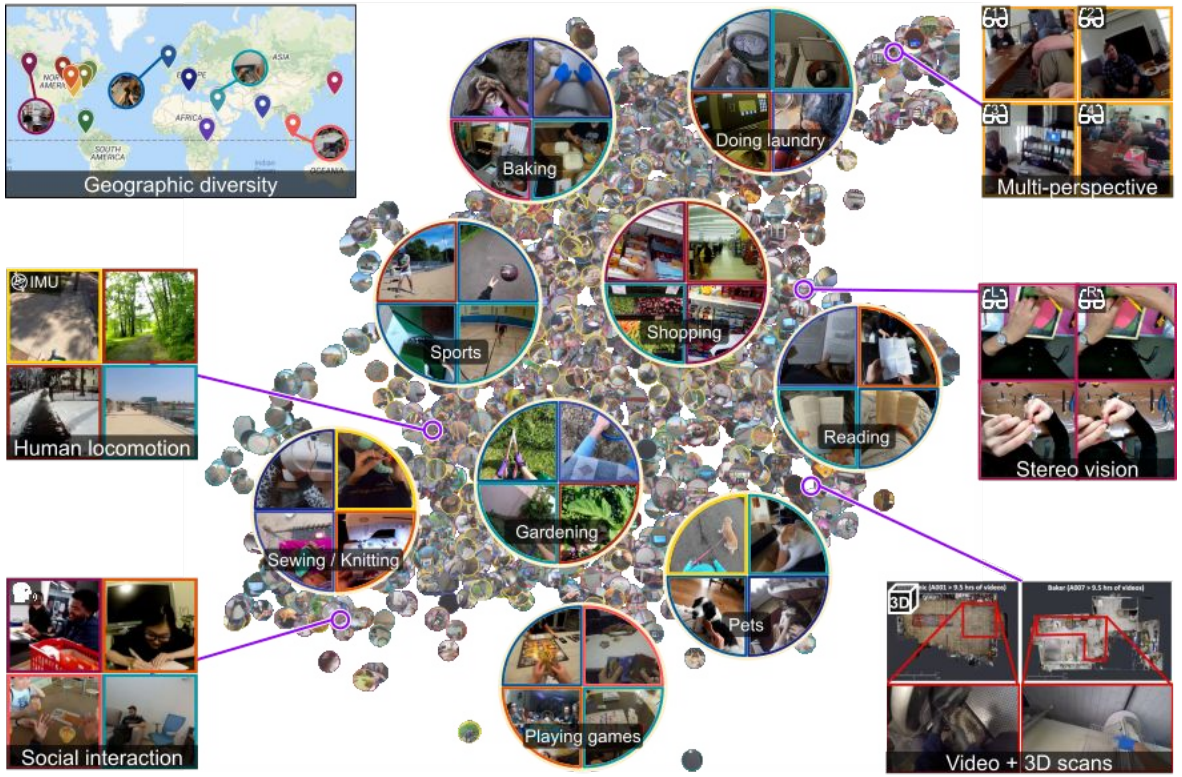
State-Space Transformers: 各シーン数フレームの入力から依存関係を捉え，長時間動画を認識するモデル.

M. M. Islam et al. “Efficient Movie Scene Detection using State-Space Transformers”, in CVPR 2023.

CVPR 2023 の動向・気付き(48/259)

Workshop: Visual Pre-training for Robotics

- Kristen Grauman(Univ. of Texas Austin / Meta AI)
  - Ego4Dのプロジェクトの説明



Ego4Dにて用いられるタスク(意図推定，物体認識など)はロボットが人間を理解してサービスを提供するためのあれこれに役立つと主張.

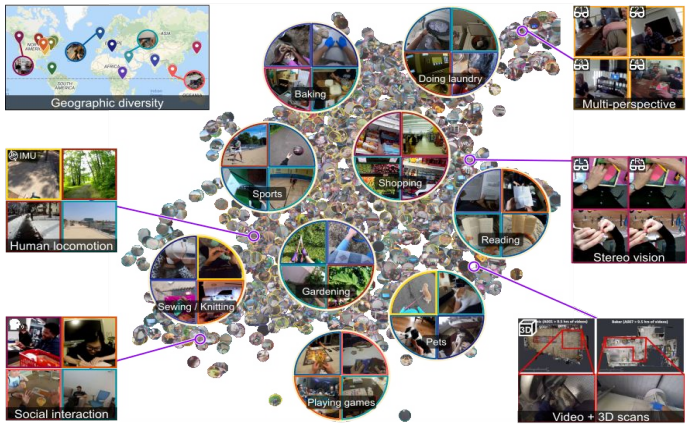
CVPR 2022でEgo4Dが提案(Best Paper Finalist)され，一年経った今回のCVPRでは一人称ビジョンの用途以外にも「ロボティクス向けの事前学習」向けであることが答え合わせされた.

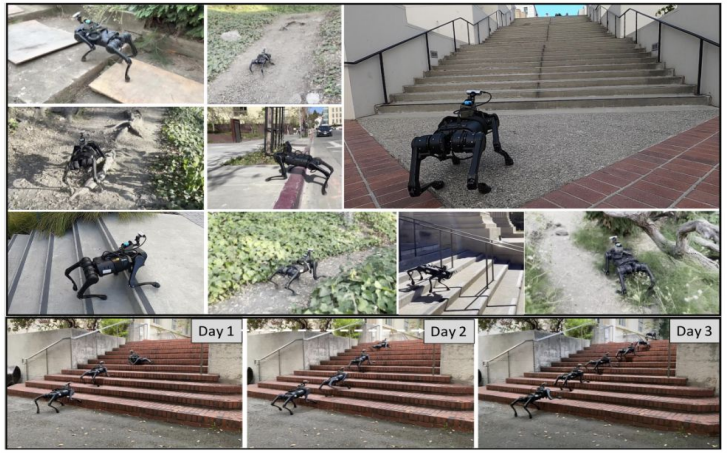
source: https://ego4d-data.org/

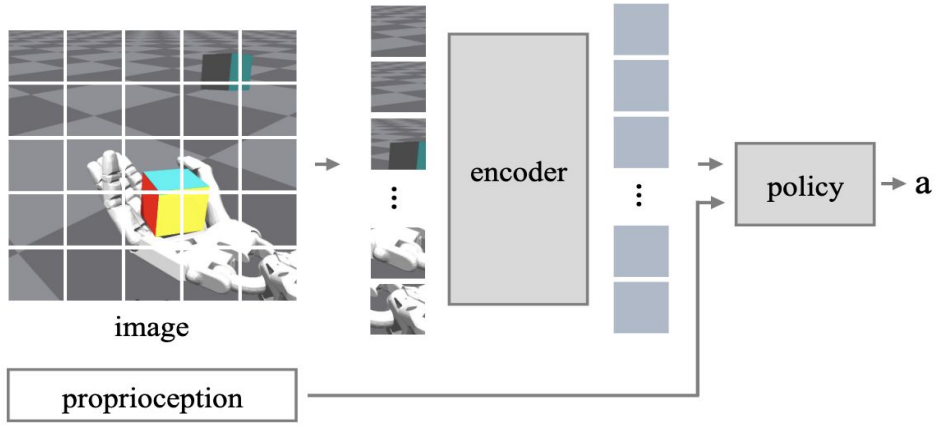
CVPR 2023 の動向・気付き(49/259)

Workshop: Visual Pre-training for Robotics

- Jitendra Malik(UC Berkeley / Meta AI)
  - 現在UC Berkeleyにて進行中の複数のプロジェクトを紹介
  - Ego4D:前ページに同じ，やはり大規模一人称ビデオはロボット事前学習に有効と強調
  - Locomotion:Visionのモダリティを追加すると4足歩行ロボットの事前学習促進
  - Masked AutoEncoder(MAE)をロボット操作に使用欠けた視点からポリシー・行動学習







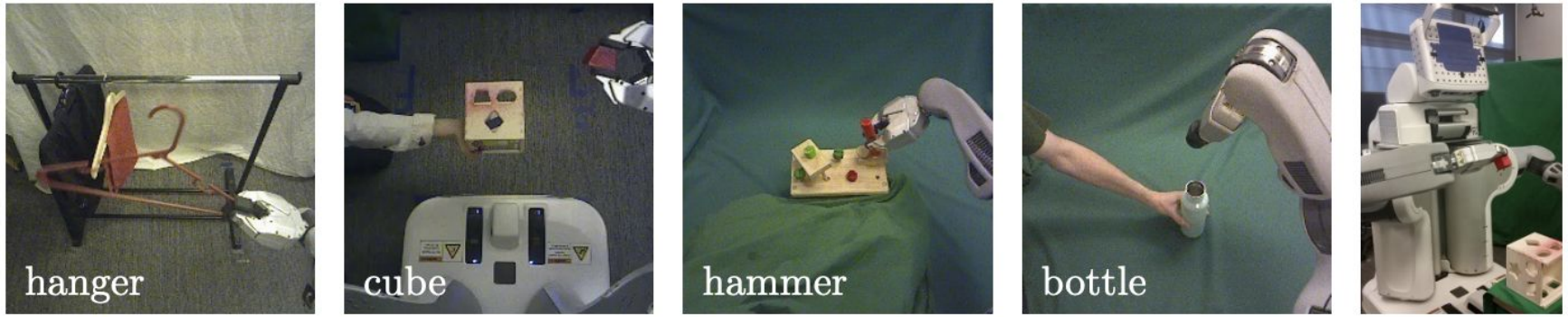
Ego4D(ロボットのための一人称ビジョンのデータセット)を始め，ロボティクス実機実装が同時進行で進んでいる

source1: https://ego4d-data.org/
source2: https://arxiv.org/abs/2211.03785
source3: https://arxiv.org/abs/2203.06173

CVPR 2023 の動向・気付き(50/259)

Workshop: Visual Pre-training for Robotics

- Jitendra Malik(UC Berkeley / Meta AI)
  - BAIR(UCBの人工知能研究所; https://bair.berkeley.edu/)にAI分野の多くの研究者がいる
  - UC Berkeleyの層の厚さにおいて，JitendraはCVの中心的な存在
  - 周囲にスタンフォード大学，シリコンバレーなどがあり研究連携やサポートのバリエーションも豊富である
  - もちろん，ロボット研究者も豊富であり，ベイエリアの連携で研究をさらに加速させている



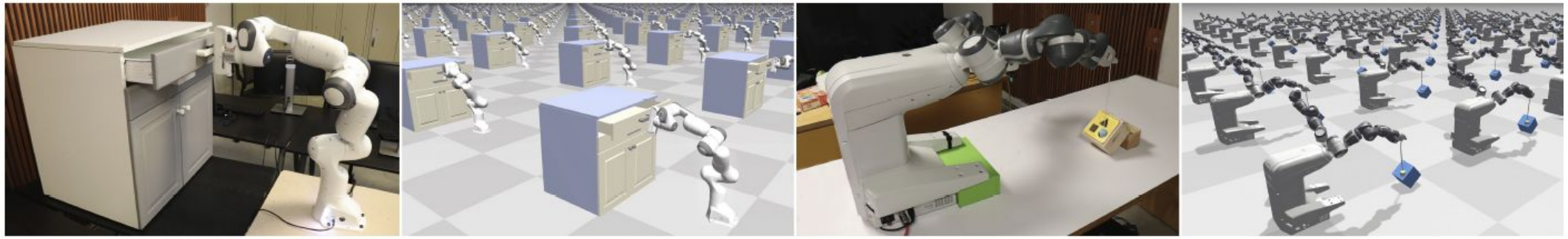
Jitendraは共著には入ってないが，UCBにてファカルティレベルの研究連携により強力な研究が生み出されている

Sergey Levine, Chelsea Finn, Trevor Darrell, Pieter Abbeel, “End-to-End Training of Deep Visuomotor Policies”, in JMLR, 2016.

CVPR 2023 の動向・気付き(51/259)

Workshop: Visual Pre-training for Robotics

- Dieter Fox (University of Washington/NVIDIA)
  - Q:ロボット制御における高次の推論は大規模基盤モデルが解いてくれるが,
    推論以前の低次のマニピュレーションをどう解決すればよいか？
  - A:大規模データ, シミュレーションの効果的な利用
    - 解きたいタスクに対するシミュレーション環境とその計算基盤を適切に備えれば最低限の量の実ロボット試行だけから十分汎化するモデルが作れる(GoogleのRT-1などとは逆の主張)
  - DexTreme(Issac Simベースの手操作シム), Curobo(分散動作生成基盤)などの高度なシミュレーション環境とその高速化などの事例を紹介
  - マニピュレーションのための事前学習に余地あり？

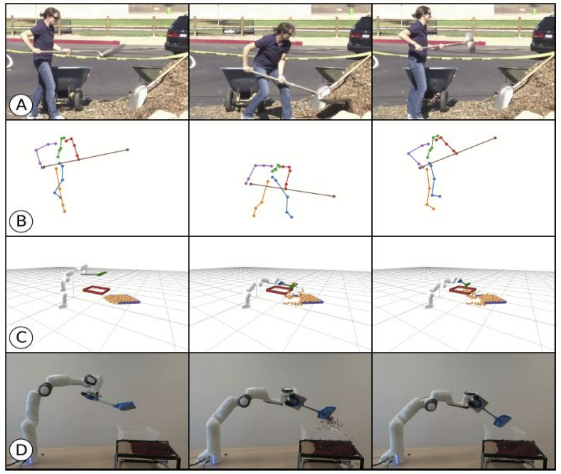


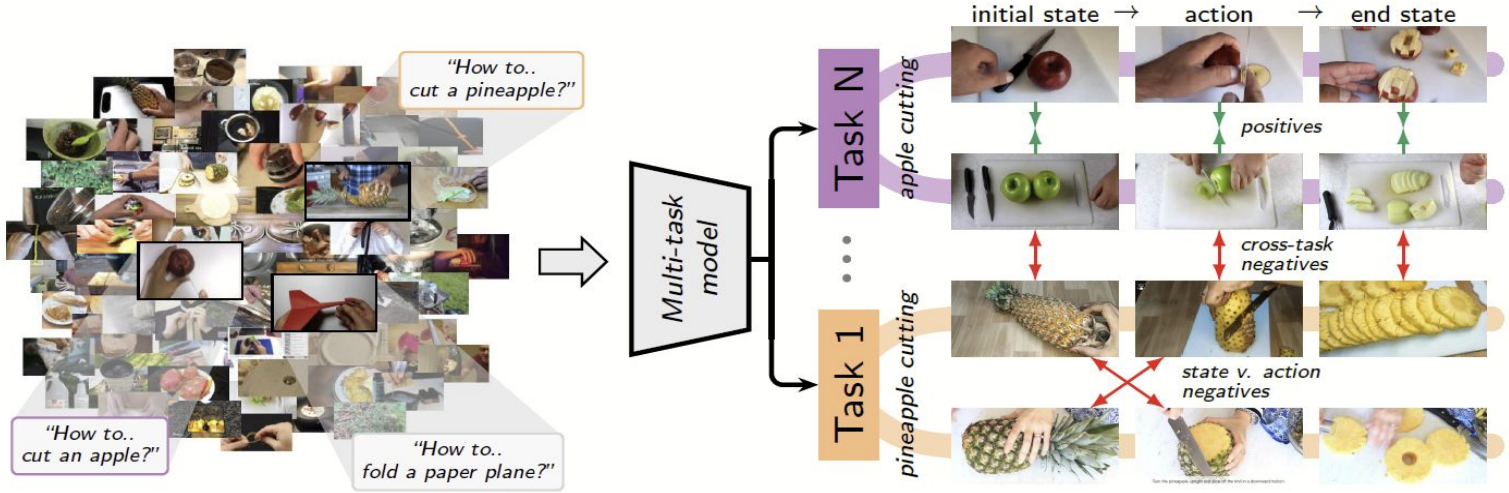
Chebotar, Y., Handa, A., Makoviychuk, V., Macklin, M., Issac, J., Ratliff, N., & Fox, D. (2019, May). Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 8973-8979). IEEE.

CVPR 2023 の動向・気付き(52/259)

Workshop: Visual Pre-training for Robotics

- Josef Sivic (Czech Technical University)
  - 人がある手順にしたがって行う行動のデモンストレーション映像(Instructional Videos)からスキルをロボットに転写するための技術を (1) 少数の高品質の試行映像からの学習 (2) 大量の(必ずしも高品質でない)インターネット映像からの学習 の両面から紹介
  - (1) のポイント:検出した人の動作をどうロボットアームの動きとアライメントするか
  - (2) のポイント:映像中の状態変化認識のために同一の意味の別インスタンスをどう教師なしで結びつけるか
  - シミュレーションとインターネット映像資源の活用→効率的なロボット学習？



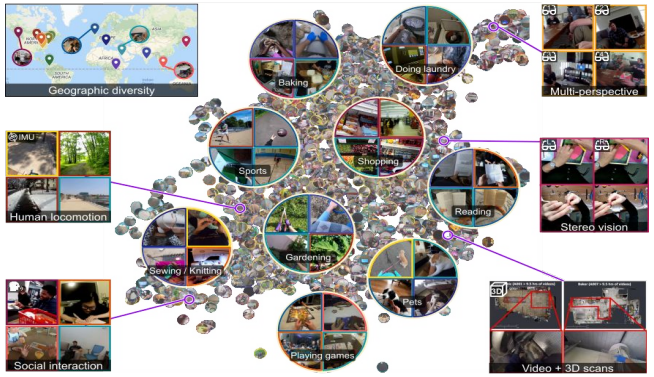


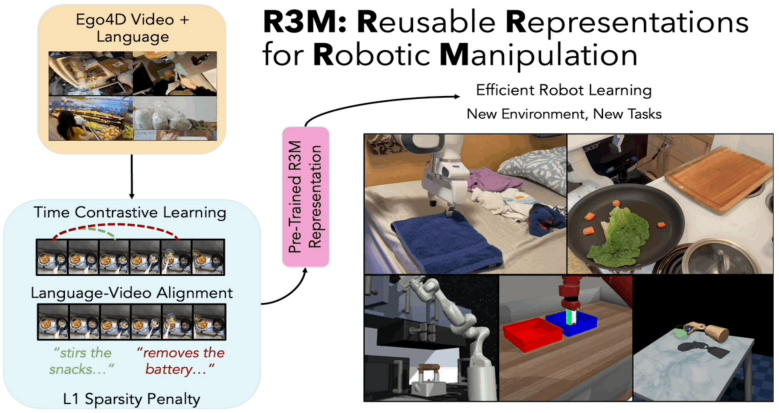
Zorina, K., Carpentier, J., Sivic, J., & Petrík, V. (2021). Learning to Manipulate Tools by Aligning Simulation to Video Demonstration. *IEEE Robotics and Automation Letters*, *7*(1), 438-445.
Souček, T., Alayrac, J. B., Miech, A., Laptev, I., & Sivic, J. (2022). Multi-Task Learning of Object State Changes from Uncurated Videos. *arXiv preprint arXiv:2211.13500*.

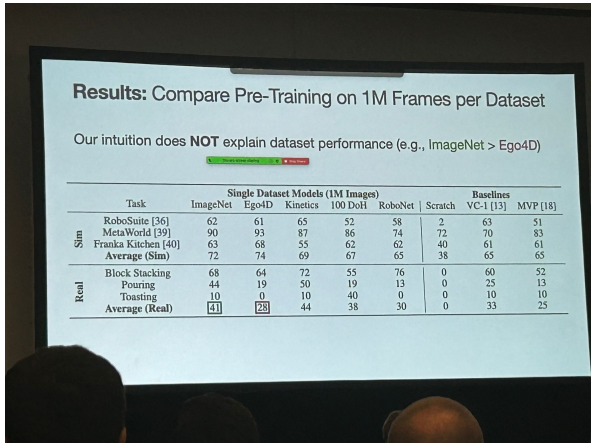
CVPR 2023 の動向・気付き(53/259)

Workshop: Visual Pre-training for Robotics

- Abhinav Gupta(CMU)
  - Roboticsにおける事前学習データとは?
    - ドメイン・多様性・スケールの三本柱をどうするか
  - これまでのスライドにある通りEgo4Dでの学習がトレンド
  - R3M等含めてEgo4Dでの検証は進んでおり，効果も実証されつつある
  - **一方で，Ego4Dが必ずしもRobotics事前学習に最適であるかどうかは疑問を抱くべき**
    - 彼いわく“100DOH or Charades might be better source of data for training robotics.”





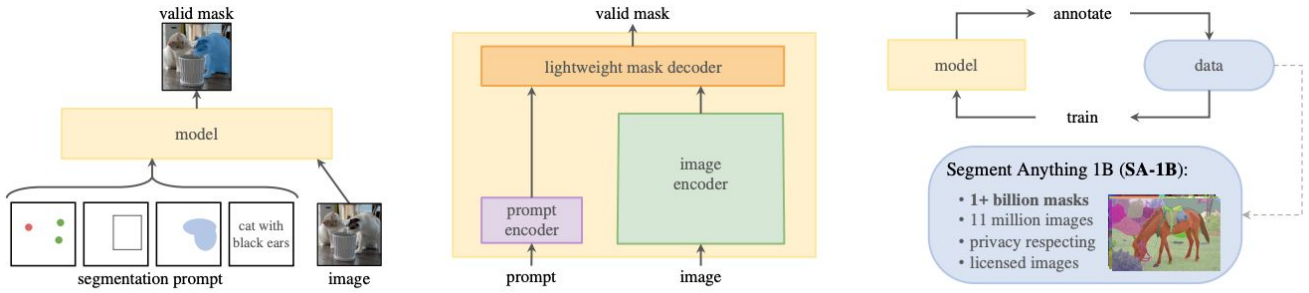


source1: https://ego4d-data.org/
source2: https://arxiv.org/abs/2203.12601

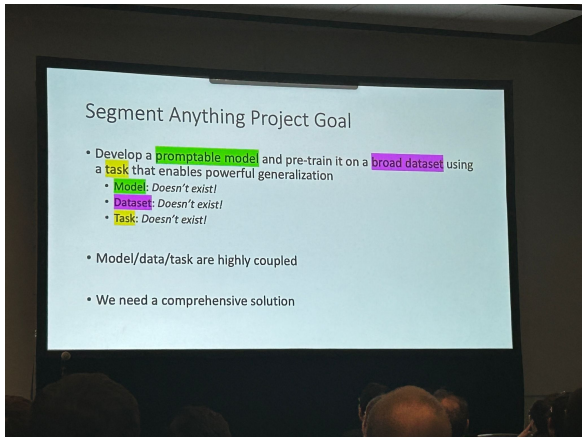
CVPR 2023 の動向・気付き(54/259)

Workshop: Visual Pre-training for Robotics

- Alexander Kirillov(Meta AI Research)
  - Segment Anythingの第一著者
  - 基盤モデルには，モデル・データ・タスクの3つが重要
    - **Segment Anythingを実現するには3要素全てが不足すぎる.「じゃあ全部やりましょう」**
  - Promptable segmentation (task): Prompt engnearingを活用したSegmentationへ
  - SAM (model): Prompt encodingを導入して，point/text等をpromptとして入力可能に
  - SA-1B: (Data): Human-in-the-Loopでアノテーション，データセット権利関係もクリアに



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

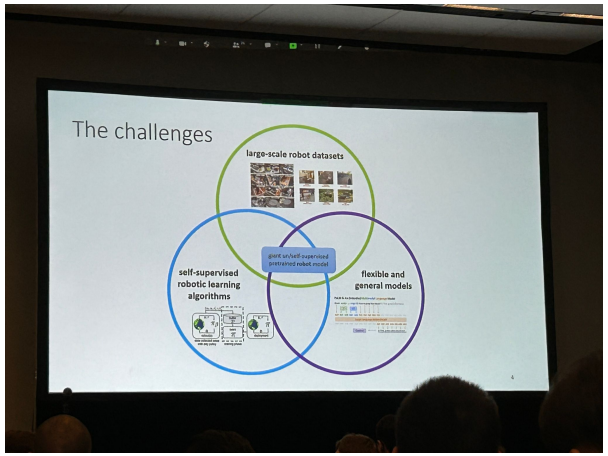


https://arxiv.org/pdf/2304.02643.pdf

CVPR 2023 の動向・気付き(55/259)

Workshop: Visual Pre-training for Robotics

- Sergey Levine(UC Berkeley/Google Deep Mind)
  - ArmFarmの第一著者. Deep Learning×Roboticsにおけるスター研究者
  - **CV基盤モデルの流れをロボティクスでも!! RT-1以降の基盤モデル動向は彼にも注目!!**
  - データ・学習戦略・モデルの3つが重要
    - データ: Bridge Data V2
    - 学習戦略: Pre-training for Robotics (PTR)
    - モデル: Visual Navigation Transformer (ViNT)
  - 最後は上記の3要素をどう統合するか.



https://arxiv.org/pdf/1603.02199.pdf

CVPR 2023 の動向・気付き(56/259)

Workshop: Visual Copy Detection Workshop

- 2023 Video Similarity Challlengeを開催
  - 動画像中に画像のコピーが含まれるか認識・検出する問題
  - 動画像中にコピーが含まれるか認識 / 動画像中のコピーがどこにあるかも検出の2課題
  - 著作権が含まれる画像が動画にコピー&ペーストされて使用され続けていることを解決するためにコンペティションが行われている

Descriptor Track

1. WeChat CV
2. Friendship First Competition Second
3. LINE CVL

Matching Track

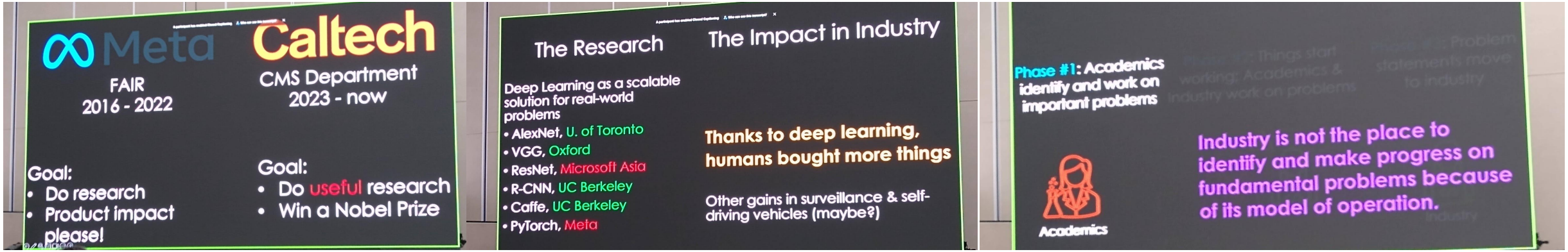
1. WeChat CV
2. Friendship First Competition Second
3. LINE CVL

日本からはLINE Corp.のComputer Vision Laboratory(CVL)が世界第3位入賞を果たした

CVPR 2023 の動向・気付き(57/259)

Workshop: Scholars & Big Model

- Georgia Gkioxari
  - Meta AI -> Caltechに移って教員メンバーになる
  - メタではプロダクトに有用な研究を行う，カルテクでは有用な研究を行いノーベル賞を取るように言われている(>_<)
  - アカデミア：次の新しい問題を考えて，提起している
  - インダストリ：問題としては新しいことがない，BigTechは大規模化・SmallTechはモデル使用で汎化させて問題を解いている

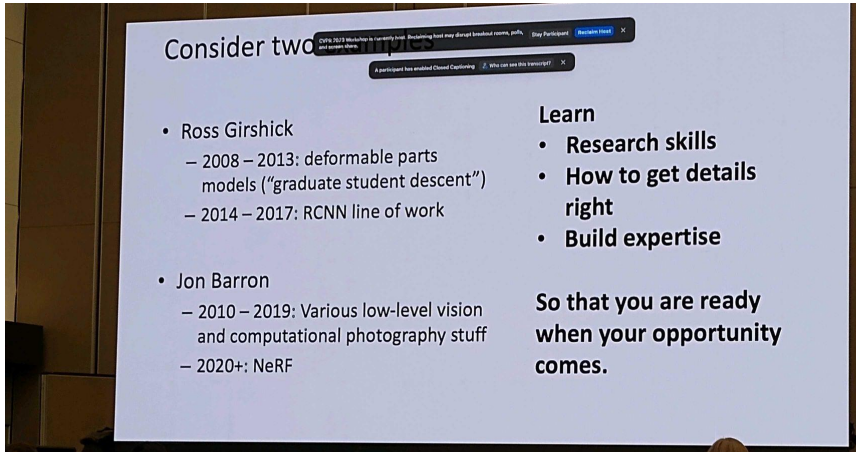


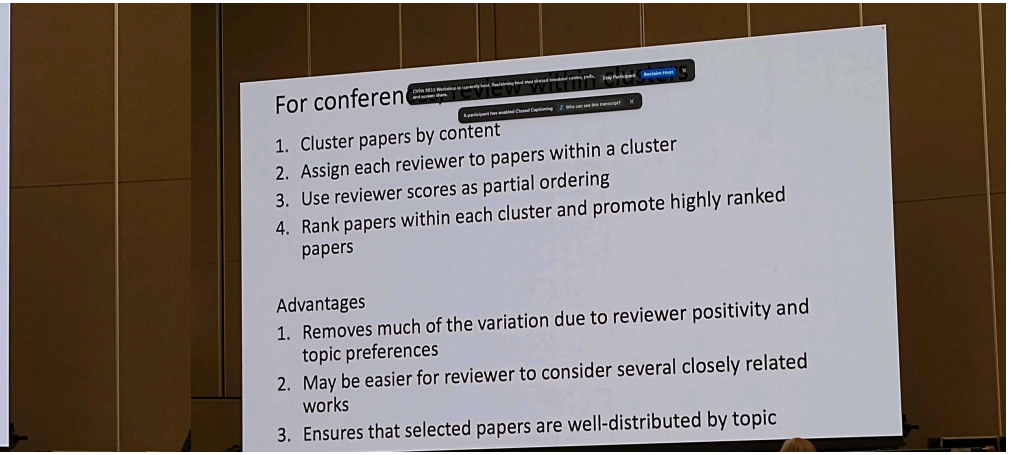
source: 講演時にスライドを撮影

CVPR 2023 の動向・気付き(58/259)

Workshop: Scholars & Big Model

- Derek Hoiem
  - 論文出るの早すぎ！しかもみんな同じ手法を使っている
  - 新しいフレームワークや問題提起できる人を増やすべき
  - 教育や(研究者としての)トレーニングを賞賛すべき
  - アウトプットの心配する前に学生は学べ！

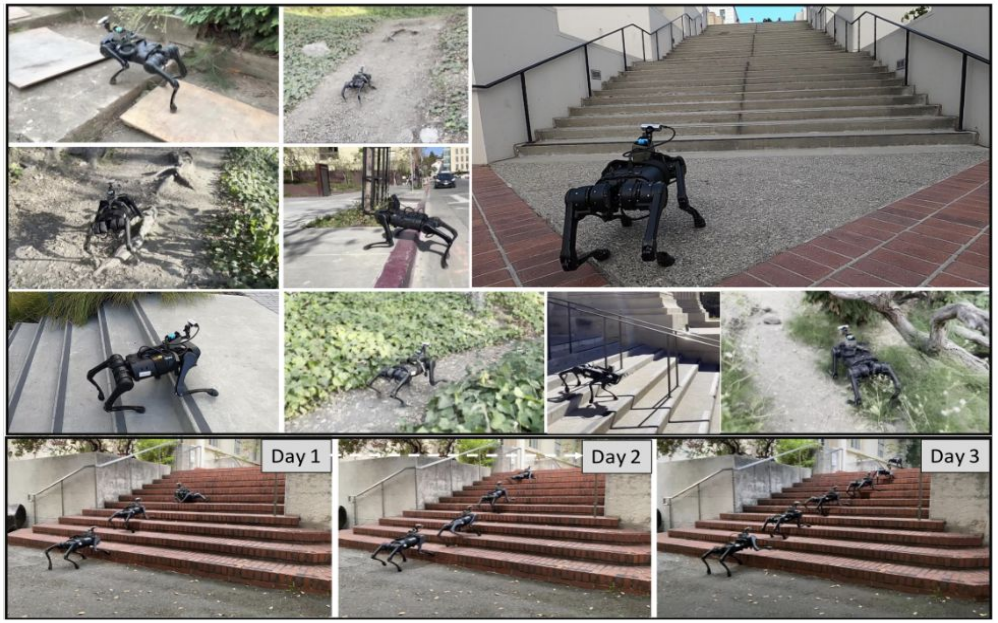




CVPR 2023 の動向・気付き(59/259)

Workshop: Scholars & Big Model

- Jitendra Malik
  - アカデミアは多くのチャンスとつながりがある
  - アカデミアは新しいアイディアを試せるので，どんどん挑戦していこう
  - Locomotion に関する前例の無い研究の例
  - → リスクをとって新しい研究をしよう
  - → 成功例が次の学生の励みになる



現在まで制御側でなんとか対応しようとしてきた4足歩行の問題も，視覚の事前学習を入れたらうまく行った，という例

source: https://arxiv.org/abs/2211.03785

CVPR 2023 の動向・気付き(60/259)

Workshop: Scholars & Big Model

- 所感
  - 確かに何万人も同じ問題をやらなくても良い
  - 深層学習の黎明期にImageNetの画像認識に多くの人が取り組み，2012/2015のチャレンジ以外は年間数%を向上させるため頑張っていた
  - 「幅広い知識と狭い範囲でも良いので特化した専門性」は学生に限らず重要
  - あとは体力，何と言っても知的・肉体的な体力で押し切る！

“To invent, you need a good imagination and a pile of junk.”
– Thomas A. Edison の精神

CVPR 2023 の動向・気付き(61/259)

論文の完成度が向上

- 数年前(2020年以前):「精度向上」して，手法として成立すれば採択
- 2023年:「精度向上+α」がかなりの頻度ある(αは計算コスト・パラメータ数削減など)，精度を同等以上にするのは当たり前でいかに付加価値を高いレベルで付けるかの勝負

- 「とりあえず精度向上」して，手法っぽく確立させるだけなら結構時間と頭を使えばできるようになってきた(？)ということか

→ 分野の査読経験値が向上してきて，熟練の査読者も増えてきた

CVPR 2023 の動向・気付き(62/259)

トップ研究者の体力・筋肉はヤバい！

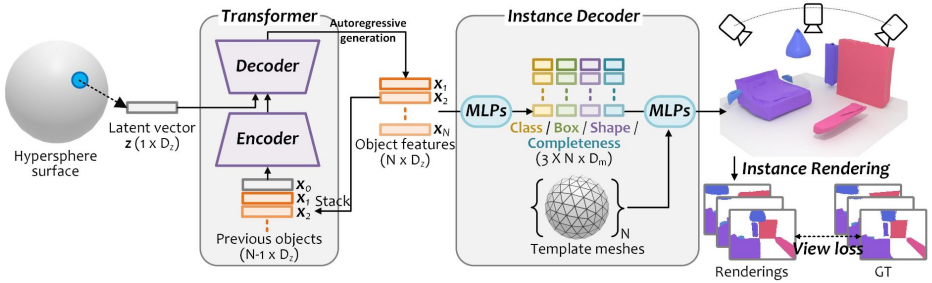
- 研究には体力が必要
  - 「知的体力が必要」みたいなことを金出武雄先生の本*に書いてあったが，最近競争が激化するAI分野においてはホントに体力がないとついて行けなくなった
  - CVPRなどAI分野の情報収集を短時間で行って，研究まで実施ためには，体力・筋肉の差がそのまま影響する構図になっている？？？
  - 実際，北米・欧州の有名な先生はいつも話しているし，忙しなく動いている(夜遅くまで懇親会していたのに翌朝4時からミーティングがあるという場面も)

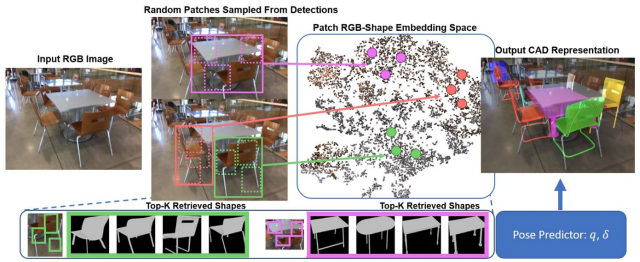
***独創はひらめかない:「素人発想，玄人実行」の法則 ― 金出武雄**

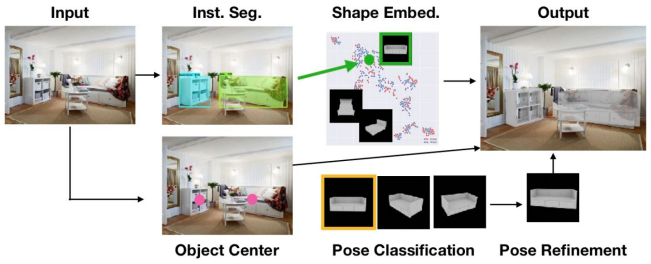
CVPR 2023 の動向・気付き(63/259)

Workshop: Structural and Compositional Learning on 3D Data

- **Speaker: Angela Dai**
  - ScanNetの提案者．TUMにて3D認識系の研究
- **Title: Structured Parametric Models for 3D Perception**
  - 3Dデータの取得は大変．そこで，**Learning 3D Prior without 3D data**
  - 画像のみを使って3Dシーンをどのように理解するか，また3D Priorをどのように獲得するか
    - Mask2CAD, Patch2CAD, Learning 3D Scene Priors with 2D Supervision





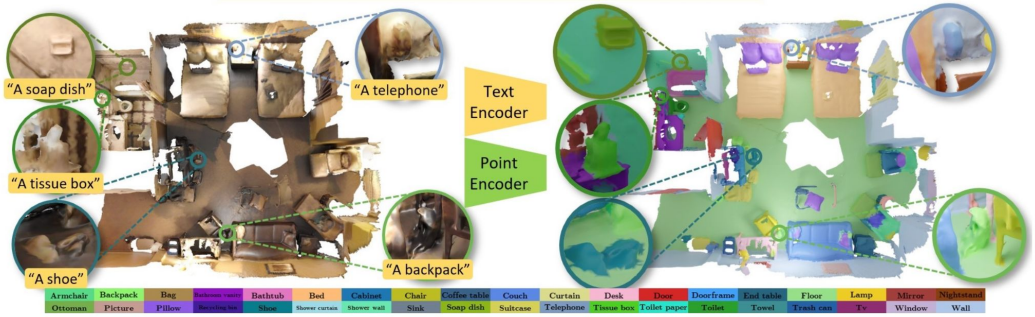


source1: https://yinyunie.github.io/sceneprior-page/
source2: https://www.3dunderstanding.org/papers/2020/kuo2020mask2cad/
source3: https://www.3dunderstanding.org/papers/2021/kuo2021patch2cad/

CVPR 2023 の動向・気付き(64/259)

Workshop: Structural and Compositional Learning on 3D Data

- **Speaker: Angela Dai**
  - ScanNetの提案者. TUMにて3D認識系の研究
- **Title: Structured Parametric Models for 3D Perception**
  - **What's next??**
    - Point Cloud & Language!
      - カテゴリが限定的な3DデータセットをLLMでどう解決していくか！
    - Rethinking ScanNet
      - 今のScanNetではデータ質・量の面で改善が必要. ScanNet++の構築が急務
      - 提案者た 識を次のステージへ

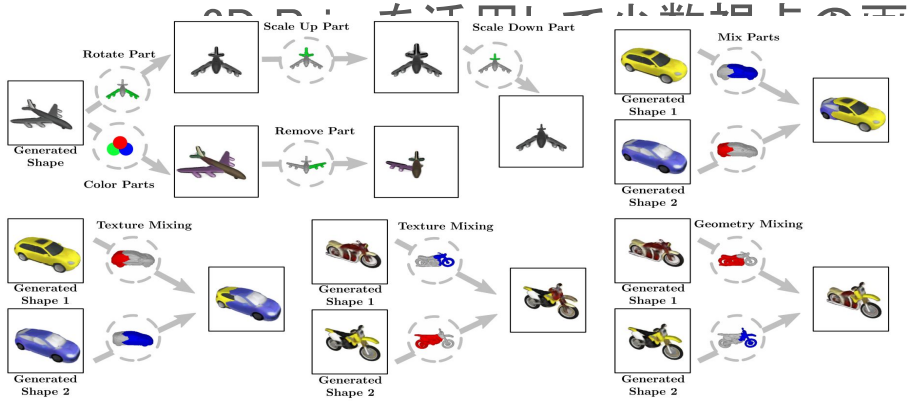


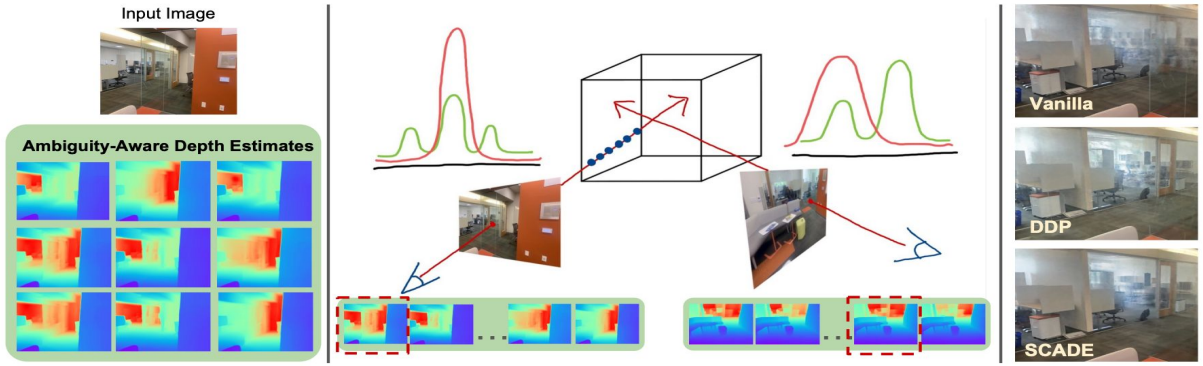
https://rozdavid.github.io/scannet200

CVPR 2023 の動向・気付き(65/259)

Workshop: Structural and Compositional Learning on 3D Data

- **Speaker: Mikaela Uy**
  - ScanObjectNNの提案者. Stanford Univ.にて3D認識・生成系の研究
- **Title: Towards Controllable 3D Content Creation by leveraging Geometric Priors**
  - PartNeRF
    - 3D教師を必要とせずにオブジェクトパーツの編集・テクスチャ生成を可能に!!
  - SCADE
    - 3D Prior を活用して少数視点の画像から3Dシーンを生成



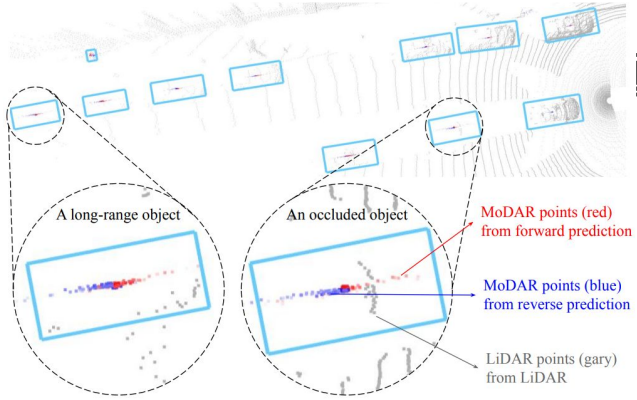


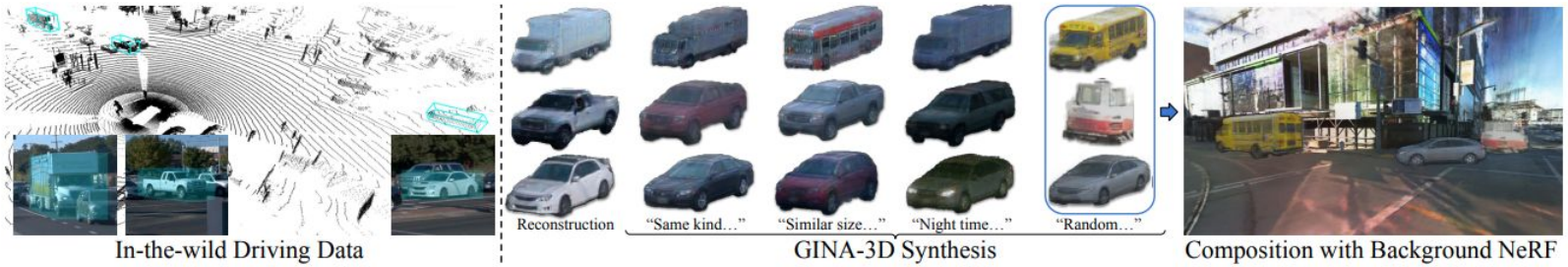
https://arxiv.org/pdf/2303.13582.pdf
https://arxiv.org/pdf/2303.09554.pdf

CVPR 2023 の動向・気付き(66/259)

Workshop: Learning 3D with Multi-View Supervision (3DMV)

- **Speaker: Charles Qi**
  - PointNet/VoteNetを始めとして，三次元点群深層学習の第一人者
- **Title: 3D Perception from Multi-X**
  - MoDAR
    - 数秒後の三次元点群を予測し，その三次元点群を三次元物体検出に利用
  - GINA-3D



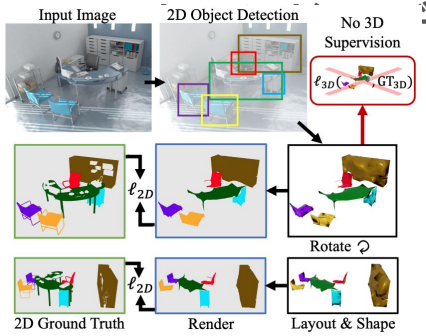


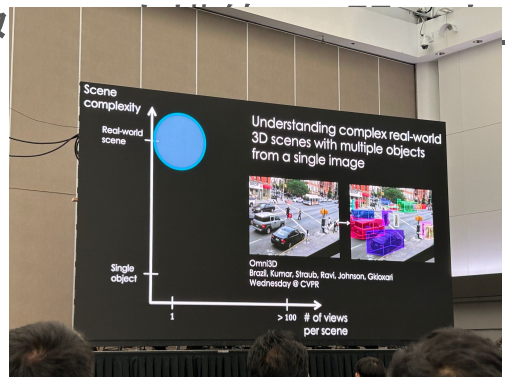
https://arxiv.org/pdf/2306.03206.pdf
https://arxiv.org/pdf/2304.02163.pdf

CVPR 2023 の動向・気付き(67/259)

Workshop: Learning 3D with Multi-View Supervision (3DMV)

- **Speaker: Georgia Gkioxari**
  - Mesh R-CNN/Mask R-CNN等の著者．CVPR 2021ではPAMI Young Researcher Awardを獲得したスター研究者．最近，Meta AI (産業界)からCaltech (学術界)へ移籍
  - ちなみに，6/19 (月)Workshopでは少なくとも3つの招待講演を掛け持ち!!
- **Title: 3D Recognition in the Wild With and Without 3D Supervision**
  - 3D教師ラベルを使用せずに実世界を認識・理解するには?
    - やはり，3Dデータの収集・アノテーションに対する膨大なコストが必要である部分が根本的問題
  - 画像から3D再構成/物体検出 → 仮想カメラで投影&学習 → 入力画像に3D



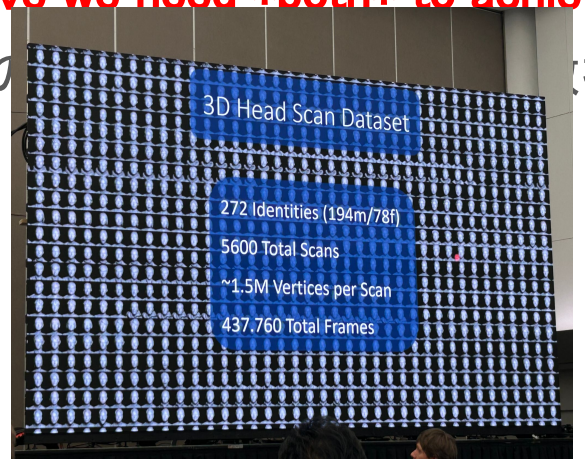


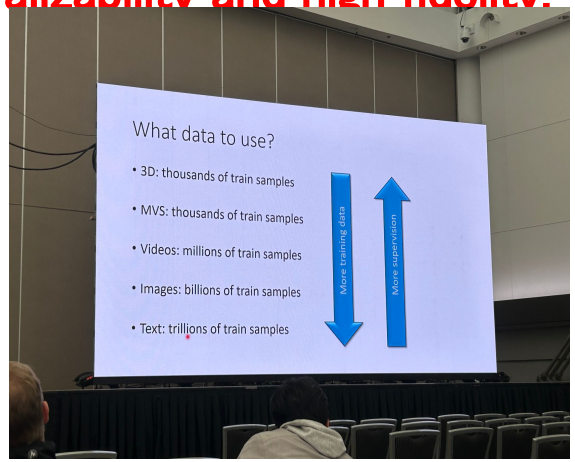
https://arxiv.org/pdf/2206.07028.pdf

CVPR 2023 の動向・気付き(68/259)

Workshop: Learning 3D with Multi-View Supervision (3DMV)

- **Speaker: Matthias Niessner**
  - TUMの先生．3次元再構成をはじめとして3D系における大先生
  - ちなみに，すごい筋肉ムキムキで半袖短パンで講演
- **Title: Photorealistic Reconstruction with Neural Parametric Models**
  - 大量データ&弱教師 or 少量データ&強教師のどちらが良い？
    - **I believe we need *both* to achieve generalizability and high fidelity.**
  - 3D再構成の 提案





https://twitter.com/MattNiessner/status/1667335181152423937?s=20

CVPR 2023 の動向・気付き(69/259)

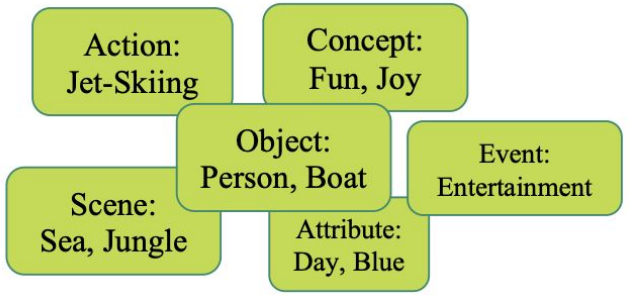
Workshop: Holistic Video Understanding workshop

- 1つの動画に多様なラベル(行動，背景，物体など)を付与したデータセットを　公開しているチームのワークショップ
- 動画の多様な情報を学習するためにはどうしたら良いかについて議論
- 全体的な動画の認識の応用について議論

研究の動向としては，，

過去

マルチタスクやマルチラベルを使ったHolistic Video Understanding



現在

文章を使ったHolistic Video Understanding

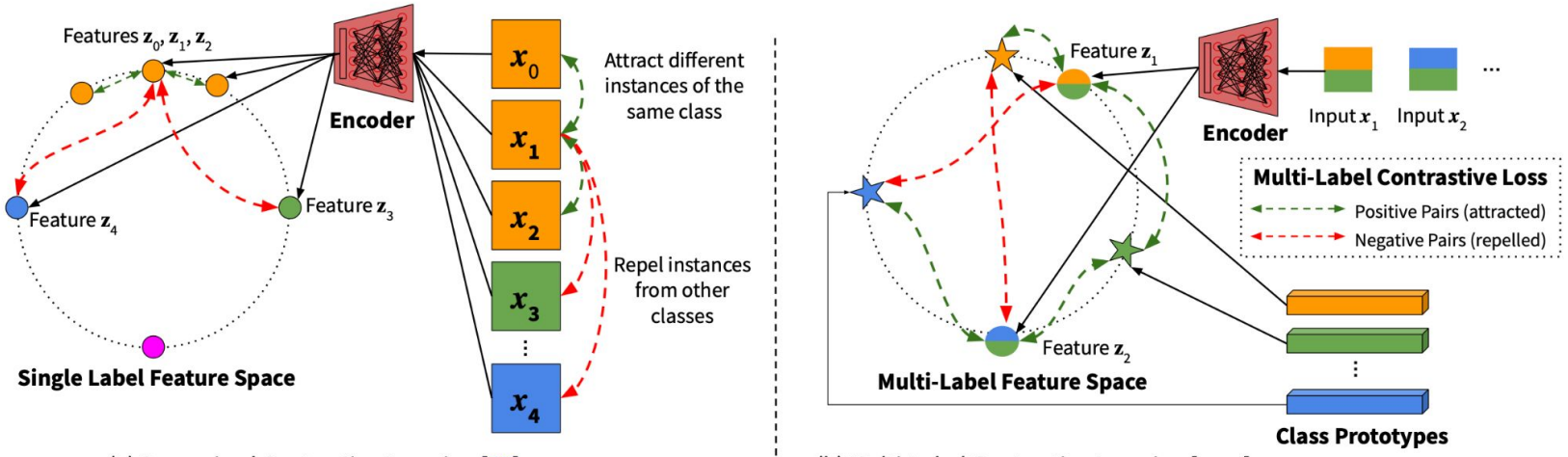
このワークショップを見た感じでは．．

On a beautiful blue day, a joyful person was having fun jet-skiing on the sea near the jungle, their laughter echoing from the boat nearby, creating a pure sense of joy.

CVPR 2023 の動向・気付き(70/259)

Workshop: Oral発表1 Class Protptype based contrastive learning for classifying multi-label

- オンラインビデオから子供向けの教育的なコンテンツを識別する新しいアプローチを提案
- 複数のラベルに関連する微細なサンプルを処理できる新たなクラスプロトタイプに基づくSupervised contrastive learningを用いて，ビデオが複数の教育的内容を含むことができ，内容のクラスが視覚的に似ている場合でも分類する細かい多ラベル分類問題として定義した



(a) Supervised Contrastive Learning [26]　(b) Multi-Label Contrastive Learning [ours]

CVPR 2023 の動向・気付き(71/259)

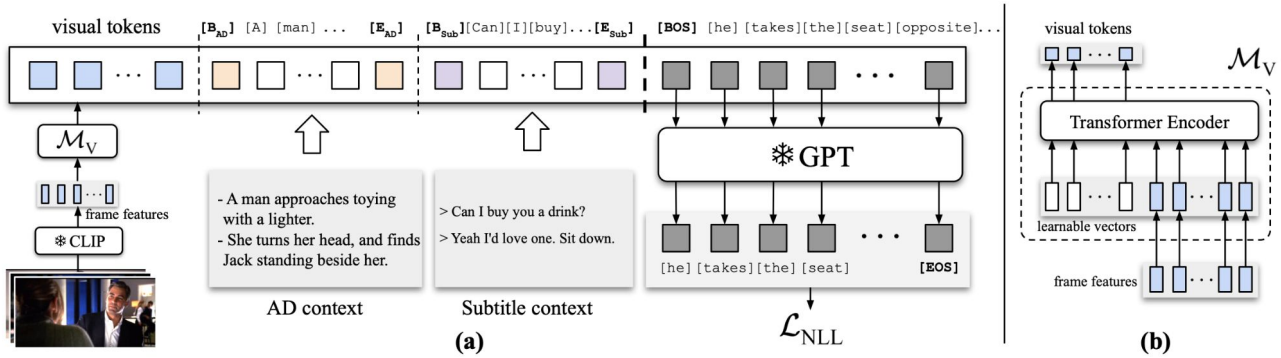
Workshop: Oral発表2 AutoAD:movie description incontext

- 映画から自動的に音声説明(Audio Description: AD)を生成するモデルを開発し，評価
- GPTやCLIPのような事前学習済みのモデルを活用し，視覚的に制約されたテキスト生成のために2つのモデルを橋渡しするマッピングネットワークを学習
  - 事前学習モデルGPTとCLIPの組み合わせは，視覚的に条件付けられたテキスト生成(映画からの音声説明(AD)の生成)に対して有効
  - 映画クリップ，前のクリップからのAD，および字幕からの情報を含む，より広範な文脈の使用が，質

Context AD:
As Karen stares groomly out of the window, a man approaches toying with a lighter. She turns her head, and finds Jack standing beside her.

Subtitles:
> Can I buy you a drink?
> Yeah I'd love one. Sit down.

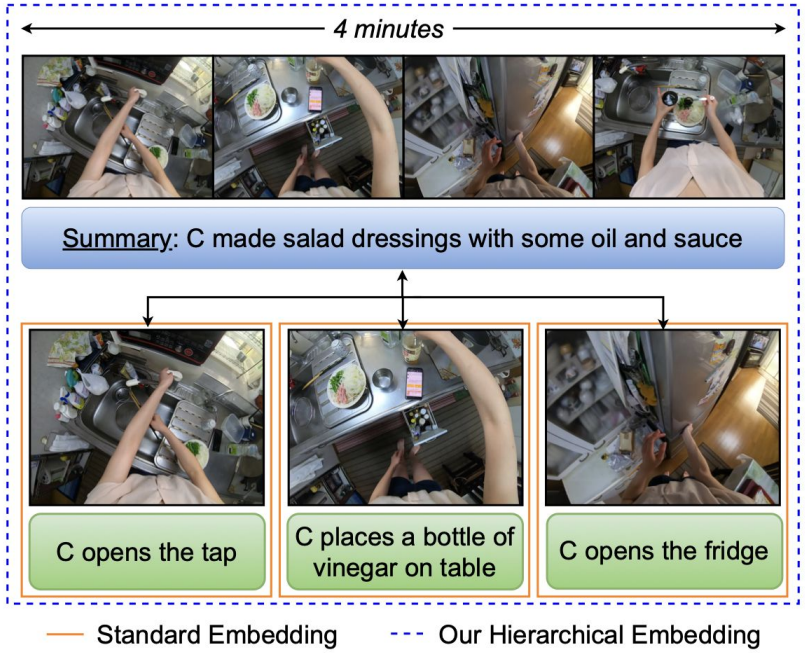
Target AD: He takes the seat opposite, then places his lighter on the table

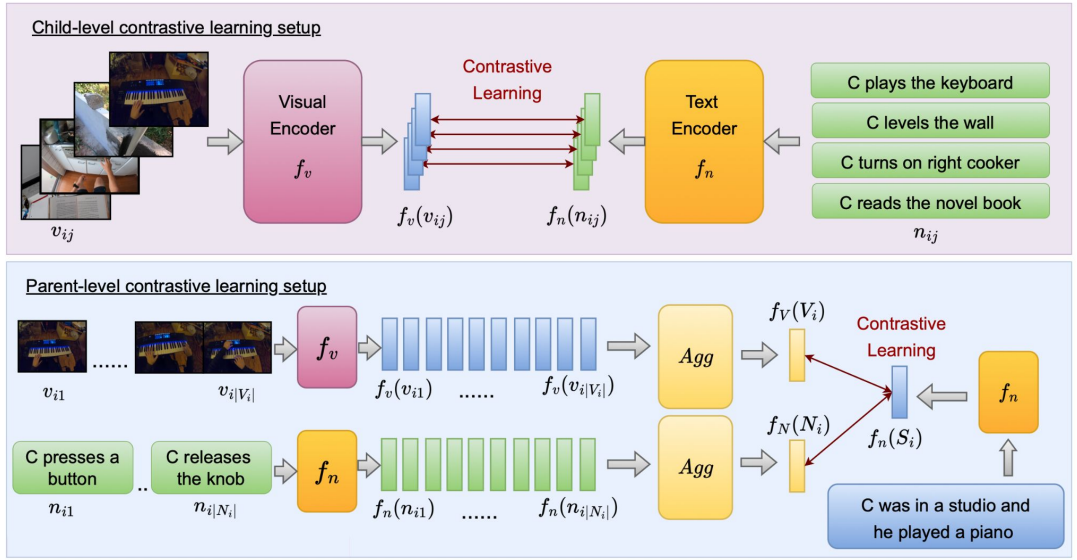


CVPR 2023 の動向・気付き(72/259)

Workshop: Oral発表3 HierVL: Learning Hierarchical Video-Language Embeddings

- 長期的・短期的なビデオとテキストの関連性を同時に学習する新しい階層的動画-言語埋め込みモデル"HierVL"を提案
  - 結果，モデルは特定のクリップが何を表しているのか(短期的な理解)，そして全体のビデオが何を意図しているのか(長期的な理解)を理解する能力を獲得
  - これにより，モデルはさまざまなビデオ理解タスク，特に長期的なビデオモデリングを必要とするタスクで，より良いパフォーマンスを達成することが可能に



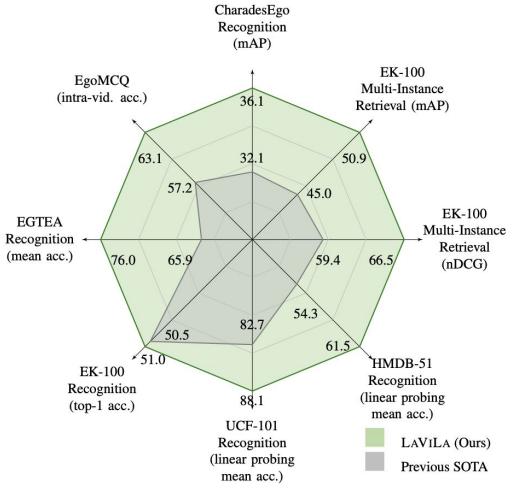


CVPR 2023 の動向・気付き(73/259)

Workshop: 招待公演:Learning holistic visual representation with minimal supervision

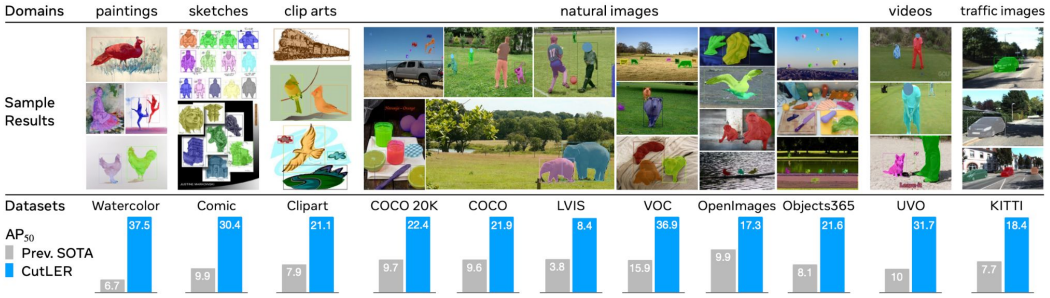
- Webからは大量のデータを取得するのは簡単だが，これらのデータにholisticなラベルを付与するのは大変
  - そこで，この招待公演では以下の論文を紹介
    - LaViLa:Language model augmented Video-Language pre-training
    - CutLER: Cut and Learn for Unsupervised Object Detection and Instance Segmentation
    - ImageBind: One Embedding Space To Bind Them All

LaViLa



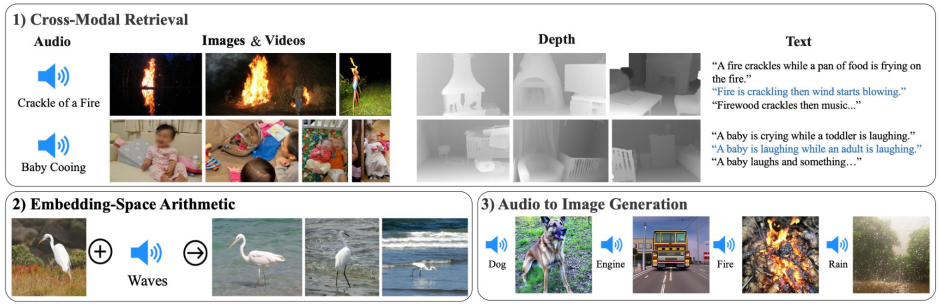
ナレーターとなる大規模言語モデルの
スーパービジョンを用いてビデオ言語表現を
学習することで，多くの一人称および
三人称のビデオ理解タスクにおいてSoTA

CutLER



人間の監視なしに学習させたCutLERモデルを用いたゼロショット教師なし物体検出とインスタンス分割を行う．標準的な検出APboxを使用してモデルを評価する．CutLERは，ビデオフレームなどの多様な画像領域にまたがる様々なベンチマークで，強力な性能を発揮する．

ImageBind



異なるモダリティの埋め込みを統合してクロスモーダル検索を可能にし，新たなセマンティクスを形成し，音声から画像への変換を行うというマルチモーダル機能を提供

CVPR 2023 の動向・気付き(74/259)

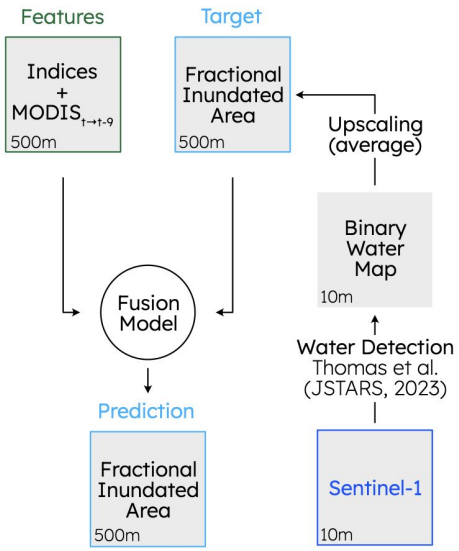
Workshop: EarthVision workshop

- 地球観測とリモートセンシングは，コンピュータビジョン，機械学習，信号/画像処理の融合により，空中や宇宙から収集したデータを利用して地球表面のプロセスについての情報を提供する研究分野
- これは，位置情報サービス，オンライン地図，大規模監視，3D都市モデリング，自然災害の予測と対応，気候変動監視などの多岐にわたるアプリケーションに使用され，大量のデータ解析には高度な自動化が必要
- この分野は，国連の「持続可能な開発目標」の多くに直接関連し，地球観測技術と機械学習の進歩は，自然災害監視，都市成長，森林破壊，気候変動などの社会的課題の解決にも貢献している

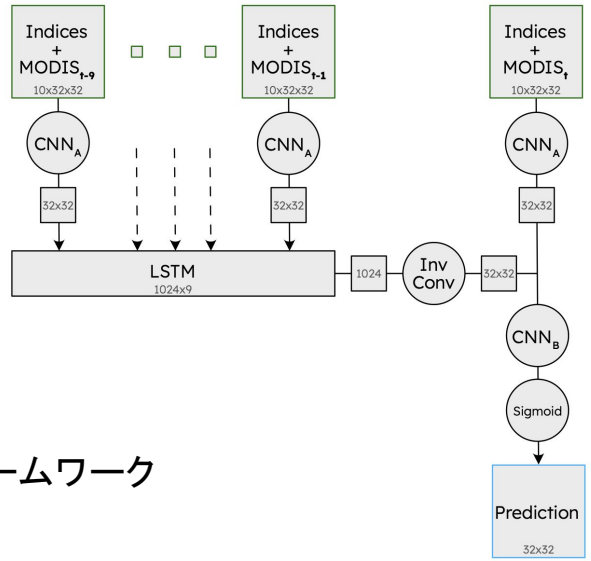
CVPR 2023 の動向・気付き(75/259)

Workshop: Inferring the past: a combined CNN-LSTM deep learning framework to fuse satellites for historical inundation mapping

- Sentinel-1から派生した洪水エリアとMODISデータを統合するための深層学習フレームワーク，具体的にはCNN-LSTMを開発
  - 衛星データは，洪水影響地域をリアルタイムで監視し評価する能力を持つため，緊急対応や災害管理の努力において重要な情報源となる
  - 衛星データは，洪水のトレンドやパターンを時間を通じて分析するための歴史的情報を提供することができ，長期的なレジリエンシー計画やリスク管理戦略の策定を支援
  - Sentinel-1衛星とMODIS時間系列を統合することで，指数ベースの洪水保険指数における各センし，高解像度データ利用可能性の両方の利点を活る



モデリングのフレームワーク



深層学習アーキテクチャの模式図

CVPR 2023 の動向・気付き(76/259)

Workshop: Seasonal Domain Shift in the Global South: Dataset and Deep Features Analysis

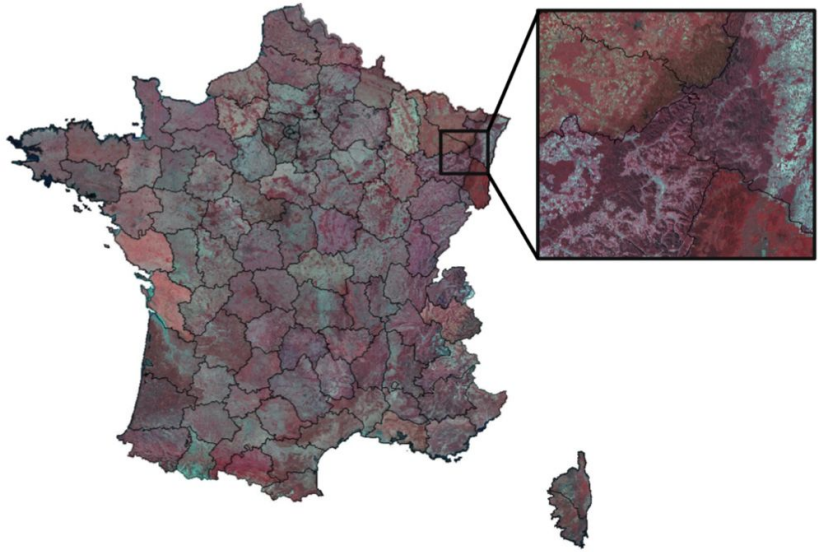
- 季節変動によるドメインシフトを考慮に入れた航空画像からの土地利用分類のためのデータセットを提案し，色，形，テクスチャの画像特徴がシーン分類の精度にどのように影響するかについて詳細に調査
- 特徴の抽出において，Gabor特徴の事前知識を利用し，大きな受容野を実装したアーキテクチャを提案し，それが浅いもしくは深いネットワークでの分類パフォーマンスを向上させることを示した

CVPR 2023 の動向・気付き(77/259)

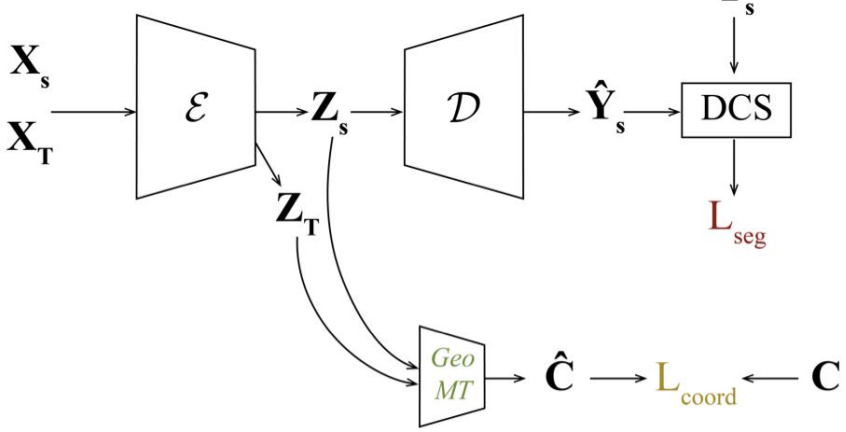
Workshop: GeoMultiTaskNet: remote sensing unsupervised domain adaptation using geographical coordinates

- 地理座標を利用して高解像度リモートセンシング画像のセマンティックセグメンテーションにおける教師なしドメイン適応を実現する新しい軽量モデル，GeoMultiTaskNetを提案
  - 教師なしなためアノテーションコストが削減されている
  - ドメイン適応を用いるので，別地域のデータの利用(ドメインシフトが可能)

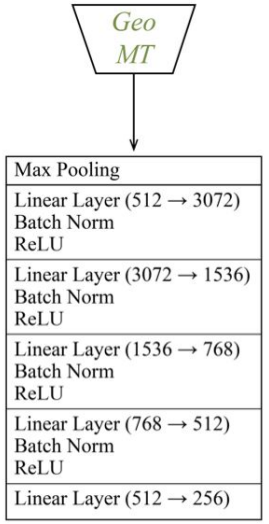
ート



ドメイン間の航空画像のラジオメトリックな不一致
表示されるバンドは，近赤外線，赤色，緑色のスペクトル情報を合成したもの



GeoMultiTaskモジュールと
Dynamic Class Samplingモジュール

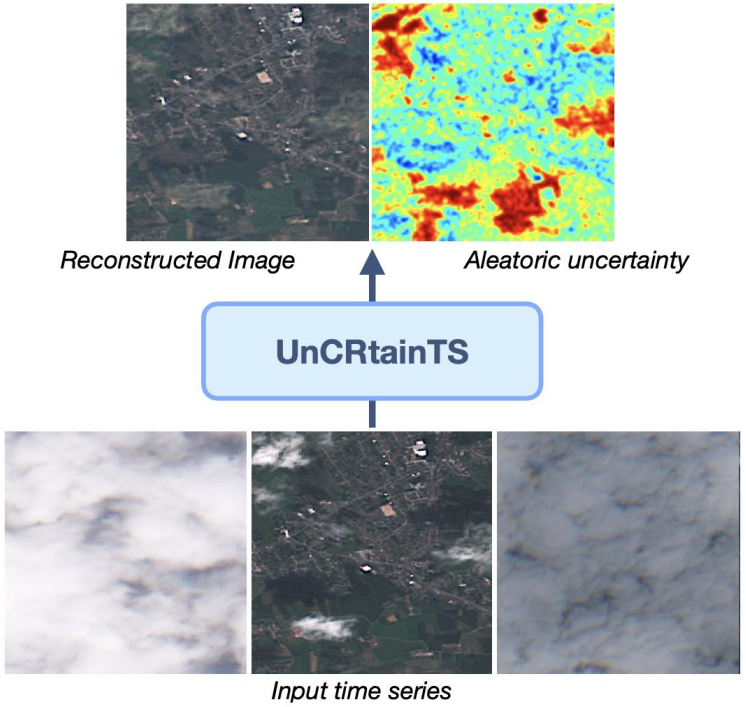


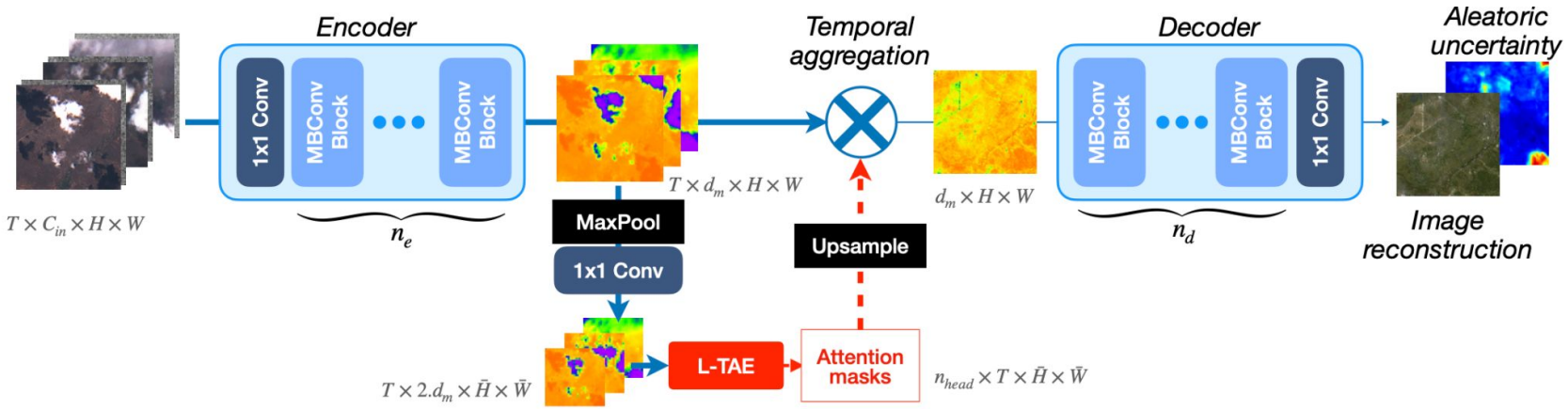
GeoMultiTaskモジュールの構造で，入力と出力のサイズが示されている

CVPR 2023 の動向・気付き(78/259)

Workshop: UnCRtainTS: Uncertainty Quantification for Cloud Removal in Optical Satellite Time Series

- 新たなAttention baseベースのアーキテクチャと多変量不確実性予測を組み合わせた，多時点の雲除去手法「UnCRtainTS」を導入し，画像再構築のSoTAを達成
  - UnCRtainTSは雲やヘイズなどによる遮蔽を取り除くための手法
  - 特に光学衛星画像の多時点雲除去に適用される
    雲除去は画像の品質と解釈可能性を向上させるとともに，限られたアノテーションしか利用できない

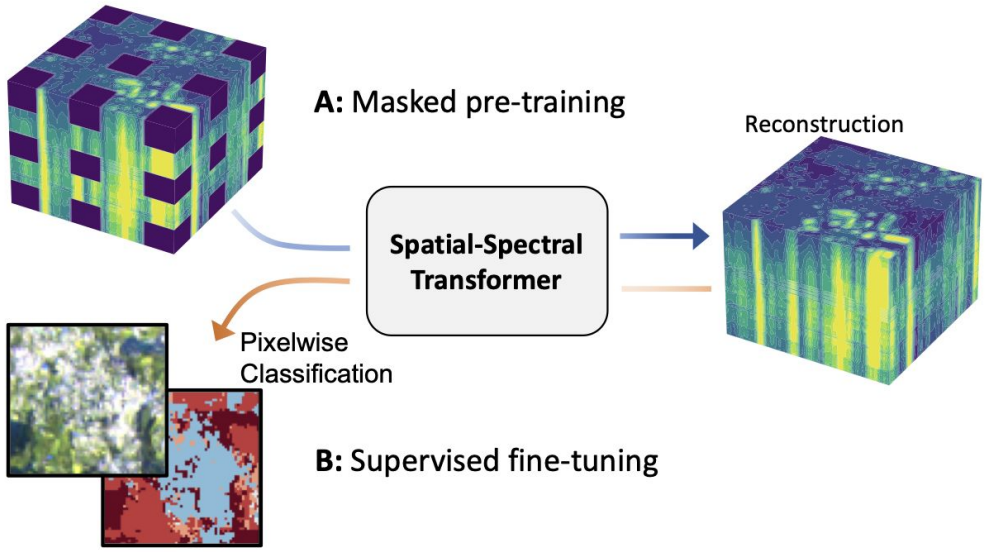




CVPR 2023 の動向・気付き(79/259)

Workshop: Masked Vision Transformers for Hyperspectral Image Classification

- Hyperspectral Imageを用いたVision Transformerの学習方法を検討
  - ブロックワイズパッチ埋め込み
  - 空間-スペクトル自己アテンション
  - スペクトル位置埋め込み
  - マスク自己教師付き事前トレーニング
  - 土地被覆分類の精度に著しい改善を発見



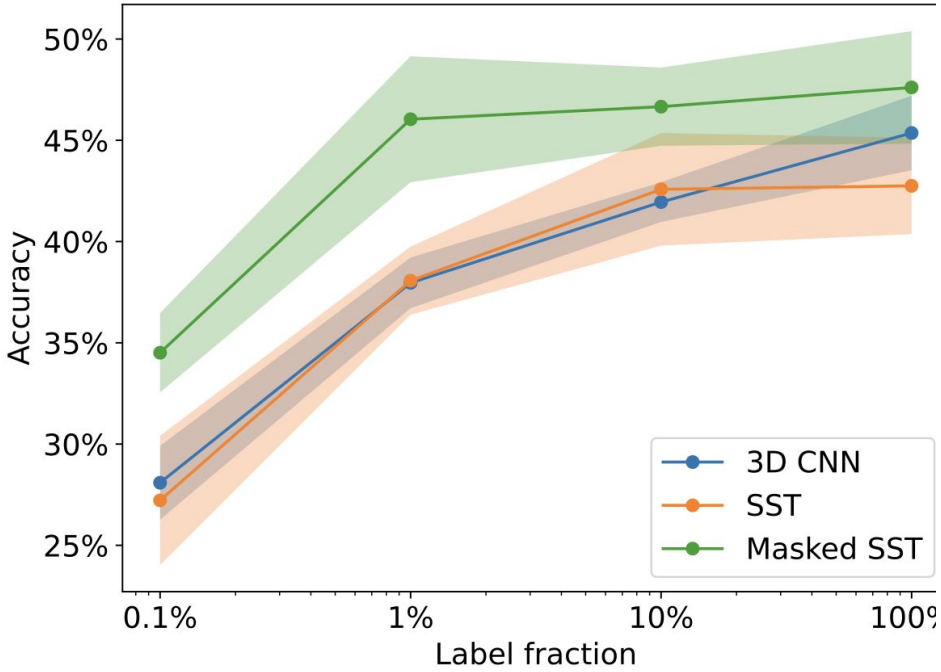


Figure 7. Performance of 3D-CNN baseline [25] and spatial-spectral transformer models trained on varying fractions of the Houston2018 dataset.

CVPR 2023 の動向・気付き(80/259)

Workshop: 2nd Workshop on Learning with Limited Labelled Data for Image and Video Understanding

- 招待公演：Reducing supervision to Enable Detailed Video Understanding
  - Speaker：Hazel Douthy

本講演では，ビデオコンテンツをよりきめ細かく理解するために，限られたラベル付きデータを利用する必要性について紹介する．特に，利用可能なデータを有効に活用することで，必要な監視をさらに減らすドメイン適応 [CVPR 2022, ArXiv] と自己監視 [ECCV 2022, ArXiv] における我々の最近の研究に焦点を当てる．

Bio

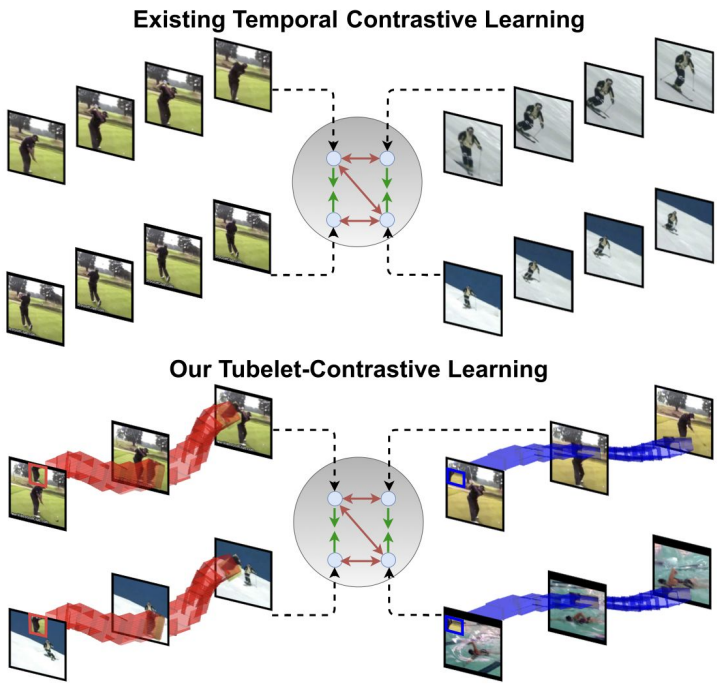
Hazel Doughtyは，Cees Snoek教授のもとで働くアムステルダム大学のポスドク研究者．英国ブリストル大学にてDima Damen教授の指導のもと，博士課程を修了．博士課程では，Inria Parisの客員研究員として数ヶ月を過ごす．9月にライデン大学の助教授として着任予定．研究テーマは映像理解，特にきめ細かい理解と不完全な監視下での学習である．

CVPR 2023 の動向・気付き(81/259)

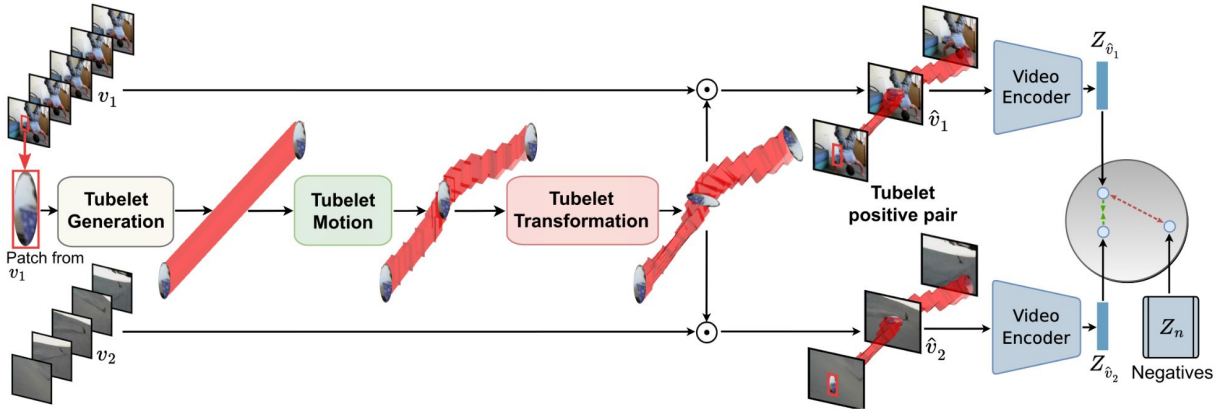
Workshop: 2nd Workshop on Learning with Limited Labelled Data for Image and Video Understanding

- 招待公演：Reducing supervision to Enable Detailed Video Understanding
  - 論文紹介1：Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization (CoPR' 23)
    - 人工的に生成された動き("Tubelets"と呼ばれる)を使用して，ローカルなモーションダイナミクスに焦点を当てたビデオ表現を自己教師ありで学習する新たな方法を提案

動認識のダウンストリー



Tubelet-Contrastive Positive Pairs下は，模擬チューブレット内部の時空間運動ダイナミクスを共有するだけで，時間コントラストペア．上は高い空間バイアスに悩まされる．チューブレットを対比させることで，データ効率がよく，一般化可能な映像表現が得られる．

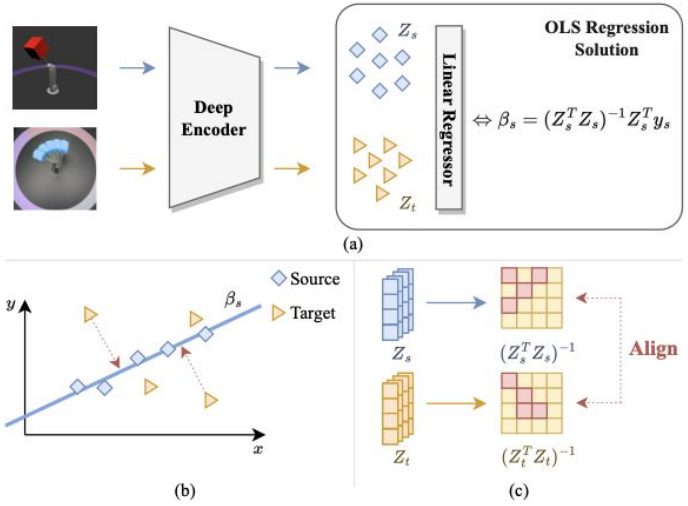


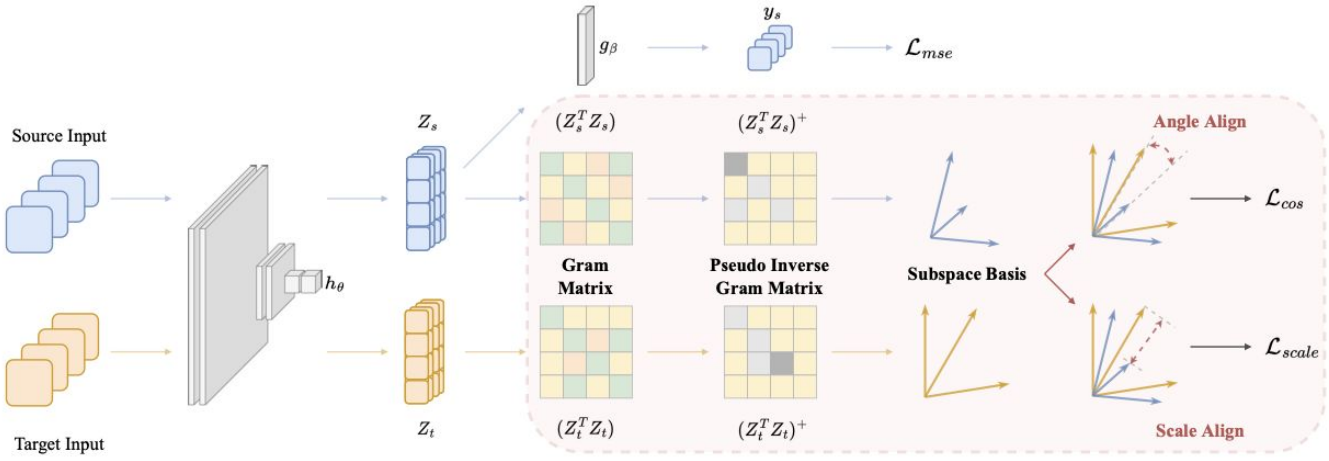
Tubelet-Contrastive Learning

CVPR 2023 の動向・気付き(82/259)

Workshop: DARE-GRAM : Unsupervised Domain Adaptation Regression by Aligning Inverse Gram Matrices

- 教師なしドメイン適応回帰問題に対する新しい視点を提供し，特徴の逆グラム行列を整列させることでソースとターゲットドメイン間のギャップを埋める新しい手法DARE-GRAMを提案
  - 新規性は，従来の特徴空間の整列に代わり，特徴の逆グラム行列を整列するという新しい視点を提供している点
  - 一般化最小二乗(ordinary least squares, OLS)解の中に存在する逆グラム行列と，グラム行列が特徴相関を捉える能力にヒントを得ている

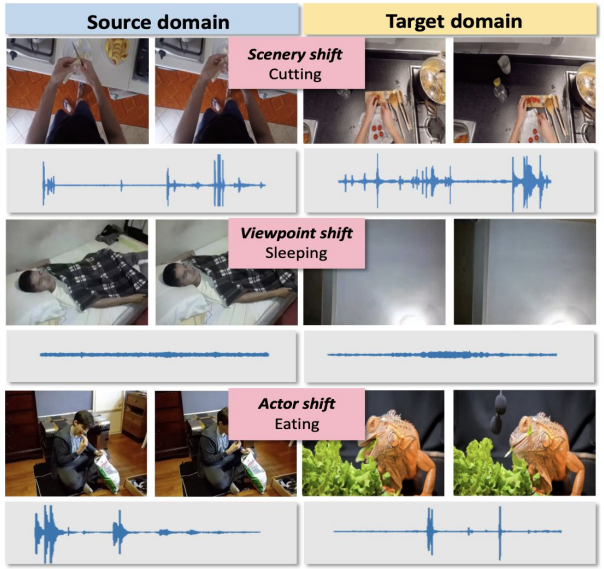




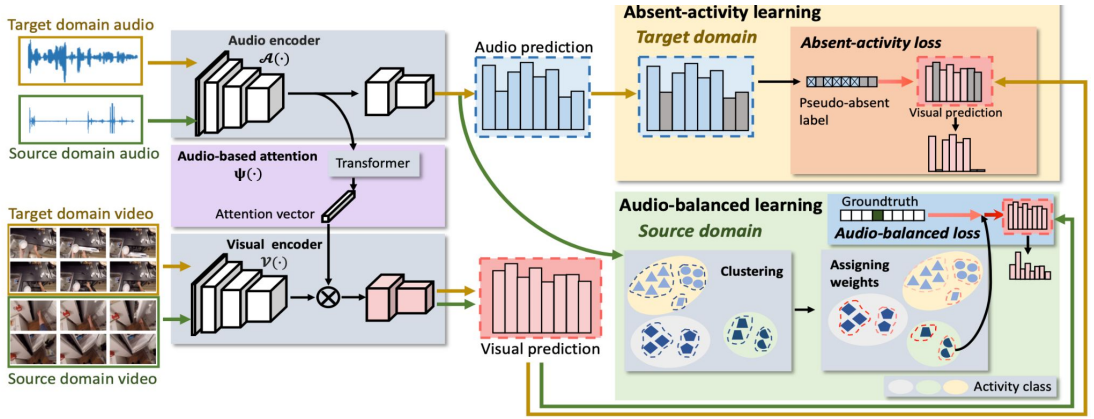
CVPR 2023 の動向・気付き(83/259)

Workshop: 2nd Workshop on Learning with Limited Labelled Data for Image and Video Understanding

- 招待公演：Reducing supervision to Enable Detailed Video Understanding
  - 論文紹介2：Audio-Adaptive Activity Recognition Across Video Domains (CVPR' 22)
    - 風景やカメラ視点の変更，または俳優の変更によるドメインシフトの下での活動認識を行うために，ドメイン間での変動が少なく，信頼性の高い指標となる活動音を利用した音声適応型エンコーダと学習方法を提案
    - さらにドメイン固有の特徴を排除し，ドメイン不変の活動音を活用した認識器を提案し，新たなタスクを導入してその効果を示し



風景やカメラの視点及び役者が変わることで発生するDomain shiftの活動を音の力を借りて認識

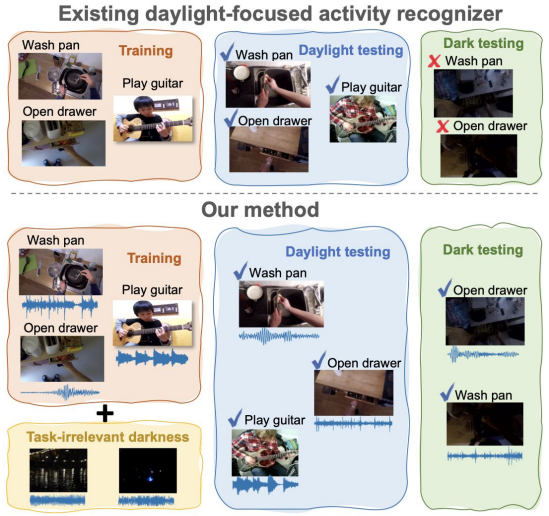


Audio-adaptive encoder for activity recognition under domain shift.

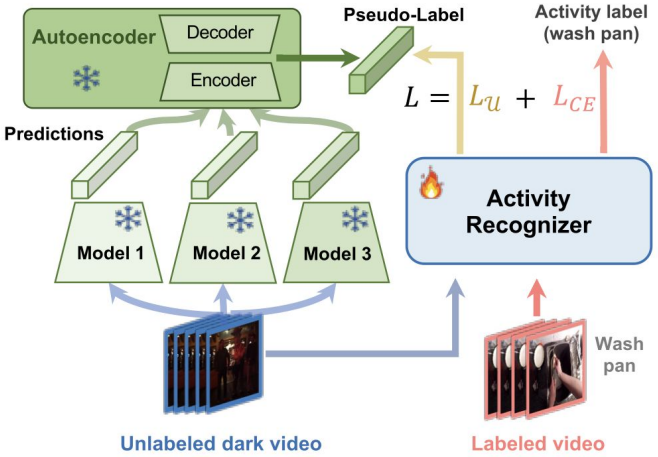
CVPR 2023 の動向・気付き(84/259)

Workshop: 2nd Workshop on Learning with Limited Labelled Data for Image and Video Understanding

- 招待公演：Reducing supervision to Enable Detailed Video Understanding
  - 論文紹介3：Day2Dark: Pseudo-Supervised Activity Recognition beyond Silent Dayligh (CoPR' 22)
    - ラベル付けされていない暗闇のビデオを活用する擬似教師あり学習戦略と，照明に応じてオーディオビジュアルの統合方法を変更する「闇の認識」を導入した新しい活動認識モデルを提案
    - 提案したモデルが低照明環境でも効果的にActivityを認識できること，また日中の遮蔽に対してもロバストであることを確認



Activity認識装置は明るい映像に重点を置いているため，暗いところでは分布のずれや低コントラストのために失敗する．
本手法は，ラベルのないタスクに無関係な暗い映像を使用し，照明に影響されない音と外観を適応的に融合させることで，暗闇に適応する．

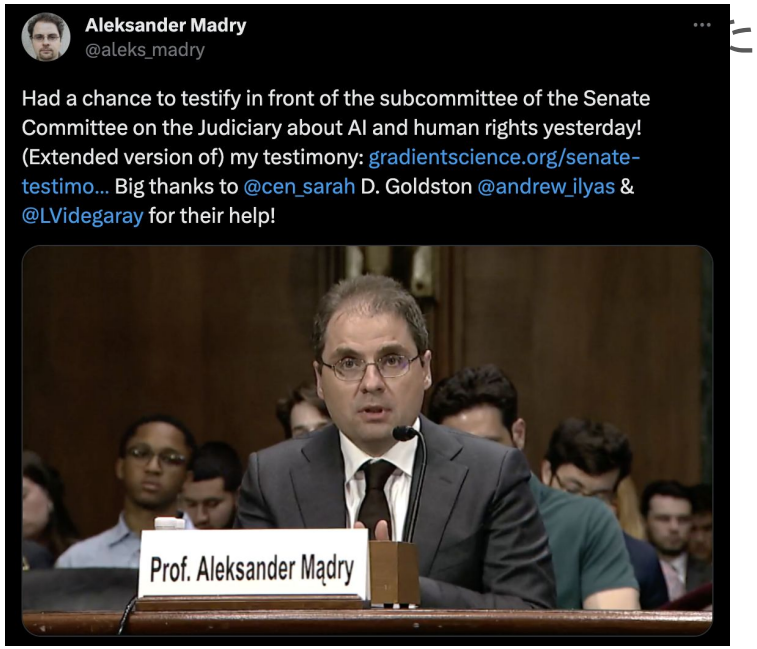


Pseudo-Supervised Day2Dark Learning

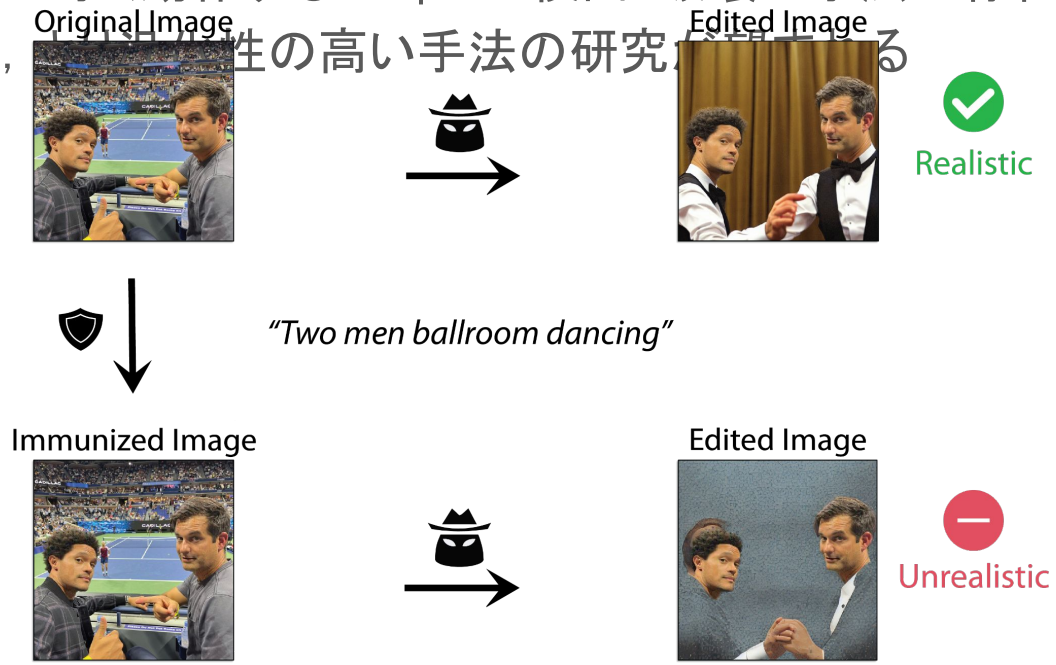
CVPR 2023 の動向・気付き(85/259)

Workshop: Workshop of Adversarial Machine Learning on Computer Vision: Art of Robustness

- 生成モデルの発展と共にDeepfakeの被害が実社会で急速に広がっている
  - Deepfakeを使用した脅しのメール(blackmail)の被害は(少なくともアメリカ国内では)既に実社会のあらゆる場所で発生しているとのこと
  - 特定の生成モデルに対して上手く動作するDeepfake検出・破壊の手法は存在するが社会実装としては[illegible]ため,[illegible]性の高い手法の研究[illegible]る



https://twitter.com/aleks_madry/status/1669044821779156992



https://gradientscience.org/photoguard/

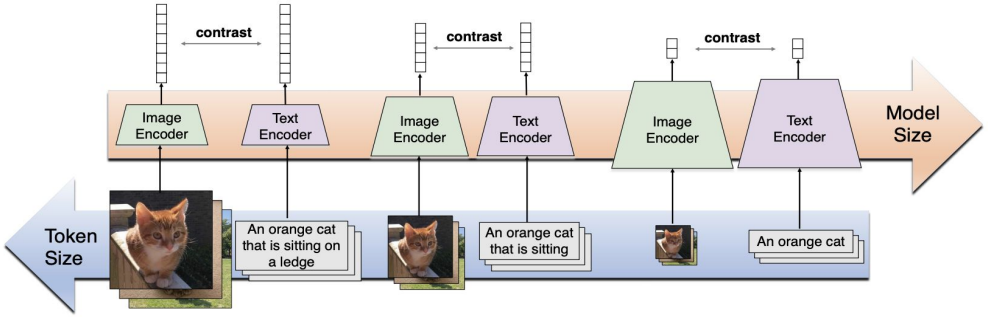
CVPR 2023 の動向・気付き(86/259)

Workshop: Workshop of Adversarial Machine Learning on Computer Vision: Art of Robustness

- 限られた予算でCLIPを学習するのは可能か？-> “頑張れば可能”

: An Inverse Scaling Law for CLIP Training: https://arxiv.org/abs/2305.07017

- CLIPはOut of distributionに強くロバスト性の研究においても重要
- OpenCLIPのようにオープンソースのCLIP再現実装もあるが，学習に必要なコストが高く限られた予算では学習が現実的ではない？
- しかし，CLIPはencoderのサイズが大きいほど，少ない画像・テキストで効果的に学習出来る経験則(Inverse scaling law)を発見した



Models	Training Epochs	GPU Hours	zero-shot classification						zero-shot retrieval			
			IN-1K	IN-V2	IN-A	IN-R	ObjectNet	IN-Sketch	COCO		Flickr30k	
									image	text	image	text
OpenAI-B/32, Our Eval	32	4600	63.3	55.9	31.6	69.3	44.2	42.3	30.4	50.2	58.9	77.6
OpenAI-B/16, Our Eval	32	10700	68.3	61.9	49.9	77.7	55.3	48.2	33.1	52.4	62.1	81.9
OpenAI-L/14, Our Eval	32	50800	75.5	69.8	70.8	87.8	68.9	59.6	36.5	56.4	65.3	85.1
OpenCLIP-B/32, Our Eval	32	4600	62.9	55.1	21.7	73.4	48.9	49.4	35.3	52.6	61.7	79.0
OpenCLIP-B/16, Our Eval	32	10700	67.1	59.6	33.2	77.9	51.5	52.4	38.3	55.4	65.5	83.3
OpenCLIP-L/14, Our Eval	32	50800	72.8	65.4	46.5	84.9	59.9	59.6	43.0	59.7	70.3	87.6
CLIPA-B/16 (I50,T16)	6+0.36	444	63.2	55.6	26.8	73.2	44.3	48.7	35.2	53.1	58.3	75.3
CLIPA-L/16 (I17,T16)	6+0.36	628	67.8	60.4	38.3	81.2	52.8	56.4	40.1	58.4	64.0	81.5
CLIPA-L16 (I37,T8)	6+0.36	826	69.3	61.7	43.6	84.0	55.4	58.7	39.8	56.8	67.5	81.9

CVPR 2023 の動向・気付き(87/259)

Workshop: Workshop of Adversarial Machine Learning on Computer Vision: Art of Robustness

- LLMの頑健性・信頼性の向上は社会的にも急務なタスク
  - ChatGPTを初めとするLLMを利用したサービスの急速な社会実装が進む一方で，Jailbreak等のLLMに対する攻撃対策は追いついていないのが現状

CVPR 2023 の動向・気付き(88/259)

Workshop: Women in Computer Vision

- 招待公演: Dr. Devi Parikh
  - 生成モデルの様々なモダリティへの拡張(テキスト, 画像, 動画, 音声, 3Dなど)
  - 紹介論文: MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and GENeration (ECCV22)
    - ゲーム動画から言語・ビデオ・音声のデータセットを自動構築するスの

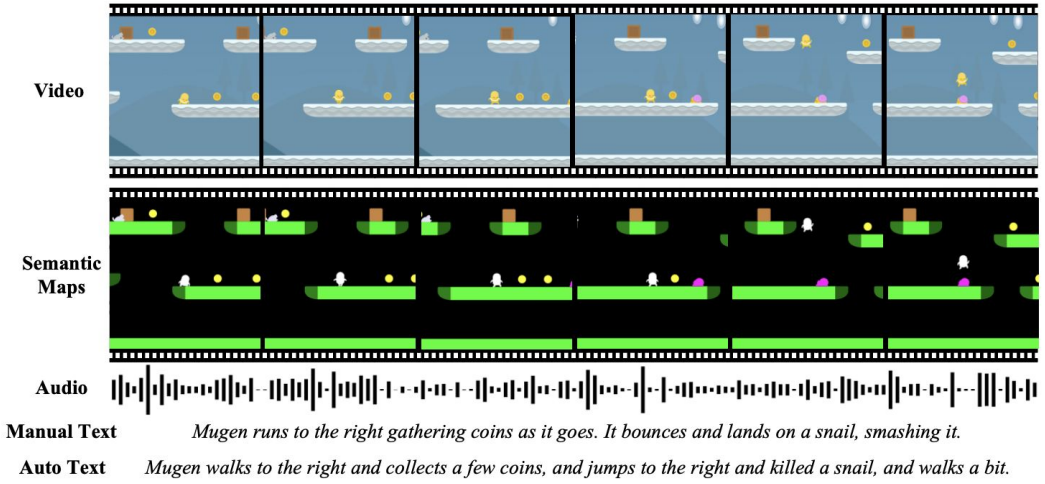


Fig. 1: An example from our dataset. For each 3.2s video clip, we have rich annotations including accurate semantic maps, synchronized audio, manual text collected from human annotators, and auto-text generated based on certain rules.

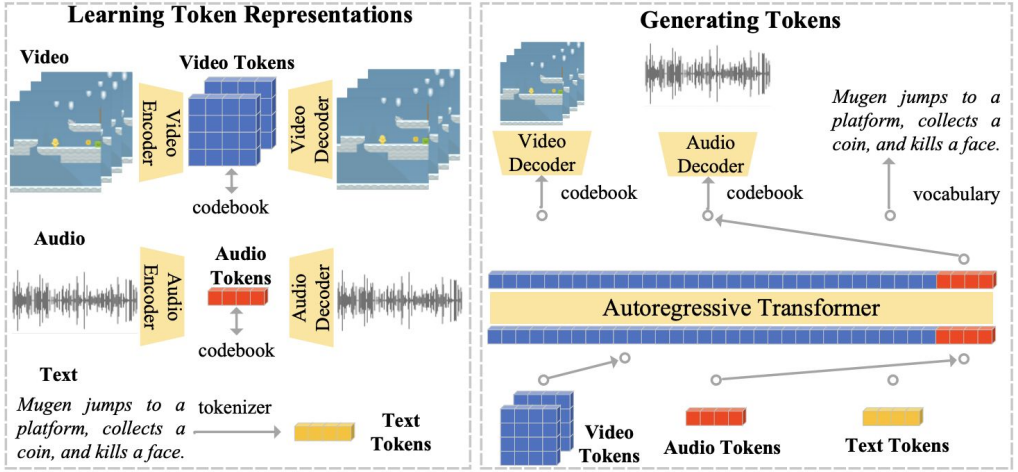


Fig. 3: A unified framework for generation between every pairs of modalities. The right part shows an example of video-to-audio generation.

MUGANより引用

CVPR 2023 の動向・気付き(89/259)

Workshop: Women in Computer Vision

- 招待公演: Dr. Judy Hoffman
  - 冗長性を取り除くモデル開発
  - 紹介論文1: Token Merging: Your ViT But Faster
    - ViTに入力するトークンを組み合わせる.(プルーニングよりも高速)
    - ViTのattention kye(K)の類似度を使用して,トークンが類似するものをマージする.

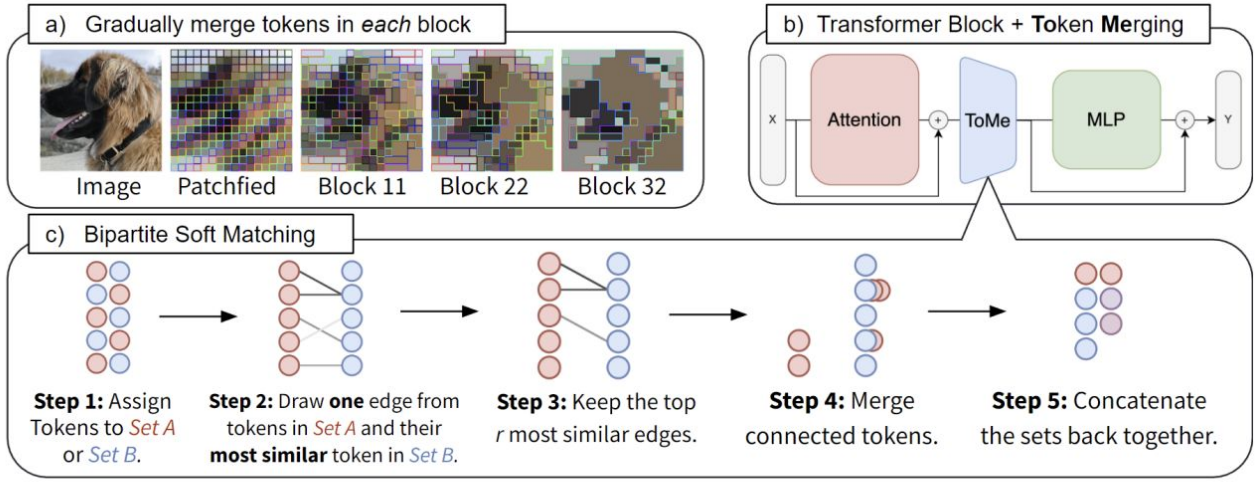
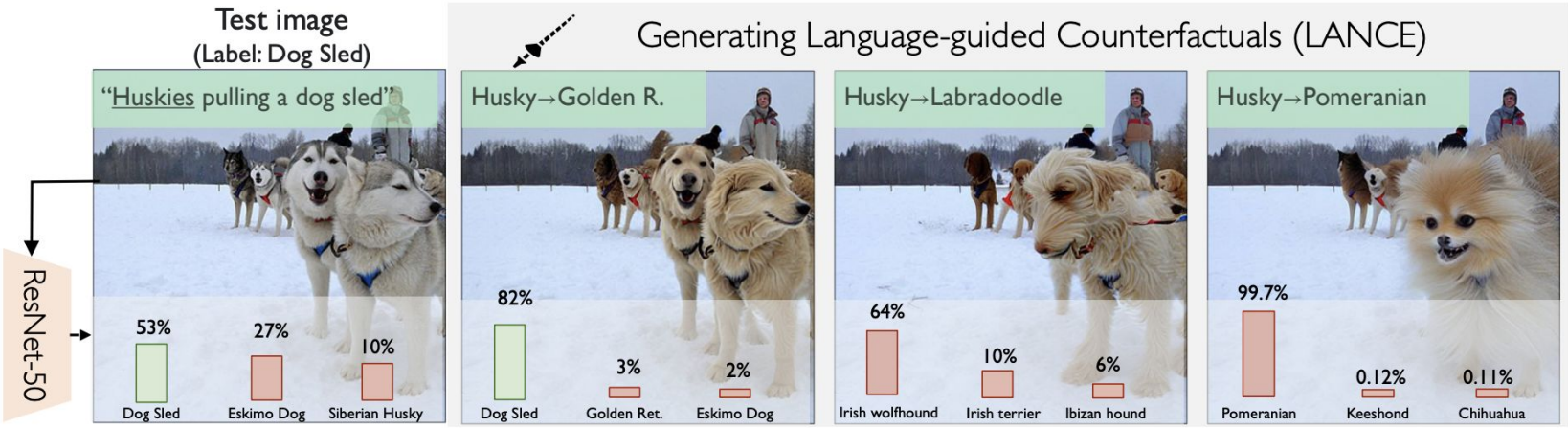


Figure 1: **Token Merging.** (a) With ToMe, similar patches are merged in each transformer block: for example, the dog's fur is merged into a single token. (b) ToMe is simple and can be inserted inside the standard transformer block. (c) Our fast merging algorithm, see Appendix D for implementation.

CVPR 2023 の動向・気付き(90/259)

Workshop: Women in Computer Vision

- 招待公演: Dr. Judy Hoffman
  - 紹介論文2: LANCE: Stress-testing Visual Models by Generating Language-guided Counterfactual Images
    - もとのテスト画像からそれの説明文を生成し，一つのAttributeを変更し(例: ハスキー→ポメラニアン)，それをまた生成モデルにいれることで，新しいテストデータを生成する.
    - 提案手法で生成されたテストデータを使うことで，ImageNetのテスト結果が低下(つまりタスクを難しくすることに成功)

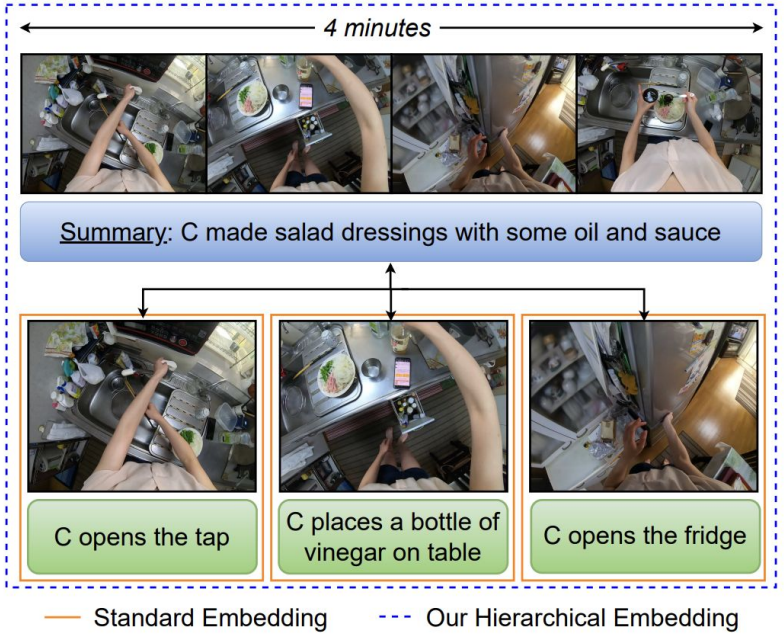


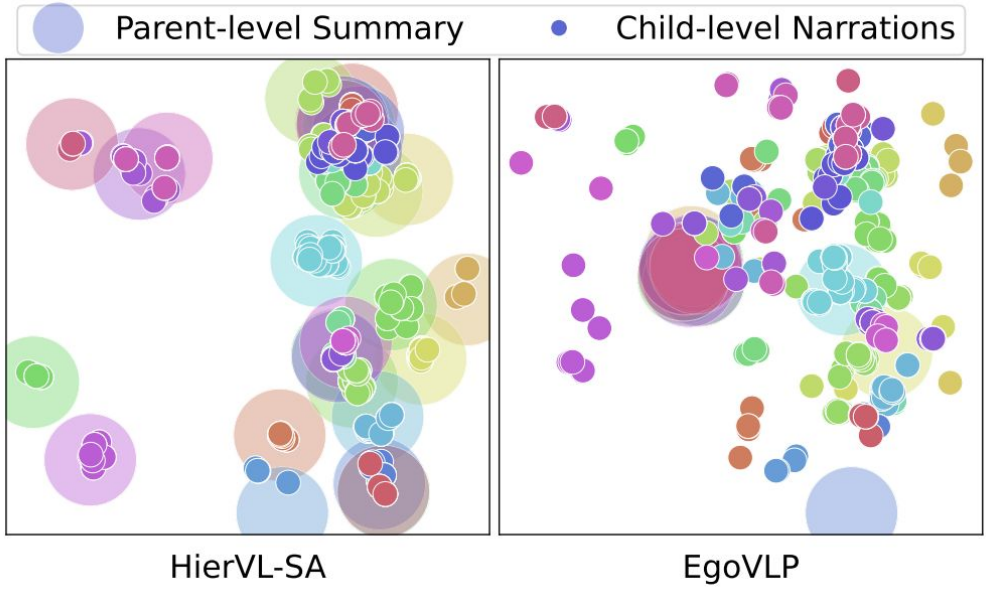
することに貢献できる.

CVPR 2023 の動向・気付き(91/259)

Workshop: Women in Computer Vision

- 招待公演: Dr. Kristen Grauman
- 1人称視点研究の紹介(ロボットへの応用や，拡張現実に向けて)
- 紹介論文: HierVL: Learning Hierarchical Video-Language Embeddings (CVPR23)
  - 行動の説明(what)と要旨(why)の間に階層構造を見出す
  - 長い時間の動画説明に，短い動画からのリッチなreasoningを使用することで，より高度な





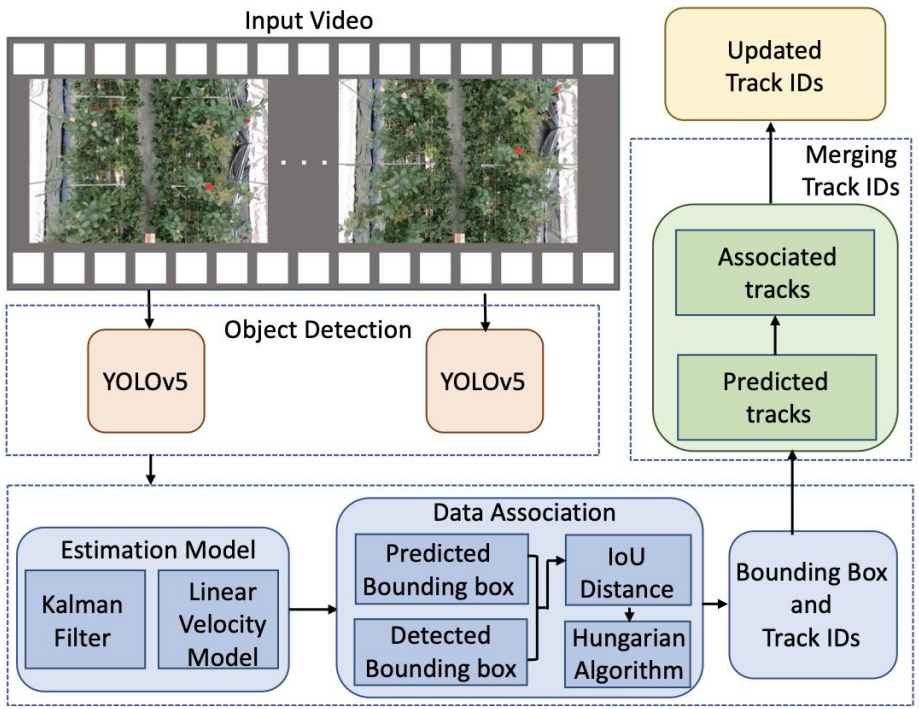
CVPR 2023 の動向・気付き(92/259)

Workshop: Women in Computer Vision

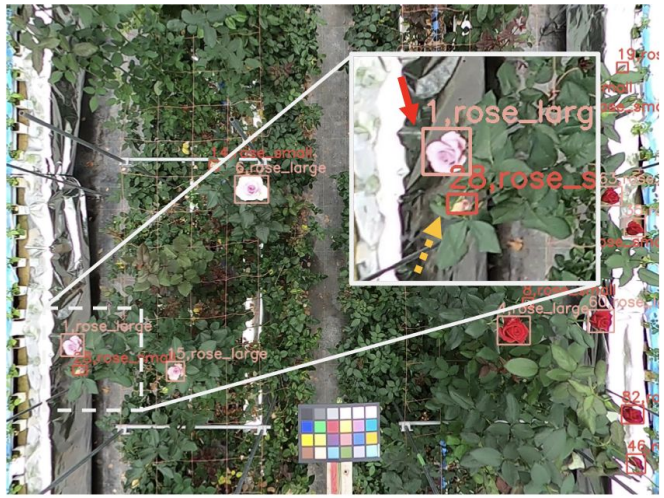
- A dataset and system for automated rose growth monitoring

(Risa Shinoda , Ko Motoki, Kensho Hara, Hirokatsu Kataoka, Ryohei Nakano, Tetsuya Nakazaki, Ryozo Noguchi)

  - バラの自動モニタリングシステムの開発
  - 葉によるオクルージョンが激しいバラの成長ステージ別検出を可能にする．物体検出＋物体追跡＋オリジナルの回帰モデルを使用することで，さまざまな角度からバラを検出する．



Frame 119

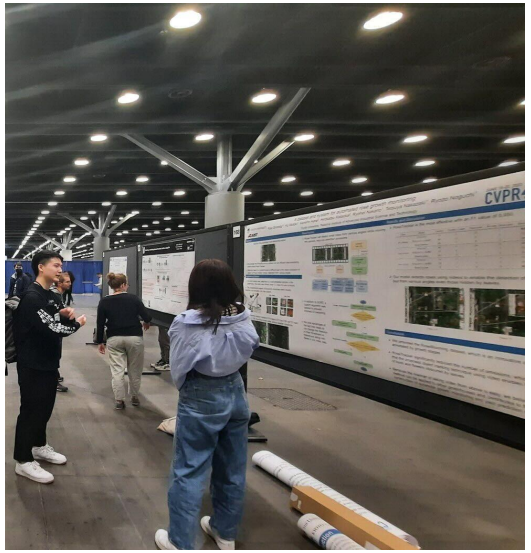


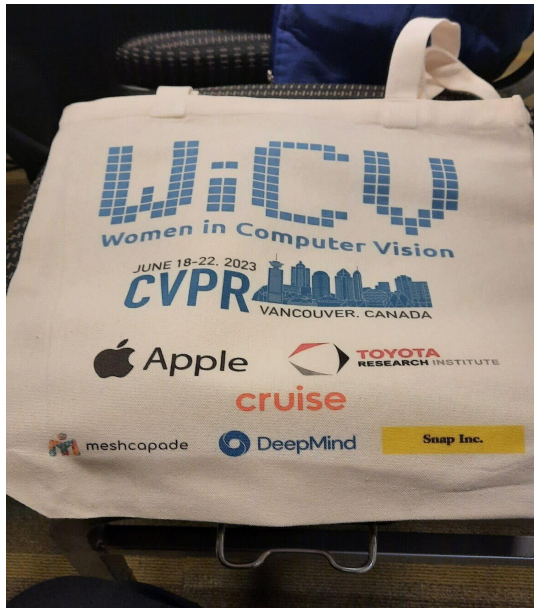
Frame 70

CVPR 2023 の動向・気付き(93/259)

Workshop: Women in Computer Vision

- WiCV初参加であったが，コミュニティの連携が強いと感じた
- 著者はディナーに招待され，招待講演の著者や，他の採択者と話せる





（WiCV Twitterより写真を拝借）

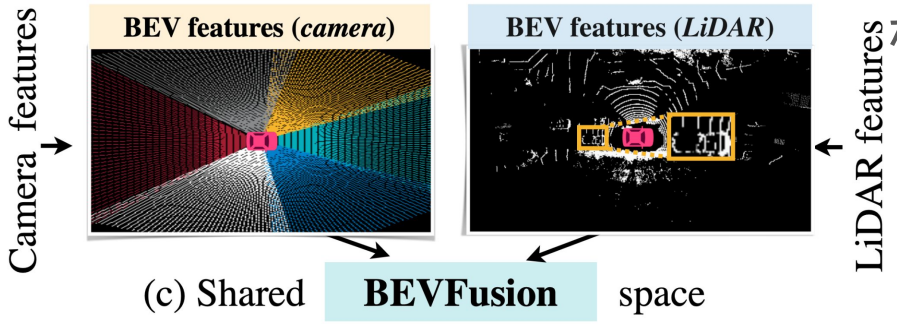
CVPR 2023 の動向・気付き(94/259)

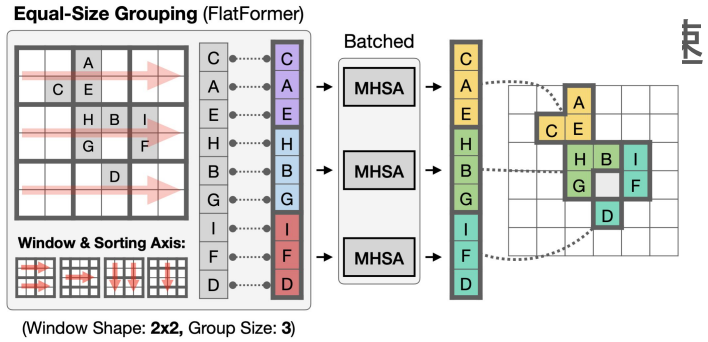
Workshop: Efficient Deep Learning for Computer Vision (inv. talk)

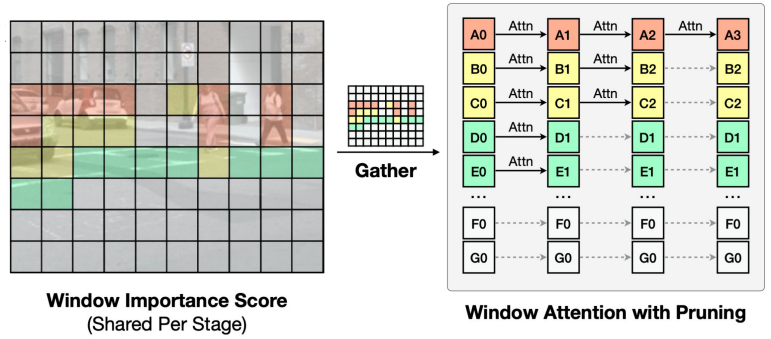
- Efficient 3D Perception for Autonomous Vehicles

  (Speaker: Zhijian Liu)

  - 背景：自動運転車における3次元物体認識のレイテンシ
  - 手法：
    - BEVFusion：様々なセンサーから得られる情報を鳥瞰図(BEV)の空間に統合し，主要なview transformationを40倍高速化
    - FlatFormer：点群変換における空間的な規則性に着目してAttention計算を工夫することで，sparse CNNより高精度かつ高速で，エッジGPU上でリアルタイム性能を達成
    - SparseViT：スパース性を考慮したアダプテーションを導入し，進化的探索を適用することで，







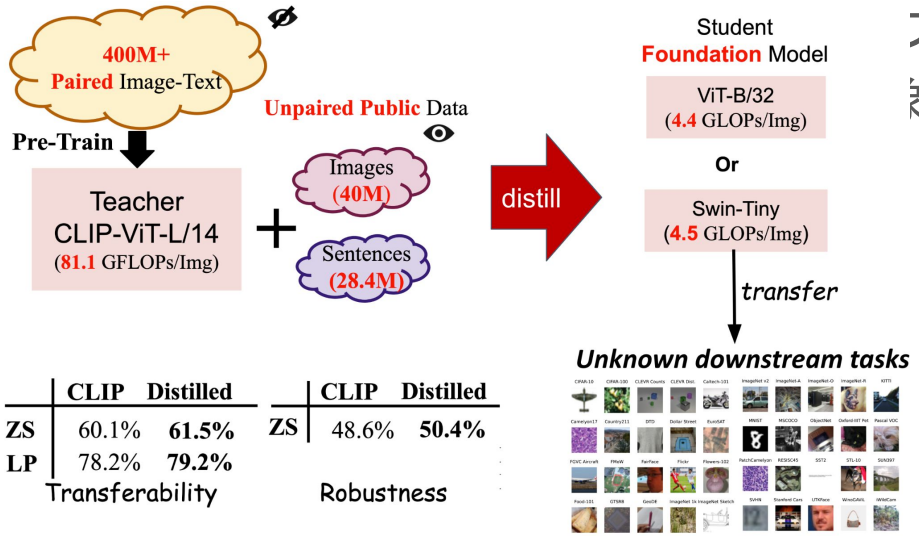
CVPR 2023 の動向・気付き(95/259)

Workshop: Efficient Deep Learning for Computer Vision (inv. talk)

- CLIP for Less: Efficient Training and Adaptation of Foundational Models

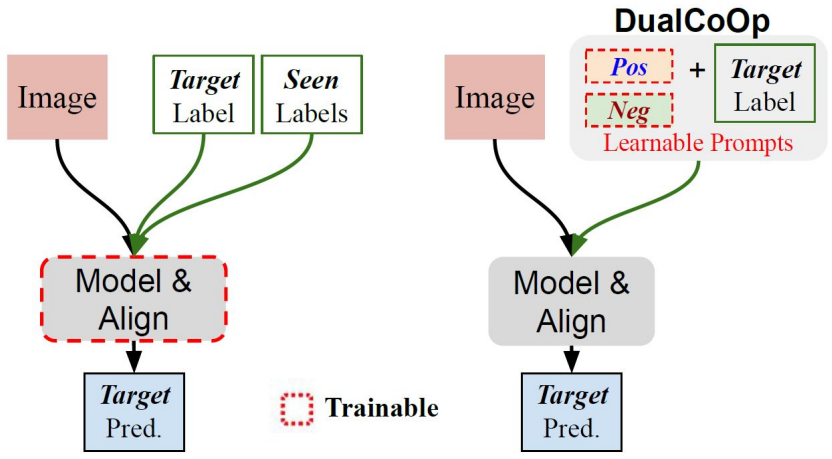
  (Speaker: Kate Saenko)

  - 背景：CLIP登場以降のfine-tuning→zero-shotへのトレンド変遷
  - 手法：
    - DIME-FM：CLIPを蒸留し，比較的小規模なモデルかつUnpairdなデータセットで高水準なゼロショット性能を達成



プトの学習(p

認識タスクを



(b) Previous Zero-Shot MLR (c) Limited-Annotation MLR (Ours)

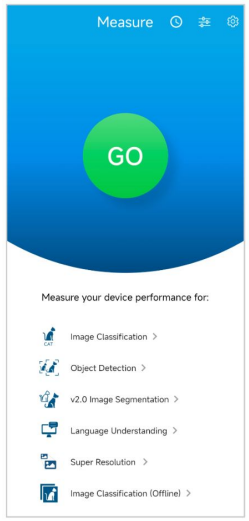
CVPR 2023 の動向・気付き(96/259)

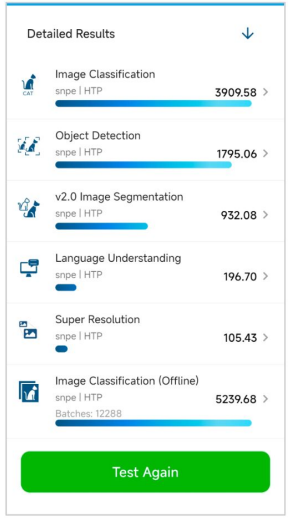
Workshop: Efficient Deep Learning for Computer Vision (inv. talk)

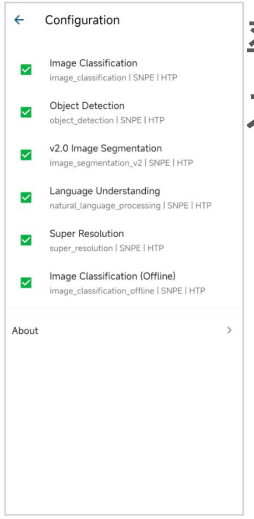
- Mobile Super Resolution: MLPerf App – Benchmarking and Challenges

  (Speaker: David Tafur & Sanghyun Son)

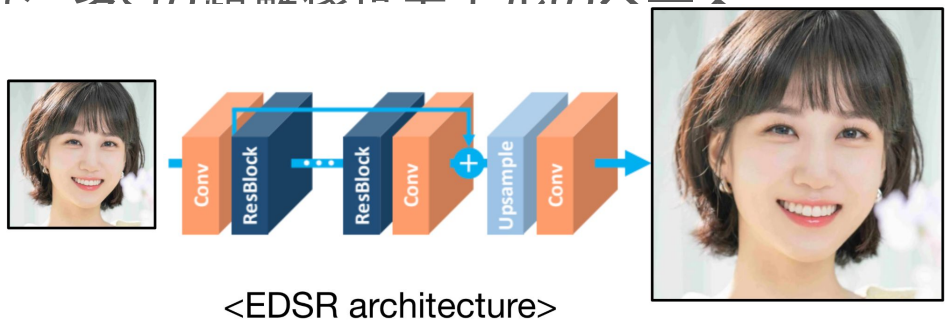
  - 背景:MobileAIのパフォーマンス評価の難化(デバイス, ソフトウェア, 設定の差異)
  - 手法:
    - MLPerf Mobile App:統一的にMobileAIのパフォーマンス比較を行えるBenchmarkを公開し, 公正な研究を促進
    - EDSR:超解像度タスクのBenchmarkとして用いられたモデル







率的で 多くの超解像度モデルのベース
スク

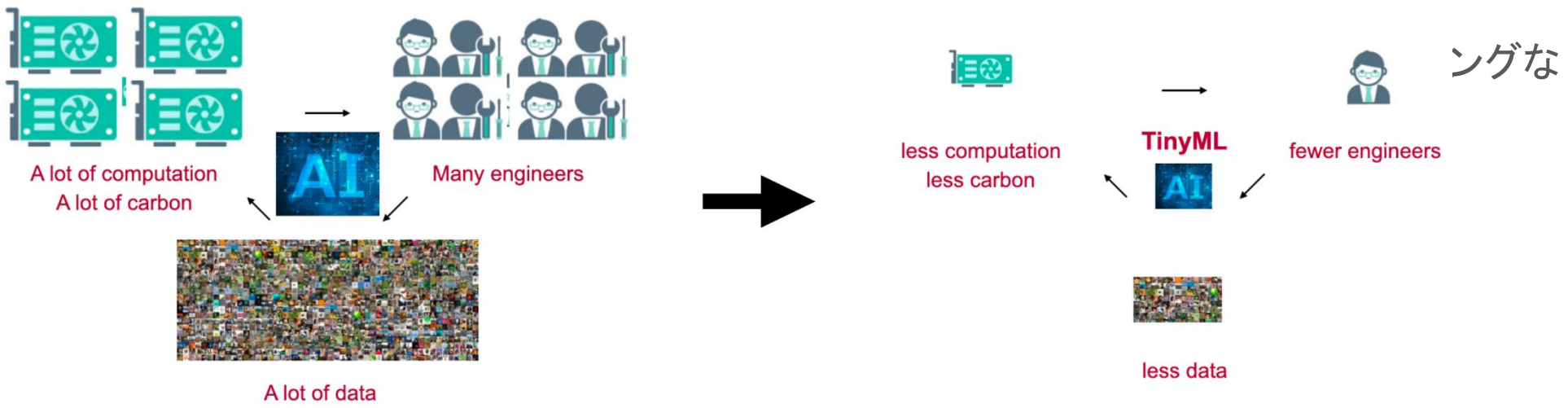


<EDSR architecture>

CVPR 2023 の動向・気付き(97/259)

Workshop: Efficient Deep Learning for Computer Vision (inv. talk)

- Efficient Deep Learning Computing with Sparsity
  (Speaker: Song Han)
  - 背景:深層モデルの大規模化による学習/推論コストの増大
  - 手法:
    - 生成モデルの高速化
      - FastComposer, GAN Compression, DiffAugment, Anycost GANs
    - LLMのメモリ効率化



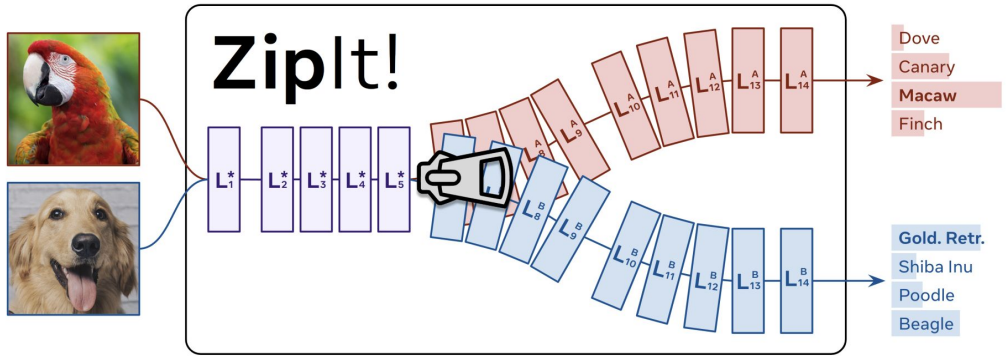
ングな

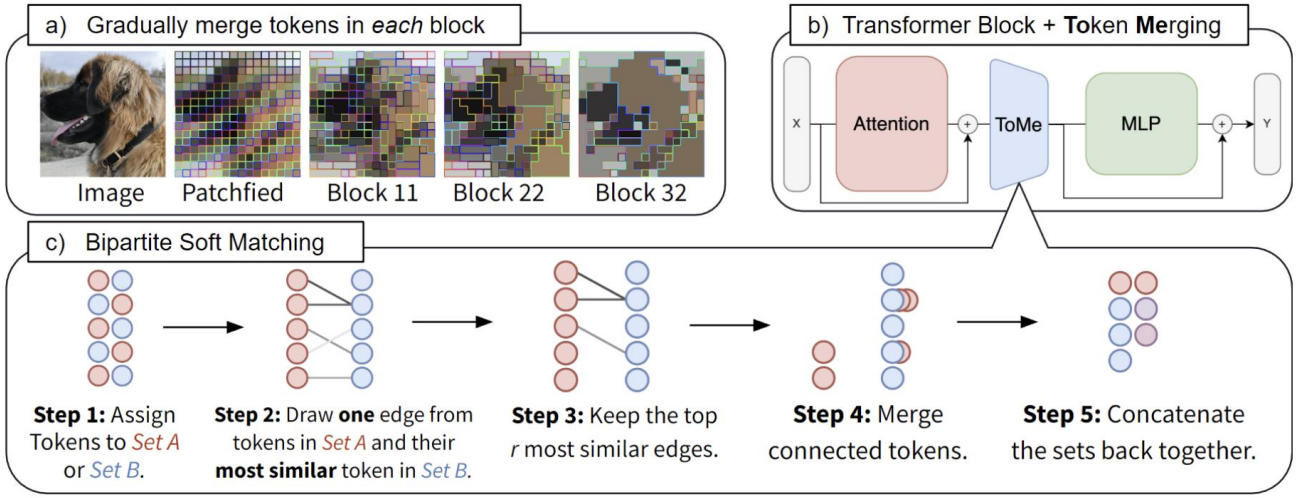
CVPR 2023 の動向・気付き(98/259)

Workshop: Efficient Deep Learning for Computer Vision (inv. talk)

- Increasing Efficiency by Reducing Redundancy

  (Speaker: Judy Hoffman)
  - 背景:深層モデルの大規模化による学習/推論コストの増大
  - 手法:
    - ZipIt!:複数の事前学習済みモデルを統合し,
      再学習することなく統合された汎化性を発揮する単一モデルを生成
    - ToME:処理前に類似したトークンを見つけてマージすることで
      回帰計算を明示的に削減し,

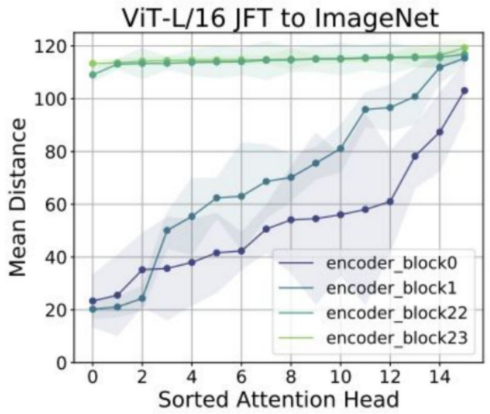


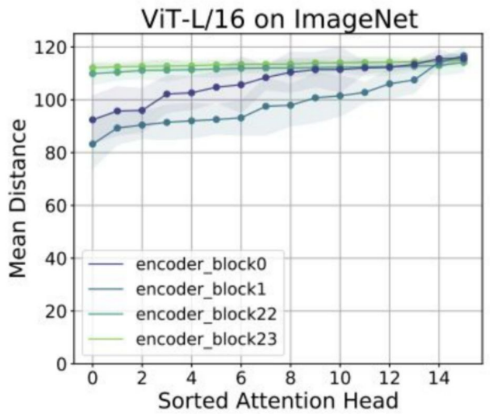


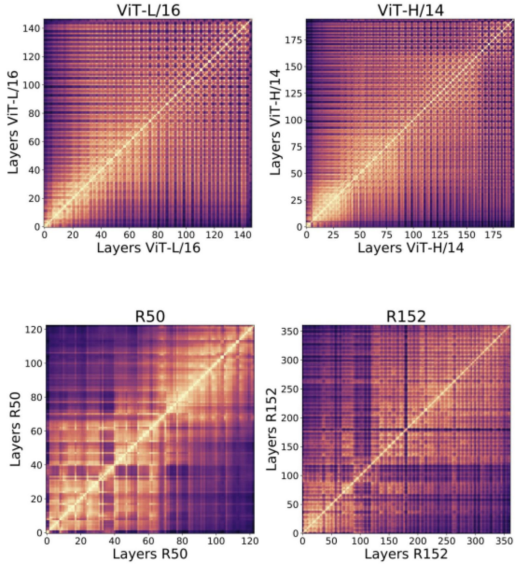
CVPR 2023 の動向・気付き(99/259)

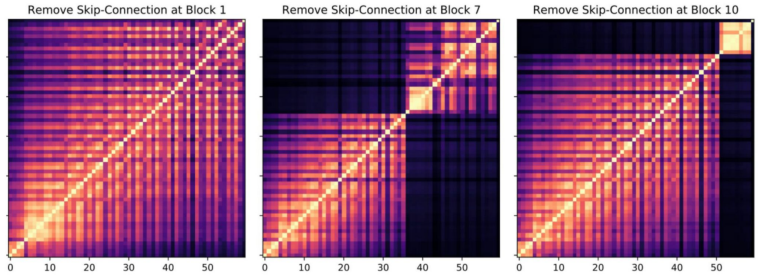
Workshop: All Things ViTs: Understanding and Interpreting Attention in Vision (1/4)

- Probing Vision Transformers
  - Mean attention distance (MAD)
    - 画像のピクセル間の関係性の強さを，ピクセルの物理的な距離とAttention値を掛け合わせ導出
    - 大規模データセットで事前学習すると，局所的特徴と大域的特徴の双方をカバー可能
      - 高層では大域的特徴のみをカバー，低層では双方をカバー
  - モデル中の層同士の表現空間の類似性を比較
    - ViTは全層通して均一，CNNは低層と高層で分離
    - Skip ConnectionがViTの表現空間(





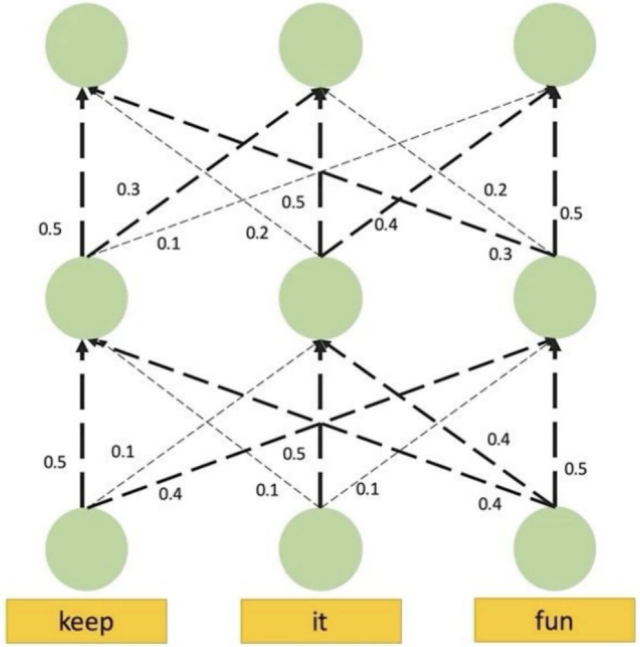


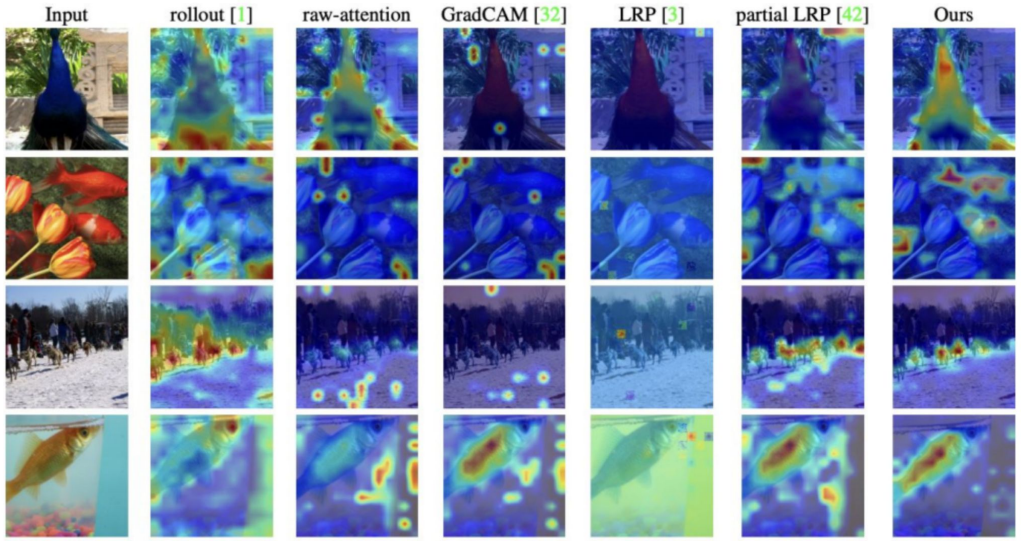


CVPR 2023 の動向・気付き(100/259)

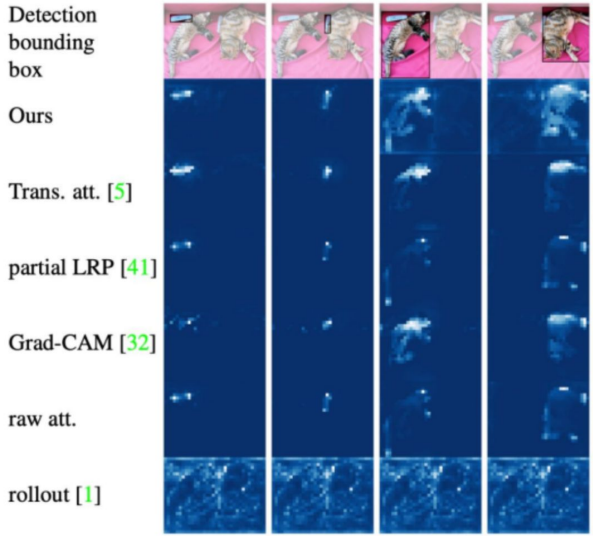
Workshop: All Things ViTs: Understanding and Interpreting Attention in Vision (2/4)

- Explaining Transformers
  - Transformerの大規模化/複雑化によって，モデルの説明性の担保が困難に
  - Transformerに合わせた説明手法が台頭
    - Attention Rollout：コンテキスト集約のためにlayer wiseにAttention mapの行列を乗算
    - Attention Flow：コンテキスト集約のためにlayer wiseにAttention graphの最大流問題を解く
    - TiBA，GAE：Grad-CAMのようにgradientを用いてAttention以外のネットワークも考慮





result of TiBA



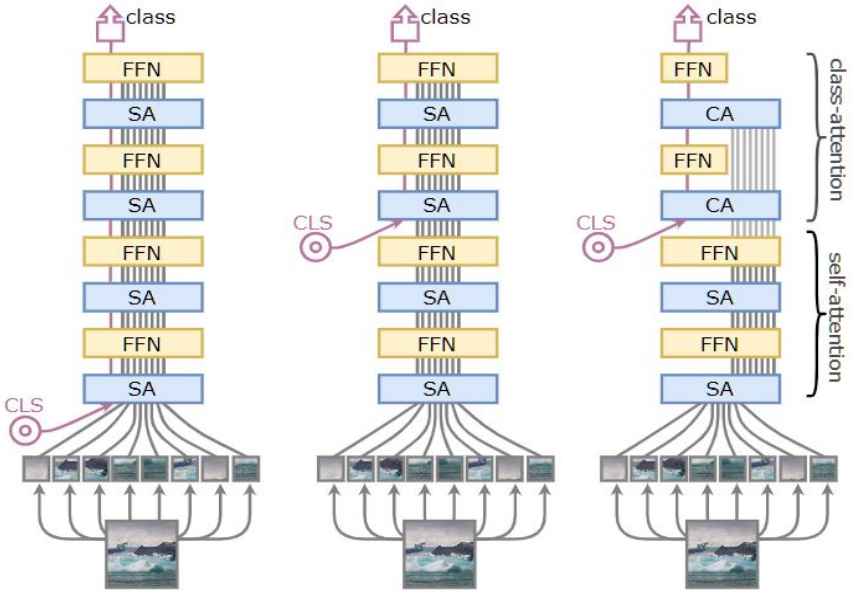
result of GAE

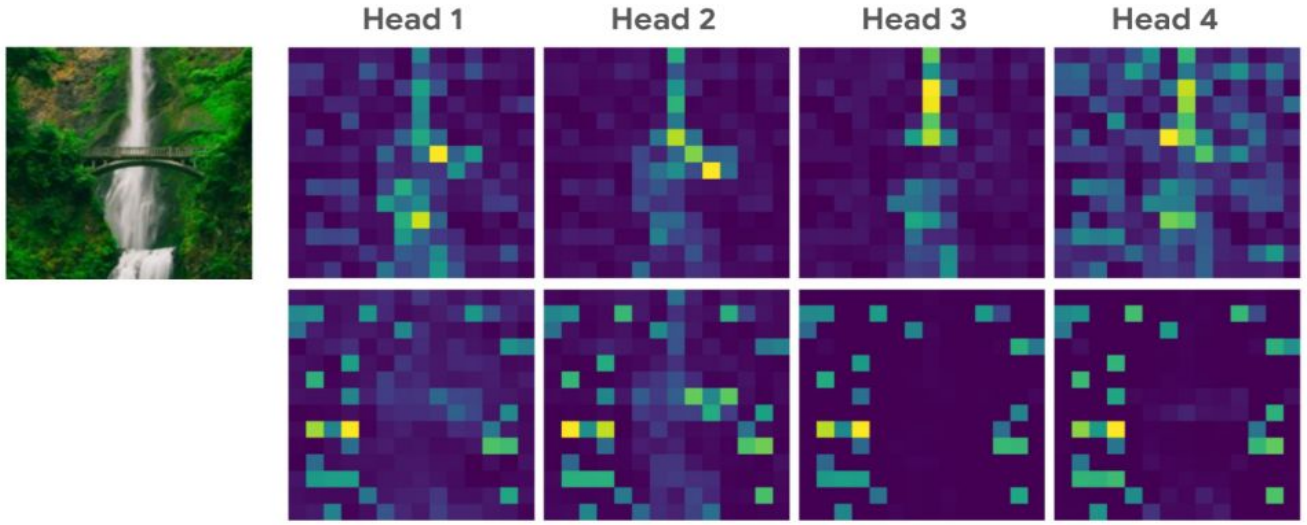
CVPR 2023 の動向・気付き(101/259)

Workshop: All Things ViTs: Understanding and Interpreting Attention in Vision (3/4)

- Attention as explanation (CaiT)
  - CLSを挿入する際にパッチ埋め込みを凍結することで計算量を節約
  - CaiTはクラス-アテンション段階でクラストークンとパッチとの間に相互に作用
  - 分類の判断が行われる画像の主要部分に対応し，興味のあるオブジェクトに明確に注目

判断根拠を示す一つの指標として利用可能





CVPR 2023 の動向・気付き(102/259)

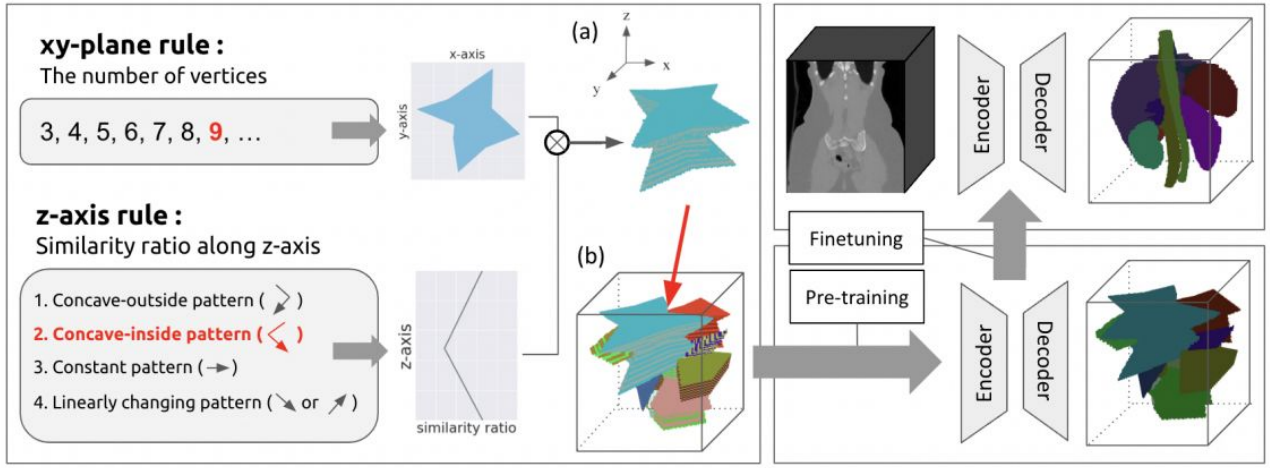
Workshop: All Things ViTs: Understanding and Interpreting Attention in Vision (4/4)

- Attention for Downstream Tasks (Latent Diffusion Models)
  - エンコーダは，入力を埋め込み空間にマッピング
  - 潜在空間には拡散モデルが適用
  - →学習コストが低く，推論を高速化

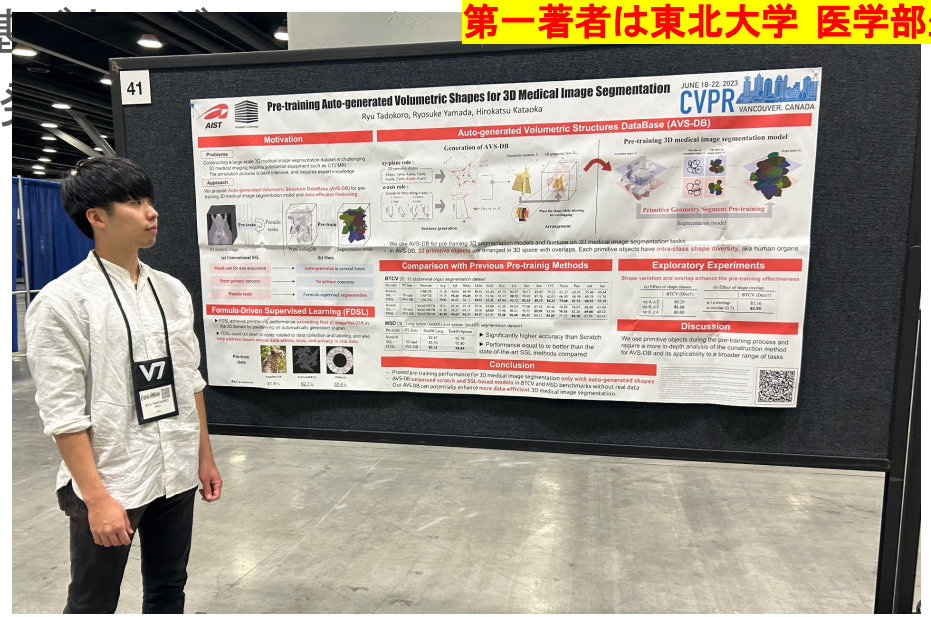
CVPR 2023 の動向・気付き(103/259)

Workshop: Efficient Deep Learning for Computer Vision (poster)

- Pre-training Auto-generated Volumetric Shapes for 3D Medical Image Segmentation
  - 目的:3D医療セグメンテーションのための事前学習データ&方法を提案
  - 手法:
    - AVS-DB:xy平面とz軸の異なる規則に基づいて3次元構造を生成,
      複数オブジェクトを組み合わせることで事前学習用3Dデータを生成
    - 学習方法: 規則に応じて与えられたカテゴリラベルに基



第一著者は東北大学 医学部生!!



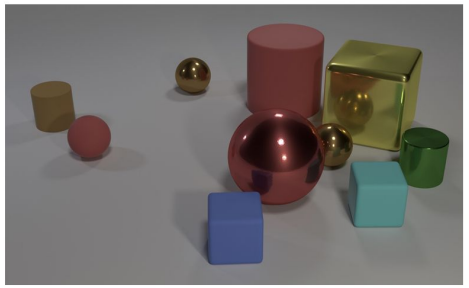
https://openaccess.thecvf.com/content/CVPR2023W/ECV/papers/Tadokoro_Pre-Training_Auto-Generated_Volumetric_Shapes_for_3D_Medical_Image_Segmentation_CVPRW_2023_paper.pdf

CVPR 2023 の動向・気付き(104/259)

Workshop: O-DRUM Workshop @ CVPR 2023

Open-Domain Reasoning Under Multi-Modal Settings

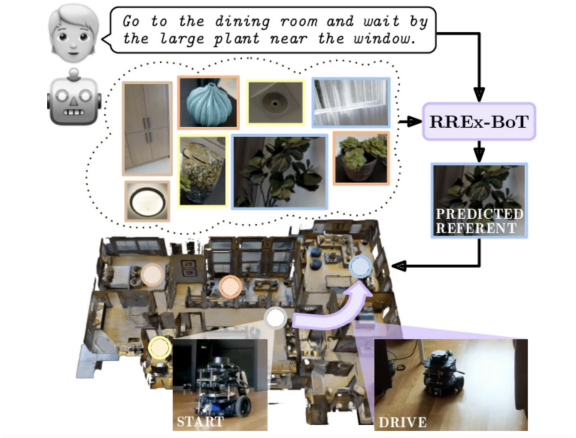
- **Jiajun Wu氏による招待公演**: ***Concept Learning Across Domains and Modalities***
  - 視覚的推論にはCLEVRなどシミュレータによるデータが用いられることが多い
    - 実画像とのドメインの違いをどう考慮するか？
  - 視覚的推論にもさまざまなタスク:従来はドメイン特化の手法
    - 2D question answering, 3D refferring expression, temporal sequence reasoning, robot manipulatiionなどのタスク



Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

2D question answering



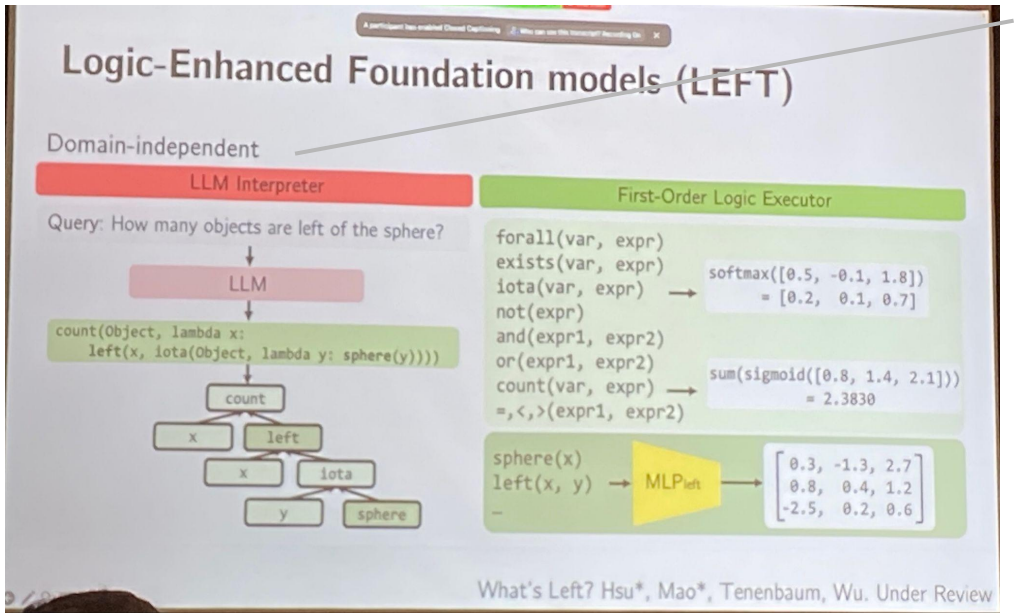
3D refferring expression

CVPR 2023 の動向・気付き(105/259)

O-DRUM Workshop @ CVPR 2023

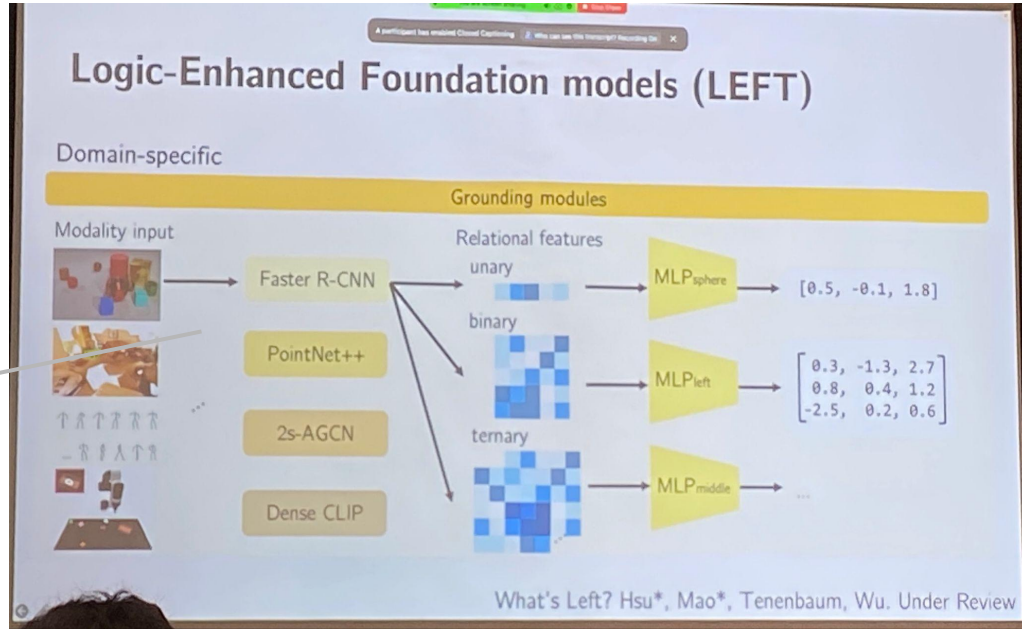
Open-Domain Reasoning Under Multi-Modal Settings

- **Jiajun Wu氏による招待公演**: ***Concept Learning Across Domains and Modalities***
  - 今後の動向としては, 全てのタスクを同時に解くLogic-Enhanced Foundation Model(LEFT)
    - 基盤モデルにより各タスクにおいても性能向上



LLMの利用
:ドメイン不変に

ドメイン特化に
下流タスクに利用



CVPR 2023 の動向・気付き(106/259)

Workshop on Computer Vision for Microscopy Image Analysis (1/3)

- 顕微鏡画像解析に焦点を当てたワークショップ

顕微鏡画像においても基盤モデルの流れ

- 病理画像解析における基盤モデルの構築
  - 招待公演:James Zou (Stanford University)
    - PLIP: Leveraging medical Twitter to build a visual-language foundation model for pathology AI

- 医学分野におけるSegment Anything Modelの利用
  - 招待公演:Dong Xu (University of Missouri Columbia)
    - Enhancing SAM's Biomedical Image Analysis through Prompt-based Learning

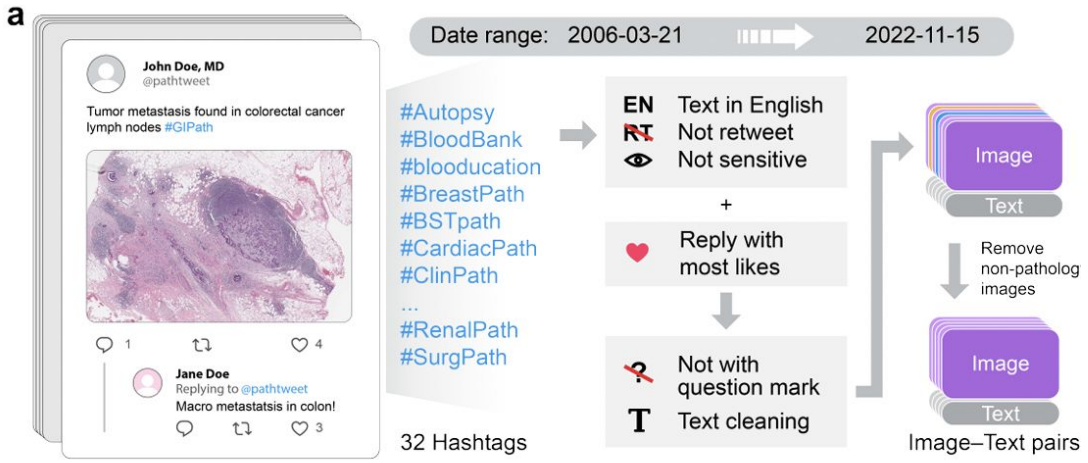
CVPR 2023 の動向・気付き(107/259)

Workshop on Computer Vision for Microscopy Image Analysis (2/3)

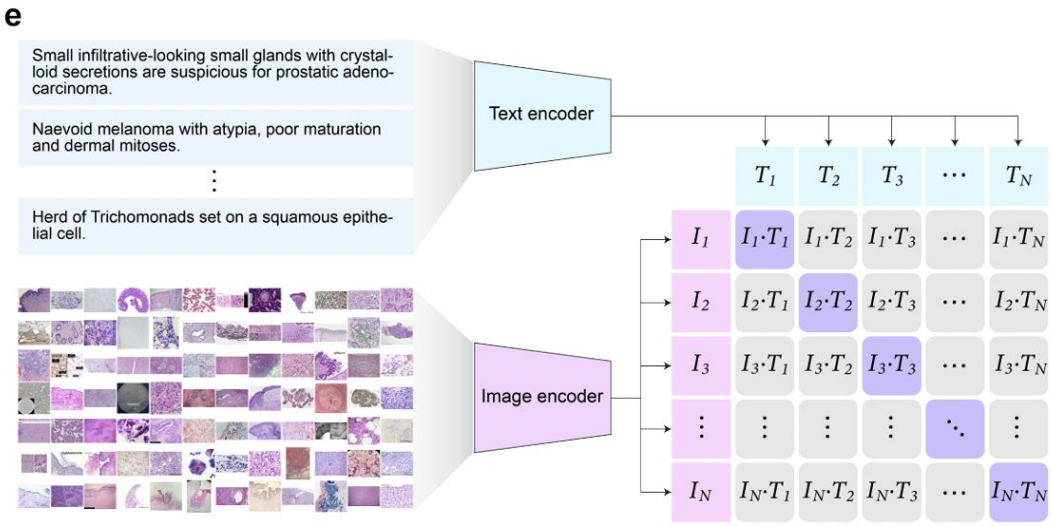
- **招待公演:James Zou (Stanford University)**

PLIP: Leveraging medical Twitter to build a visual-language foundation model for pathology AI

- Twitterから病理画像-テキストのペアを収集して, 対象学習することにより基盤モデルを作成→データクレンジングが重要
- zero-shot, linear probing, text-to-image retrieval, image-to-image retrievalなどの下流タスクを想定
- 将来的な展望はspatial omics:ゲノム情報を空間的に予測



twitterからデータ収集→クレンジング



CVPR 2023 の動向・気付き(108/259)

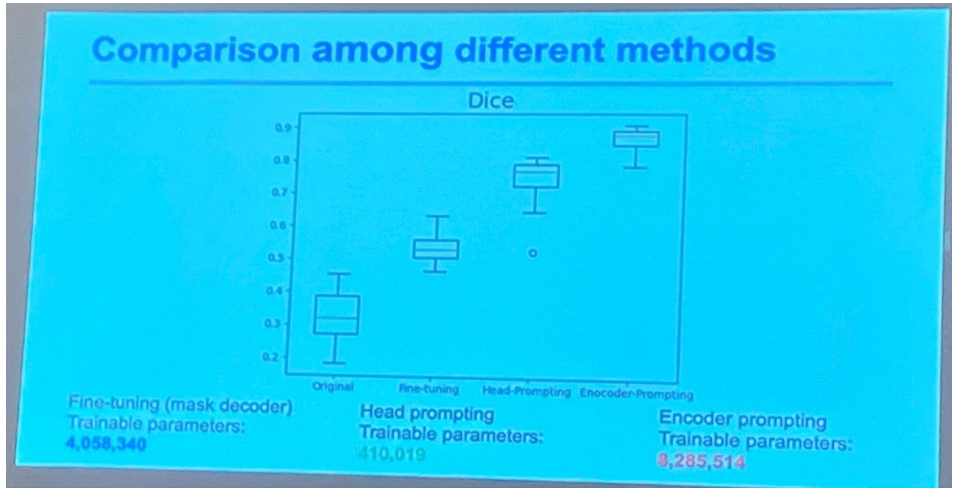
Workshop on Computer Vision for Microscopy Image Analysis (3/3)

○ 招待公演:Dong Xu (University of Missouri Columbia)

タイトル:Enhancing SAM's Biomedical Image Analysis through Prompt-based Learning

- 医学教科書の図のOCRによる解析への利用
  - SAMでfigureに対してsegmentation

❑ 顕微鏡細胞画像に対するSAMの応用
  - ❑ SAMをいかに上手く利用するか？(full finetune / prompt tuning / adapter tuningなど)について色々検証している
  - ❑ full finetune→decoderの特定の層のみ学習→入力でprompt tuning→encoderにおけるadapter tuningの順番で性能が向上
  - ❑ 今後の展望としては, 別のprompt tuning試したり, chain-of-thoughtでprompt最適化したり

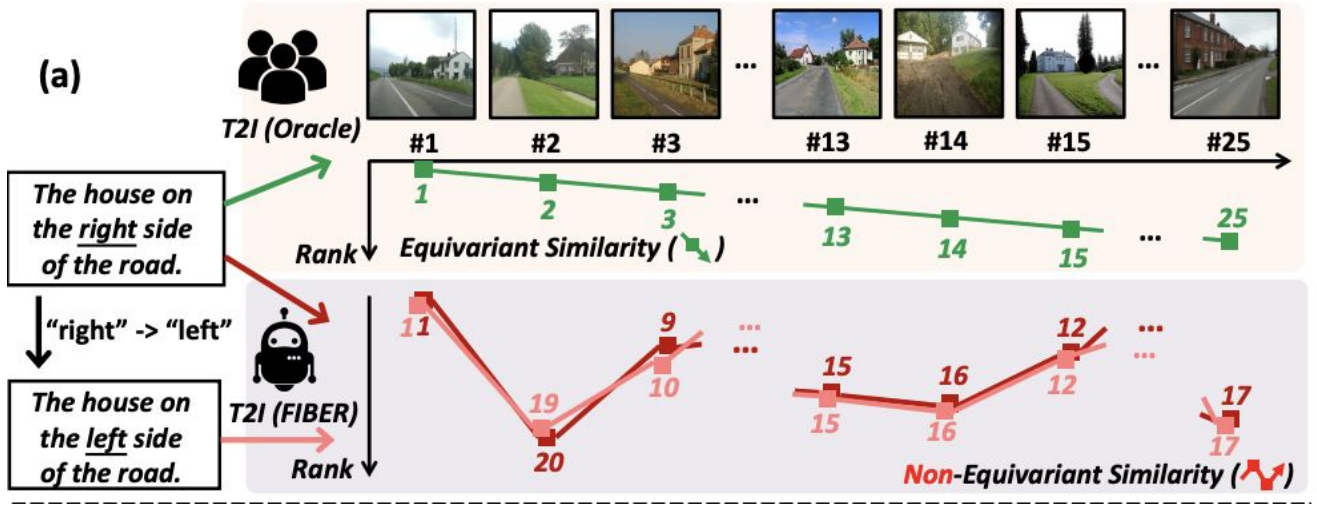


SAMの利用方法で性能比較

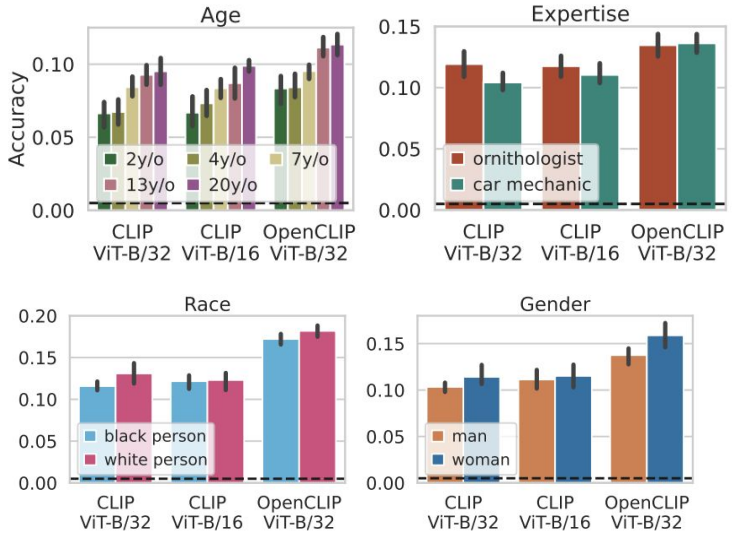
CVPR 2023 の動向・気付き(109/259)

Workshop: New Frontiers in Visual Language Reasoning: compositionality, prompts, and causality

- LLM, Multi-Modal LLM(例CLIP)などの分析に関する研究が多い
  - 例1:CLIPの学習済み特徴はdisentangledで, equivariantではない(左下).
  - 例2:LLMにGender, Age, Ability別でPromptをRequestすると, 解答能力が異なってくる. 使い側の情報となり, LLMの学習バイアスも解析できる.
  - 新しい指標などで, 現在のMultimodal LLMがCompositionality的にもないがあることを示した. (次のページ)



Equivariant Similarity for Vision-Language Foundation Models (arXiv, 23)

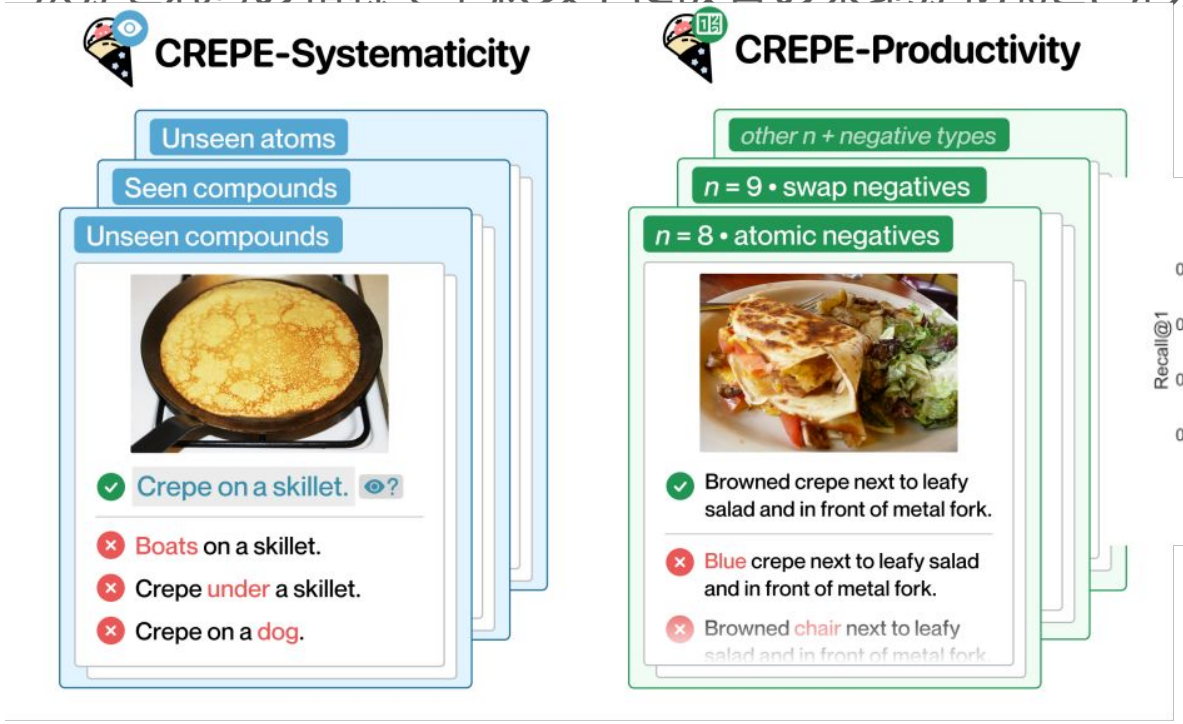


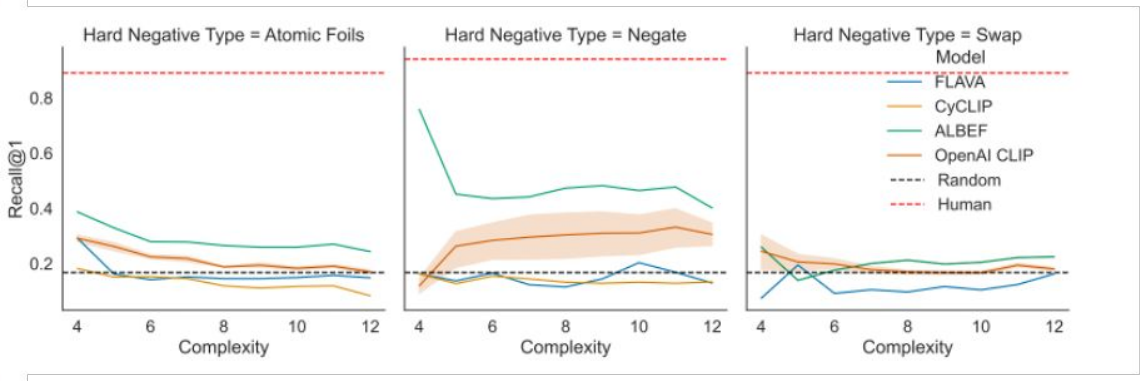
CLIPが各年齢・専門性・人種・性別を真似して生成したテキストの精度差

CVPR 2023 の動向・気付き(110/259)

CREPE: Can Vision-Language Foundation Models Reason Compositionally?(CVPR 2023 highlight)

- 既存のVision and LanguageモデルのSystematicity(SeenとUnseenの概念，組み合わせの理解)とProductivity(異なる難易度を持ったHard examplesに対しての理解)を評価するベンチマークデータセットと評価指標を提案.(左下)
- CREPEではText-to-Image retrievalを評価タスクとして網羅的な既存手法分析を行った．評価の結果により既存手法がこれらの指標で予想以上に改善の余地があることを示した.(右図)

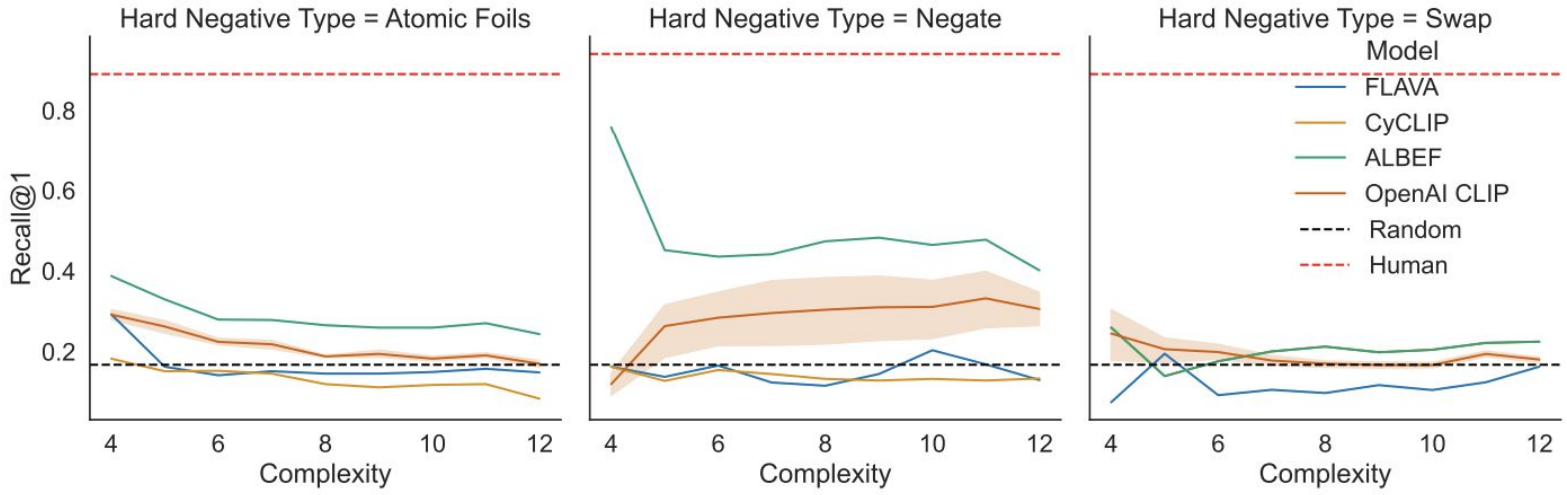




CVPR 2023 の動向・気付き(111/259)

Workshop: New Frontiers in Visual Language Reasoning: compositionality, prompts, and causality

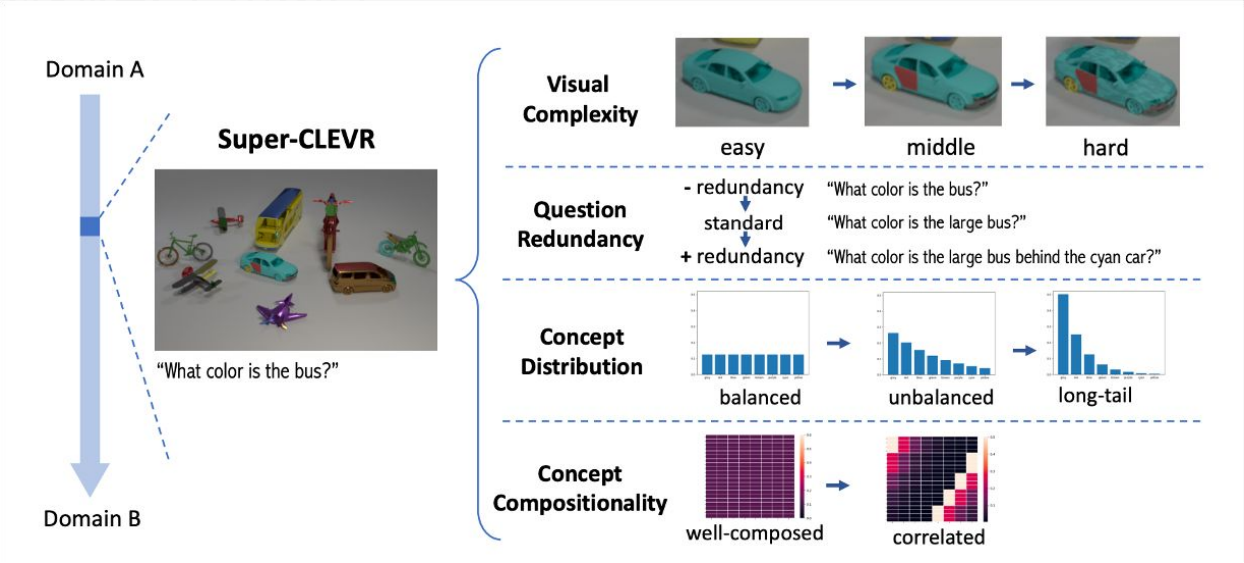
- Multi-modal LLMとSymbolization(例:Neural Module Network系)
  - CLIPなどを含めて，既存のMulti-modal LLMのCompositionalityが想像よりも高くない問題がある.（下図，次の2スライド）
  - 重要性1:解釈性，問題が出た際の分析可能性が高くなる.
  - 重要性2:Unseen概念へのtransfer能力が高くなる.
  - 重要性3:Compositional 問題において性能が良い(次ページのSuper-CLEVRを参照)
  - 研究例:仕組み上から解釈可能なcausal foundation AI



CVPR 2023 の動向・気付き(112/259)

Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning(CVPR 2023, highlight)

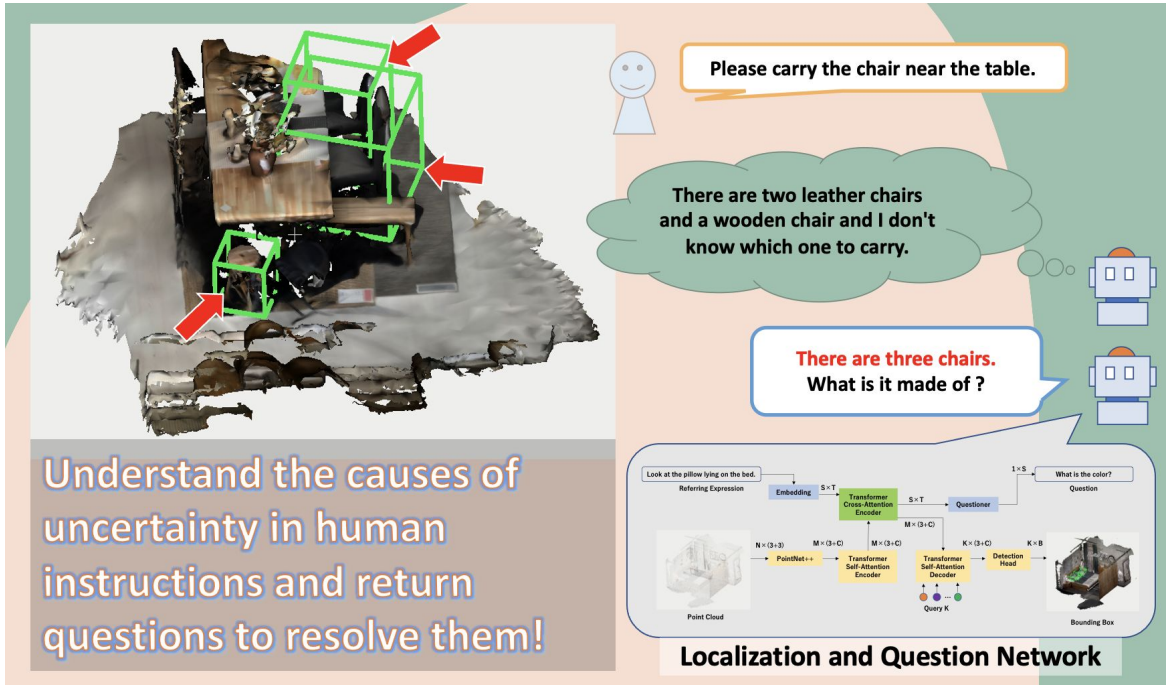
- VQAモデルがOut-of distributionへの適応性やドメインの汎化性能が低い傾向になる. しかし, これらを他の要素から独立して評価することが困難.
- 上記を詳細的に分析・評価するためのデータセットSuper-CLEVRを提案. 特に, モデルが視覚複雑性,質問回答の一致性, コンセプトの分布, コンセプトのcompositionalityで評価をした.
- 既存手法を分析した上, Modularity(Neural Module的な手法)と probabilistic uncertaintを結合した手法が最も高い性能を得られることを示した.

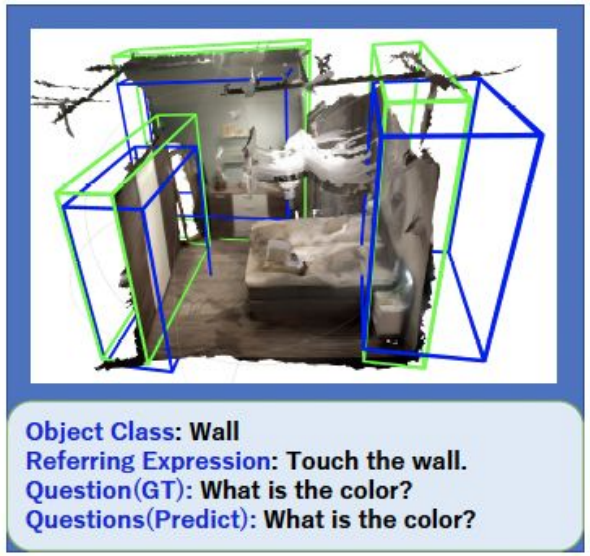


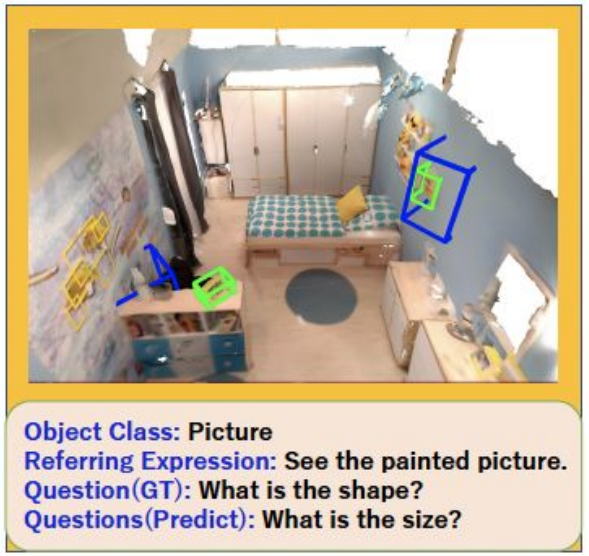
CVPR 2023 の動向・気付き(113/259)

Question Generation to Disambiguate Referring Expressions in 3D Environment

- 既存のObject Groundingタスクでは入力文章の情報が不十分である場合が考慮されていない.
  - 参照表現の曖昧性を緩和するための質問生成タスクを提案.
- 3次元シーングラフをベースとしてテンプレートベースの参照表現と質問テキストのペアを自動生成
- Transformerベースの手法により三次元点群と参照表現テキストを入力すると曖昧性の原因となる候補物体のBBOXと曖昧性を緩和する質問文を出力



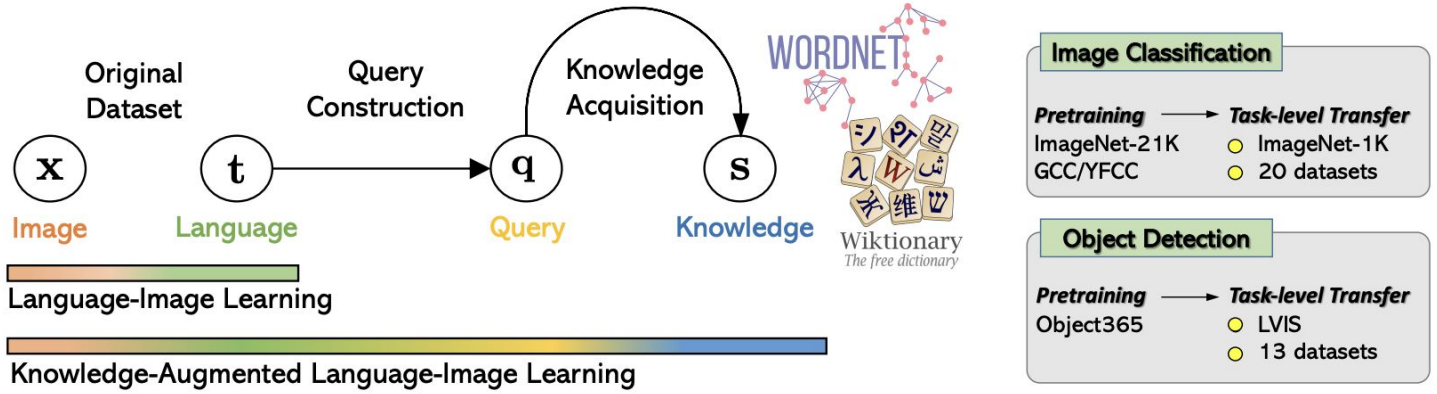


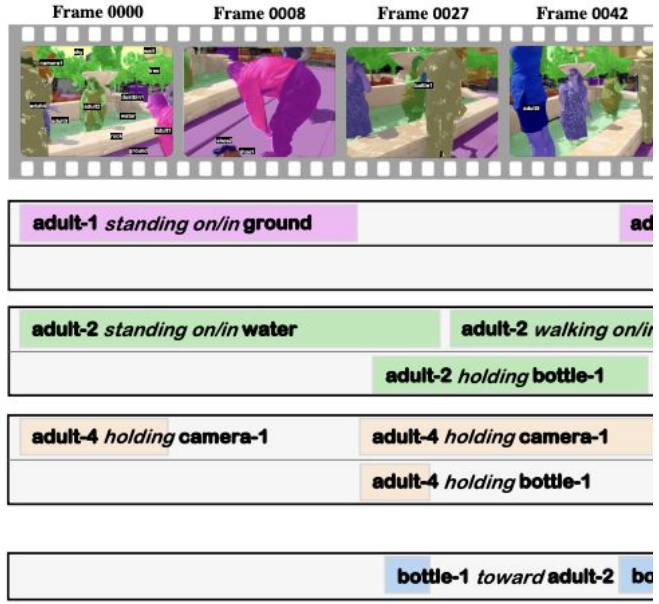


CVPR 2023 の動向・気付き(114/259)

Workshop: New Frontiers in Visual Language Reasoning: compositionality, prompts, and causality

- Multi-modal LLMと構造化Representation／Knowledgeの結合
  - 構造化されたKnowledgeがcross dataset/ domainのtransferへ応用可能
    - 特にfine-grainedで，Few-shotにおいて性能が優れている(左下図)
  - Image scene graph→Video scene graph→4D scene graph(右下，次Slide)



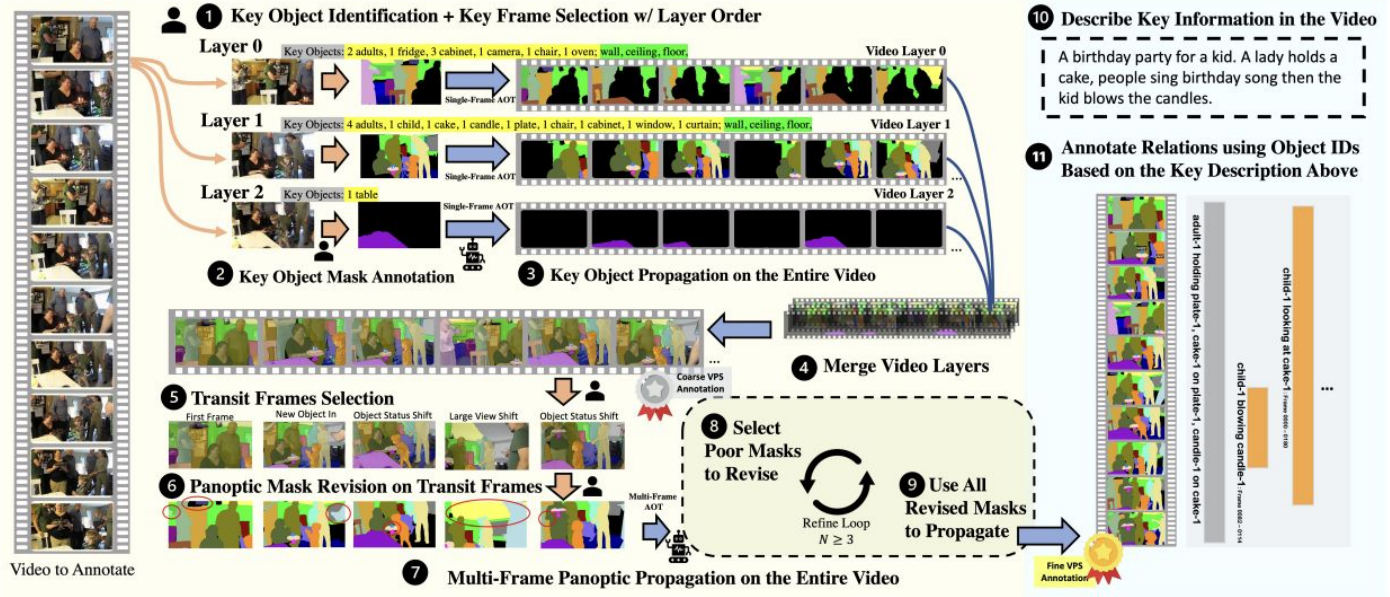


構造化知識を利用した転移学習(K-LITE, NeurIPS 2022)

CVPR 2023 の動向・気付き(115/259)

Panoptic Video Scene Graph Generation(CVPR 2023)

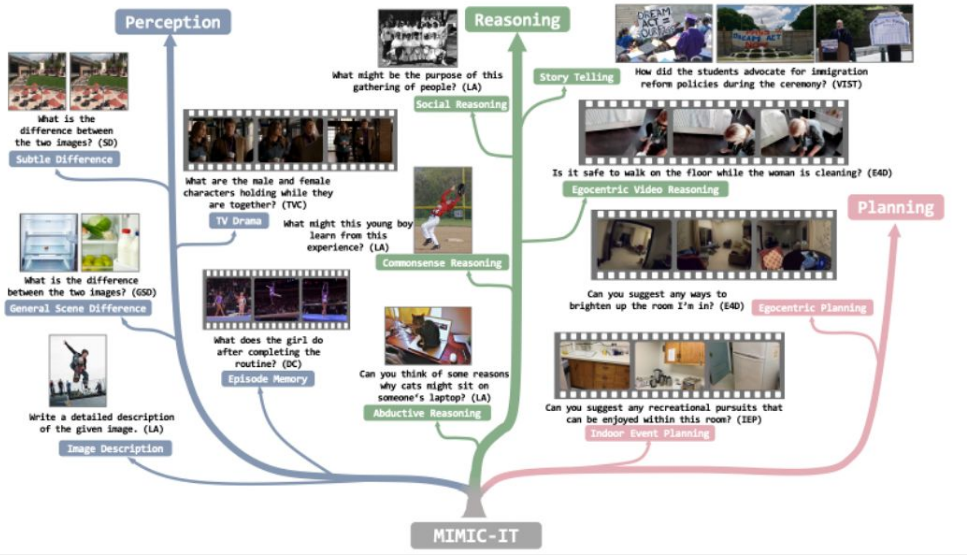
- 新しい問題設定panoptic scene graph generation(Videoのscene graph生成＋Panoptic segmentation)とそのためのデータセットを提案.
- 提案データセットは, Human labelingでデータを一部収集した上(例:ラベルなど), LLMなどを使ってVideo内容の重要内容のまとめも行った.(下図, 現在Human＋LLMでデータ作成がかなり増えてきた)
- Videoからラベルの検出・追跡, そしてラベルの識別の2段階から構成するモデルも提案.
- タスクの新規性が薄いかもしれませんが, 論文の完成度・手法の結果がとても高い.



CVPR 2023 の動向・気付き(116/259)

MIMIC-IT: Multi-Modal In-Context Instruction Tuning(arXiv2023)

- VandL モデルのタスクのための大規模instruction-response pairsデータセットMIMIC-ITを提案し，In-contex的なVandLモデルの構築を可能に.
- MIMIC-ITが2.8M instruction response pairsから構成される．また，視覚側では画像，連続画像，ビデオなどがある.
- MIMIC-ITで学習したOtterモデルが強いIn-context visual learning能力がある.
- タスクの新規性が薄いかもしれませんが，**論文の完成度・手法の結果がとても高い.**



MIMIC-ITで扱っているタスク

CVPR 2023 の動向・気付き(117/259)

Workshop: New Frontiers in Visual Language Reasoning: compositionality, prompts, and causality

- 将来課題1 : Incontext multimodal learning
  - Multi-modal massive dataset+Foundation model.
  - Incontextで言語のみならず，あらゆるモダリティの情報をIncontextで学習
  - MIMIC-ITは視覚データ(画像・3次元データ・ビデオ)をIncontext学習し，テスト時に様々なタスクに活用できる.（前のページ）
  - 現実世界から自主的な学習と今後結合されていきそう.

CVPR 2023 の動向・気付き(118/259)

Workshop: New Frontiers in Visual Language Reasoning: compositionality, prompts, and causality

- 将来課題2:Analysis by synthesis perspectiveの実現
  - Analysis by synthesis perspective:実世界ではなく, シミュレーション世界でリアル世界のモデル化・分析をする.
  - 上記においてDiffusion モデル は有用だが, representations, compositionality, abstractionなどの面で改善する余地がある.
  - Module Network(例:Neural Module Networkなど)と結合し, 高精度のSim2Realを達成する希望がある.

CVPR 2023 の動向・気付き(119/259)

Workshop: Embodied AI

- Embodied AIタスクの目標:
  - How intelligence emerges from agents interaction with the environment

- ワークショップ2023の3つのTopics
  - Foundation Models
    - 大量なマルチモーダルデータで学習し，複数のDownstreamタスクへ
  - Generalist Agents
    - モデル的に，複数のタスクを同時に解く
  - Sim2Real Transfer
    - simulation (matterport3D, habitat, HM3D)で学習し，Real環境・ロボットでタスクを解いていく

- 上記のTopics以外に，Invited talksでよく出てくるキーワード
  - Learning from videos
  - LLM
  - Diffusion model
  - Neural Module Networks (/Neural Symbolic Network) (Best PaperのVisual Programmingも)

CVPR 2023 の動向・気付き(120/259)

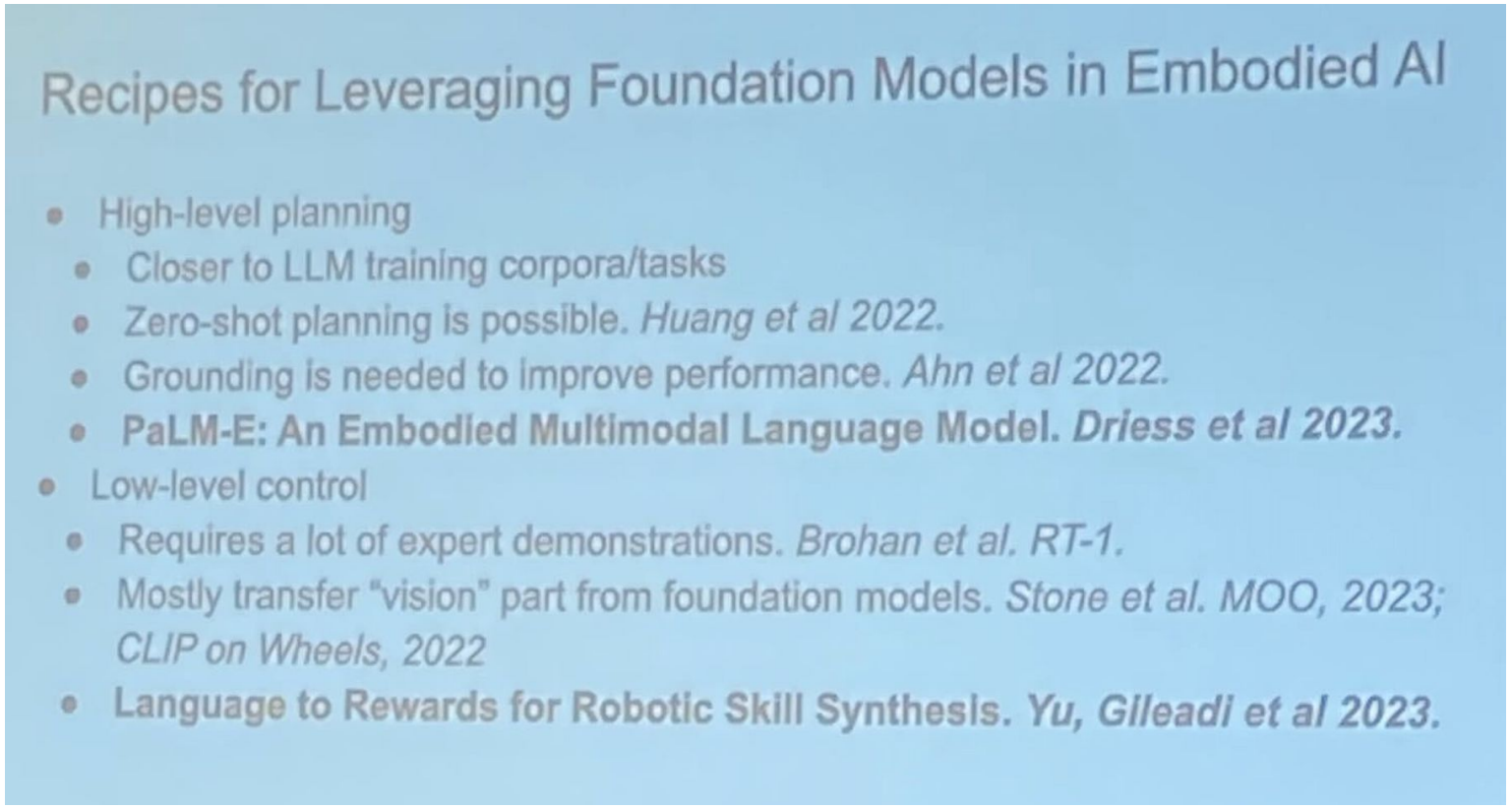
Workshop: Embodied AI

- Topics 1:Foundation Models
  - 概念:Internetデータで事前学習し, Embodied AIに適応
  - Foundation Modelsの分類
    - High level planningに適応するモデル(次のページ, PaLM-E)
    - Low-level controlに適応するモデル(次のページ, Robot skill synthesis)
  - Palm-eなどのモデルでロボティクスやEmbodied AIでも大量なマルチモーダルデータで学習し, 様々なEmergent能力が得られることが示された.
  - データ準備について:
    - これまでのhigh qualityシミュレーションデータセットなどが活用可能だが, データ作成のコストが高く, 多様なデータの作成が困難.
    - 今後は, diffusionモデル等の進化により低コストで高質なデータが作成可能.

CVPR 2023 の動向・気付き(121/259)

Workshop: Embodied AI

- Topics 1：Foundation Models
  - Embodied AI系のFoundation Modelsの例：(後ろの3つのスライドで例を挙げる)



CVPR 2023 の動向・気付き(122/259)

Workshop: New Trends in Image Restoration and Enhancement workshop(NTIRE)

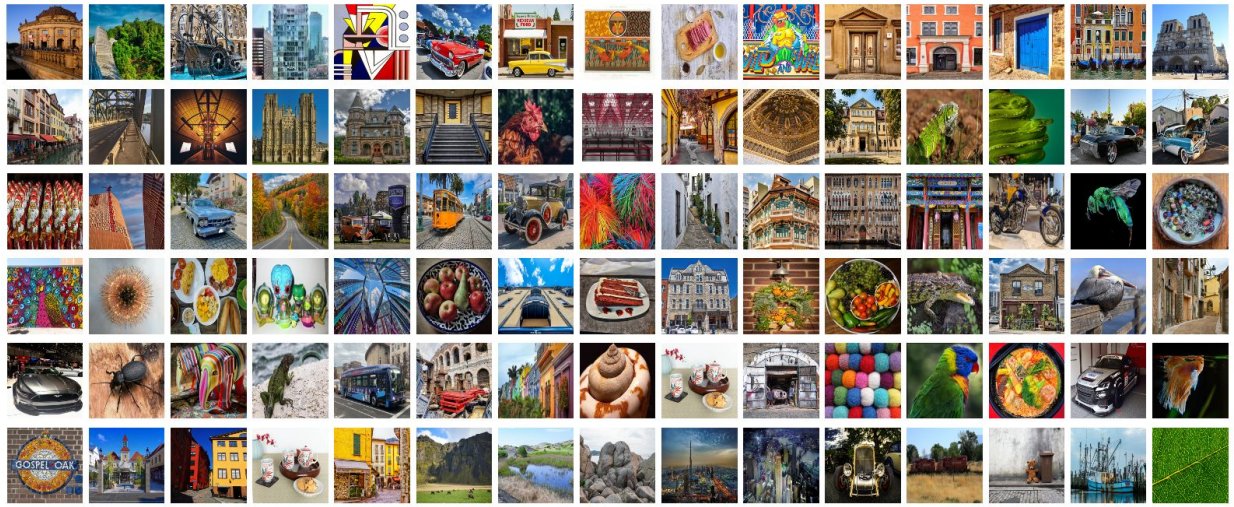
LSDIR: A Large Scale Dataset for Image Restoration

画像復元のための大規模データセット(LSDIR)を提案

- 背景
  - 画像復元における最近の研究は，ネットワークの設計が中心
  - 共通して使用されている学習データセットは2017年に提案されているため，数千枚の画像 しか含んでおらず，ネットワークの性能を最大限発揮できていない

LSDIR datasetの画像例

0枚



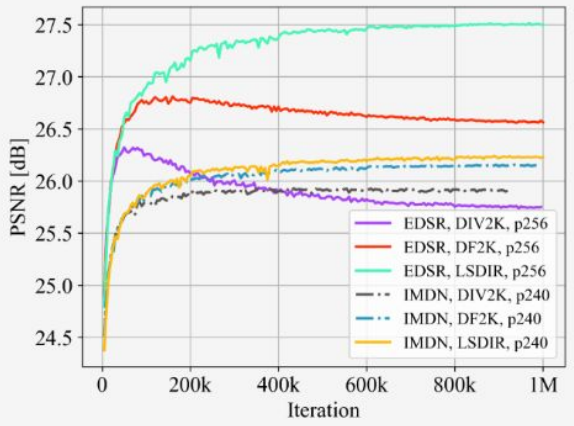
学習画像:84,991枚，検証画像:1,000枚，テスト画像:1,000枚で構成

CVPR 2023 の動向・気付き（123/259）

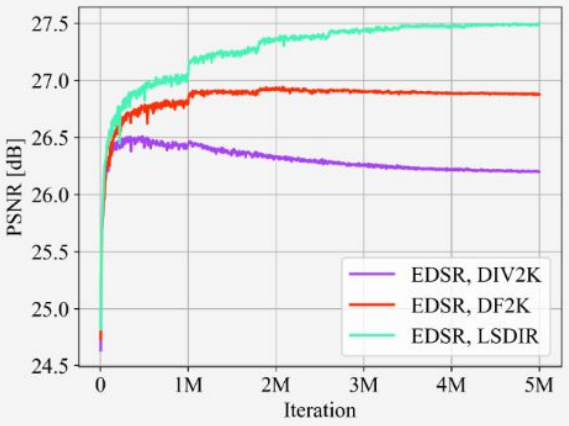
- データセット構築手順
  - 画像復元タスクにおけるデータの性質
    - 認識や検出を始めとするhigh-level visonとは異なり，画像復元などのlow-level visonではピクセル単位の品質が重要
    - 画像に含まれるブラーやノイズといったアーチファクトが学習時に悪影響を及ぼす
  - データ収集
    - オンライン写真共有サービスであるFlickrから4K画像を収集
    - ライセンス，解像度，アスペクト比，撮影日及びタグを考慮
  - データクレンジングのため，以下の施策を実施
    - ブラー検出のため，Laplacianフィルタでエッジ抽出した結果の分散を使用してフィルタリング
    - 低周波領域を多く含む画像を削除
    - JPEG圧縮ノイズやその他含まれるノイズを軽減するためのダウンサンプリング
    - アノテーターが高品質もしくは低品質か判断

CVPR 2023 の動向・気付き(124/259)

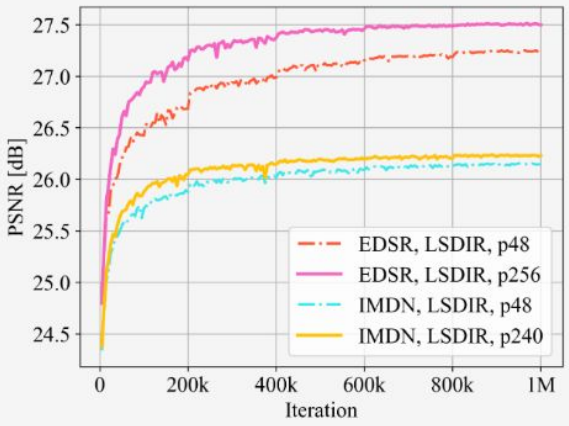
- 実験
  - 様々な画像復元タスク，モデルで学習結果を比較
  - 超解像，ノイズ除去，JPEG圧縮によるアーチファクトの軽減，ブラー除去及びデモザイキングで効果を確認



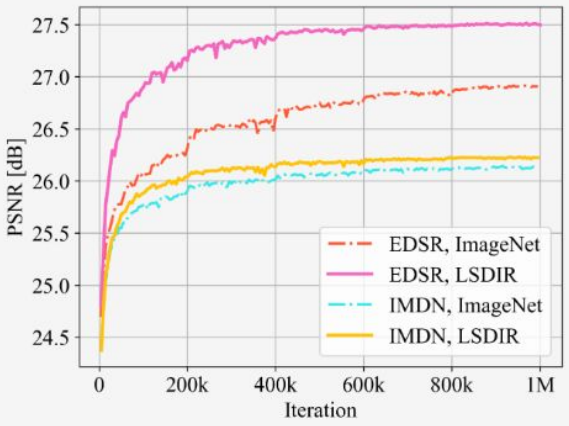
Different datasets with larger training patch size.



Different datasets with increased training iterations.



LSDIR training with different patch sizes.



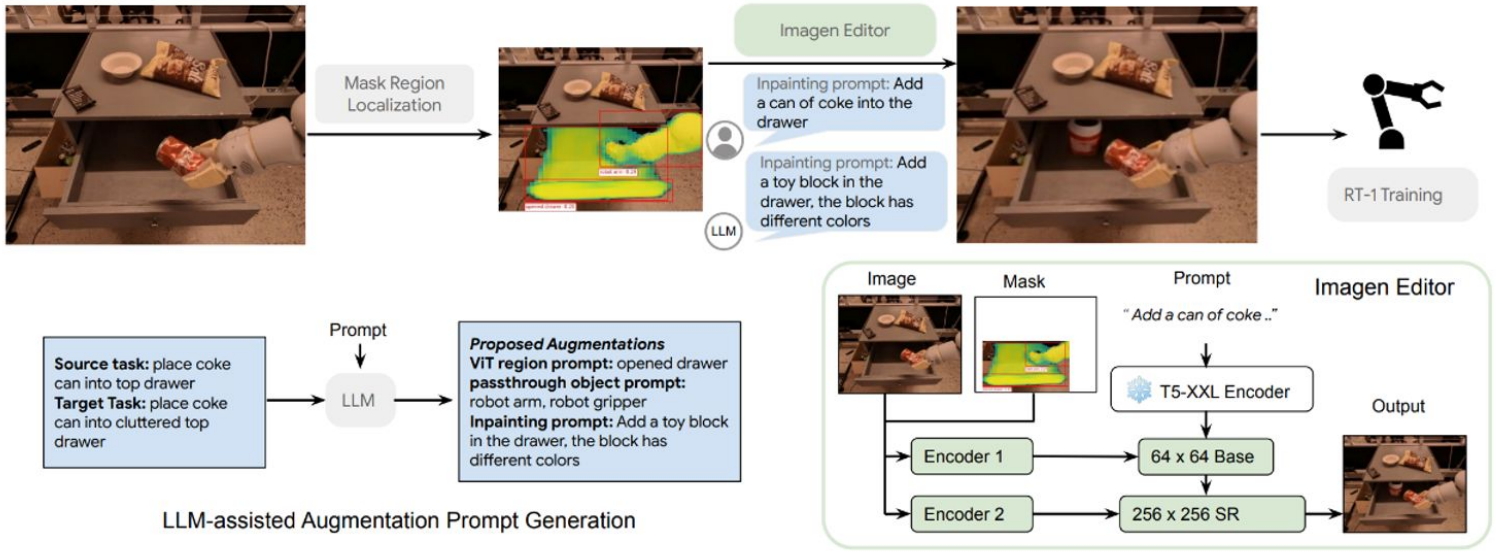
Comparison between LSDIR and ImageNet

パッチサイズと学習時間を増大させて学習を行うことで様々な画像復元タスクにおいて性能向上を達成

CVPR 2023 の動向・気付き(125/259)

Scaling Robot Learning with Semantically Imagined Experience (Google 2023)

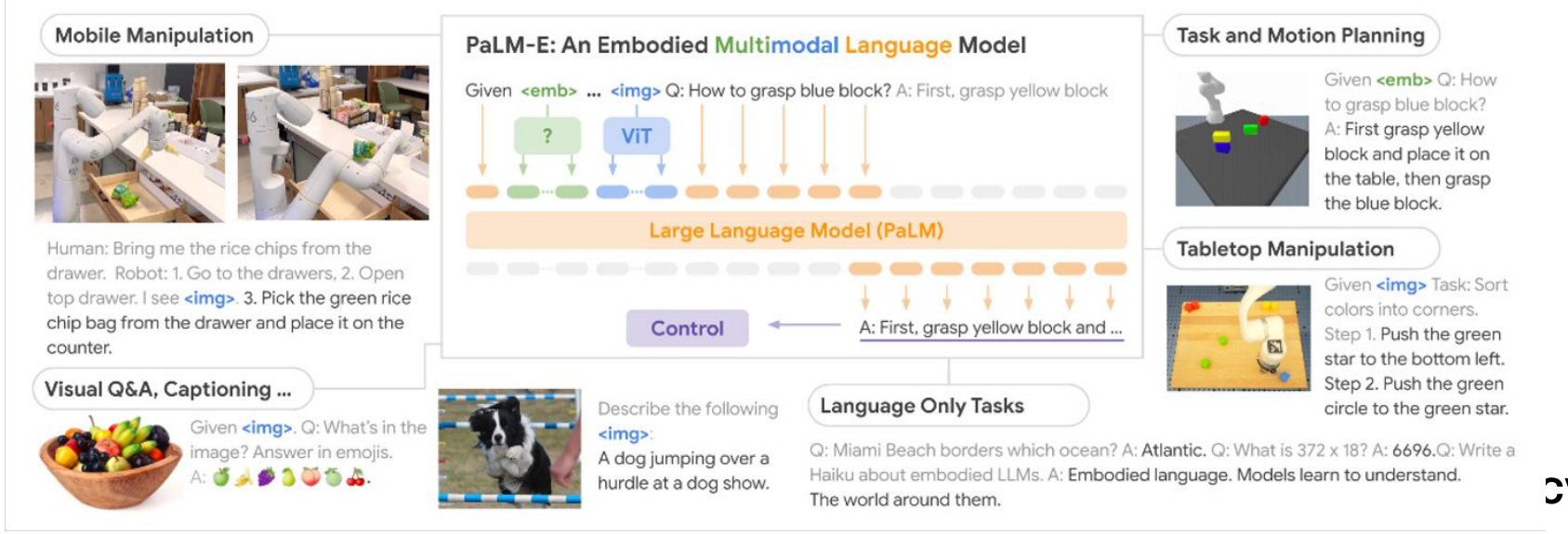
- ロボットの学習には大規模学習データが必要だが，ロボットのデータを得るのが難しいためデータ準備のコストが高い.
- text-guidedなdiffusion modelsでロボット学習のデータ拡張を行う手法を提案し，自動的にロボットデータを準備せずに大規模データの構築に成功.
- 様々なDownstreamタスクで上記の学習の有効性を示した.



CVPR 2023 の動向・気付き(126/259)

Palm-e: an embodied multimodal language model(Google 2023)

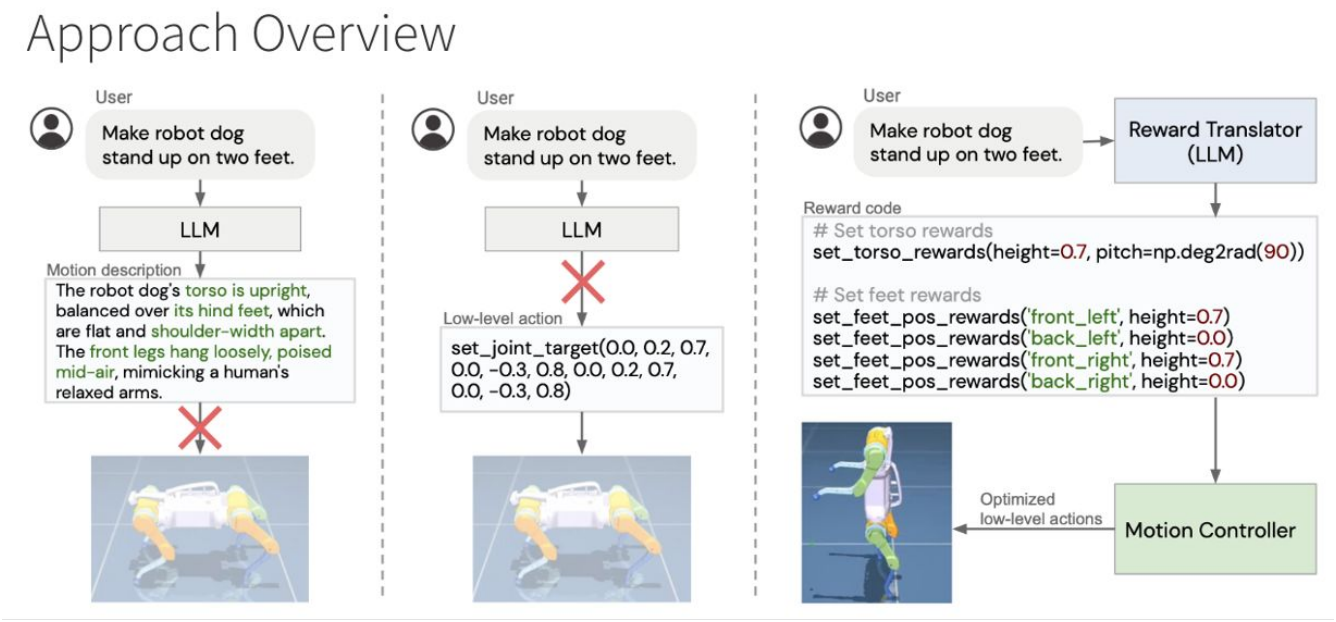
- RoboticsのFoundation model Palm-eの提案. Palm-eはマルチモーダルの入力から, Transformer構造で言語に変換し, さらにControlモデルでロボットの実際の行動計画・動作の執行まで行う.
- Palm-eではscaling効果を示し, 様々なemergent abilitiesがある:zero-shot multimodal chain-of-thought, complex multi-image reasoning, …
- Internetデータから実環境のロボットまでのpositive transferを示した.



CVPR 2023 の動向・気付き（127/259）

Language to rewards for robot skill synthesis（Google 2023）

- LLMsなどをロボットに活用し，比較的にhigh levelタスクにおいての性能が示されてきた.
- しかし，robot controlなどのLow level問題には適応しにくい問題点が残っている.（ハードウェア依存，LLMsで表現しやすい特徴量表現の選択など）
- ここでReward functionをLLMsとモデルControlの中間表現として，LLMsを用いてReward functionのパラメータを定義・最適化可能にした.



CVPR 2023 の動向・気付き(128/259)

Workshop: Embodied AI

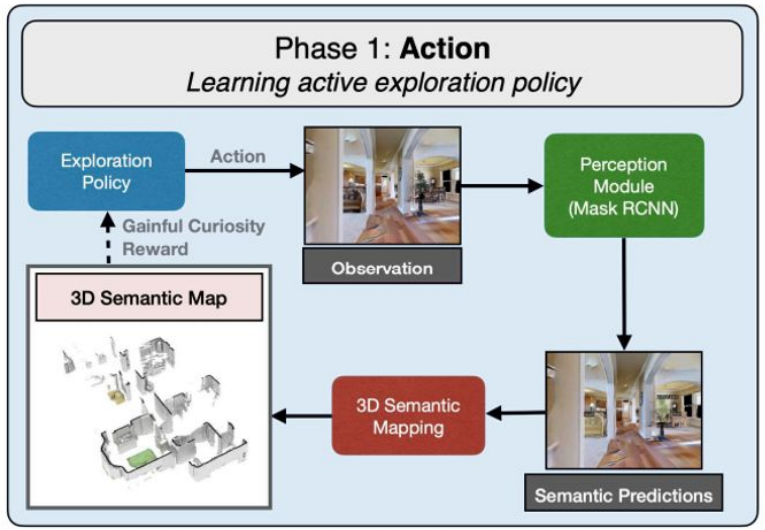
- Topics 2: Generalist Agents
  - 概念：1つのAgentで様々なタスクを解ける.
  - アプローチ（他のトピックでも通用？）:
    - Self-supervised embodied action learning (SEAL)（次のスライド）
    - Modular Models（FILM）（次のスライド）
      - Step-by-stepで大きな問題を分解して解く
    - Grounding Language for multimodal generation：FROMAGe（次のスライド）
      - 言語から知識を得る

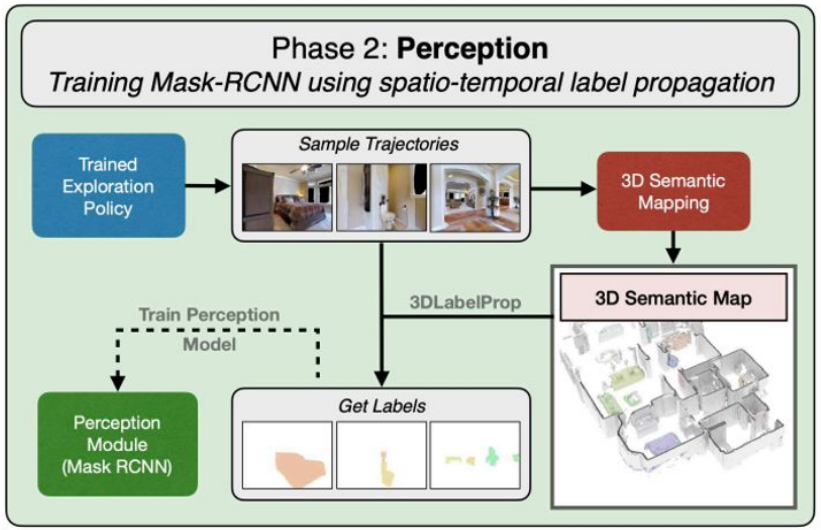
  - Generalist AgentsのためのEmbodied AIタスクも提案された
    - ALFRED＋TEACh（13タスク？）
    - ManiSkill（比較的に低レベルなロボットコントロール）

CVPR 2023 の動向・気付き(129/259)

SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency (NeurIPS, 2021)

- インターネットデータから追加データなしでRoboticsタスクへ適応する研究.
- 提案手法SEALはインターネットデータからActive explorationポリシーを学習する. Explorationで得られる観測からセマンティックマップを作成し(3D consisencyによりself-attentionで行う).
- 上記のSEAL手法が, ActionとPerceptionのclose-loopを実現し, 追加データなしでexplorationと3次元マップ推定を行える.

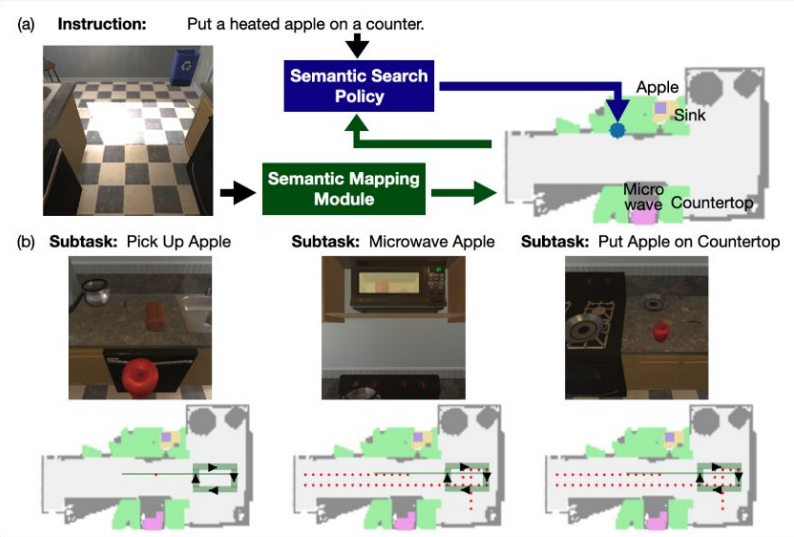




CVPR 2023 の動向・気付き(130/259)

FILM: Following Instructions in Language with Modular Methods (ICLR, 2022)

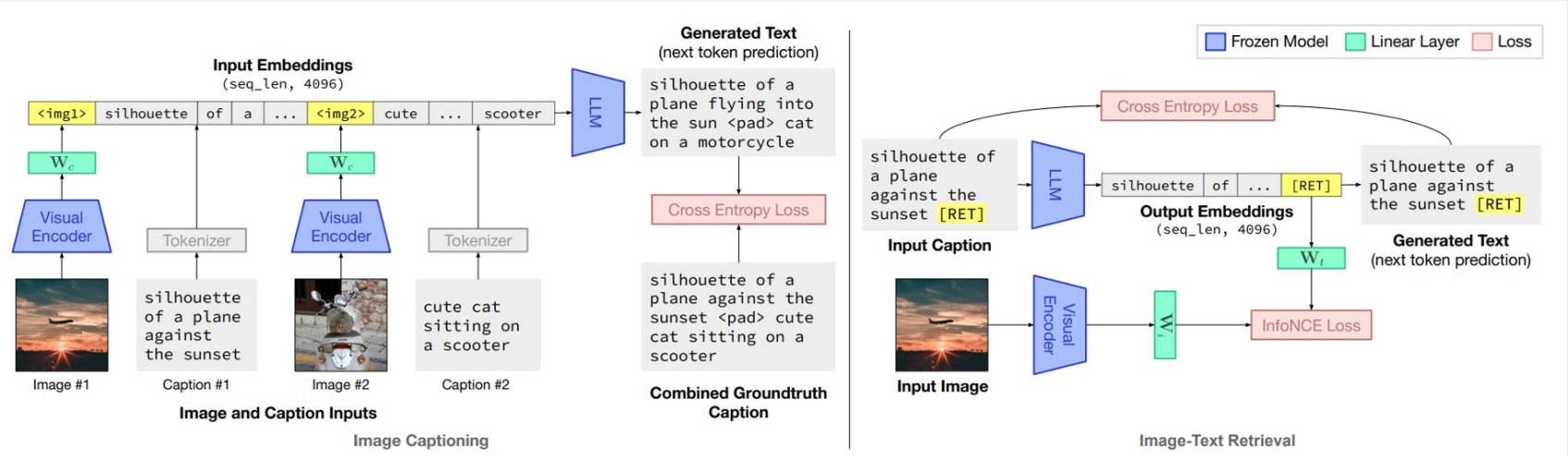
- Modularベースのstructured特徴表示を利用したInstruction Followingモデル.
- 提案手法は具体的に，まず自然言語のInstructionとEgocentric画像フレームの入力から，Semantic top-downマップを更新しつつ，言語からSemantic Search Policyで環境のexplorationを行う．そして，Search goalに届くまで上記のプロセスを更新していく.
- ModularベースのEmbodied AIモデルは最近増えてきて，VQAなどだけではなく，Roboticsタスクでも性能を示した.



CVPR 2023 の動向・気付き(131/259)

Grounding Language Models to Images for Multimodal Inputs and Outputs (ICML, 2023)

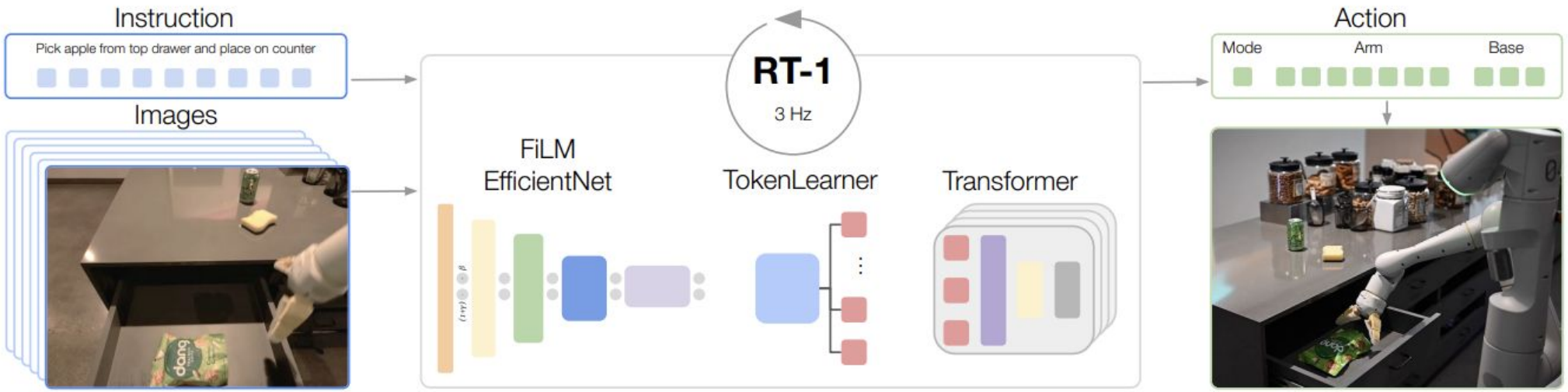
- 言語と画像両方を理解・生成可能なMultimodal LLM手法の提案.
- 提案手法が言語のみで学習したLLMを画像ドメインに拡張.
- 提案手法では言語モデルの部分をFreezeし, 入出力層のみで再学習して, Multimodal的なデータを扱う.
- 提案手法がZero-shotでImage retrievalやMultimodal dialogueなどを行える.



CVPR 2023 の動向・気付き(132/259)

RT-1: robotics transformer for real-world (Google, 2022)

- Generalist agentsの研究一例:
- Roboticsの学習のための大規模, Task-Agnosticデータセットの提案.
- Scaling効果を検証するため, Transformerベース手法Robotics Transformer—RT1を提案.
- RoboticsにおいてもScaling効果や, 大規模データセットでScaling効果や下流タスクへの適応性が検証された.



RT1の概念図

ワークショップ全体で何回も出たモデル

CVPR 2023 の動向・気付き(133/259)

Workshop: Embodied AI

- Topics 3: Sim2Real Transfer
  - あらゆる領域で高い質のデータが備ってきた
  - Stable Diffusion modelなどで今後さらに高質なデータ生成
  - Sim2realの環境整備系研究も続出：次のSlideのBenchBotが例

  - Foundational 的なリアル環境で動くEmbodied AI(ロボットなど)のため：
    - Learning at scales
    - Suitable training data (特徴：大きい，リアル)
    - Test and fine-tuning in real world

  - 残っている問題点：
    - Low level physics はよくIgnored
    - Geared toward benchmarking, not Sim2Real
    - Sim training limited to rather specific skills

CVPR 2023 の動向・気付き(134/259)

BenchBot environments for active robotics (BEAR): Simulated data for active scene understanding research (arXiv, 2023)

- Active sceneでのEmbodied AIのためのシミュレーターの提案.
- 既存のシミュレーターと比べて, BenchBotはシームレスで実環境Robot応用と繋げられる.(direct sim-to-real transfer of robot agent code)

BenchBot画像例

Environment/interface	Sim type	GT 3D BBox map	Scene variations	Agent control	Sim agent interface	Real robot interface
iTHOR/AI2THOR (Kolve et al. (2017)/Kolve et al. (2017))	Full sim	Yes	Randomise (materials, lighting and colour)	DBA (active)	AI2THOR	None
RoboThor/AI2THOR (Deitke et al. (2020)/Kolve et al. (2017))	Full sim w. real counterparts	No[a]	All envs are vars of 1 apartment. Walls and objects change	DBA (active)	AI2THOR	AI2THOR w. PyRobot bridge
MP3D/HabitatLab (Chang et al. (2017)/Savva et al. (2019))	3D scan	Yes	None	DBA (active)	Habitat lab	Habitat lab w. PyRobot bridge
Replica/Habitat lab (Straub et al. (2019)/Savva et al. (2019))	3D scan w. refine	Yes	6 variations of 1 scene (object pose)	DBA (active)	Habitat lab	Habitat lab w. PyRobot bridge
Gibson/Gibson (Xia et al. (2018)/Xia et al. (2018))	3D scan	No	None	DBT	Gibson	Gibson w. ROS bridge
iGibson/iGibson (Shen et al. (2020)/Shen et al. (2020))	Full sim based on real	No[a]	Randomisation (object models and materials)	DBT	iGibson	iGibson w. ROS bridge
BEAR/BenchBot	Full sim	Yes	Day + night lighting and objects added/removed for 5 scenes	DBA (passive and active)	ROS	ROS

既存研究との比較

CVPR 2023 の動向・気付き(135/259)

Workshop: Embodied AI

- Topics 3: Sim2Real Transfer
  - あらゆる領域での Sim2Real：**ISAAC Sim**
  - NVIDIA の大規模ロボットシミュレーションアプリで，多種類の合成データ生成に使える
  - Manipulating,Navigation
  - Human Simulation
  - 室内の工場環境・倉庫などで活用できる材料操作

Human Simulation

工場環境

Manipulation

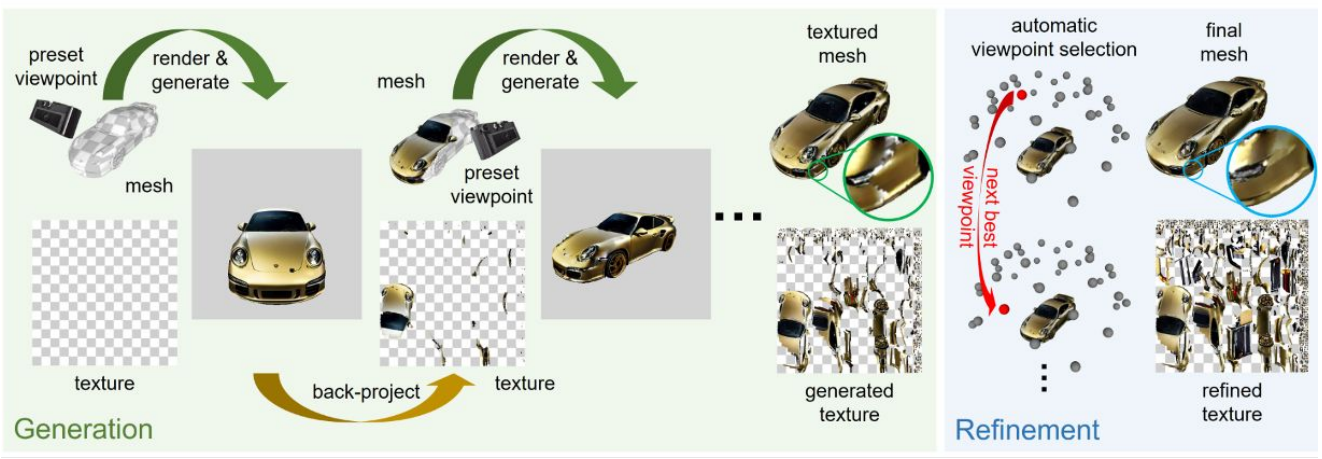
Navigation

CVPR 2023 の動向・気付き(136/259)

Text2Tex: Text-driven Texture Synthesis via Diffusion Models (arXiv, 2023)

- 高質な3次元テクスチャーを生成する手法Text2Tex.
- テキストから3次元メッシュの高質テクスチャを生成するモデル提案.
- 提案手法が既存のDiffusionモデルをInpaintingに使って, 段階的徐々にテクスチャーの生成を行う

結果例



提案手法

CVPR 2023 の動向・気付き(137/259)

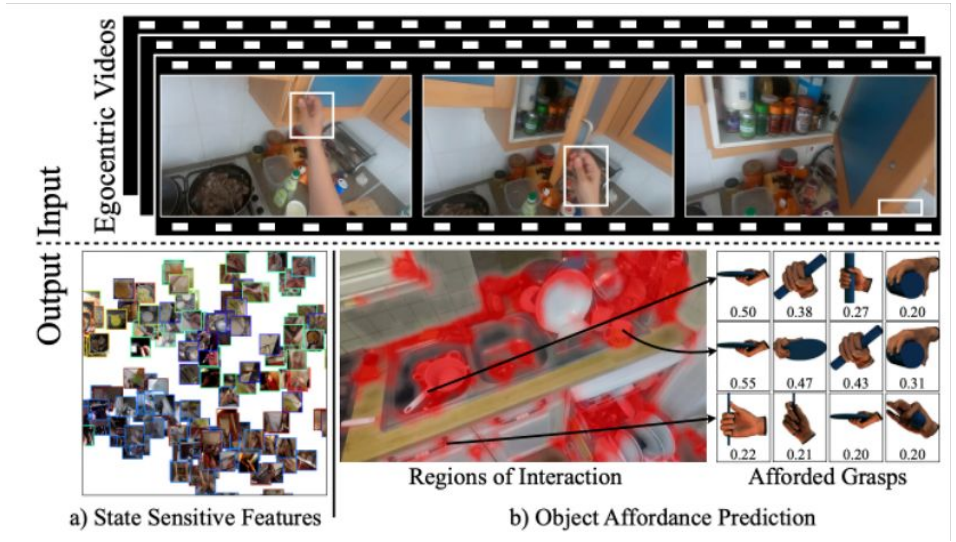
Workshop: Embodied AI

- Learning from videos (1/2): Saurabh Gupta先生
  - 特にEgo-centric videosから学習
  - 何を学習する:
    - どうやって物体とInteractする
      - 接触位置, どうやって, 接触したらどうなる, など(次のスライド)
  - 難しい点:
    - Embodiment Gap;
    - negative examples が少ない;
    - Interactionのための学習信号が少ない／直接ではない
  - 手段:
    - Learning at different abstract levels

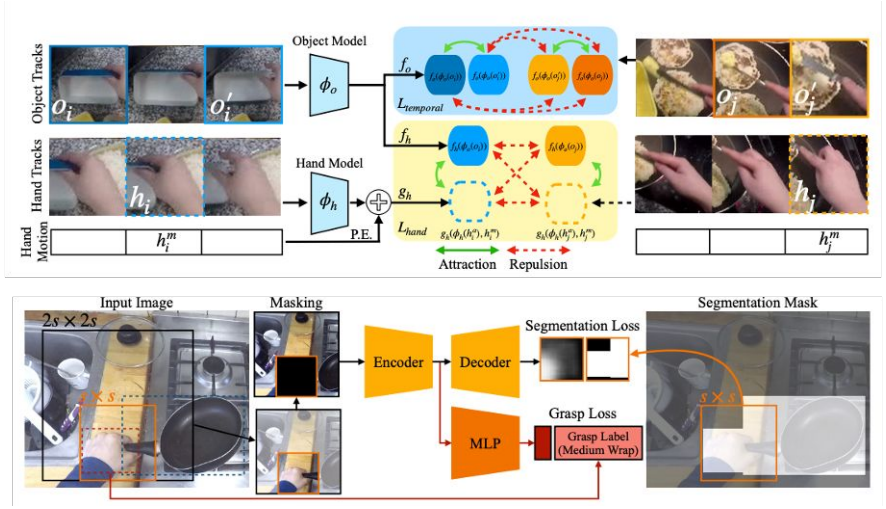
CVPR 2023 の動向・気付き(138/259)

BeHuman Hands as Probes for Interactive Object Understanding (CVPR, 2022)

- Egocentricビデオから撮影された手と物体のInteractionから，物体の把持，インタラクションなどの知識を学習する手法を提案.
- 手と物体の一致性を学習するTemporal SimCLR手法と，コンテキストを予測する(Self-supervised)仕組みで物体affordanceを学習する.（右）
- Epic-kitchensデータセットで提案手法でstate-sensitive 情報・物体の把持位置と把持方法などの学習に有効性を示した.



アイデア概念図

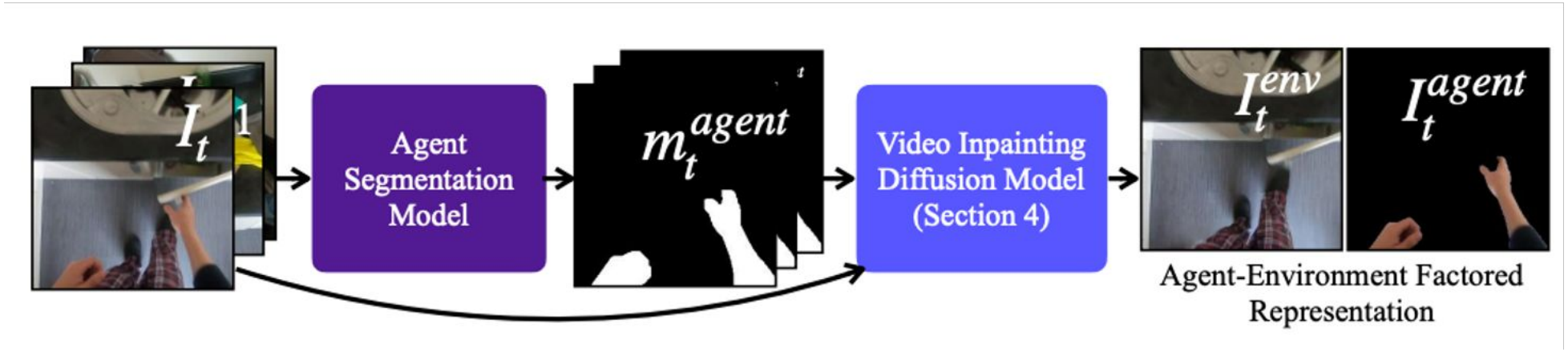


手法：Temporal SimCLR（上），Affordance（下）

CVPR 2023 の動向・気付き(139/259)

Look Ma, No Hands! Agent-Environment Factorization of Egocentric Videos (arXiv, 2023)

- Egocentric動画から，手の表現を学習することは様々なrobotic manipulationタスクにおいて重要．しかし，手が他の物体を遮蔽してしまい，把持などに重要な情報が失ってしまう場合がある．
- 上記を改善するため，Egocentric動画から，手と環境の特徴表現を分解した特徴表現学習の手法を提案．
- 上記の手法のコアな部分は：diffusion モデルを使ったvideo inpainting.
- 手法がシンプルだが，既存手法より良い手の表現学習ができることを示した．



提案手法

CVPR 2023 の動向・気付き(140/259)

Workshop: Embodied AI

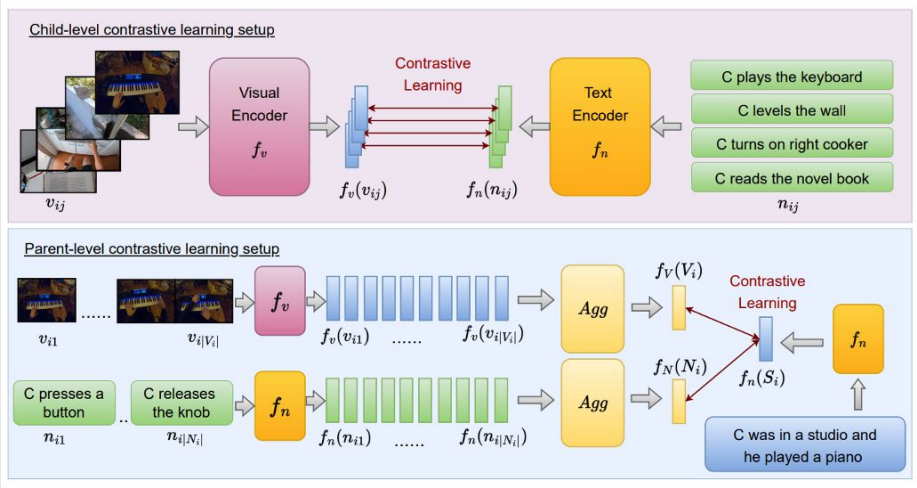
- Learning from videos (2/2):Kristen Grauman先生
  - 概念:Human-centric environmentから学習するため，動画像データが有用.
  - 特にEgo-centric videosから学習
  - Ego-ExoのAlignmentも検討
  - Ego-exoからのactivity contextの学習も検討
  - 何を学習する:
    - Hierarchical video-language learning (Ego-4D，Kumar2023次のスライド)
    - Ego-exo alignment from unpaired data(次のスライド)
    - Activity context(complex sequential behaviorなど)
    - Human-like grasping

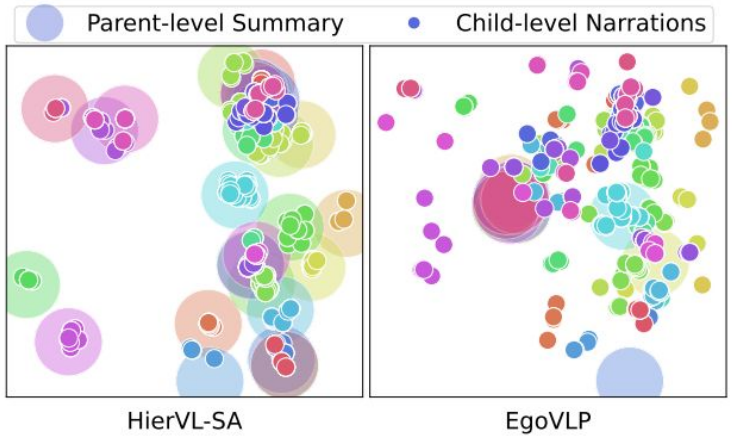
CVPR 2023 の動向・気付き(141/259)

HierVL: Learning Hierarchical Video-Language Embeddings (CVPR, 2023)

- 既存のビデオ認識の手法は数秒間の動画間とテキストの関係性を学習している．そこで，LongTerm動画認識性能が限られる．
- ここで，階層的なVideo-Language embeding手法を提案し，Long termとShort term両方を統一したモデルで扱えるようにした．また，ビデオクリップレベルとトータルビデオレベルのContrastive Lossを適応した．
- 短いビデオクリップで細かく何が発生したかを学習し，ビデオ全体でWhyに関してReasoningできるようにした．

提案手法



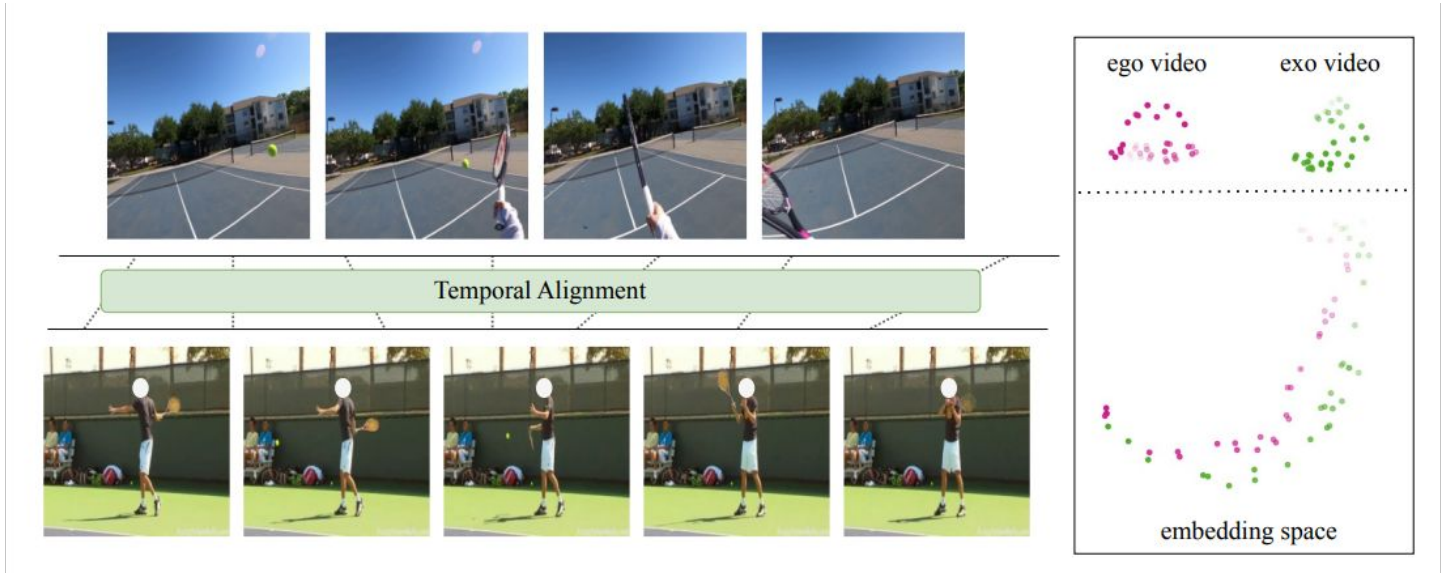


T-SNE分析

CVPR 2023 の動向・気付き(142/259)

Learning Fine-grained View-Invariant Representations from Unpaired Ego-Exo Videos via Temporal Alignment (arXiv, 2023)

- Self-superviseでExoとEgo動画のアラインメントを行う手法の提案.
- 学習では, Unpaired動画から学習可能(時間的に一致性が特にない, 環境も同一ではない).
- 提案手法は, objectとhandをフォーカスし, エンコーディングをする. また, Reversedビデオをネガティブサンプルとして Contrastive学習する.



提案タスクと学習済み

Embedding

CVPR 2023 の動向・気付き(143/259)

Workshop: Embodied AI

- ワークショップ2023の10個のChallengesのその(5/10):
  - Navigation and Understandingタスク
    - RxR-Habitat(continous space, room cross room navigation)
    - Multi-On (multi-object navigation, 物体の順序が異なる)
    - SoundSpaces(音源や音信号からnavigation)
    - Robot Vision Scene Understanding (室内環境のnavigationとscene,object recognition)
      - scene change も対応
    - Habitat (image goal navigation, object goal navigation)
      - Open-vocabularyも一部対応

CVPR 2023 の動向・気付き(144/259)

Workshop: Embodied AI

- ワークショップ2023の10個のChallengesのその(10/10):
  - Interaction and Rearrangementタスク
    - AI2-Rearrangement
      - 2つのシーンの変化を認識し，物体を復元する
    - ALFRED+TEACh
      - Generalist Embodied AIのための13タスクを同時に学習する設定
    - DialFRED
      - Dialog-enabled embodied instruction following
    - ManiSkill
      - Generalist Embodied AIのためのGeneralizable Skillタスク.
      - ロボットの可視化(例:Manipulation)が容易にできる.
    - TDW-Transport
      - Rearrangementタスク. WSまでに未公開

CVPR 2023 の動向・気付き（145/259）

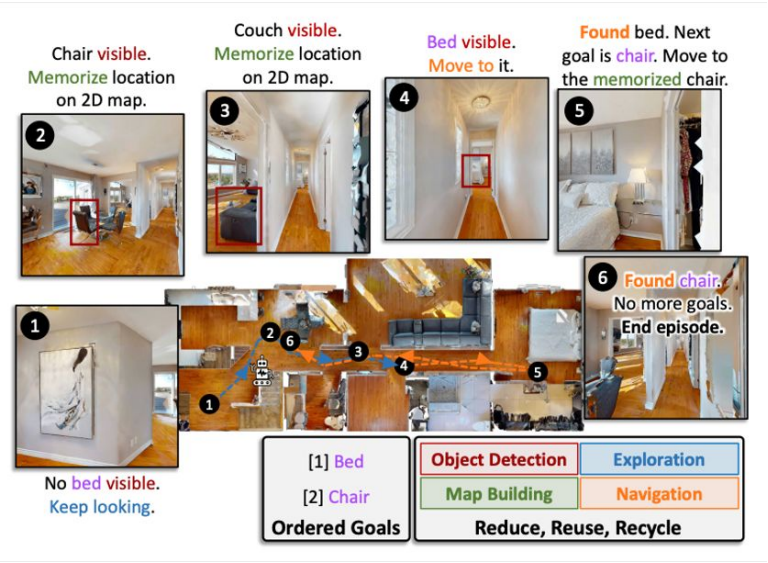
Workshop: Embodied AI

- あらゆるEmbodied AIの手法からみた手法のトレンド：
  - LLMsを使用し，Low-level認識から高レベルPlanningを行う
    - 比較的にLow-level: manipulation skillsなど
    - 比較的にHigh level: Task and motion planningなど
    - LLMsを使ってlong horizon planningを行う
  - Structuredな構造を使用
    - 例：GridMM（RxR-Habitatタスク優勝）では3種類のGraph構造を使用
  - 高性能のObject detection ¥Segmentationの事前学習モデルの使用
  - Module network (symbolic network)系の導入（次のページ）

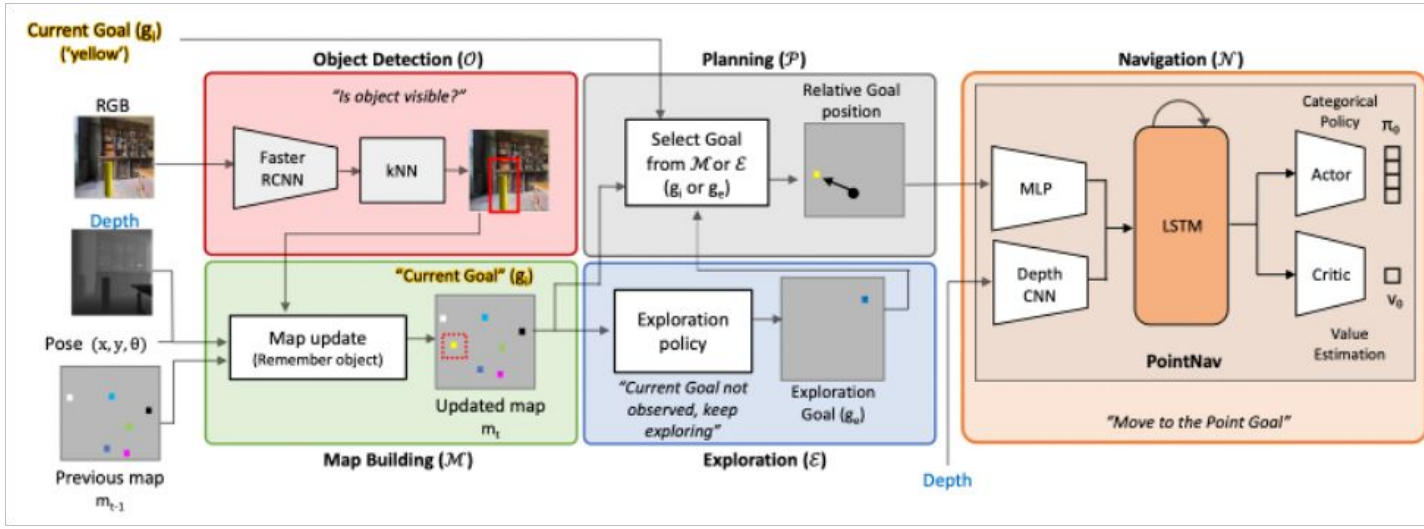
CVPR 2023 の動向・気付き(146/259)

Reduce, Reuse, Recycle: Modular Multi-Object Navigation (arXiv, 2023)

- Multi-object navigationタスクのためのNeural Module系手法の提案.
- 提案手法は5つのパーツ:物体検出, マップ生成, Exploration, Planning, navigationから構成され, 入力されたGoal(複数の対象物体)から, 複数のモジュールが連動して, Step-by-stepでタスクを解く.



実行する際の概念図



手法図

CVPR 2023 の動向・気付き（147/259）

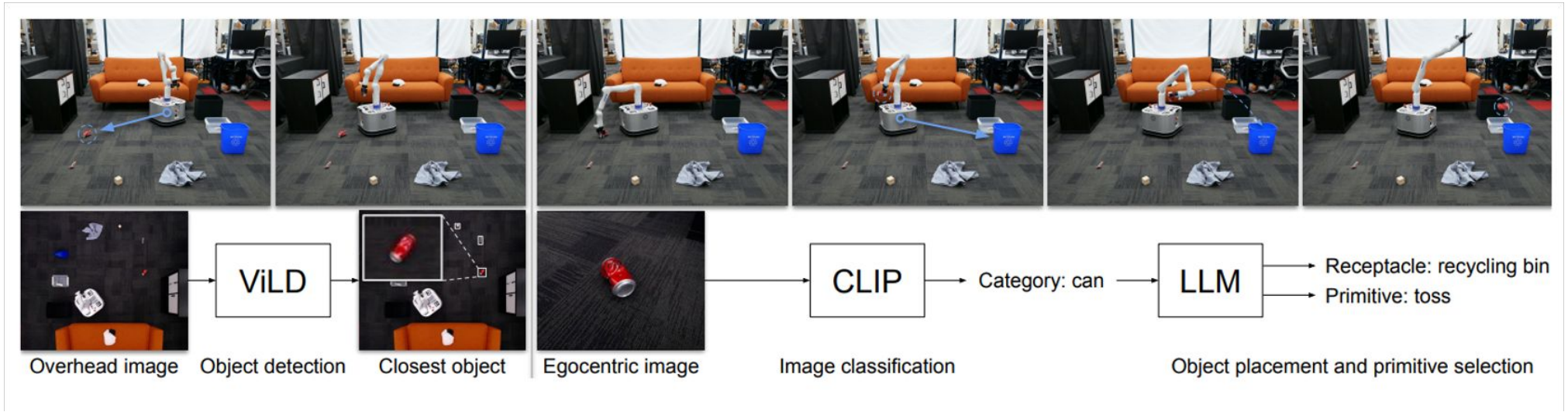
Workshop: Embodied AI

- 今後重要と思われる他のトピック:
  - Dynamic scene recognition
    - 変化する室内環境の理解が重要
    - 今回の10つのChallengeの中に2つはChange scene recognition
  - In contextでembodied Interaction
    - すでに一部やられているか，純粋な言語ベース／Vision +languageベースよりまだまだ検討する空間がある.
  - Diffusion モデルを導入されたFoundational EAI
    - 学習データ生成
    - Future prediction / planning
    - In context 学習可能にするなど
  - Personalized的な情報の学習（次のスライド）

CVPR 2023 の動向・気付き(148/259)

TidyBot: Personalized Robot Assistance with Large Language Models (arXiv, 2023)

- 言語指示や対話の形式で，Personalized化されたRobot assistance(物をピックアップし，Rearrangementする)できる手法の提案.
- また，Personalized情報は少ない学習examplesで人間とインタラクションし，人間の物置に関してのPreference Priors情報を学習する.
- 手法がシンプルだが，問題設定は重要でこれまでの検討が少なかった.



CVPR 2023 の動向・気付き（149/259）

Workshop: Omnilabel（1/2）

https://sites.google.com/view/omnilabel-workshop-cvpr23/overview

- カテゴリー名を超えたより複雑なラベル空間を推論するWS
- より汎用的にラベルの推論ができているかを評価するため，**Omnilabelデータセット**を用意
- 以下の3種類のラベルによってオブジェクトをさまざまな概念から分類できる
  - 広い範囲から狭い範囲まで全ての階層に対する名詞ラベル（動物＞犬＞コーギー）
  - 属性・機能・関係など，複数のカテゴリのオブジェクトをまとめて抽象化するラベル
  - 上記の2種類のラベルから記述される文章

CVPR 2023 の動向・気付き(150/259)

Workshop: Omnilabel (2/2)

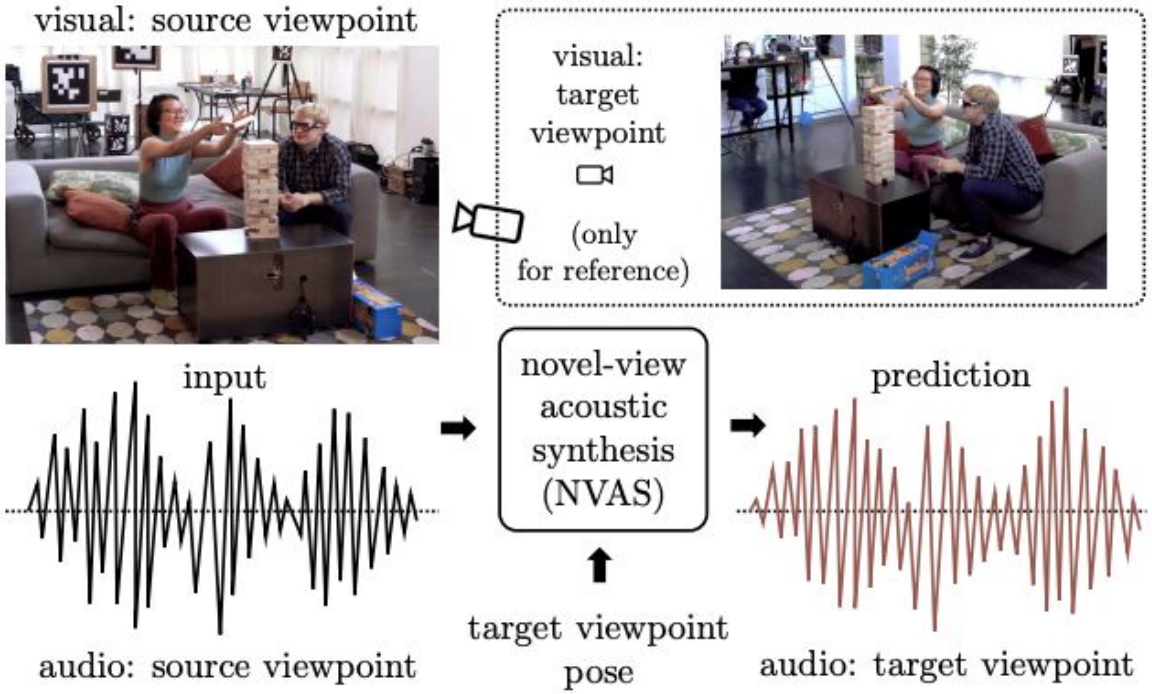
https://sites.google.com/view/omnilabel-workshop-cvpr23/overview

- NVIDIAはピクセルアノテーション無しで，テキスト画像間の対象学習によるセグメンテーション手法GroupViT(CVPR 2022)やDiffusionモデルとCLIPを組み合わせたパノプティックセグメンテーション手法ODISE(CVPR 2023)を紹介.
  - Diffusionモデルは今後識別タスクへの応用が増えていきそう
- その他には異なるデータセット間でのラベルの関係性理解タスクや，既存データセットよりもFine-GrainedなVQAデータセットなどが紹介されていた.
  - 既存モデルがどれほど汎用性があり，詳細を理解できるかはそれぞれのタスクで検討する必要がありそう

CVPR 2023 の動向・気付き(151/259)

音響情報&シーン理解

- Workshop “Sight and Sound” (1/2)
  - “Novel-View Acoustic Synthesis” (CVPR2023)
  - マイクの位置と向きを変化させた際に収音される音がどのように変化するのか予測
  - 反響特性を部分

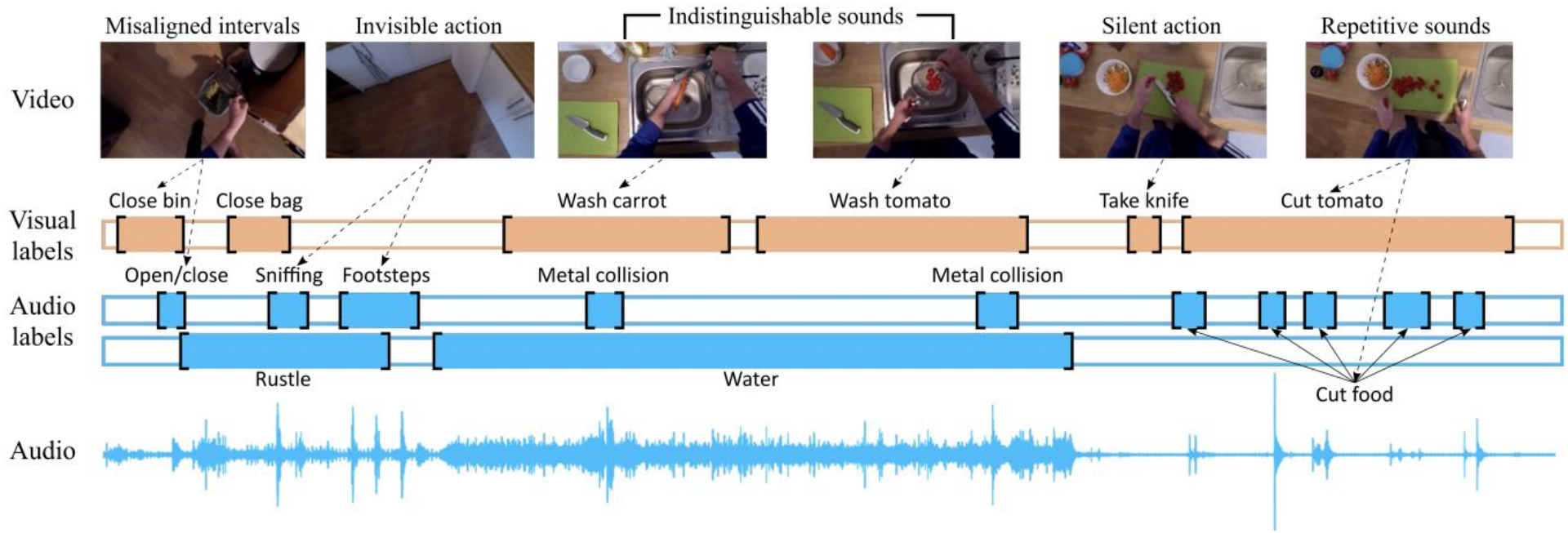


Chen, Changan, et al. "Novel-view acoustic synthesis." CVPR 2023.

CVPR 2023 の動向・気付き(152/259)

音響情報&シーン理解

- Workshop “Sight and Sound” (2/2)
  - “Epic-Sounds: A Large-Scale Dataset of Actions that Sound” (ECCV2023)
  - アクションセグメンテーションのアノテーションは視覚情報に影響を受けると結果に一貫性がなくなることがある
  - EF



Huh, Jaesung, et al. "Epic-Sounds: A Large-Scale Dataset of Actions that Sound." *ICASSP 2023*

CVPR 2023 の動向・気付き(153/259)

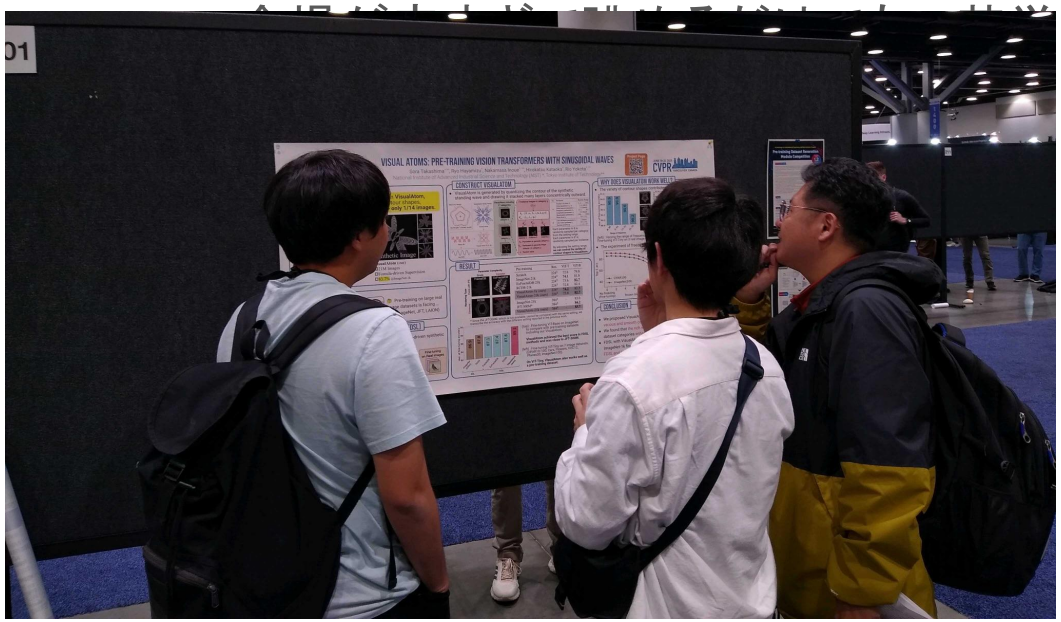
音響情報&シーン理解

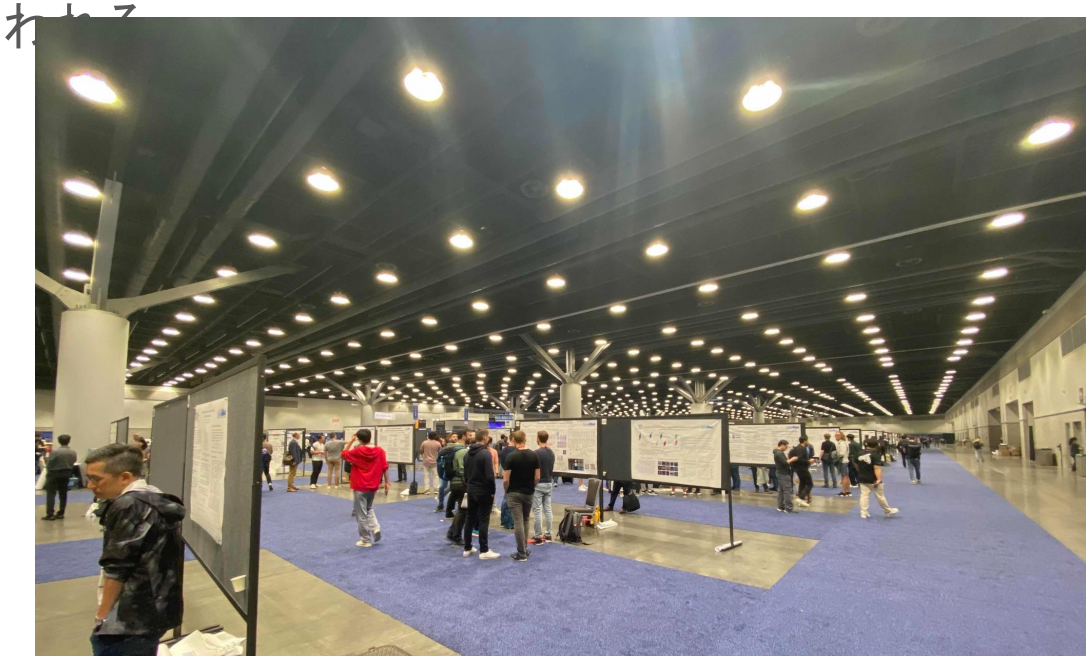
- マルチモーダルに注目が集まる中で音声&画像の研究が散見
  - Sound to Dance
    e.g., “EDGE: Editable Dance Generation From Music”
  - Sound to Speech
    e.g., “Generating Holistic 3D Human Motion from Speech”
  - Sound to 3D position
    e.g., “Seeing With Sound : Acoustic Beamforming for Multimodal Scene Understanding”
  - Sound to Image
    e.g., “Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment”

CVPR 2023 の動向・気付き（154/259）

ポススターセッション

- 3日間で2359本ものポスター発表を開催
  - ポスター発表は午前と午後に分けて開催
  - ポスターセッションの時間は2時間程度
  - その時間の間に約400本ものポスター発表が行われる



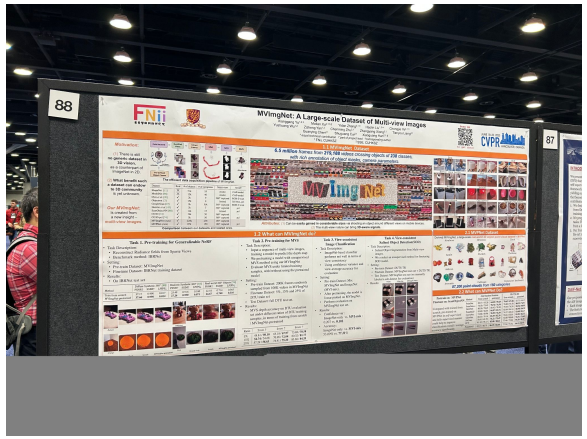


研究に関するまとめを一部紹介

CVPR 2023 の動向・気付き(155/259)

3Dビジョンのデファクトスタンダードな事前学習データセットへ！

- **MVImgNet: A Large-scale Dataset of Multi-view Images**
  - 事前学習用の大規模3Dデータセットが存在しない
  - NeRFを利用して大規模3Dデータセット MVImgNetを構築
    - 動画: 219,188, フレーム: 650M, クラス: 238
  - 3D点群物体認識において事前学習効果を発揮!!!
  - 著者曰く, データセット構築まで約1年かかったらしい.
  - 以下のURLからダウンロード可能

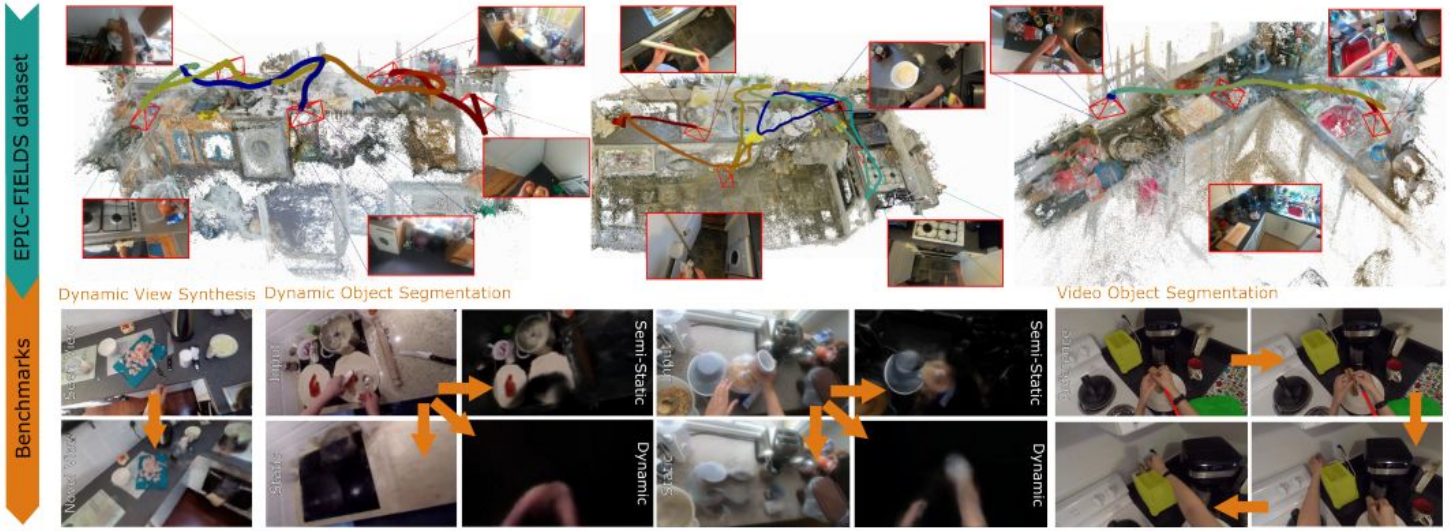


https://gaplab.cuhk.edu.cn/projects/MVImgNet/

CVPR 2023 の動向・気付き(156/259)

EPIC-KITCHENSがNeural Fields表現として拡張!!

- **EPIC Fields Marrying 3D Geometry and Video Understanding**
  - 3D幾何学と動画理解の融合への挑戦！
  - EPIC-KITCHENSを拡張し，人間と物体のインタラクションを3D情報込みで理解
  - EPIC-Fieldsデータセット
    - フレーム: 19M, 時間: 99h, Location: 45 kitchens
  - 以下URLからダウンロード可能!! Non-commercial license

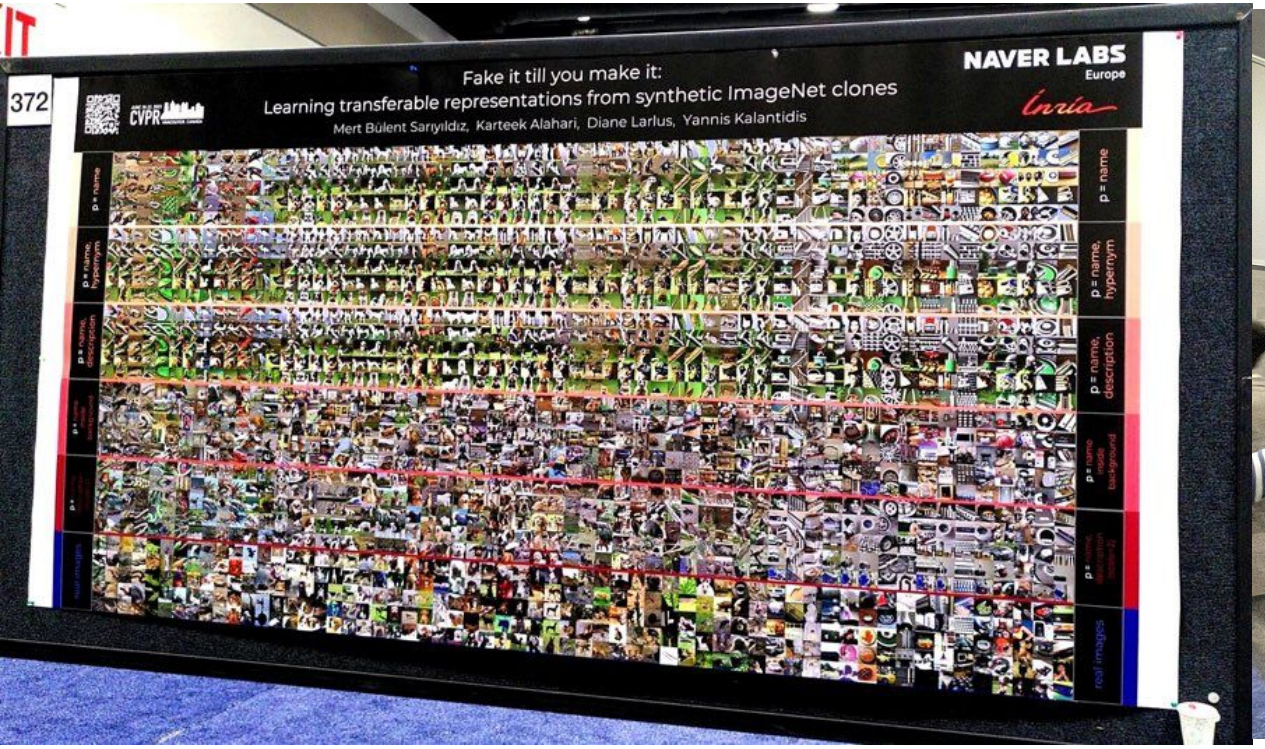


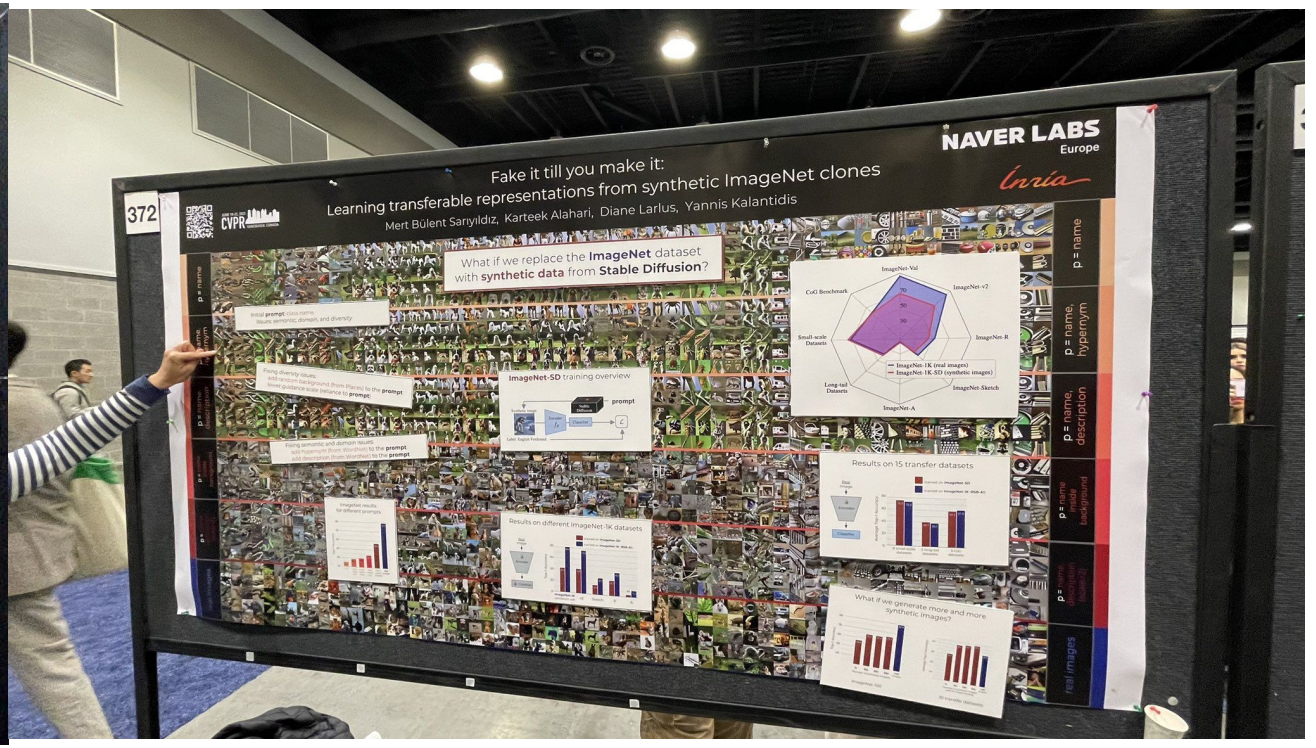
https://arxiv.org/pdf/2306.08731.pdf

CVPR 2023 の動向・気付き（157/259）

NAVER LABのポスターがすごい！

- 人の目を引くために著者らが作成したをポスター1面に記載
- 必要な実験結果などの情報は後から付け足す形式
- 聞きに来る人が多く隣のポスターに入り混む事に



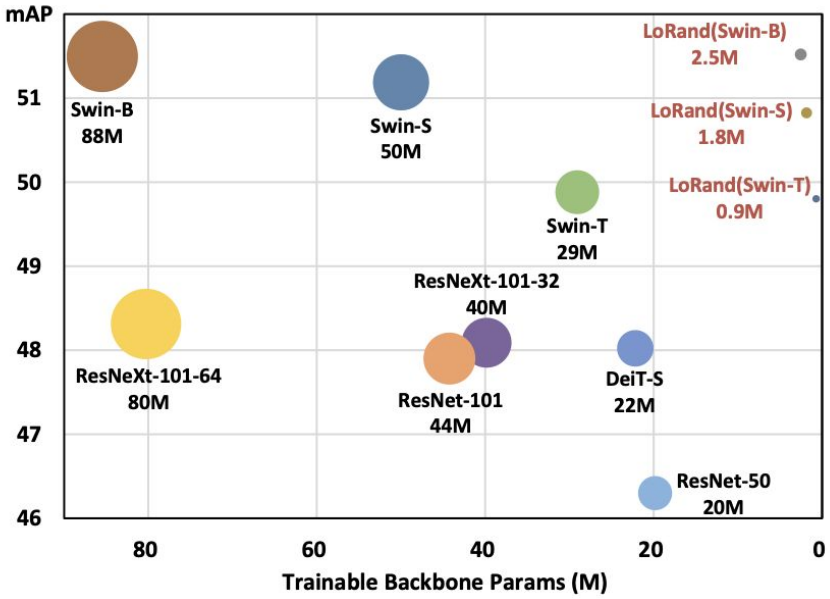


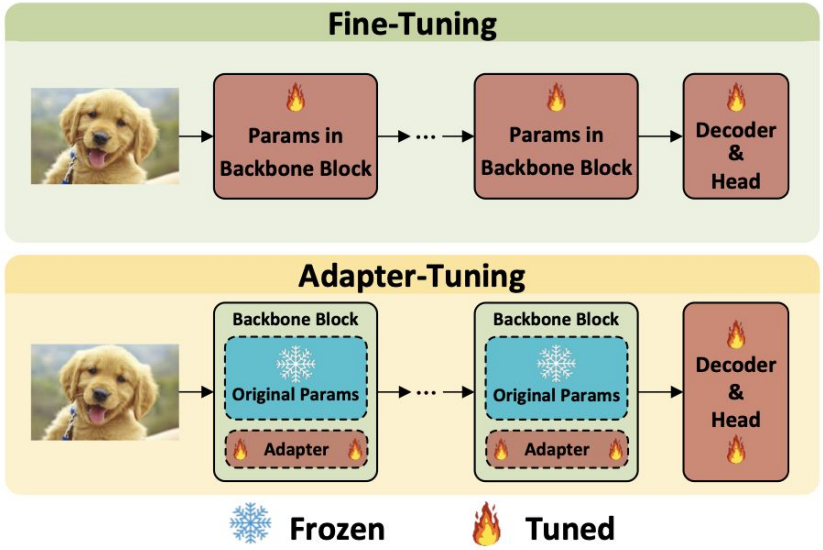
CVPR公式Twitterより引用
https://twitter.com/CSProfKGD/status/1671311393189167106?s=20

CVPR 2023 の動向・気付き(158/259)

1% VS 100%: Parameter-Efficient Low Rank Adapt

- **低ランク合成を用いることでパラメータ効率の良いAdaptation方法 Low Rank Adapter(LoRand)を提案**
  - 元のバックボーンパラメータを固定したまま，少数のパラメータだけを訓練することで効率的なモデル微調整が可能に
  - Fine-tuning中のバックボーンパラメータは大部分が冗長であり，それらはLoRandの追加パラメータ(1.8%～2.8%)で置き換えることができることを示した．これはモデルの効率性とストレージの節約につながる．

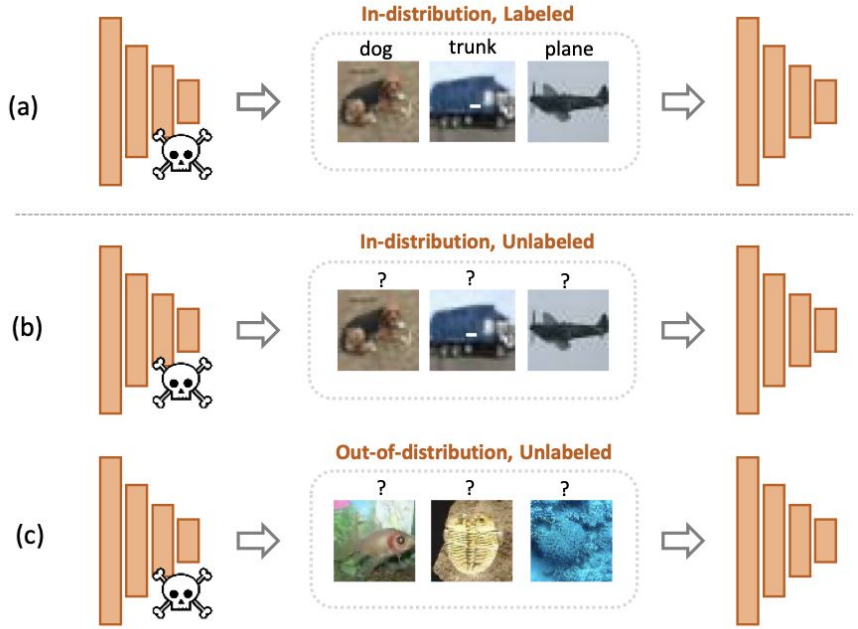




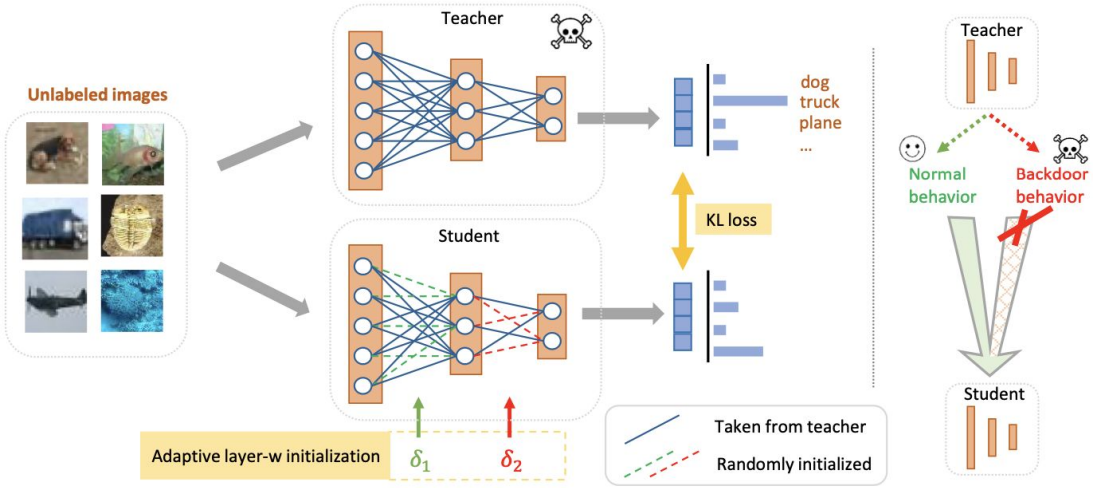
CVPR 2023 の動向・気付き(159/259)

Backdoor Cleansing with Unlabeled Data

- ラベル付データを必要としないバックドア攻撃から防御するための新しい方法を提案
  - 知識蒸留とレイヤーごとの重み再初期化を組み合わせ，疑わしいネットワークのバックドア行動を効果的に排除
  - 一方で，ネットワークの通常の動作にはほとんど影響を与えないことが確認
  - ラベル付きで訓練された最先端の防御方法と同等のパフォーマンスを示した



これまでの研究では，バックドアを浄化するためにラベル付き流通内データ(a)を使用していたが，本研究ではラベル付けされていない分布内データ(b)または分布外データ(c)を使用する．



CVPR 2023 の動向・気付き（160/259）

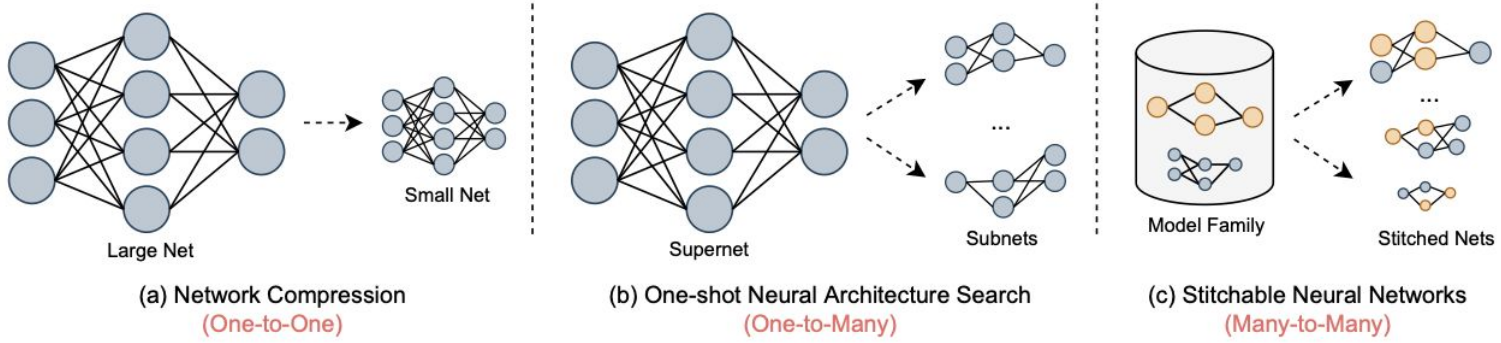
GENIE: Show Me the Data for Quantization

- 高品質な量子化ネットワークを効率的に生成するためのFramework「GEINE」を提案
  - データ生成（GENIE-D）とモデルの量子化（GENIE-M）の二つのサブモジュールから構成
  - GENIE-Dは，学習済みのモデルから知識を抽出し，それを用いて実データセットがない状況でも強力な量子化モデルを生成するための合成データセットを作り出す新たなフレームワーク
  - GENIE-Mは，合成データセットに対して量子化モデルを最適化し，量子化パラメータを同時に最適化する新たな手法を提供するフレームワーク

CVPR 2023 の動向・気付き(161/259)

Stitchable Neural Networks

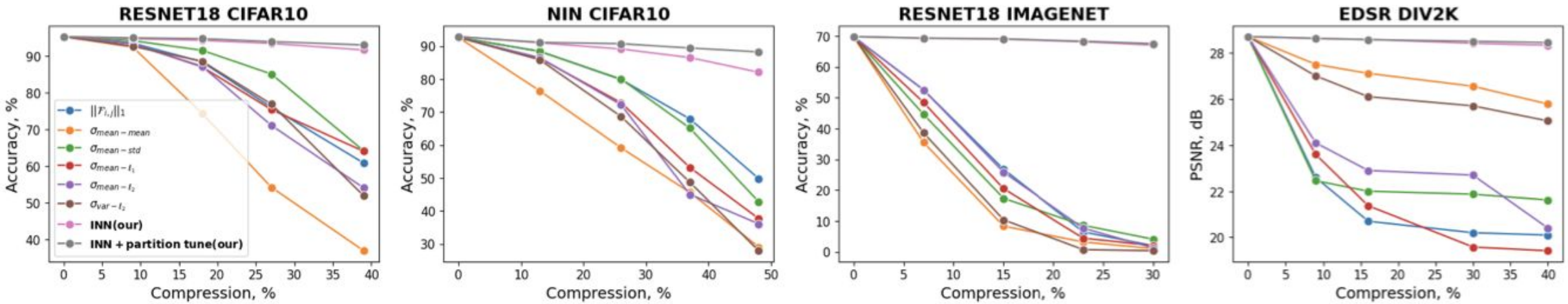
- 既存の学習済みモデル群(モデル族)を効率的に組み合わせて実行時に動的な精度と効率性のトレードオフを達成する新たなフレームワーク，"Stitchable Neural Networks (SN-Net)を提案
  - 既存の学習済みモデルの各ブロックや層を分割し，それらを新たな層(ステッチング層と呼ばれる)で結合
  - このステッチング層は，一つのアンカーモデルから別のアンカーモデルへの活性化マッピングを行う役割を持つ
  - これにより，あるアンカーモデルとその隣のアンカーモデルとの間にパフォーマンスの「パス」を生成し，そのパス上を動的に移動することで，異なる精度と効率性のトレードオフを実現



CVPR 2023 の動向・気付き(162/259)

Integral Neural Networks

- フィルタとチャネルの次元に沿った連続的な表現を使用し，任意のサイズへのディスクレティゼーションと構造的な剪定が可能な新しいNNクラス，Integral Neural Networks (INNs) を導入
  - その性能を従来の離散的なDNNと比較すると同等で，しかも高い剪定率でもほとんどパフォーマンスが低下せず，追加のファインチューニングが不要であることを実証

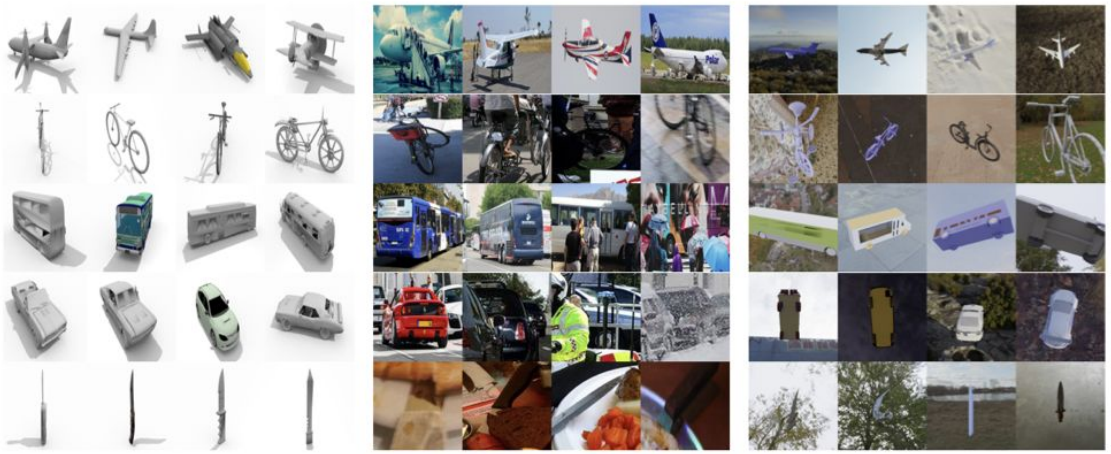


Cifar10上のResNet-18，Cifar10上のNINアーキテクチャ，ImageNet上のResNet-18，Div2k検証セット上の4x EDSR，compressionは削除されたパラメータの割合

CVPR 2023 の動向・気付き(163/259)

A New Benchmark: On the Utility of Synthetic Data with Blender for Bare Supervised Learning and Downstream Domain Adaptation

- 新たな大規模な合成から現実データのベンチマークS2RDAを提案
- 3Dレンダリングによる合成データを用いた新たな学習方法の提案
- 一般化への影響を調査するための画像形成要因の広範な評価，合成データと実データのpre-training間のtransferabilityを比較する新たなアプローチ

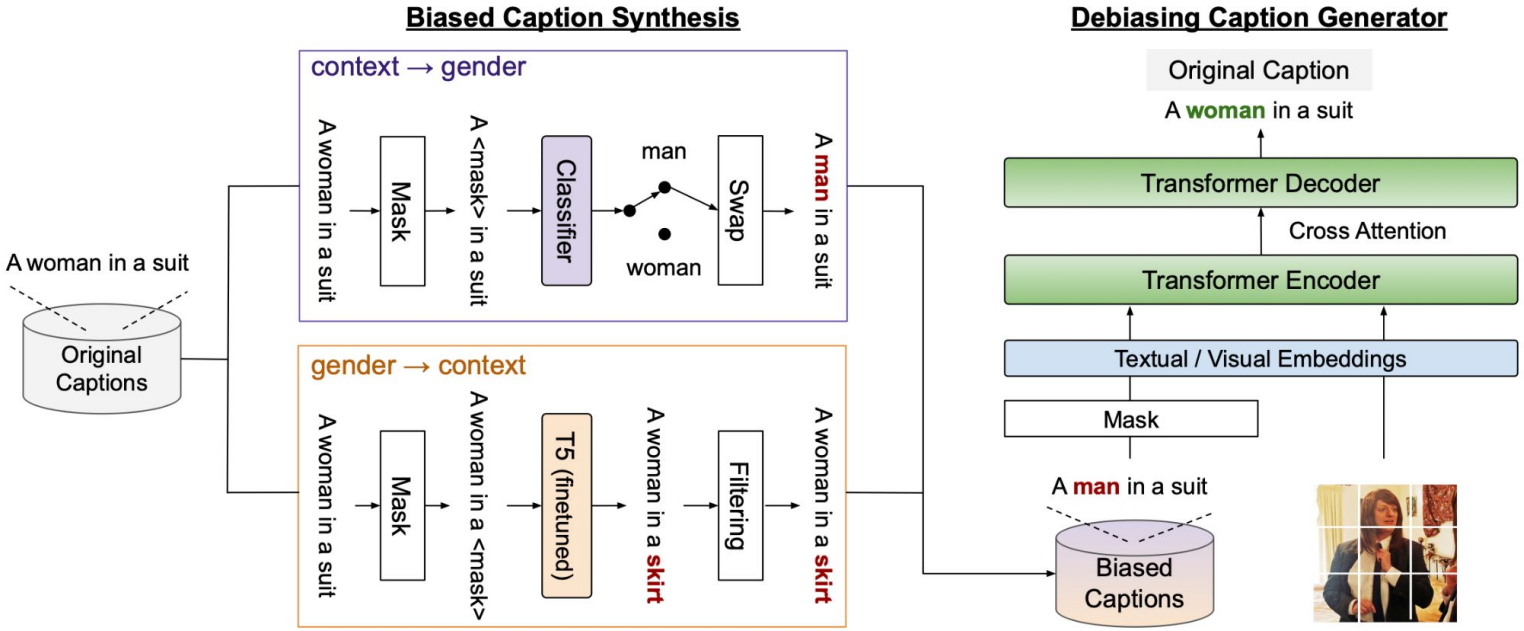


VisDA-2017のトレーニング領域(左)と検証領域(中央)のサンプル画像と，本論文で合成したデータ(右)

CVPR 2023 の動向・気付き(164/259)

Model-Agnostic Gender Debiased Image Captioning

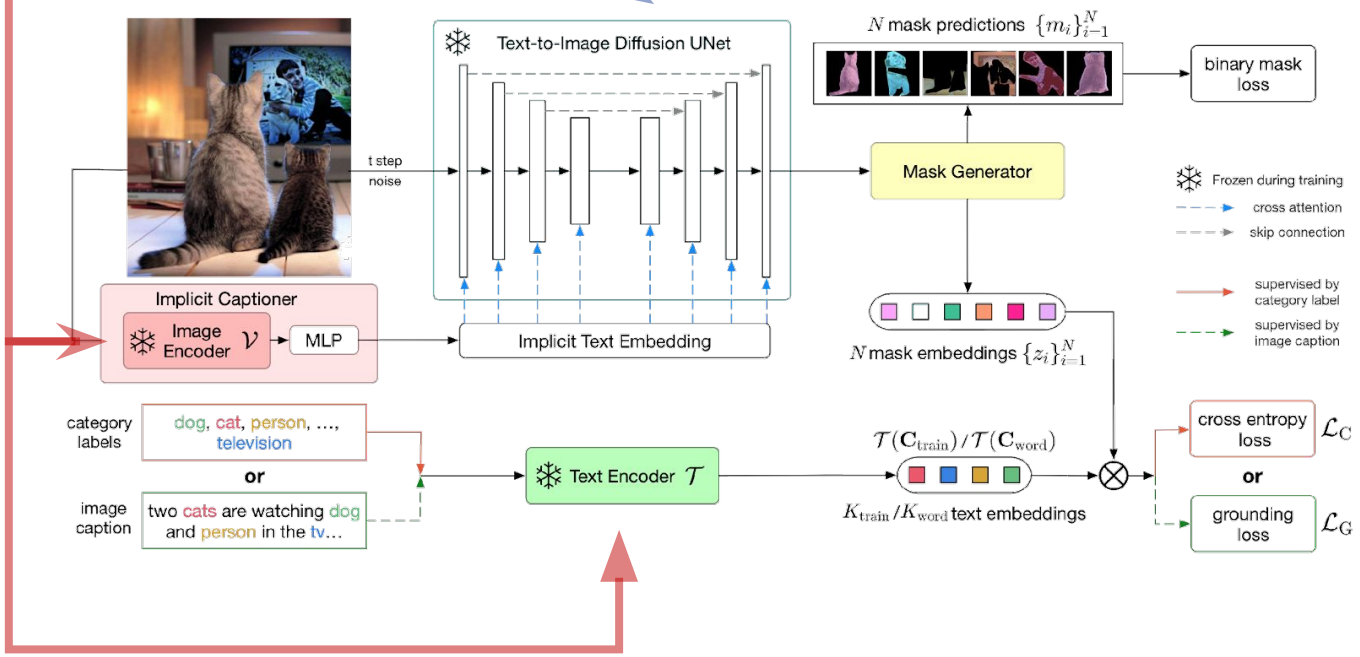
- 画像キャプショニングモデルにおけるジェンダーバイアスを軽減するための新しいフレームワーク，LIBRAを提案
  - LIBRAは，ジェンダーバイアスが含まれるキャプションを合成し(Biased Caption Synthesis)，その後，これらのバイアスを軽減するキャプション生成器(Debiasing Caption Generator)を訓練
  - LIBRAの効果は，様々な画像キャプショニングモデルと評価指標に対して確認された



CVPR 2023 の動向・気付き(165/259)

Highlight: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

- Stable Diffusionを特徴量抽出器として使用する動き？
  - 背景: 拡散モデルはtextとimageとのCross-Attentionを取る
    → 言語で表現されるような概念を潜在空間上で捉えている可能性がある
  - CLIPとStable Diffusionを使ったopen-vocab.なsegmentationモデルODISEを提案



中間層をk近傍法でクラスタリング
→ SDがmaskを捉えている可能性

CVPR 2023 の動向・気付き(166/259)

Tree Instance Segmentation with Temporal Contour Graph

(Adnan Firoze, Cameron Wingren, Raymond A. Yeh, Bedrich Benes, Daniel Aliaga)

- 密集している木を上空撮影した画像からインスタンスセグメンテーションを行う
- データセットの構築，新しいセグメント手法に貢献
  - 4つのデータセットを提案: Synthetic, Forest A(著者らが収集), Forest B,C(Google Earthから収集)
  - Limi 必要とする.

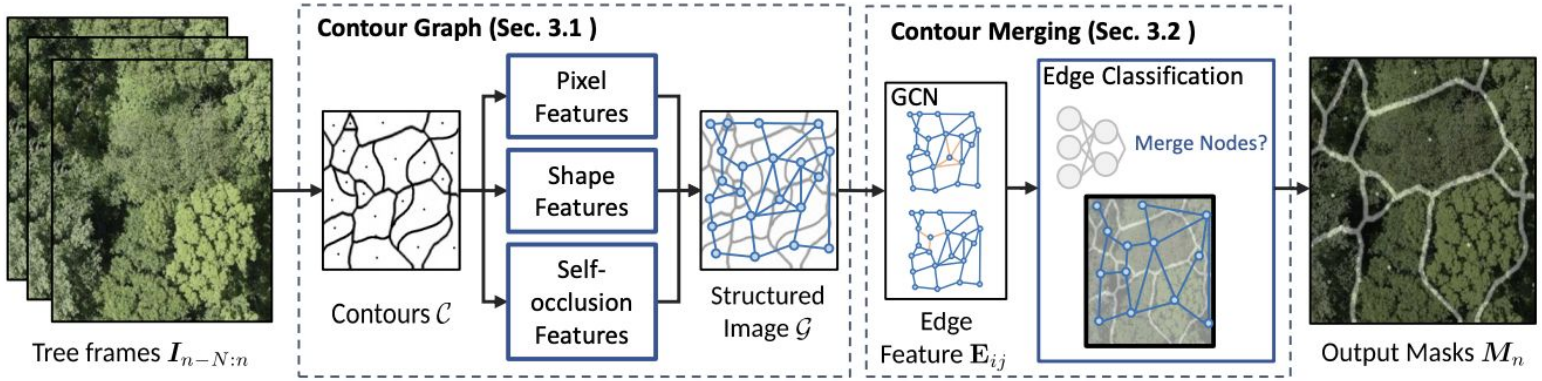


Figure 2. **Pipeline:** The input image sequence is analyzed. Initial contours and features are detected and organized into a contour graph that is refined by merging edges and nodes, resulting in the final output mask.

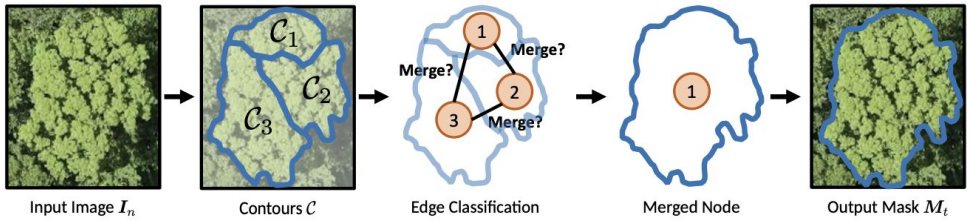


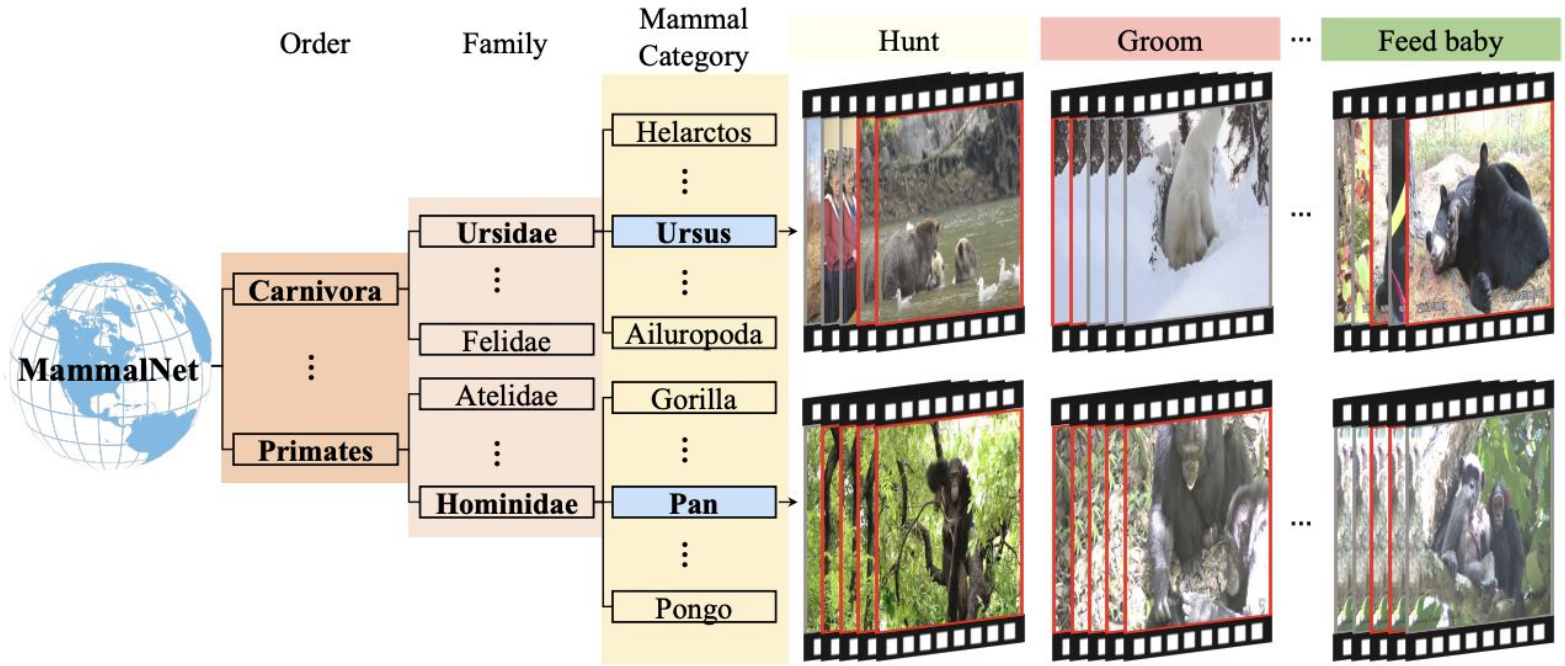
Figure 5. **Merging Process:** Illustrative presentation of our process to/from contour and graph spaces to merge noisy contours using edge classification (detailed in Sec. 3.2).

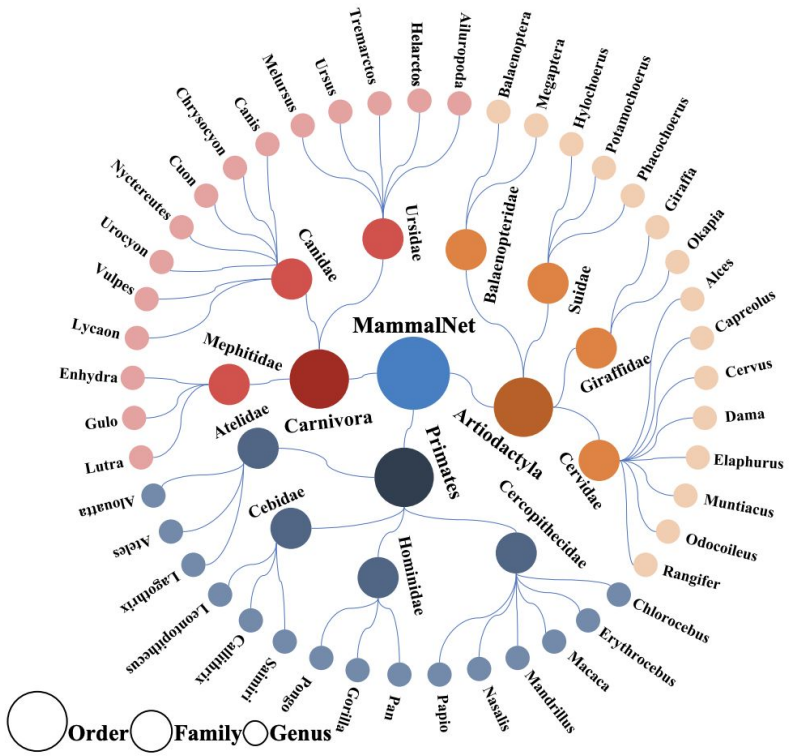
CVPR 2023 の動向・気付き(167/259)

MammalNet: A Large-scale Video Benchmark for Mammal Recognition and Behavior Understanding

(Jun Chen, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, Mohamed Elhoseiny)

- 173の哺乳類のカテゴリに対し，12種類の一般的な行動説明を加えたデータセットを提案．収録時間は539時間にのぼる．
- 動画からの動物の種類推定，行動推定に貢献するデータセ





cvpaper.challenge

CVPR 2023 の動向・気付き(168/259)

PartSLIP: Low-Shot Part Segmentation for 3D Point Clouds via Pretrained Image-Language Models

(Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, Hao Su)

- Low-shotでのPart Segmentaitonを事前学習済みの画像-テキストモデル(GLIP)を使って実現するPartSLIPを提案
- セグメンテーションはSemanticでもInastanceでも可能
- 3Dのpointcloudを様々な角度から取得した2D画像へと変換することで，GLIPとの接続を可能にした

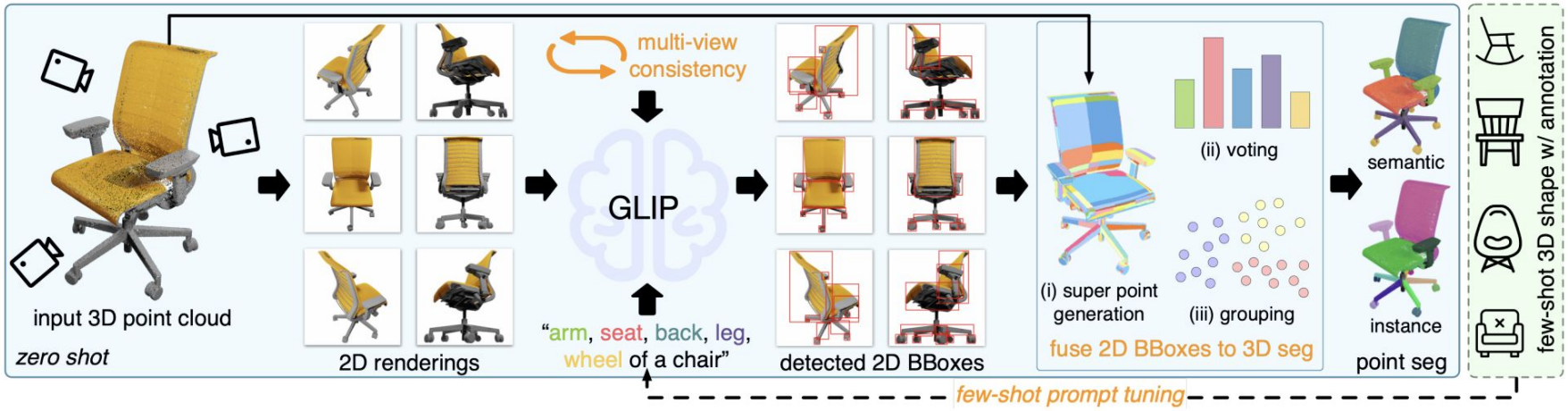


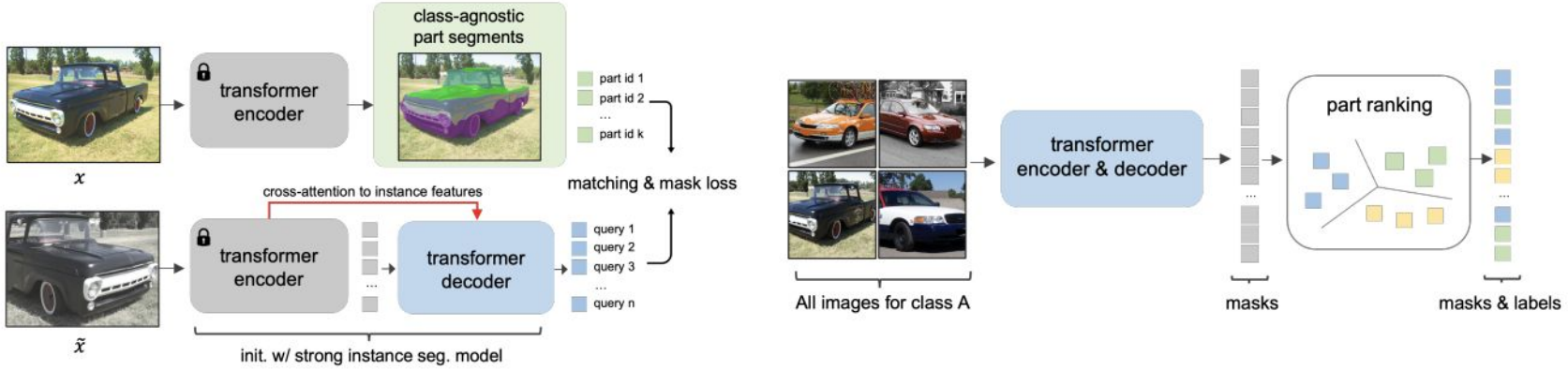
Figure 2. The figure shows our overall pipeline. Our proposed components are highlighted in orange.

CVPR 2023 の動向・気付き(169/259)

PartDistillation: Learning Parts from Instance Segmentation

(Jang Hyun Cho, Philipp Krähenbühl, Vignesh Ramanathan)

- Low-shotでのPart Segmentaitonを事前学習
- データセット内の全ての画像に対してpixcel-levelの特徴量からPart Segmentationを行うことで，本当に正しいPartの定義を学習させていく
- その後，
- Part Se



(a) First stage: *Part-proposal learning* **(b)** Second stage: *Part ranking*

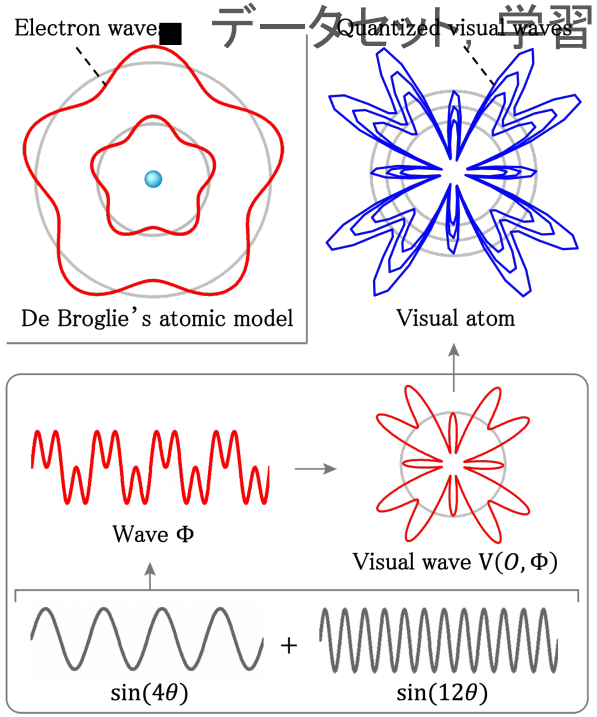
Figure 2. Overview of PartDistillation. **Left**: In the first stage, a transformer encoder produces instance segmentation feature which we group into class-agnostic part segments, *part proposals*, as described in Sec. 4.2. We then train a separate transformer decoder bootstrapped from these part segments and improved through self-training. **Right**: In the second stage, we assign part labels for all part-regions in a class by clustering across dataset and ranking by the density estimates of the clusters. We call this process *class-specific part ranking*.

CVPR 2023 の動向・気付き(170/259)

Visual Atoms: Pre-training Vision Transformers with Sinusoidal Waves

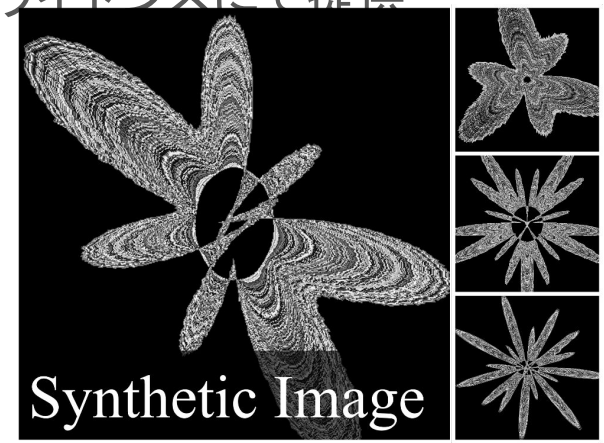
(Sora Takashima , Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, Rio Yokota)

- 新規FDSLが，約1/14のデータサイズでJFT-300Mに迫る事前学習効果を達成
  - FDSL：数式ドリブンの自動生成人工画像のみで事前学習を行う手法
  - 様々な振動数の定常正弦波を合成することによって，従来手法よりも多様かつ滑らかな曲線を表現可能となり，FDSL手法中でSOTAを達成
  - データセット，学習済みモデルを商用利用可能なライセンスにて提供



JFT-300M
- 300M Images
- Human Supervision
- 84.2% @ImageNet-1k

1/14 size



VisualAtom (our)
- 21M Images
- Fomula-driven Supervision
- 83.7% @ImageNet-1k

Pre-training	Res.	ViT-T	ViT-B
Scratch	224^2	72.6	79.8
ImageNet-21k	224^2	**74.1**	81.8
ExFractalDB-21k	224^2	73.6	**82.7**
RCDB-21k	224^2	72.8	82.4
VisualAtom-1k (ours)	224^2	**74.2**	82.3
VisualAtom-21k (ours)	224^2	73.8	**82.7**
ImageNet-21k	384^2	-	83.0
JFT-300M*	384^2	-	**84.2**
VisualAtom-21k (ours)	384^2	-	**83.7**

CVPR 2023 の動向・気付き(171/259)

医療画像ではどのような研究が？

- 目立つのは病理画像解析・CT/MRI画像セグメンテーション・X線画像に対する読影レポート生成
  - **病理画像解析**
    - AMIGO: Sparse Multi-Modal Graph Transformer with Shared-Context Processing for Representation Learning of Giga-pixel Images
    - Histopathology Whole Slide Image Analysis With Heterogeneous Graph Representation Learning
    - Weakly supervised segmentation with point annotations for histopathology images via contrast-based variational model
    - Benchmarking Self-Supervised Learning on Diverse Pathology Datasets
    - Task-Specific Fine-Tuning via Variational Information Bottleneck for Weakly-Supervised Pathology Whole Slide Image Classification
    - Topology-Guided Multi-Class Cell Context Generation for Digital Pathology
    - OCELOT: Overlapped Cell on Tissue Dataset for Histopathology
    - Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images
    - Hierarchical Discriminative Learning Improves Visual Representations of Biomedical Microscopy
    - RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images With Diverse Sizes and Imbalanced Categories
    - Bi-Directional Feature Fusion Generative Adversarial Network for Ultra-High Resolution Pathological Image Virtual Re-Staining
    - Interventional Bag Multi-Instance Learning on Whole-Slide Pathological Images

CVPR 2023 の動向・気付き(172/259)

医療画像ではどのような研究が？

- 目立つのは病理画像解析・CT/MRI画像セグメンテーション・X線画像に対する読影レポート生成
  - **CT/MRI画像セグメンテーション**
    - Geometric Visual Similarity Learning in 3D Medical Image Self-supervised Pre-training
    - SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation
    - Devil Is in the Queries: Advancing Mask Transformers for Real-World Medical Image Segmentation and Out-of-Distribution Localization
    - Rethinking Bayesian Deep Learning Methods for Semi-Supervised Volumetric Medical Image Segmentation
    - Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation
    - Orthogonal Annotation Benefits Barely-Supervised Medical Image Segmentation
    - Rethinking Few-Shot Medical Segmentation: A Vector Quantization View
    - Ambiguous Medical Image Segmentation Using Diffusion Models
    - Fair Federated Medical Image Segmentation via Client Contribution Estimation
    - Pseudo-Label Guided Contrastive Learning for Semi-Supervised Medical Image Segmentation
    - Neuralizer: General Neuroimage Analysis without Re-Training

CVPR 2023 の動向・気付き(173/259)

医療画像ではどのような研究が？

- 目立つのは病理画像解析・CT/MRI画像セグメンテーション・X線画像に対する読影レポート生成
  - **X線画像に対する読影レポート生成**
    - METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens
    - Interactive and Explainable Region-Guided Radiology Report Generation
    - KiUT: Knowledge-Injected U-Transformer for Radiology Report Generation
    - Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing
  - **その他**
    - Image Quality-Aware Diagnosis via Meta-Knowledge Co-Embedding
    - Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction
    - Directional Connectivity-Based Segmentation of Medical Images

CVPR 2023 の動向・気付き(174/259)

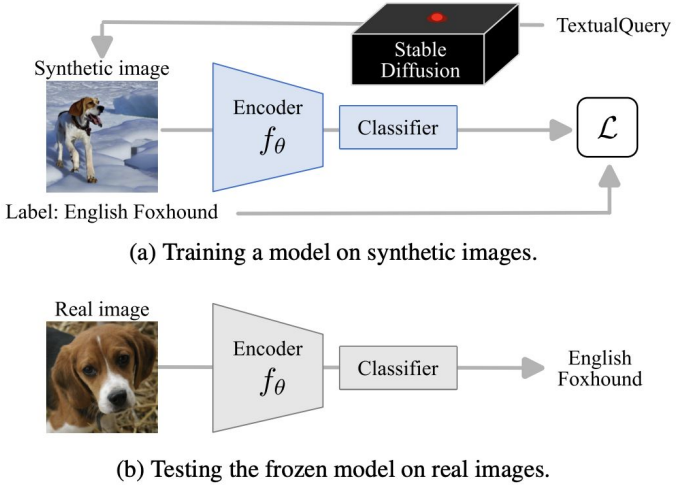
近年急速な発展を見せるデータセット蒸留も多くみられる

- データセット蒸留は, テスト精度を維持しつつデータサイズを小さくする手法
  - Accelerating Dataset Distillation via Model Augmentation
    - 従来のデータ蒸留では, 蒸留時に用いたアーキテクチャに過剰適合してしまう問題があった.
    - メタ学習のouter-loopごとに, 用いるアーキテクチャをランダムに変えることで過剰適合の問題を軽減し, 性能向上
  - Generalizing Dataset Distillation via Deep Generative Prior
    - 従来の手法は蒸留画像の各ピクセル値を直接最適化するため計算コストが大きく, 高い解像度・データ枚数への蒸留が困難であった.
    - 生成ベースで蒸留画像を最適化することで, 上記の問題を解決
  - Minimizing the Accumulated Trajectory Error To Improve Dataset Distillation
    - CVPR2022で提案されたSOTA手法Trajectory Matchingにおける問題を軽減し, 性能向上
  - Slimmable Dataset Condensation
    - 従来のデータセット蒸留では, 最初に固定したデータ容量から変更するには再度データ蒸留を行う必要があった
    - 用途に応じてデータ容量を変更できる手法を提案
  - Improved Distribution Matching for Dataset Condensation
    - Distribution Matchingベースの手法は他の勾配ベースの手法などと比較しても計算量が小さいものの, 性能が低いという問題点があった.
    - Distibution Matchingにおける問題点を解析し, 性能改善

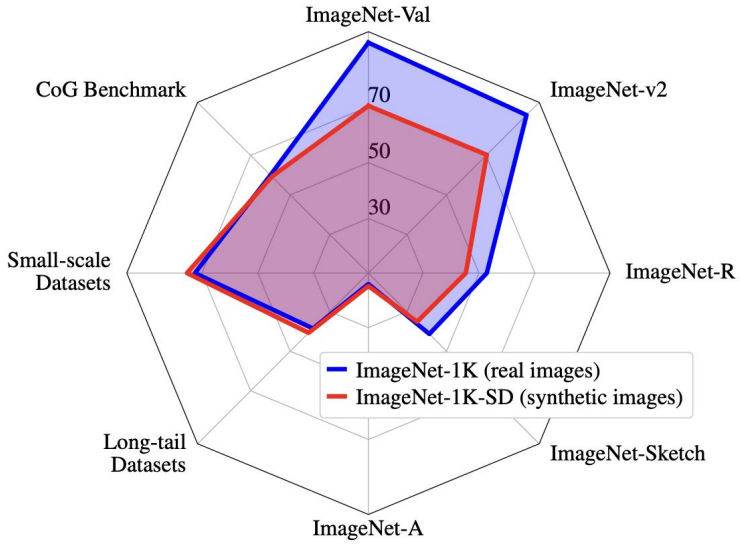
CVPR 2023 の動向・気付き(175/259)

Fake it till you make it: Learning transferable representations from synthetic ImageNet clones

- Stable DiffusionでImageNetデータセットを生成し様々な評価を行なった
  - Stable diffusionでデータ生成して学習する際の性能を網羅的に実験.
  - 良い結果：データセット規模が同じく1Kの場合で，小さいデータセットでImageNetを超える場合もある．また，Finetuneの場合ではImageNet精度を大幅に超える場合がある.
  - 他の結果：同じ規模でFrozen特徴量の場合．Scaling効果があまり観測



実験のprotocol

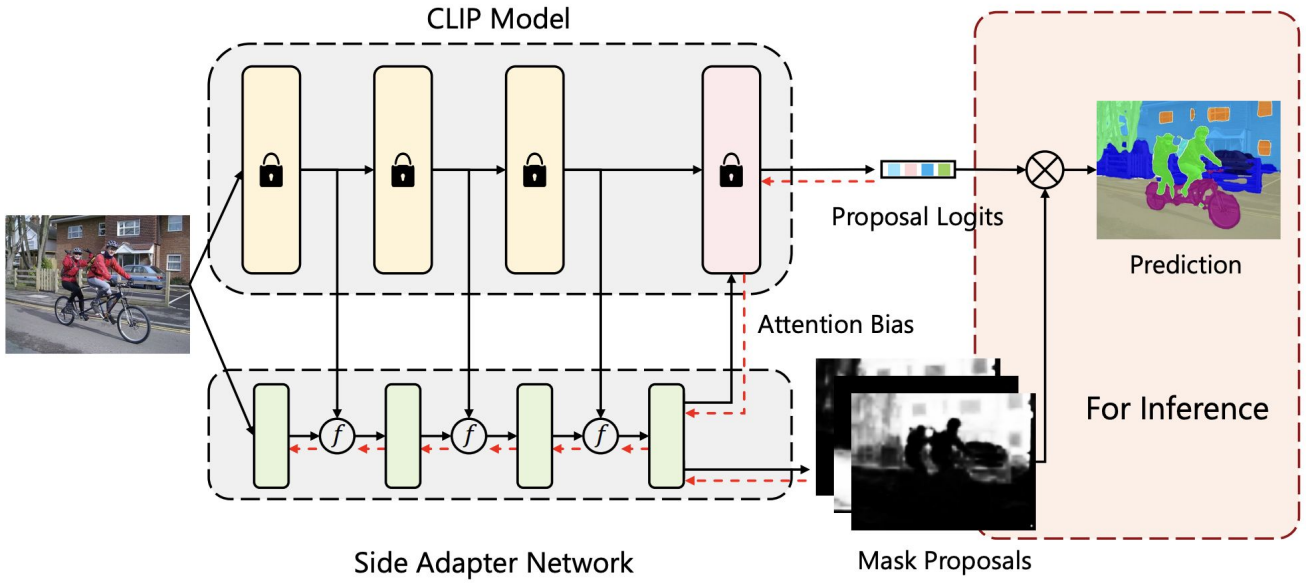


結果比較

CVPR 2023 の動向・気付き(176/259)

Side Adapter Network for Open-Vocabulary Semantic Segmentation

- CLIP(frozen)を使ったopen-vocabulary semantic segmentation手法
  - CLIPをopen-vocabularyのセグメンテーションに使ったほぼほぼ最初の文章？
  - CLIPモデルに2つのブランチを追加：mask proposalはマスク領域を生成し，attention biasを生成しCLIPモデルからmaskのクラスを推定する.
  - FrozenCLIPをベースにして既存の研究よりはモデルが軽量的で学習スピードも速い.



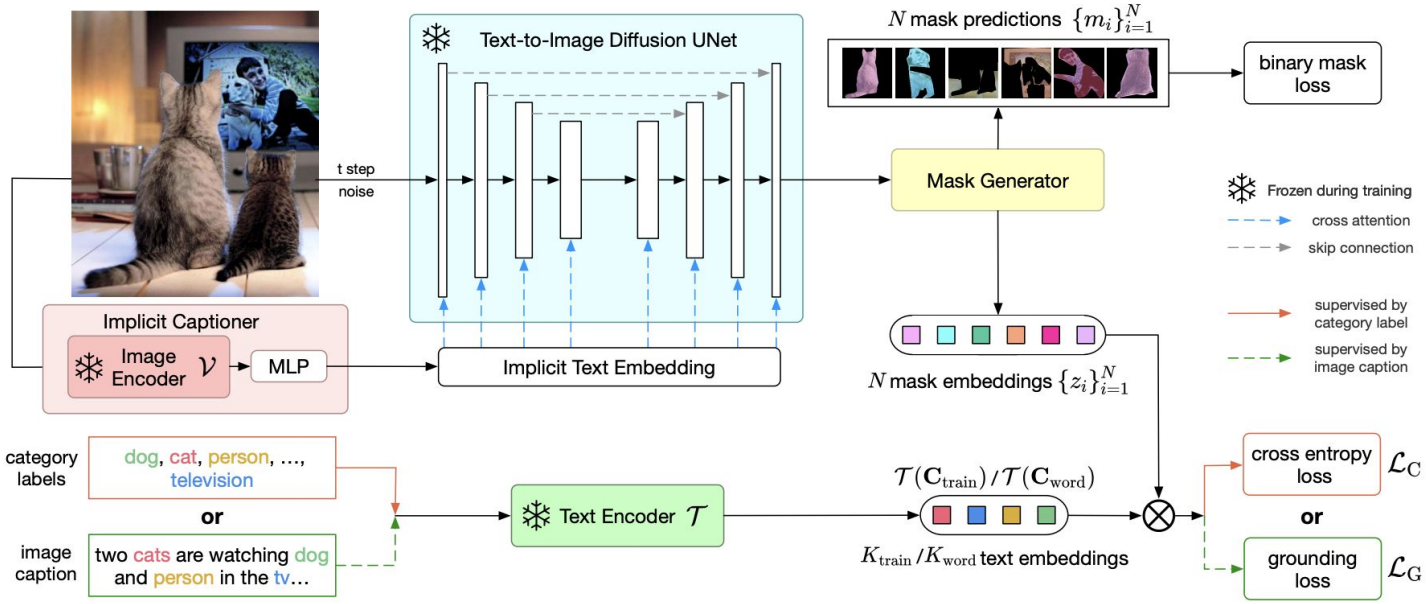
手法図

結果例

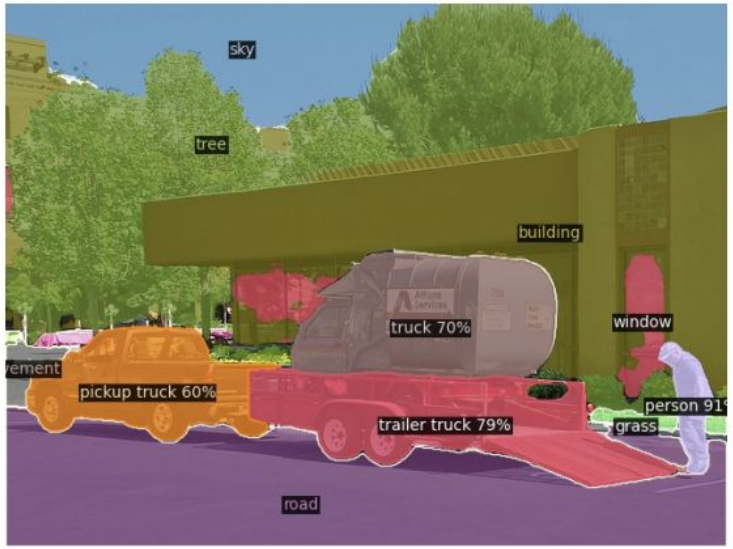
CVPR 2023 の動向・気付き(177/259)

Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

- DiffusionモデルとCLIPを使ったopen vocabulary panoptic segmentation手法
  - まず画像をキャプションモデルとMLPでImplictなテキスト特徴量に変化し，そこから学習済みのdiffusion unet から diffusion 特徴を抽出し，マスク生成器でマスクとマスクエンコーディングを生成.
  - マスクエンコーディングと学習済みモデルで抽出したテキストエンコーディングの類似性をベースに評価.
  - CVPR2023では類似したモデルで同じタスクを解くものが多数あった.



提案手法

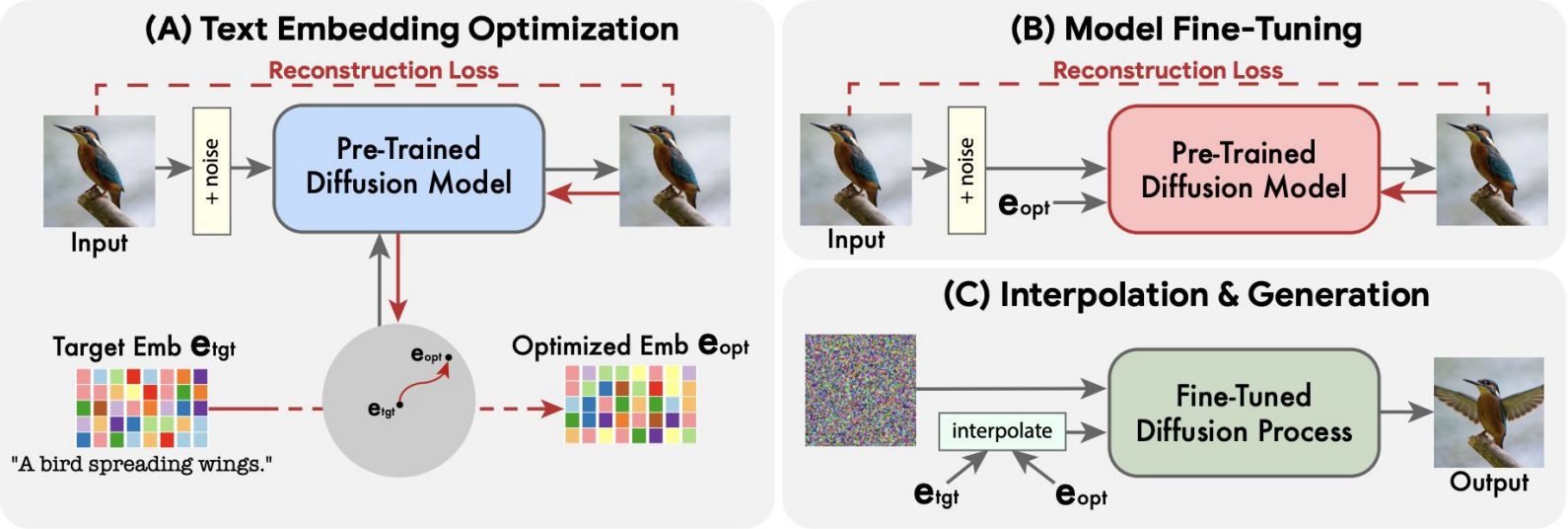


結果例

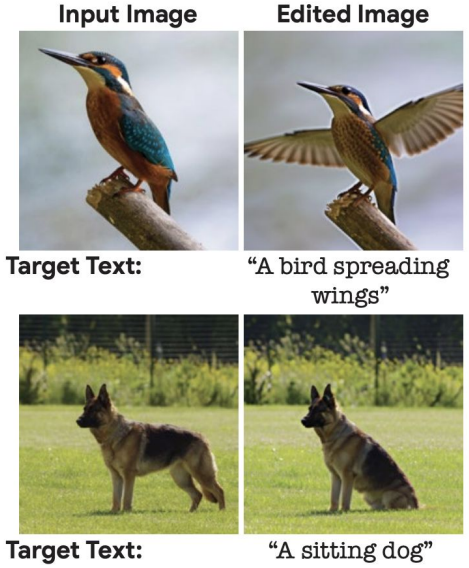
CVPR 2023 の動向・気付き(178/259)

Imagic: Text-Based Real Image Editing with Diffusion Models

- Diffusion modelを使ったテキストベース画像編集の研究
  - 既存研究でうまく扱えなかった”non-rigid”編集や,複数の物体編集などを可能にした.
  - 提案手法は学習済みのtext-to-image diffusionモデルを利用する. 具体的に, まず入力画像とターゲットテキスト両方にアラインしたテキストエンベディングを生成する. その後に, オリジナル画像のアピアランスをキャプチャーしながら, diffusion modelを再学習しターゲット画像を生成する.



提案手法

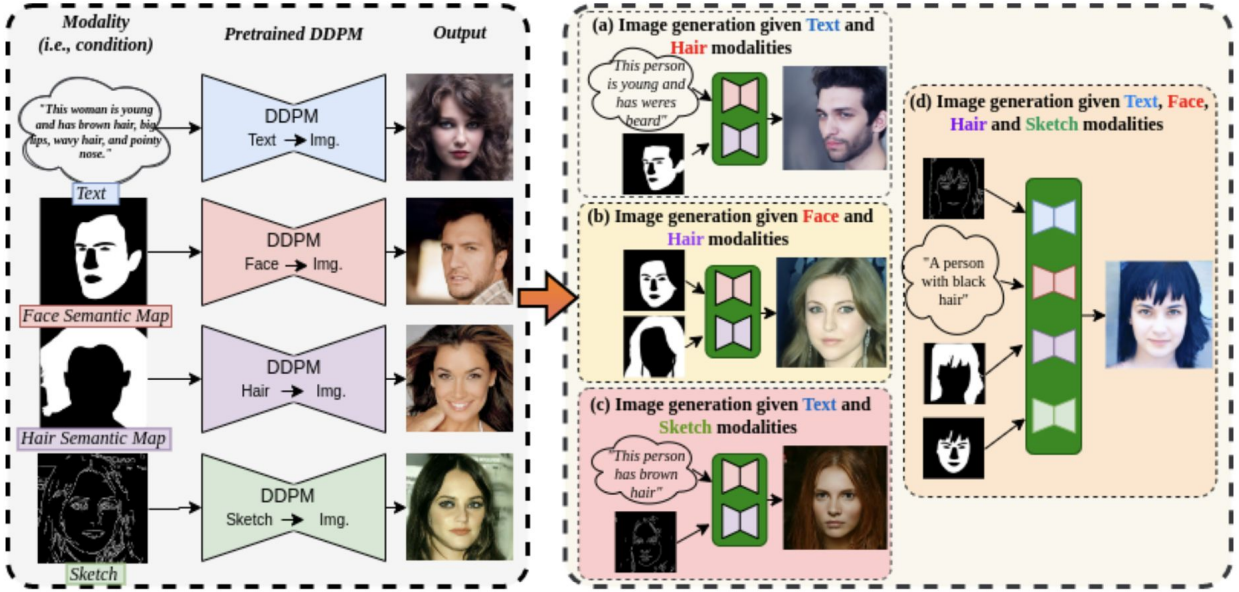


結果例

CVPR 2023 の動向・気付き(179/259)

Unite and Conquer: Plug and Play Multi-Modal Synthesis Using Diffusion Models

- Diffusionモデルをマルチモーダルconditioned画像生成に適応した
  - 複数のモーダル(言語, マスク, エッジ, セグメンテーション)の入力からそれぞれの一致性が保たれた画像生成ができる. また, 同じモダリティの複数のリソースも合成可能(背景と全景から画像合成など).
  - 提案手法は複数のソースで学習されたdiffusionモデルからsampling timeを調整し, 最終の生成を足し合わすことで学習せずに生成できる. (モデル理解できていない詳細は論文を参照)

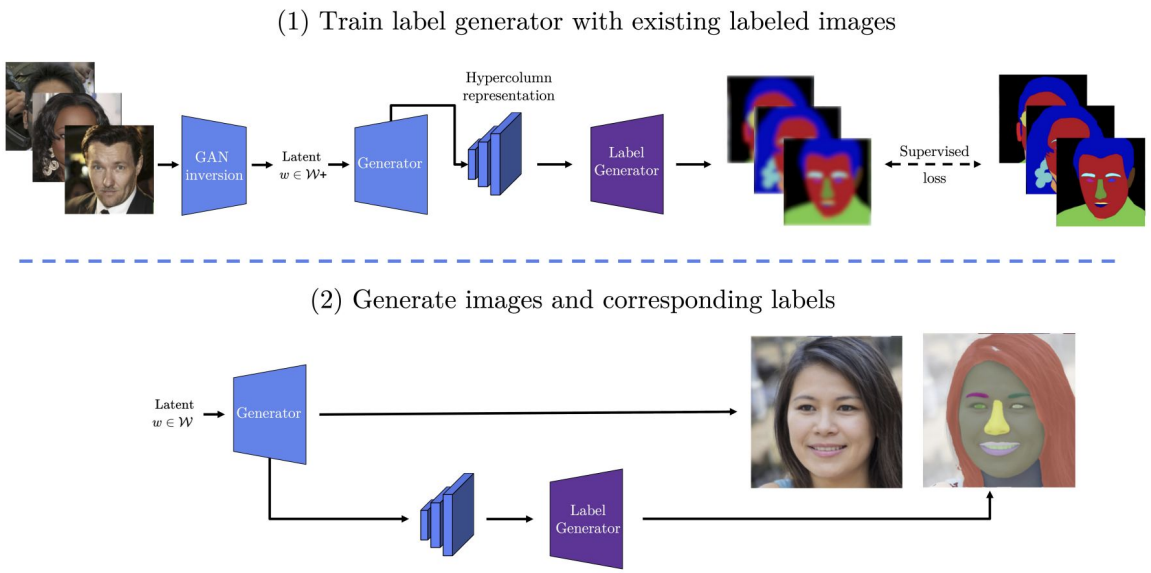


生成の概念図

CVPR 2023 の動向・気付き(180/259)

HandsOff: Labeled Dataset Generation With No Additional Human Annotations

- 少ないデータセットexampleから学習データセットを自動生成する手法を提案
  - 追加human label使わずに50程度のexampleから学習データを無限？に自動生成できる.
  - 提案手法はgan inversionを使ってdataset generation用のパラメータを学習.
  - 学習データ(dense prediction task)が少ない時にすぐにデータ生成可能で様々なところで活用可能.



提案手法

CVPR 2023 の動向・気付き（181/259）

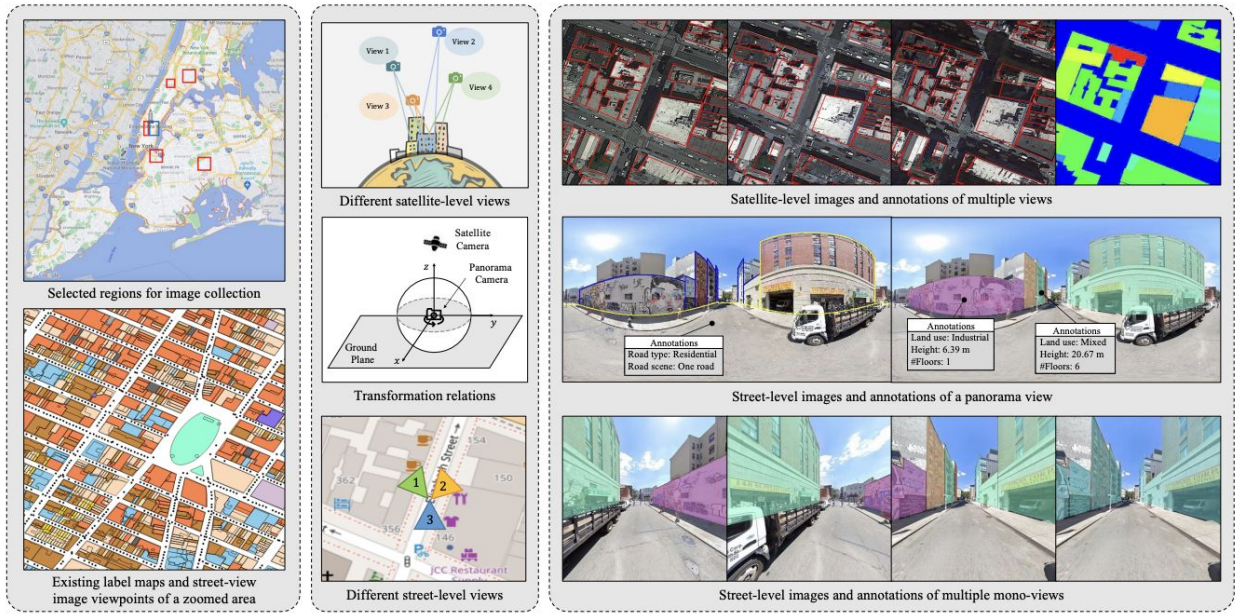
新しいデータセット

- 屋外環境データセット
  - Omnicity：屋外シーンの大規模データセット，複数視点でlayoutのアノテーション付き
- 室内環境3次元データセット：HM3D（Highlight）
  - データセットの分野においての有用性や，作成コストなどが評価された？
- 物体3次元データセット：
  - OmniObject3D（**Award nominate**）：大規模物体3次元データセット（6000+, growing）で，様々なタスクでの有用性が示された
  - Objaverse：800K＋（and growing）　3次元物体
- human poseデータセット
  - SLOPER4D：Sceneを考慮した4D Human Poseデータセット
  - BEDLAM: 4D Human Poseデータセット(詳細動作，服装の動きの細部まで)

CVPR 2023 の動向・気付き(182/259)

OmniCity: Omnipotent City Understanding with Multi-level and Multi-view Images

- 屋外のシーンの詳細理解のための大規模データセットOmniCityの提案
  - OmniCityでは屋外シーン(New York City)をベースに，多視点衛星画像，街レベルのパノラマ画像，mono-view画像から構成する.
  - これまでの屋外データセットではセマンティックセグメンテーションのアノテーションまでされるものが少なかったため，今後屋外シーンの意味理解にデータセットが活用できる.
  - OmniCityデータセットと同時に，いくつかの室外シーン理解のベンチマークも提案した.

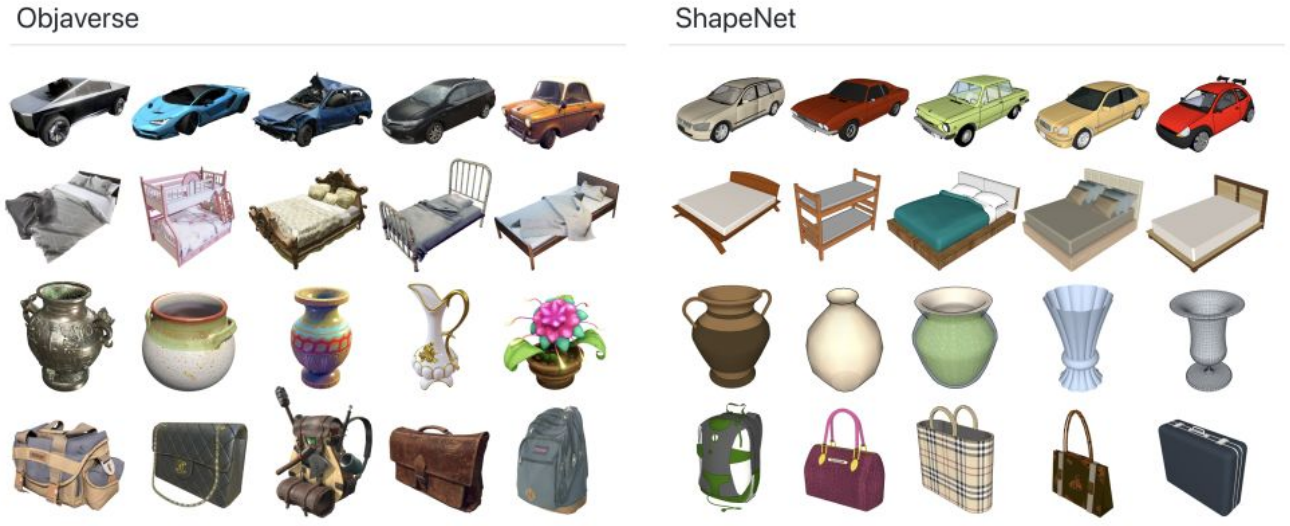


OmniCityデータセットの
イメージ図

CVPR 2023 の動向・気付き(183/259)

Objaverse: A Universe of Annotated 3D Objects

- 大規模3次元物体モデルのデータセットObjaverseの提案
  - Objaverseは818K以上の物体モデルから構成され(規模がまだ成長中), 物体モデルごとに, ラベル, 3次元モデル, キャプションから構成される. 既存の3次元物体データセットで最も規模が大きい. また, 1つの物体クラスの中のdiversityが大きい.
  - Objaverseで実験し, 3次元物体認識・2次元物体認識・Embodied AIタスクに有用であることを示し, 更にデータ



Dataset	# Objects	# Classes
YCB [5]	77	5
BigBIRD [69]	125	–
KIT [35]	219	145
IKEA [43]	219	11
Pix3D [71]	395	9
GSO [19]	1K	17
EGAD [51]	2K	–
PhotoShape [53]	5K	1
ABO [12]	8K	63
3D-Future [22]	10K	34
ShapeNet [7]	51K	55
Objaverse	**818K**	**21K**

ObjaverseとShapeNetデータセットの例の比較, 既存データセットとの比較

CVPR 2023 の動向・気付き(184/259)

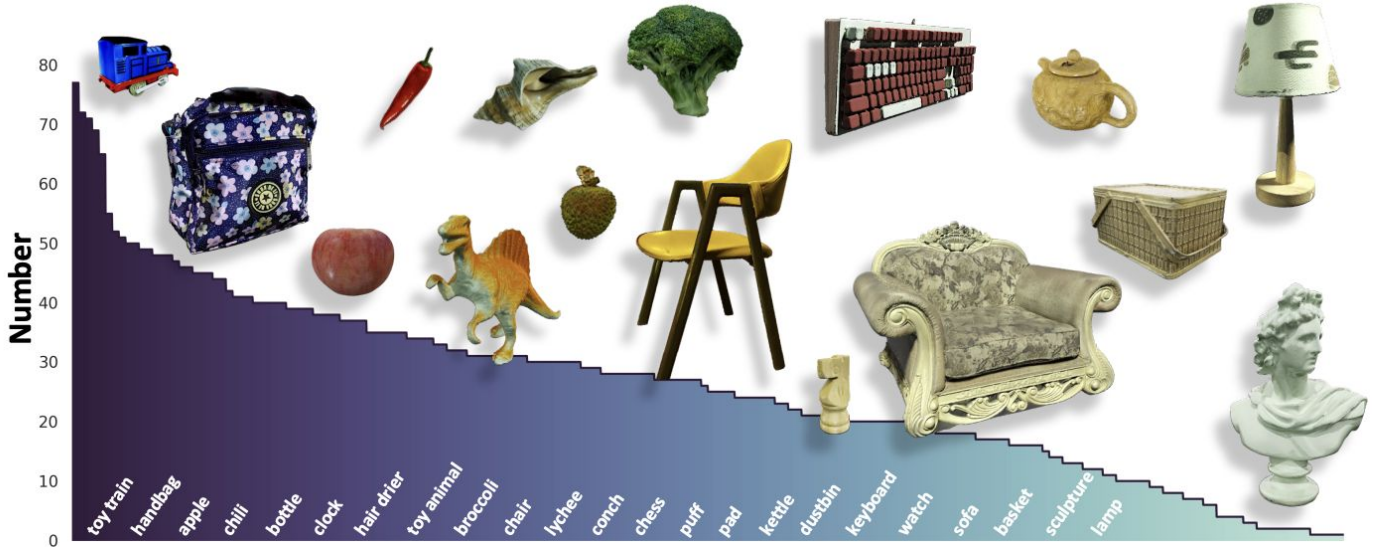
Habitat-Matterport 3D Semantics Dataset(Highlight)

- 大規模リアル環境(build)スキャンデータセットの提案
  - 950室内三次元シーンのスキャン，216セグメンテーションアノテーション付き.
  - 14,200 hours Human Effort！
  - 分野内ですでに使われていて，今後3次元室内環境の認識・生成分野のにおいて重要である.

CVPR 2023 の動向・気付き(185/259)

OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception(Award Nominate)

- 大規模リアル物体のスキャンデータセットの提案
  - 6,000 のスキャン物体データ(190カテゴリ)から構成される.
  - 物体の3次元メッシュ, 点群, 仮想・実カメラの撮影画像から構成.
  - 4つのタスクで(認識, NeRF, 再構成, 生成)既存研究の性能を向上した.

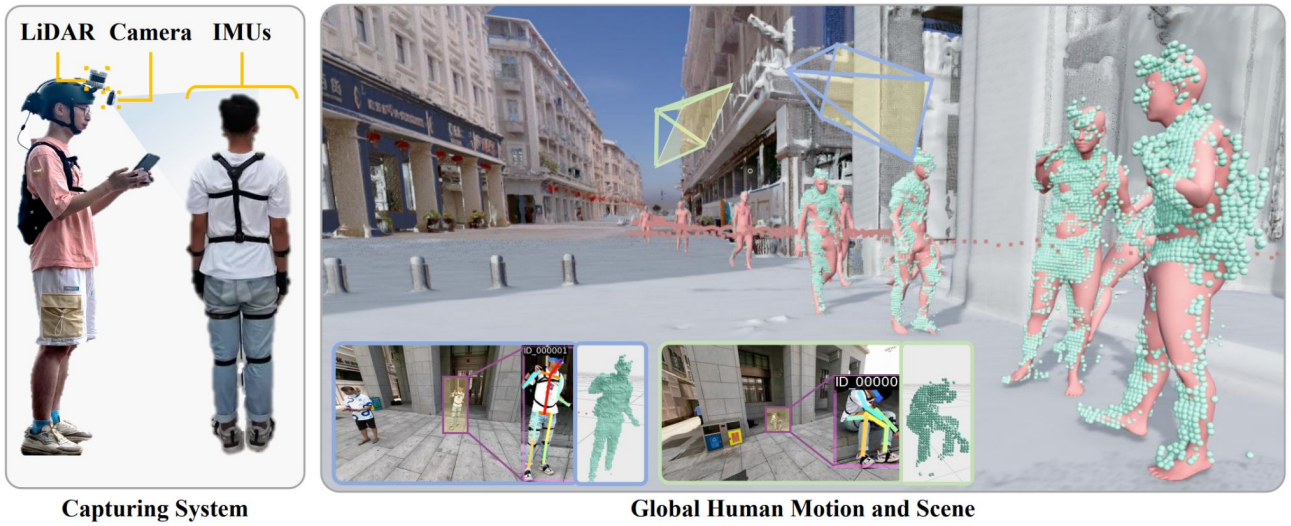


OmniObject3Dのデータ分布

CVPR 2023 の動向・気付き(186/259)

SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Environments

- 大規模リアルスキャンのhuman pose (human scene interaction)データセットの提案
  - 12 humans, 10 scenes, 15 sequence human motionから構成される.
  - 初めての屋外Urbanシーンのhuman poseデータセット.
  - 姿勢推定, シーンとのインタラクションの認識などのタスクでデータの有用性を示した.
  - 室内バージョンのデータが作成されたら, human activity認識などで活用できる.



SLOPER4Dの概念図

CVPR 2023 の動向・気付き(187/259)

BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion

- 3次元human pose生成や推定用の大規模合成ビデオデータセットの提案
  - 既存のデータセットと比較して，BEDLAMは大規模で，リアリティ性が高い(より詳細なポーズと服装が dynamic 3次元モデリングされている)．
  - BEDLAMデータセット生成が高いコストがかかって，複数の既存ソフトウェアで処理することが必要．例:Meshcapade skin textures; CLO3D 服装;Pre-simulated clothing meshes; AMASS motion.
  - BEDLAMデータセットを使って複数のリアルデータセット姿勢推定タスクで最も高い精度を達成．

提案手法で生成した例

多種類の服装とテクスチャー

CVPR 2023 の動向・気付き(188/259)

新しいタスク・アイデア

- Parts mate:物体形状構造の理解
  - Fantastic Breaks:実物体(コップなど)の割れたモデルパーツデータセット作成

- Visual Reasoning:
  - Deep Learning Smarter than Grader:大規模複数の形式のVisual Reasoningベンチマークデータと実験
  - Visual Programming:compositional reasoningに対応する手法(Best paper)

- 画像間の区別を考慮した画像Representation
  - Change-aware Satelite representation: 変化を考慮した上空画像Representation
  - GeneCIS: シンプルなモデルで既存のRepresentationのdownstream性能を大幅に向上
  - PhoneProc: 変化する室内環境Representation
  - Viusal DNA: 画像やデータセットの類似度を評価する新しいアプローチ

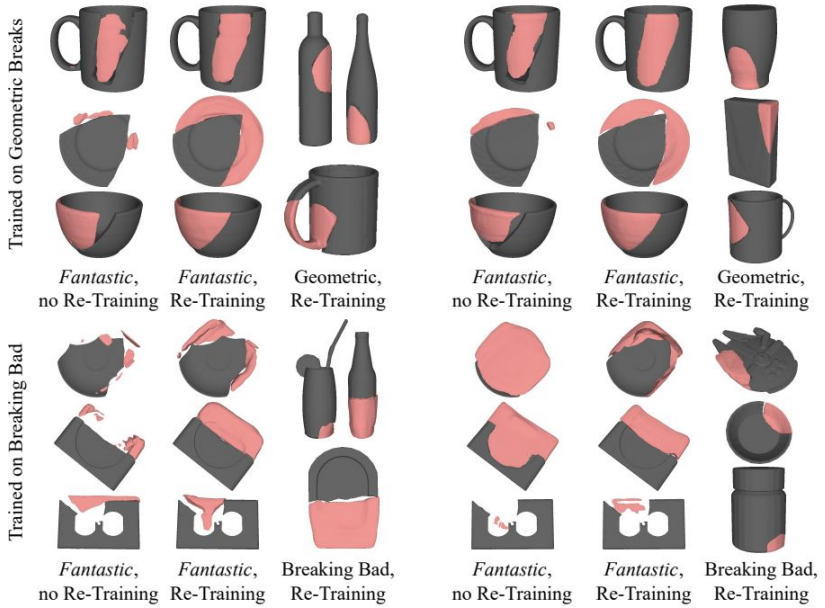
- ハイレベル意味情報の理解
  - Movie2Scenes: 映画のmeta情報を使ったcontrastive movie shot feature
  - How You Fellin:映画から感情分析
  - Make-a-story: story 画像生成
  - 概念の詳細理解:
    - CRAFT:概念のGradCAMみたいな感じ
    - Improving commonsense:Graph構造を導入
    - Learning Structured: データセットで文法拡張で学習
    - Procedure-aware pretraining: 動画にWikiHowの知識を導入

CVPR 2023 の動向・気付き(189/259)

Fantastic Breaks: A Dataset of Paired 3D Scans of Real-World Broken Objects and Their Complete Counterparts

- Shape Repairのリアルデータセットの提案，3次元モデルアノテーション付き
  - 既存のShape Repair研究ではベンチマークデータセットが少ないかつ小規模.
  - ここで実際の実物体を収集し，2分割に割って，3次元アノテーションしデータ作成した.
  - Fa[illegible]

実物体



評価例

CVPR 2023 の動向・気付き（190/259）

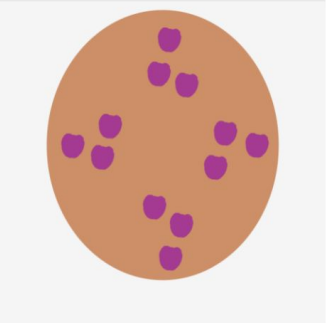
Are Deep Neural Networks SMARTer than Second Graders?

- DNNがMultimodal Algorithmic Reasoningを網羅的なベンチマーク
  - 最新のMultimodal手法（例：CLIP）や認識モデル（例：SwinTransformer）がMultimodal Algorithmic Reasoningにおいての性能評価のためのベンチマークデータセットと手法評価を行なった．結果的に，最も良い最新モデルでも小学校2年性と比べて性能が大幅に劣ることがわかった．
  - 提案ベンチマークはabstraction, deduction, generalizationを評価する101種類の問題が含まれる．
  - どうやって，人間レベル（もしくはそれ以上）で様々なReasoningできるモデルの構築がopen question.

Puzzle #2 (C)

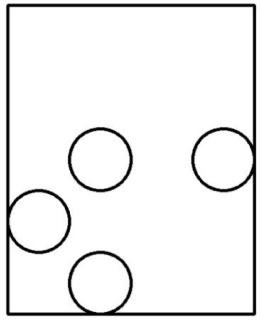
Question:In the picture, there are stars with 5 points, stars with 6 points, and stars with 7 points. How many stars that have only 5 points are there?

Puzzle #3 (C+AM)



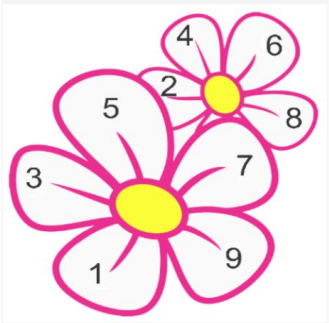
Question: The entire pie seen in the picture is divided among several children. Each child receives a piece of pie with three cherries on top. How many children are there?

Puzzle # 36 (C+S+AL+L)



Question: A man's hens lay white eggs and brown eggs. He puts eggs in the box shown in the figure. Two brown eggs cannot touch each other. At most, how many brown eggs can he put in the box?

Puzzle # 73 (C+AL+S)



Question: A number is written on each petal of two flowers. One petal is hidden. The sums of the numbers on the two flowers are equal. What number is written on the hidden petal?

提案ベンチマークの例

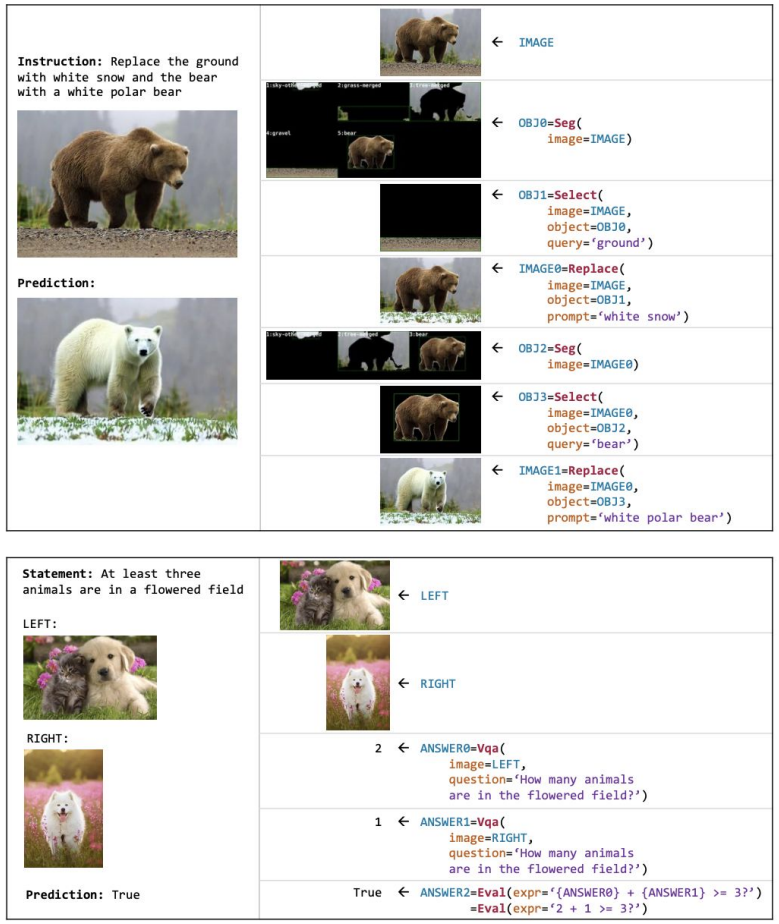
Puzzle Category →	Count	Arithmetic	Logic	Path Trace	Algebra	Measure
	Puzzle Split (PS) – Extreme Generalization Experiments					
Avg. 2nd Grader Performance	**72.8**	**81.3**	**82.2**	**81.1**	**64.5**	**90.4**
Greedy (baseline)	19.1/21.4	14.0/21.4	18.5/21.1	21.8/21.1	13.5/21.5	23.1/20.9
Uniform (baseline)	7.74/20.0	8.00/20.0	7.61/20.0	18.9/20.0	6.94/20.0	5.62/20.0
MAE + BERT	7.2/12.0	3.3/23.1	10.4/34.1	9.6/22.0	7.3/14.7	3.7/15.2
SimSiam + BERT	6.4/18.4	4.8/20.9	7.7/41.4	2.5/22.2	4.2/25.3	7.9/20.5
Swin_T + BERT	810.5/17.3	4.7/24.7	5.6/29.3	11.4/21.5	6.5/16.8	10.3/23.3
ViT-16 + BERT	9.41/22.7	5.77/26.8	6.95/25.1	4.72/18.7	5.57/15.1	8.68/21.3
CLIP	9.1/15.7	1.4/18.5	7.4/30.6	14.2/21.4	7.5/18.6	8.9/22.2
FLAVA	8.3/20.2	4.0/22.2	8.1/31.3	9.5/20.3	3.1/22.2	19.0/32.0
R50 + BERT (FT + Cls.)	10.9/18.3	6.96/15.8	12.8/20.8	19.6/19.7	7.95/15.1	16.9/26.7
R50 + BERT (FT + Reg.)	12.0/22.8	5.08/21.3	4.24/16.2	18.4/18.4	4.89/22.2	15.1/25.9
	Few-Shot Split (FS) Experiments, $m = 10$					
R50 + BERT (Cls.)	17.3/28.0	11.2/25.8	18.0/37.6	19.2/19.2	7.9/21.9	14.8/31.2
R50 + BERT (Reg.)	13.3/25.2	8.3/24.7	11.2/23.3	17.3/18.6	6.6/18.9	19.5/34.2

2nd Graderとの比較（部分的）

CVPR 2023 の動向・気付き(191/259)

Visual Programming: Compositional visual reasoning without training(Best Paper)

- in-contextでvision (/vision+language) taskを step-by-stepで解析しながら既存のモジュール(VQA, classify)を利用して解いていく仕組みの提案
  - 提案手法が4つのvision task: compositonal VQA; NLVR; 言語指示からのfactual knowledge object tagging;text-based image editingの4つのタスクの解くプロセスをin contextで学習可能なことを示した. in contextでprogram生成にはGPT3を使用した.
  - 各タスクでの結果は, 既存のstate-of-the-art task specific手法と比べて劣っている.
  - クラシックなneural symbolicをタスクレベル(VQAなどをモジュール)にして, さらに最新のLLMsを利用しながら, 完成度を高めた.
  - 根本的に, OOD的なsymbolic reasoning, そして, symbolic moduleの自動的に生成(現在が人工でデザインしている), あらゆるvisionが関わったタスクを解くにはまだopen question.

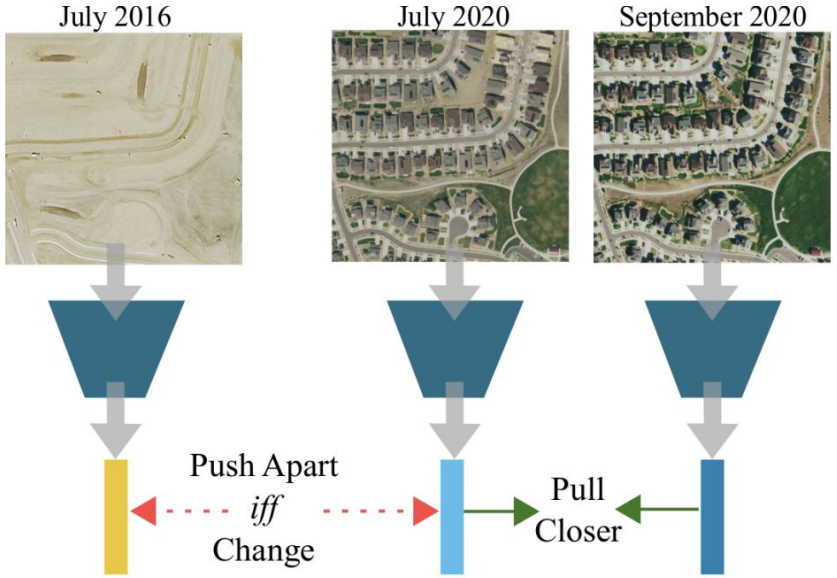


テキストベース画像編集とNeural Language Vision Reasoningタスクの結果例

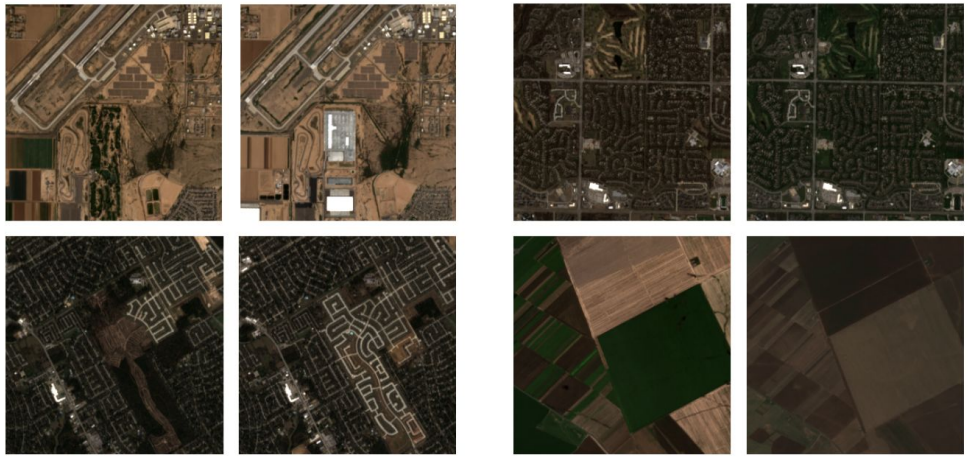
CVPR 2023 の動向・気付き(192/259)

Change-Aware Sampling and Contrastive Learning for Satelite Images

- 変化を考慮した上空画像のRepresentationの研究
  - 手法:上空画像に変化が発生した(例:レイアウト変化)場合に, Negativeとして考えるContrastive Learningを用いた手法を提案.
  - 上空画像の認識タスク, 変化認識のタスクに高い性能を実現.



手法のイメージ図

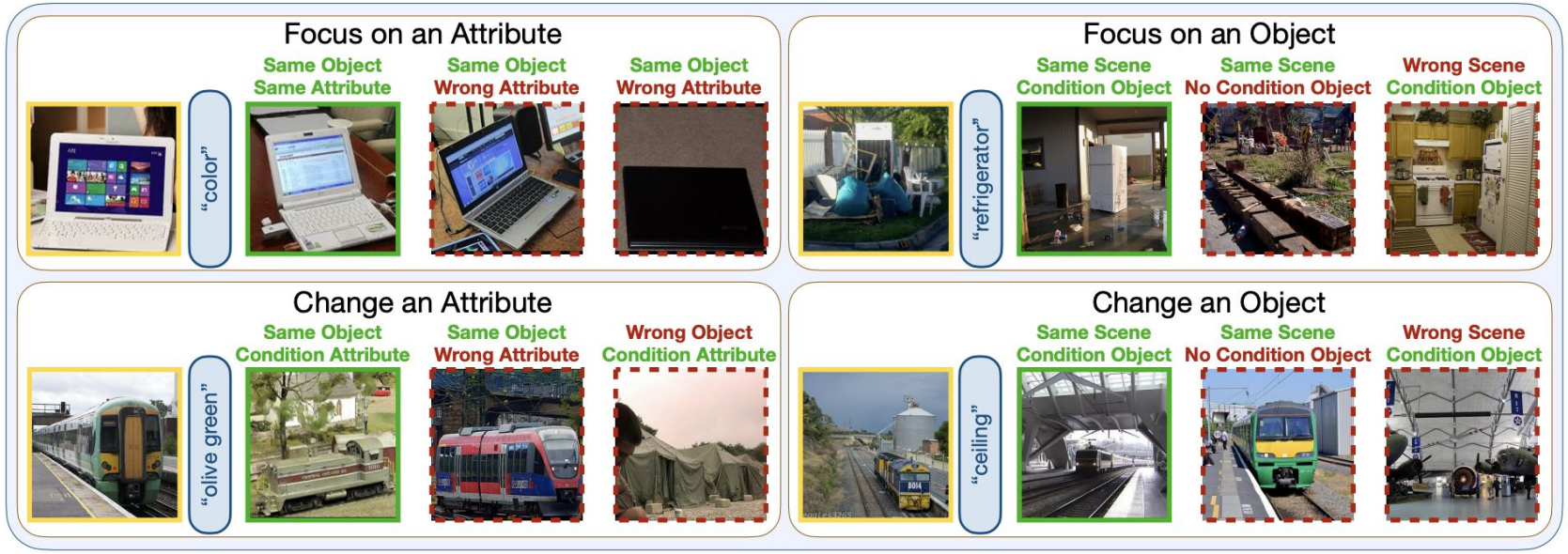


high, low change estimate

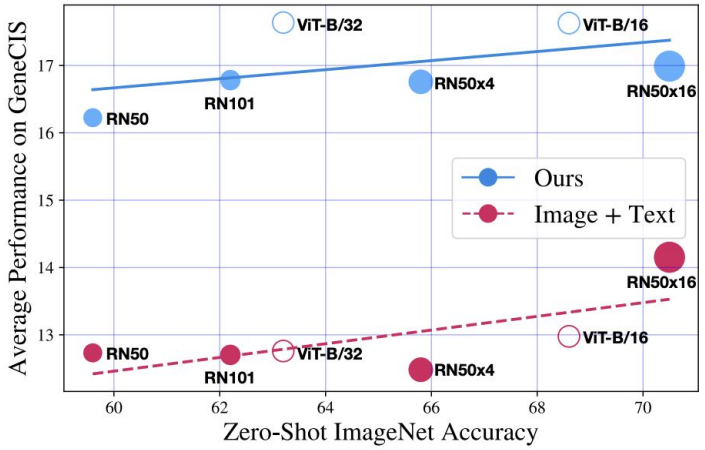
CVPR 2023 の動向・気付き(193/259)

GeneCLS: A Benchmark for General Conditional Image Similarity

- 条件付き(例:指定の物体や属性)画像と画像間の類似性をzero-shotで分析可能なタスクと手法の提案
  - 画像間の様々な粒度の類似性認識が人間のシーン理解に重要だがあまり検討されていない.
  - ここで, データセットからGraphをベースで情報をマイニングする手法を提案.
  - CLIPを用いても既存のResNetと大差がない結果を示し, 画像の類似性認識には他のモデル設計が必要であることを示した.



タスクのイメージ図

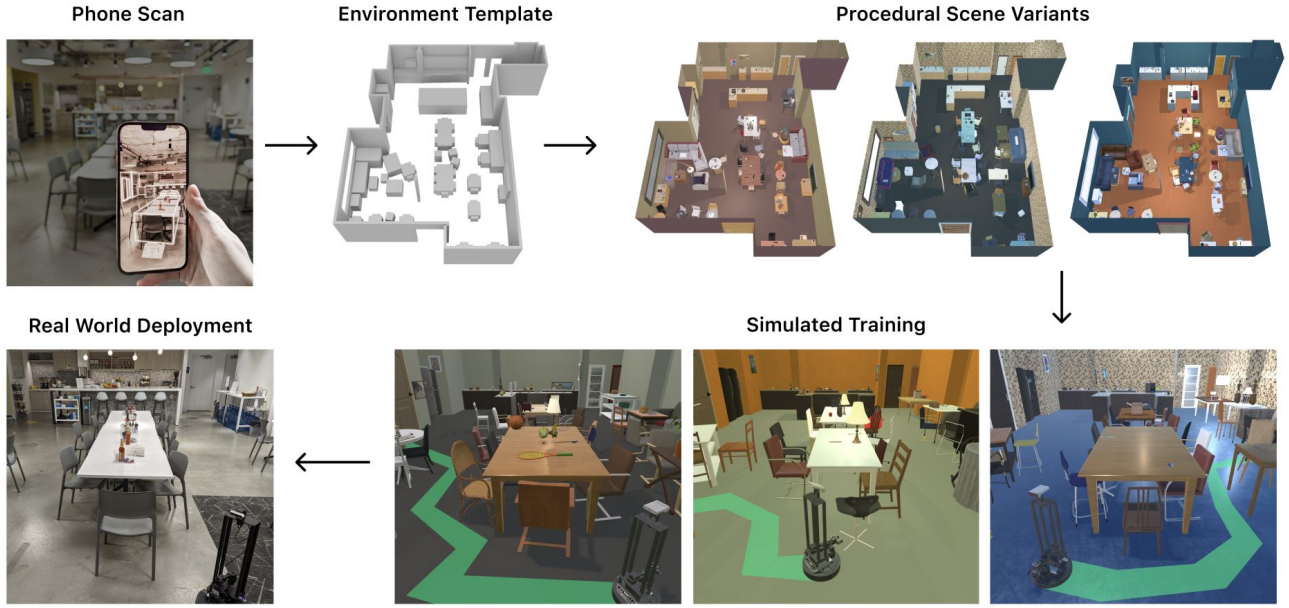


提案のマイニングを使った／使わない結果の比較

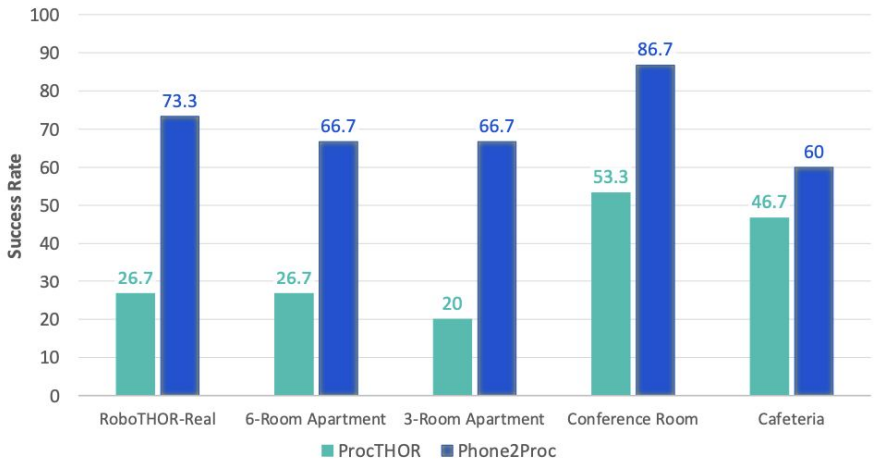
CVPR 2023 の動向・気付き(194/259)

Phone2Proc: Bringing Robust Robots Into Our Chaotic World

- 実環境で動作するロボットのためのindoor scene representationのためのデータ収集
  - 実環境でロボットが新しい家庭環境に，その環境の認識をする必要がある
  - この研究では，上記を可能にするためのフレームワークPhone2Procを提案した(左図)．提案手法はまずスマートフォンカメラ安易でシーンをスキャンし，物体のアフォーダンスを考慮したシーンの再生成を行う．また，シーンの中の物体を再配置することでその環境の新しいデータを拡張する．得られたデータで学習することで，様々なdownstream



Phone2Proc概念図



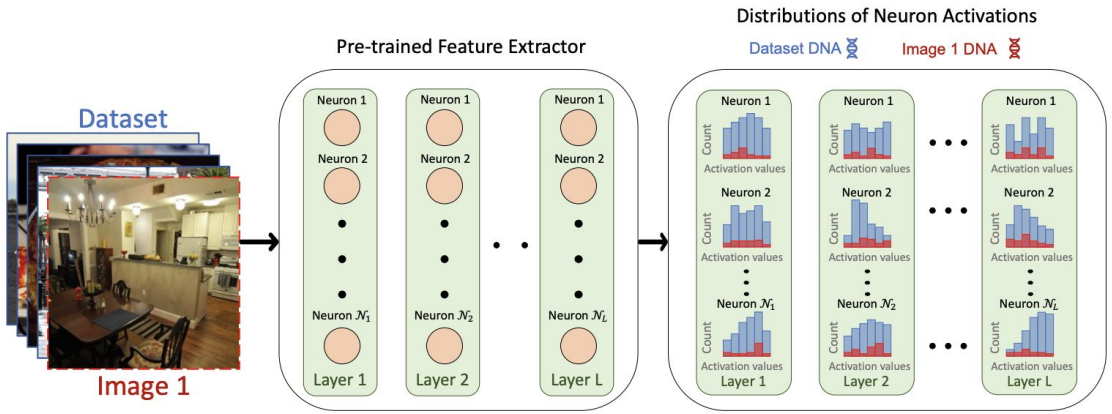
既存研究との比較

CVPR 2023 の動向・気付き(195/259)

Visual DNA: Representing and Comparing Images using Distributions of Neuron Activations

- 画像やデータセットの類似性を評価する手法Visual DNAの提案
  - データの重要性が知られているが，これまでのデータセットのVisual側の類似性を評価する研究が少ない．
  - ここで，DNNモデルの学習済みのneuronのactivationsの分布値が異なるattributesにたいてどのような変化をするかをベースに，指定されたattributes軸でデータセットや画像の類似度を評価できる(評価する際に同一画像を使う)．
  - 類似画像のretrieval, simulation datasetとreal datasetの違い評価などに活用できる．

データセット内の画像類似性評価例

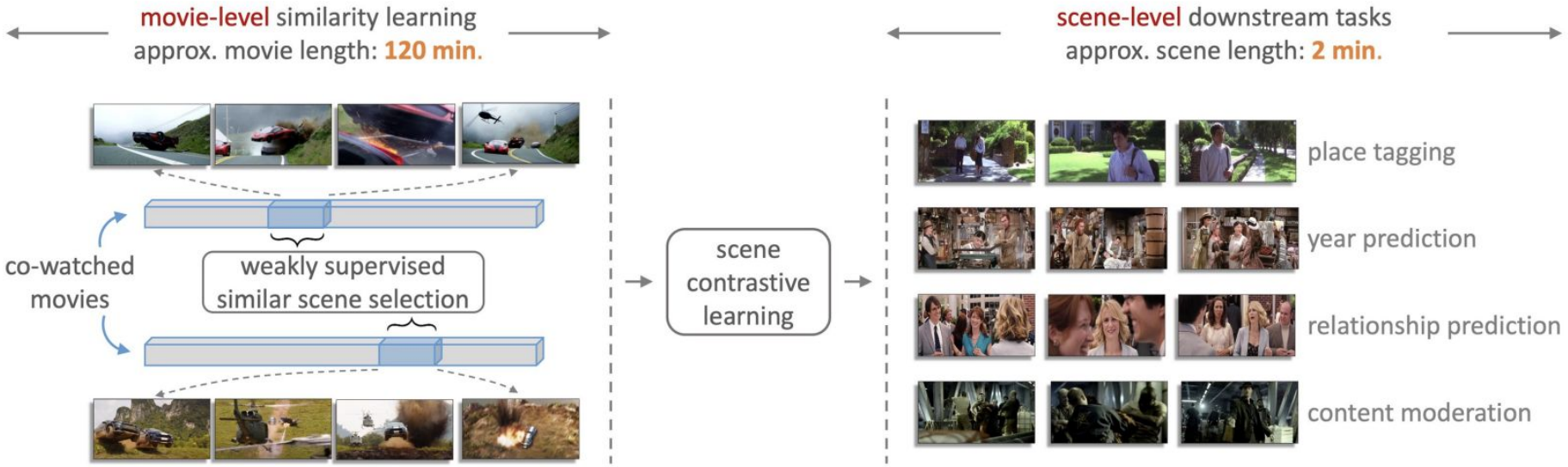


Visual DNA構築方法

CVPR 2023 の動向・気付き(196/259)

Movie2Scenes: Using Movie Metadata to Learn Scene

- Movieデータの短いshotのrepresentation学習手法の提案
  - 映画のメタデータ(ジャンル，ユーザの鑑賞履歴など)から，映画のshot(数秒程度)の類似性を評価するモデルを提案.
  - 上記の類似モデルから，movieのshotのcontrastive representationモデルを提案.
  - 複数のdownstream tasksで性能を出している.
  - movie shot理解のためのデータセットも提案　movie理解を少し加速できる？
  - n

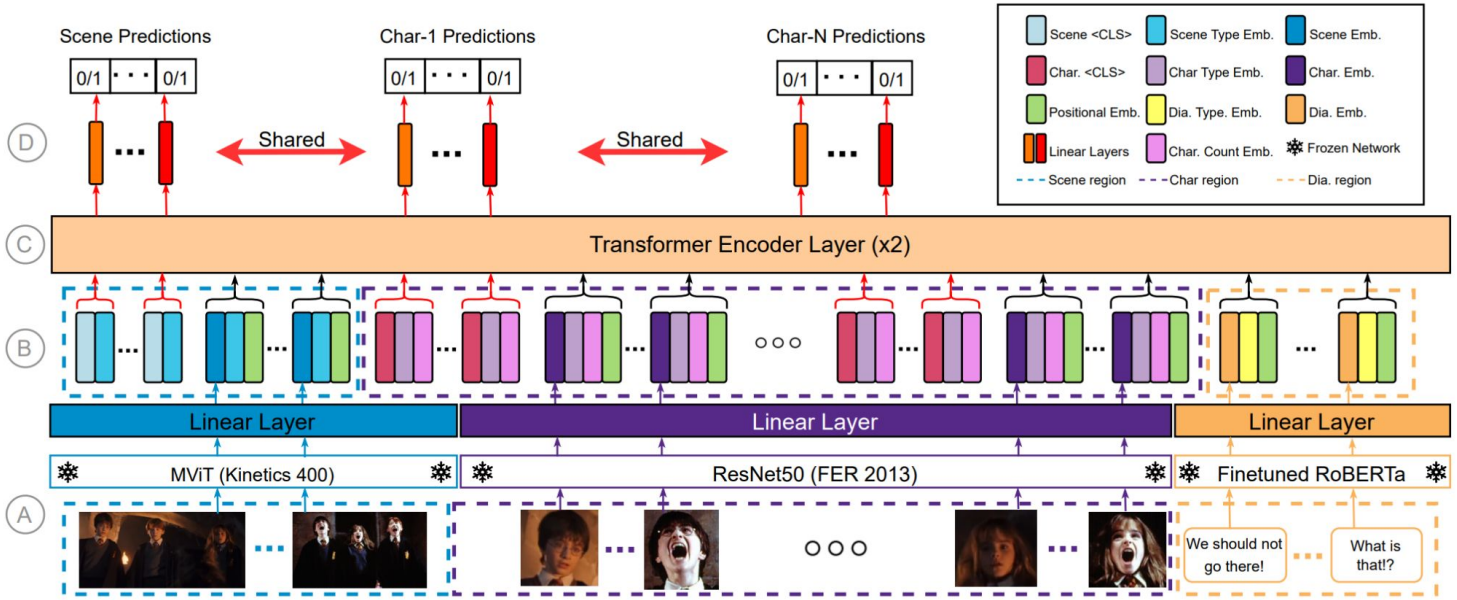


提案手法のイメージ図

CVPR 2023 の動向・気付き(197/259)

How You Feelin'? Learning Emotions and Mental States in Movie Scenes

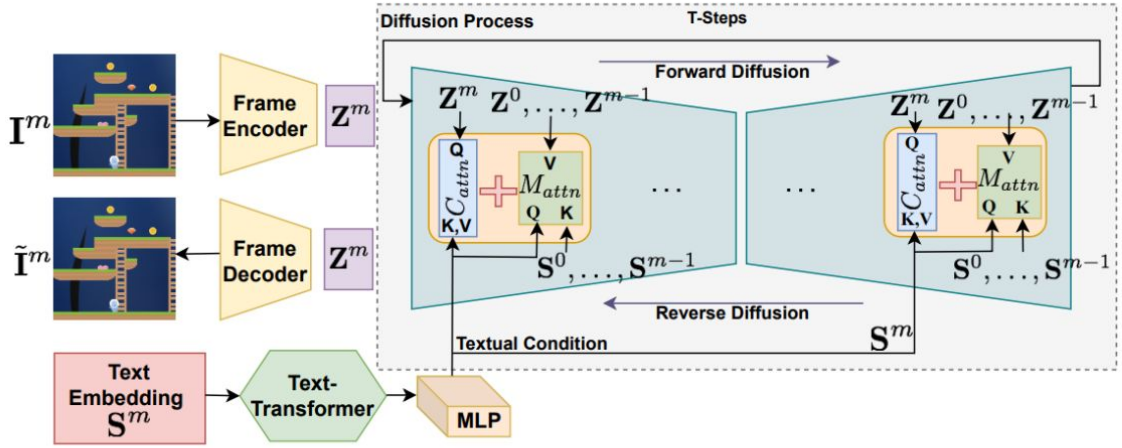
- 映画のシーンから人間の感情をマルチラベルで推定する手法を提案
  - ビデオから人間の感情をシングルラベルではなくマルチラベルで表示することを提案.
  - 提案手法は, ビデオから複数の情報(Action, Face, Dialogなど)をTransformerで扱う. これらを利用することで既存手法より高い精度を得た.
  - 将来的にはEmotionsの変化の対応が望ましい. downstreamタスクでの有用性が要検討.



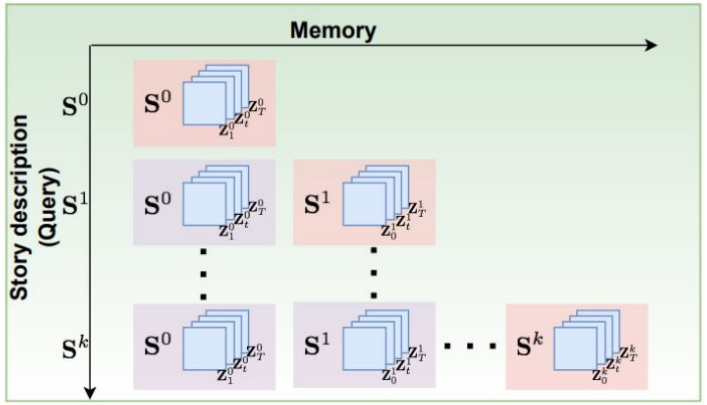
CVPR 2023 の動向・気付き(198/259)

Make-A-Story: Visual Memory Conditioned Consistent Story Generation

- Diffusionモデルを使ったストーリー(複数センテンス)画像生成手法
  - Diffusionモデルが性能高いが，複数枚の関連画像の生成に全景背景の一致性を保つ手法があまり検討されていない
  - Visual Memory構造をDiffusionモデルに導入し，全体的に背景と全景の一致性を高めた
  - また，既存データセットMUGENにキャラクタの情報やストーリーの中のテキスト情報の一致性などを評価するためのアノテーションを追加した



(a) Story-LDM architecture for conditional story generation.

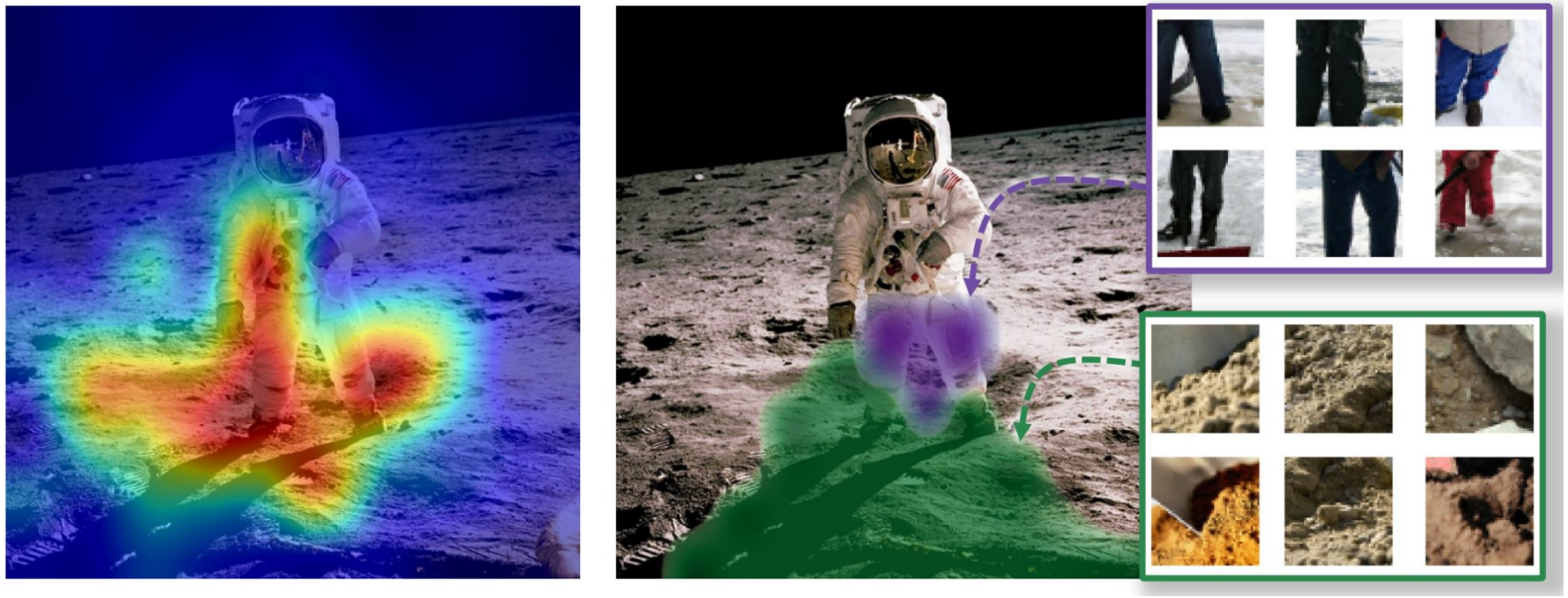


(b) Memory state with query and values for our memory-attention module.

CVPR 2023 の動向・気付き(199/259)

CRAFT: Concept Recursive Activation FacTorization for Explainability

- 認識モデルが認識する際に，”どこ”と”何”両方を可視化・説明する手法
  - 既存の可視化モデルは，attention mapを使ってモデル認識する際の注目領域のみ確認.
  - ここで，注目領域の意味情報(visual concept)も出力可能にして，モデルの認識結果の分析に活用できる.



既存の可視化とCRAFT

CVPR 2023 の動向・気付き(200/259)

Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles

- 知識グラフを使ったcommonsense reasoningの性能を高める手法の提案
  - 下記の図で示している提案手法はConceptNetで画像から知識グラフを抽出
  - 抽出された知識グラフにより学習テキストデータを拡張する
  - 拡張されたデータセットで学習することで，commonsense reasoningの能力を高めた

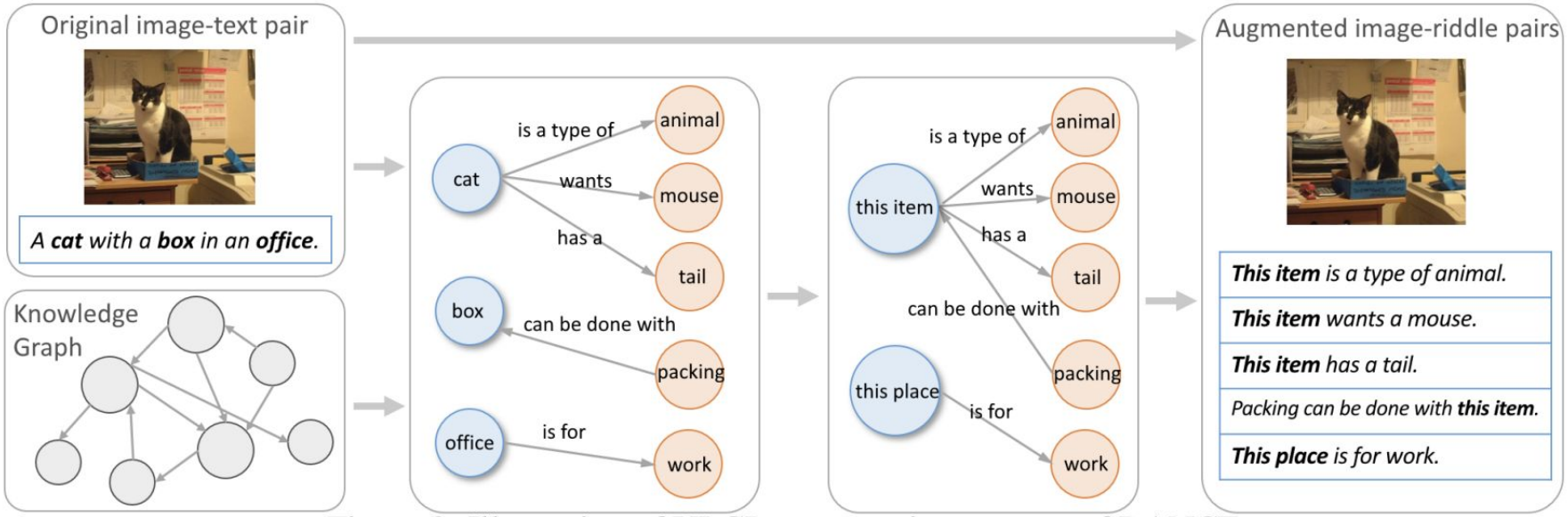
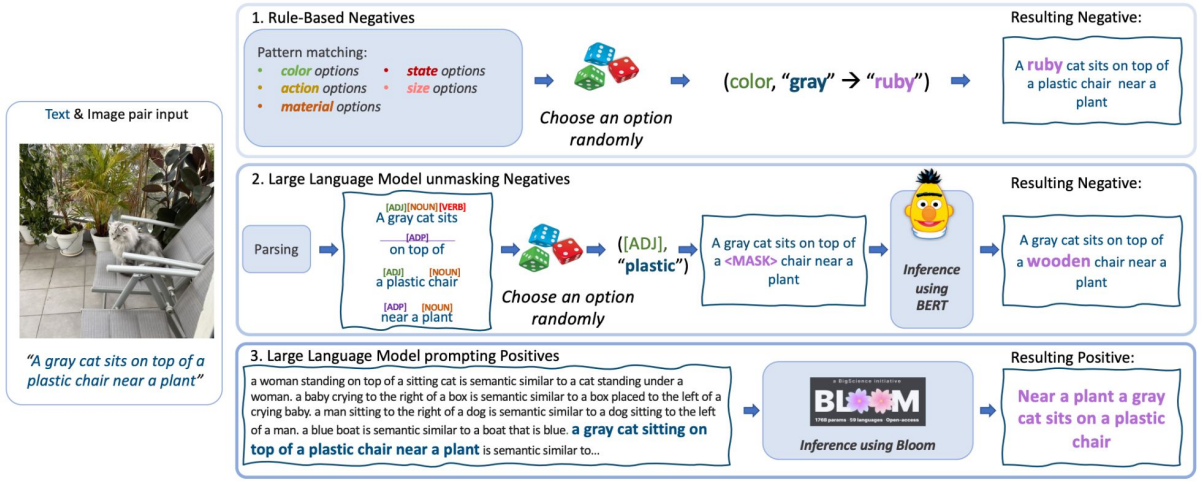


Figure 2. Illustration of VLCR construction process of DANCE.

CVPR 2023 の動向・気付き(201/259)

Teaching Structured Vision and Language Concepts to Vision and Language Models

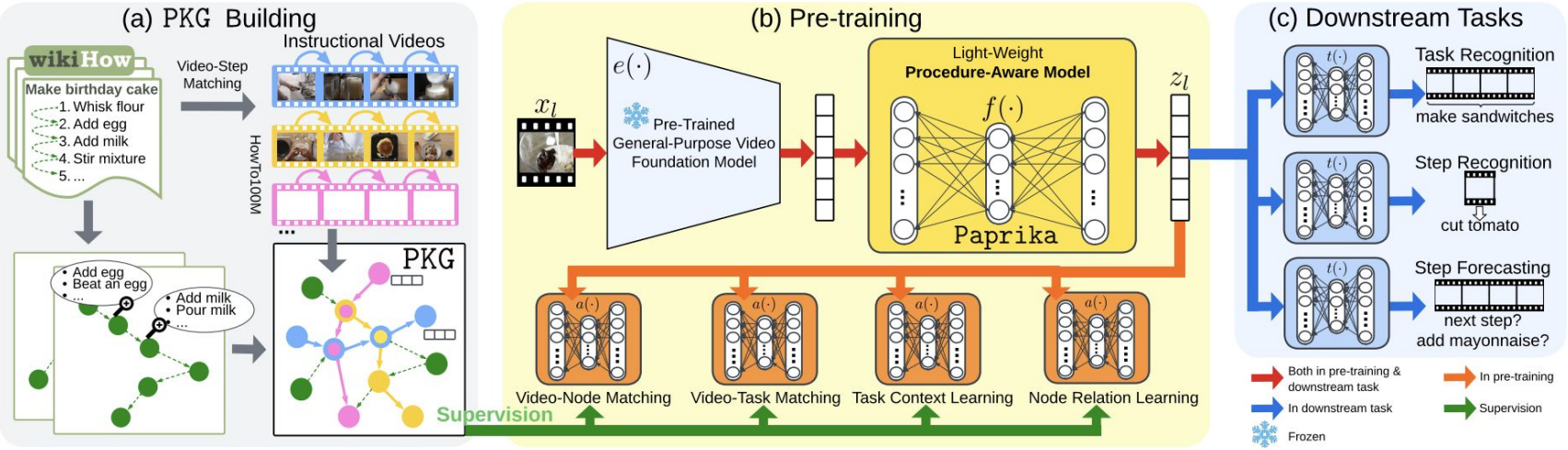
- Structured language conceptを学習するvision and languageモデルの提案
  - 既存vision and language modelが構造化された概念の理解が足りない
  - ここでstructured language concept(verb,noun,attributes,relationshipsなど)に従って学習のデータ拡張を行う手法を提案
  - 上記のデータ拡張を行うことで既存複数のvision and language model手法の性能を向上
  - 全ページの手法と高度類似. 手法よりはデータを重要視する手法が多いと感じた



CVPR 2023 の動向・気付き(202/259)

Procedure-Aware Pretraining for Instructional Video Understanding

- instructionビデオ認識のための，知識グラフを利用した手法の提案.
  - instructionビデオから，タスクを解く順番を理解するための必要な知識をunlabeledで学習することが困難であるため，ビデオからinstructionの生成学習には大量なデータが必要となる.
  - WikiHowから知識情報を抽出し，さらにそれをprocedure knowlege graph(PKG:ノードがステップで，エッジが順番を表す)の形式で表示する.
  - 学習の段階で上記のPKG情報を教師データとして，video-node, video-task matching, task context, node relationの学習をauxiliaryタスクとして学習し，これによりステップを反映したビデオ特徴表示の学習を行う.
  - 提案手法を用いて，学習された特徴量がinstruction動画認識においての有用性が検証された.



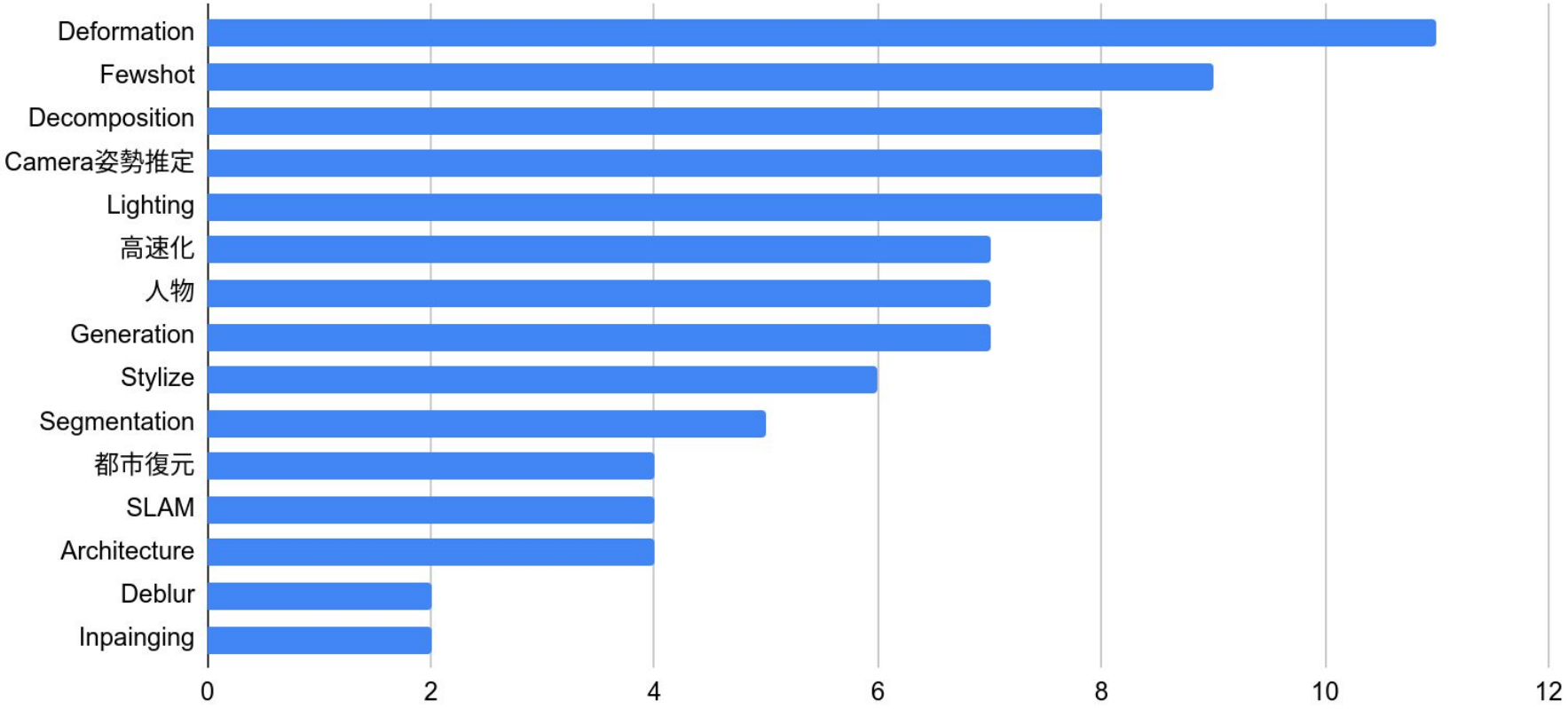
提案手法

CVPR 2023 の動向・気付き(203/259)

Neural Fields関連の動向

※個別の論文についてはこちらに要約を置いています

- NeRF関連論文の採択数が104本
  - CVPR2022の38本、ECCV2022の25本から大幅に増加



タスク別論文数

CVPR 2023 の動向・気付き(204/259)

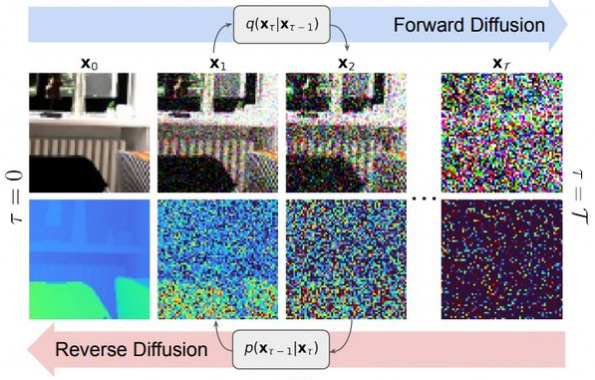
Neural Fields / Deformable

- CVPR2023では11本採択と大流行中
- 長時間・大変形化や反射光のモデル化，人物衣装・動物骨格など**より高難易度**のシーンへの活用が試みられている
- DynIBaR(Zhengqi et al.):制約のないカメラ軌道を持つ長時間の動画から写実的な新規視点合成を実現
- ReRF(Liao et al.):隣接するタイムスタンプ間の残差でモデル化することで長時間の変形を平滑に保持
- DyLiN(Heng et al.):トポロジー変化を伴うシーンでのLightFieldのモデル化
- DNRF(Kangkan et al.):単眼または少数視点映像から服を含む人物動作を復元
- MonoHuman(Zhengming et al.):人物運動に対し双方向制約を持つ変形場をモデル化
- MagicPony(Shangzhe et al.):自然環境下での動物骨格を含む復元

CVPR 2023 の動向・気付き(205/259)

Neural Fields / Few Shot

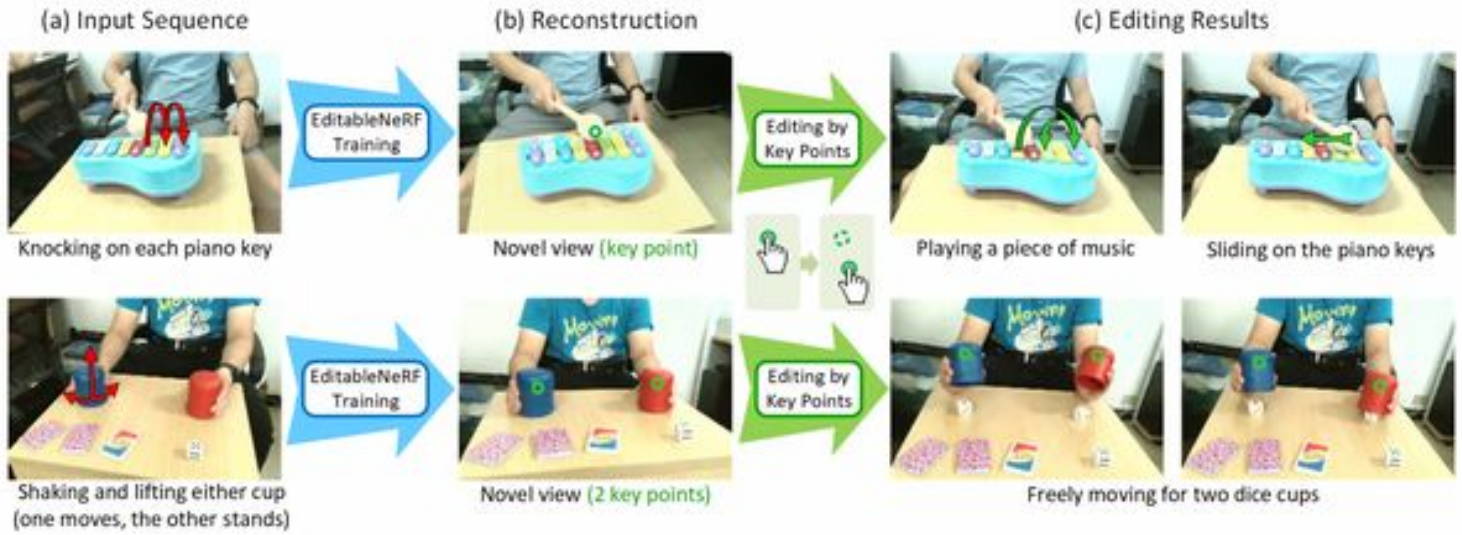
- NeRFのデータ用意に必要なコストは問題視されており，視点数を減らしても高品質を実現可能な手法が多数提案されている
- 空間平滑性を明示的に記述するアプローチも未だあるものの，Denoisingモデルなどで外部知識を使うアプローチが増加傾向
  - DietNeRFのような事前知識付与フローとして拡散モデルを使用することで**完全に未知の視点の情報**を与えられるように
  - データ数の課題から，事前知識には2D処理を用い，NeRFを**3次元的な一貫性**を持たせるための仲介に利用する方法が現状では主流
- DiffusioNeRF(Jamie et at.)：ノイズ除去拡散モデルを用いて事前知識を付与
- NeRDi(Congyue et al.)：2D DiffusionでNeRFの潜在空間中の特徴ベクトルの事前分布を設定．1枚の画像からNeRFを構成



CVPR 2023 の動向・気付き(206/259)

Neural Fields / 編集可能性

- NeRFの編集では編集指示の与え方に課題があった
  → 指示の与え方やインターフェイスに着目した研究が増加傾向
- EditableNeRF(chengwei et. al):ユーザーがキーポイント操作で編集可能な動的シーンNeRFを提案
- PaletteNeRF(Zhengfei et. al):Diffuse成分をシーン中で共有するカラーパレットの線型結合に分解し，カラーパレットの編集により外観を操作可能に
- SINE(Chong et. al):1枚の画像で神経輝度フィールドを編集



CVPR 2023 の動向・気付き(207/259)

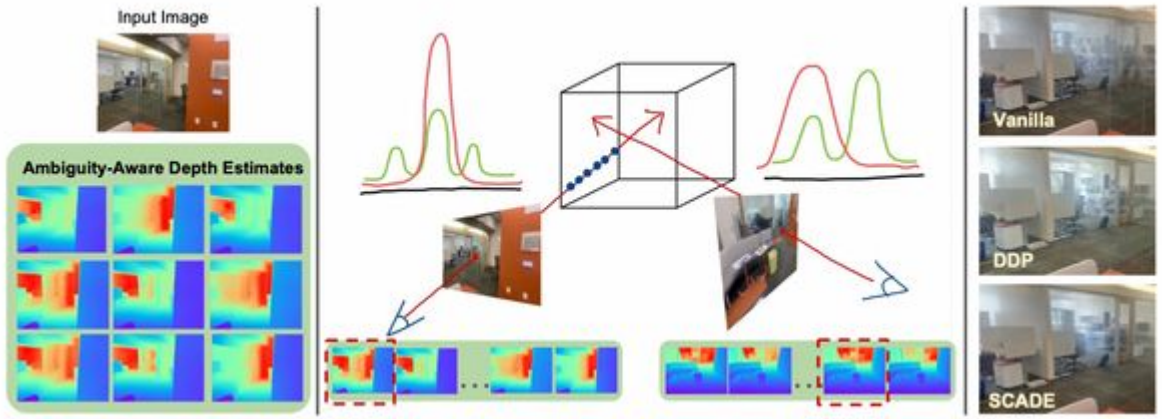
Neural Fields / カメラ姿勢推定

- NeRFでは高品質なカメラ姿勢が要求されるため，JointOptimization(iNeRF)を用いた調整や推定が探索されてきた
- BARF以降，PEを介して低周波成分を対象にした学習初期の誘導が可能になった
  →CVPR2023ではBARFの拡張が多数提案
- DBARF(Yu et al.)：汎化NeRFにBundle Adjustmentを導入
- L2G-NeRF(Yue et al.)：ピクセル単位の対応に対して微分可能なパラメータ推定ソルバを用いてGlobalなBAを実施
- BAD-NeRF(Pen et al.)：激しいモーションブラー画像や不正確なカメラポーズに強いBAを行うNeRFを提案

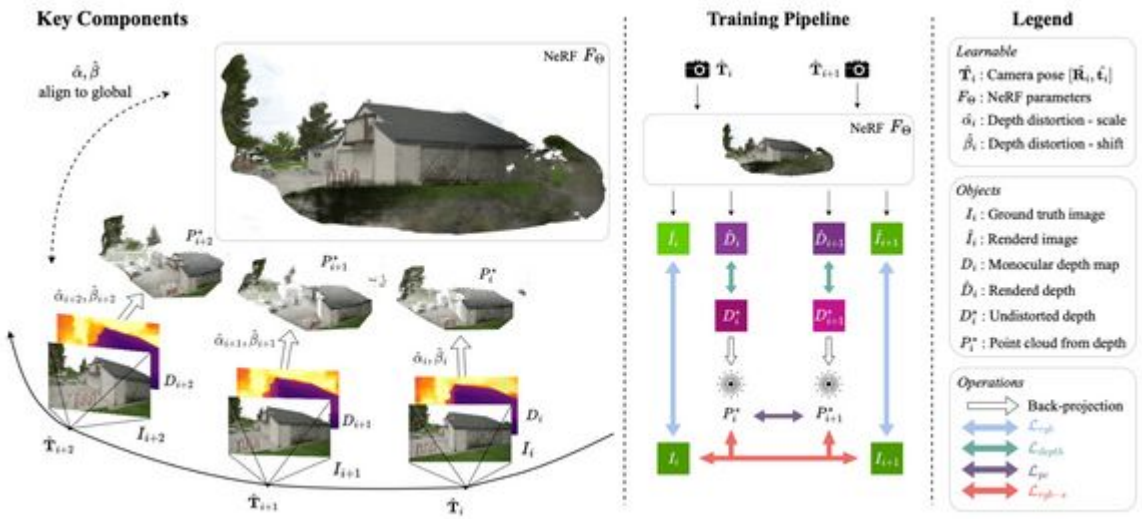
CVPR 2023 の動向・気付き(208/259)

Neural Fields / カメラ姿勢推定

- Depth値を活用した理論的な制約・事前確率を用いる手法が増加
- SCADE(Mikaela et al.):単眼デプス推定の値を事前確率として使用
- NoPE-NeRF(Wenjing et al.):カメラ姿勢の事前知識なしでNeRFとカメラ姿勢を同時最適化



SCADE

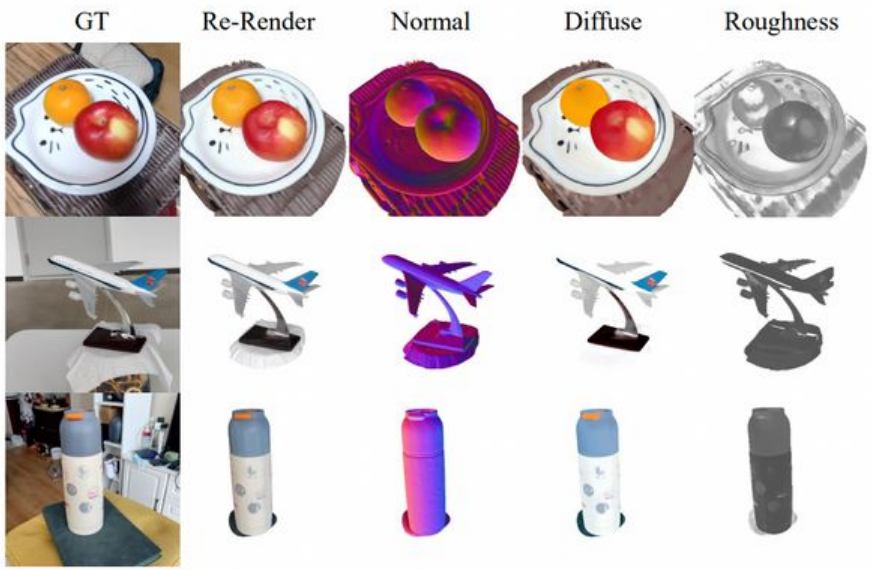


NoPE-NeRF

CVPR 2023 の動向・気付き(209/259)

Neural Fields / Lighting

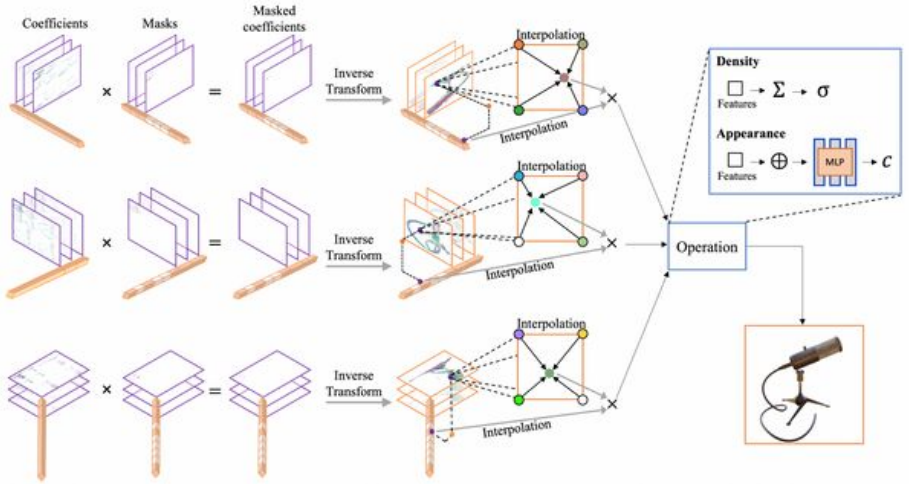
- 照明情報をRGB画像観測から復元するためには，被写体の材質推定が必要
- SDFなど距離場で**サーフェイスをパラメータ化**し，BRDFのような**境界面の反射モデル**を復元するアプローチを取るものが多い
- VDN-NeRF(Bingfan et al)：学習済NeRFにエンコードされている不変情報を抽出した正規化により，非ランバート面や動的な照明条件下のNeRFを安定化
- I2-SDF(Jingsen et al.)：SDF上の微分可能なMonte Carlo raytracingを導入
- NeFLL(Haoqian et al.)：Monte Carlo Samplingに基づくパストレーシングを導入



CVPR 2023 の動向・気付き(210/259)

Neural Fields / 高速化

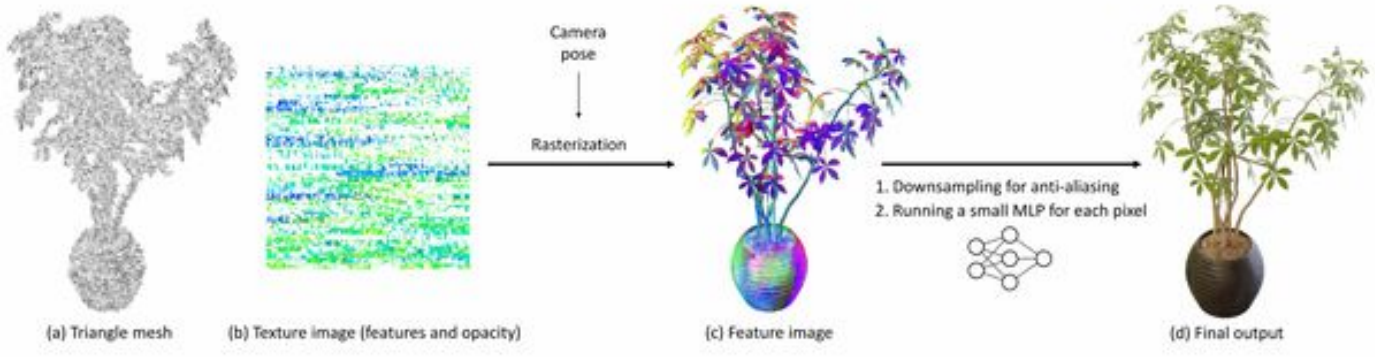
- これまでで，グリッドやハイブリッド表現による高速化が普及
  - 学習・レンダリングともに高速な一方で，モデルサイズが大きくなる点が課題
- 微分可能なパラメータの圧縮方法を用いてグリッド表現を圧縮した形で保持する方法が流行
- MaskedWavelet(Daniel et al.)：グリッドにウェーブレット変換を適用し，学習可能なマスクで保持することで密度場が疎な環境下では効率的に圧縮
- HexPlane(Ang et al.)：動的シーンを6つの特徴面へTensor分解して表現
- NeRFLight(Fernando et al.)：Decoderも分散処理とすることで低解像度グリッドで高い精度を達成

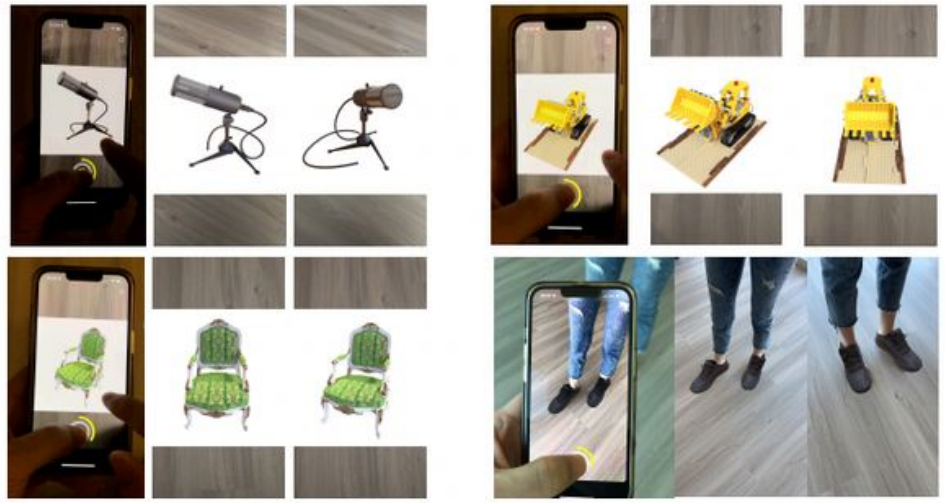


CVPR 2023 の動向・気付き(211/259)

Neural Fields / 高速化

- 負荷を小さくすることで，モバイル端末でのレンダリングが可能なNeRFが登場
- MobileNeRF(Zhiqin et al.)：通常のNeRFを訓練後，密度を2値化してメッシュを抽出し，不透明度と特徴量をテクスチャにbakeする
- Real-Time NeLF(Junil et al.)：Neural Light fieldのアプローチで高速化．MobileNeRFの15~24倍低容量

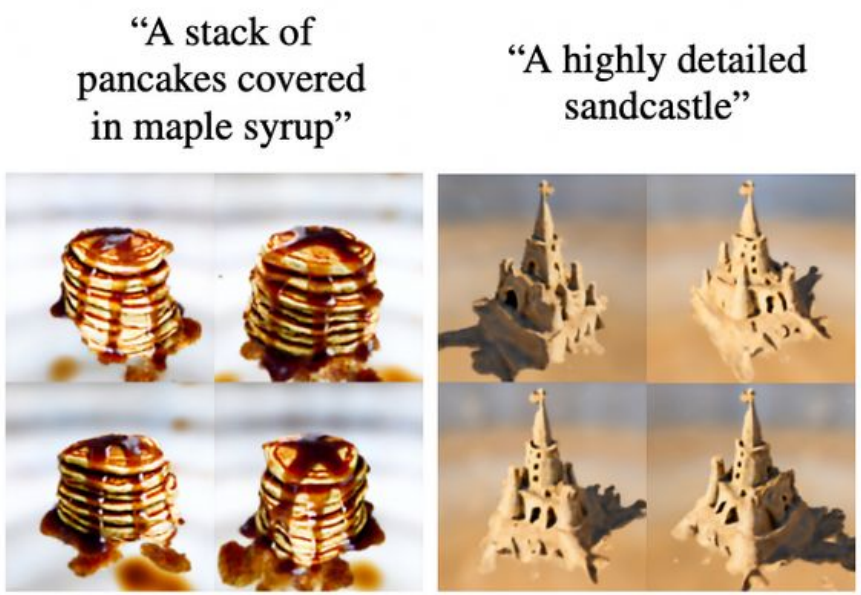




CVPR 2023 の動向・気付き(212/259)

Neural Fields / 生成

- DreamFusionを起点に，2Dの拡散モデルで生成した画像を，3次元的に整合性を持たせるための伝搬役としてNeRFを使用する研究が発展
- Text-to-3DタスクはCVPR2023以降もFantasia3Dなど急速に進展中
- Dream3D(Jiale et al.)：CLIPガイド付き3D最適化手法に，Diffusionモデルから明示的な3D形状事前分布を導入
- Latent-NeRF(Gal et al.)：RGB画像ではなくNeRFの特徴ベクトル空間を対象にしたDiffusionモデルを構築



CVPR 2023 の動向・気付き(213/259)

NeRF関連研究の動向

- NeRFで3D(or 時系列で4D)をモデル化することが一般化
  - NeRFを使うのは誰でもできる時代に
- 研究の方向性1:NeRFの適用可能性を広げる
  - より簡単に，より高速に，より制約なく
  - 特に簡単な工夫で高速化・高性能化・適用可能シナリオの拡大ができないかの試行錯誤が進められている印象
  - 既存のパラメトリックな3Dモデルも活用(人体のSMPLなどが顕著)
- 研究の方向性2: 大規模化
  - グリッドベースの特徴量記述を利用すれば都市レベルでできることがよく知られている
- 研究の方向性3: 複数シーン対応
  - シーン単位学習から複数シーン同時最適化・未知シーンへの学習結果の活用
  - 2Dやテキストの学習済みモデルをうまく利用
  - 光線や2D画像など，3Dより軽量な単位の特徴量でシーンをまたぐ

CVPR 2023 の動向・気付き(214/259)

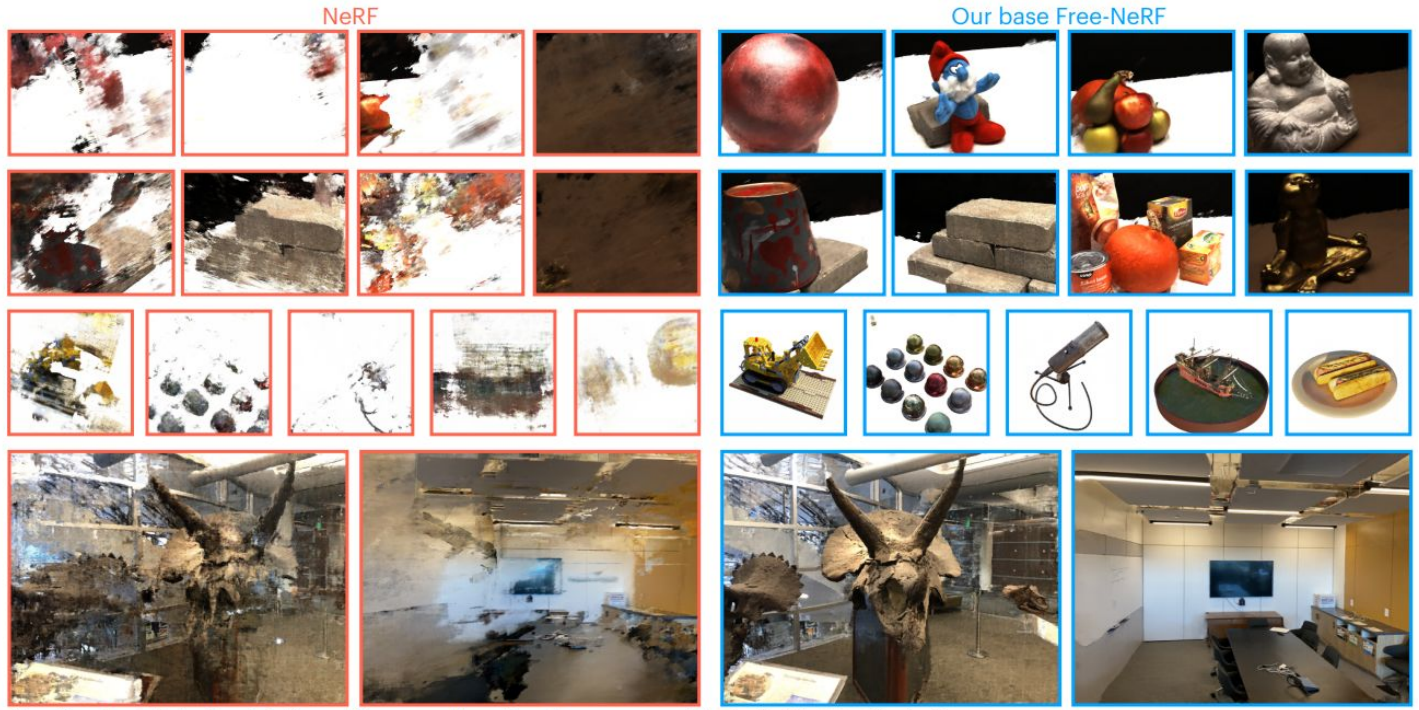
NeRF関連研究の動向

- 研究の方向性4:これまでのCV研究の応用
  - 2D CNNでの推論，パラメトリックな変形モデル，ルールベースの点群処理・画像処理など既に解けたタスクを3Dに持ち上げる道具として利用
  - NeRFを使うとうまく「3D→2D」「2D→3D」のラベル・情報伝播ができる
    - レンダリングを真面目に考えていることのメリット
- 研究の方向性5:ロボット・xRなどへの応用
  - 3Dが活用できるアプリケーション方向へ

CVPR 2023 の動向・気付き(215/259)

FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization

- Few-shotでNeRFを再構成
- アーチファクトを抑えるためにPositinal Encordingを徐々に与えるテクニックを導入
  - アイデアも実装もシンプルで確かな効果

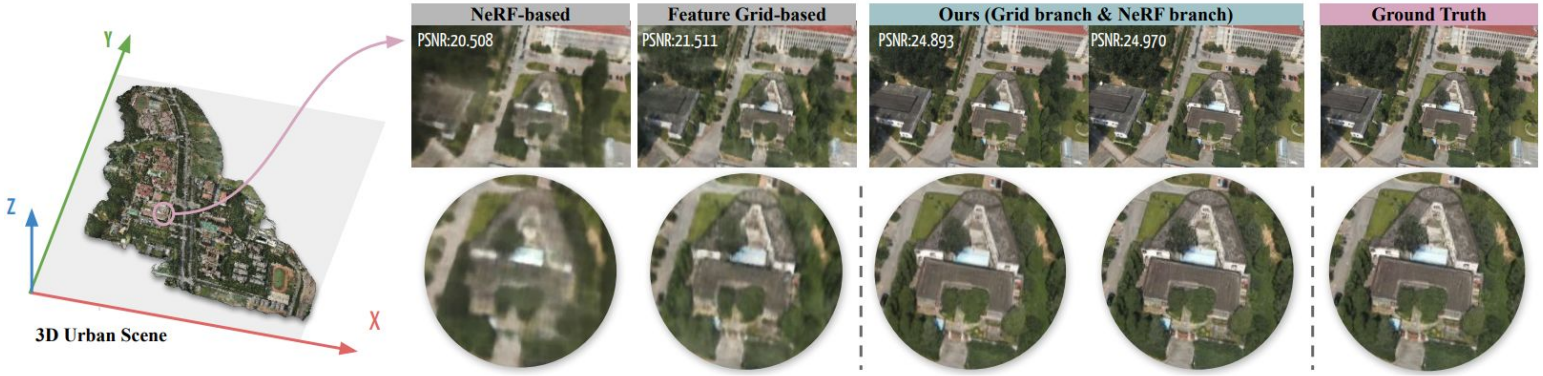


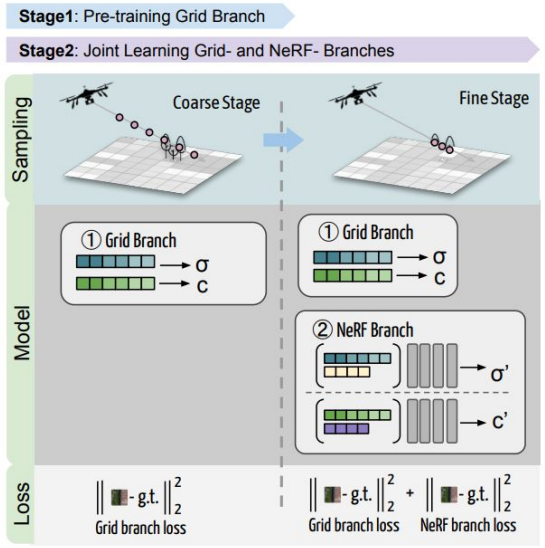
Turning the left to the right by adding ***one*** line of code: `pos_enc[int(t/T*L)+3:]=0`

CVPR 2023 の動向・気付き(216/259)

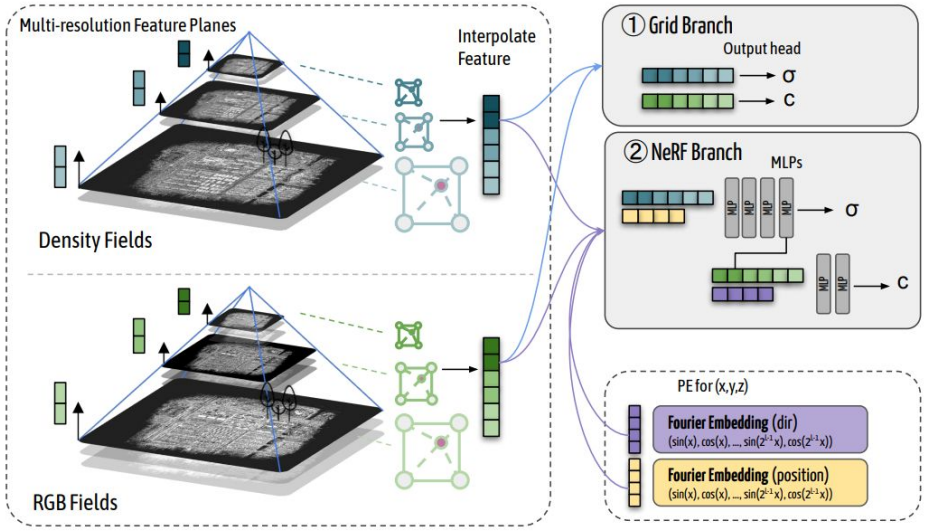
Grid-guided Neural Radiance Fields for Large Urban Scenes

- 都市レベルの大規模なNeRF
- 地面が既知であることを利用しグリッド→NeRFの順に学習





(a) Two Stage Training Scheme

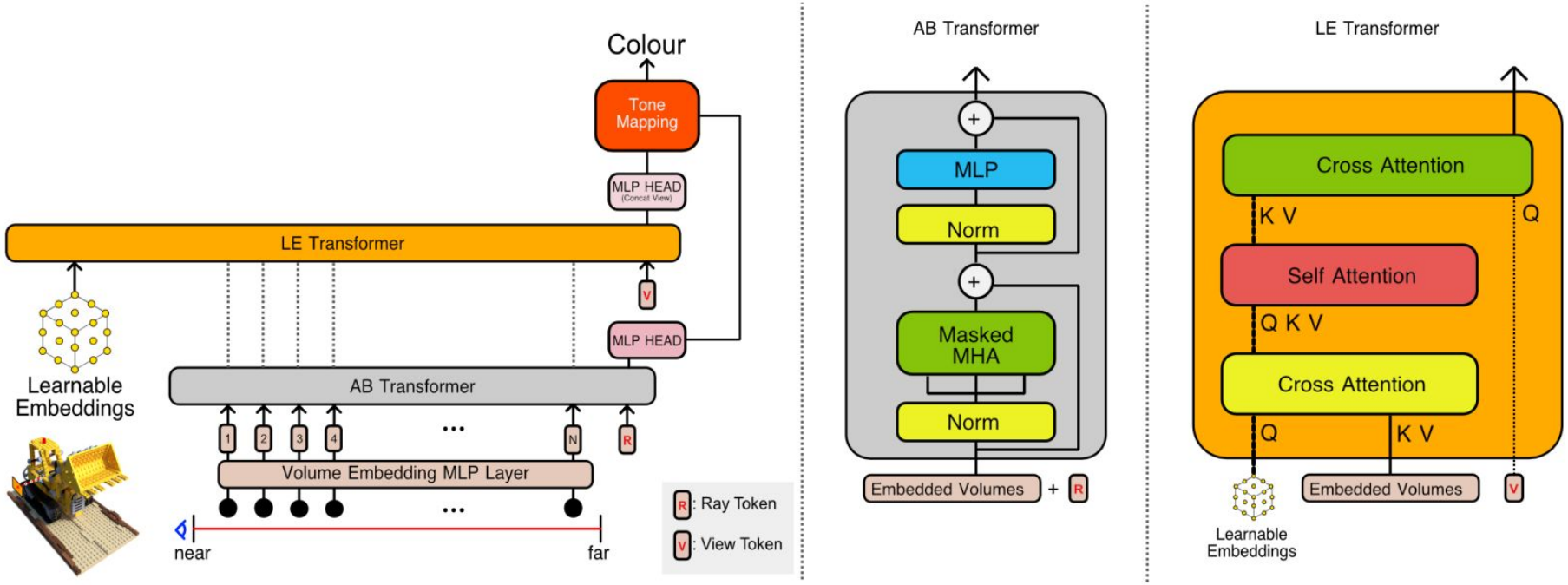


(b) Two Branch Model Design

CVPR 2023 の動向・気付き(217/259)

ABLE-NeRF: Attention-Based Rendering with Learnable Embeddings for Neural Radiance Field

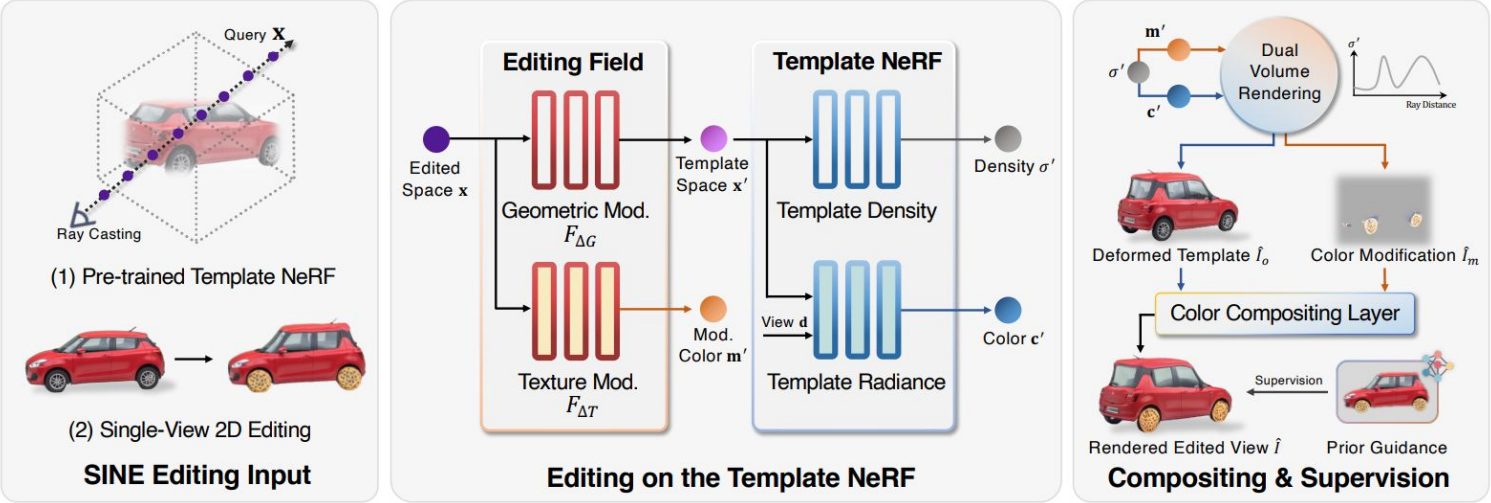
- NeRFのVolume Renderingの集約部分に二段階でTransformerを用いる
  - ある意味ではVolume Renderingのレンダリングとしての妥当性を捨てるという方向性
  - サンプル点間で相互にAttentionを貼るという発想

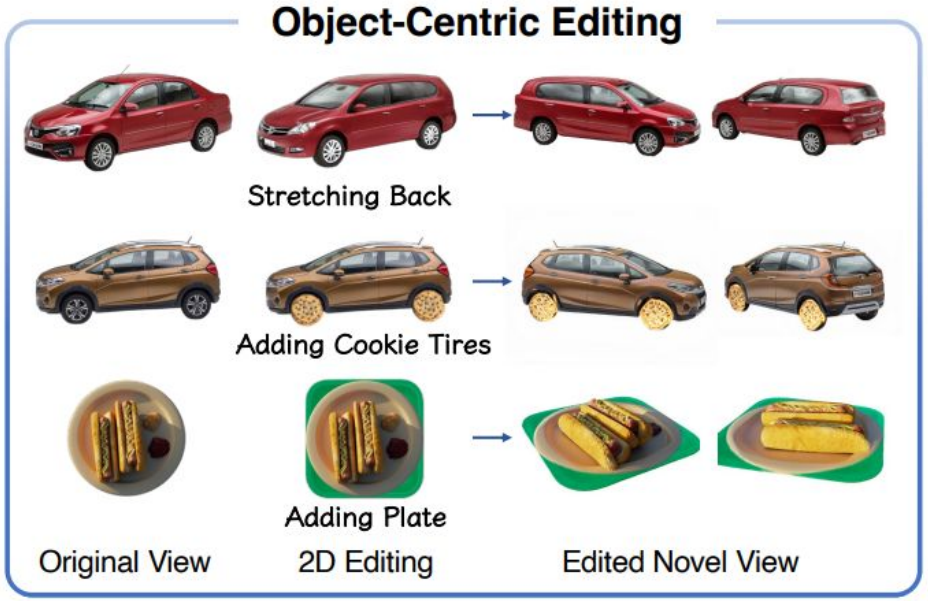


CVPR 2023 の動向・気付き(218/259)

SINE: Semantic-driven Image-based NeRF Editing with Prior-guided Editing Field

- 編集可能なNeRFの派生
- 変更前のテンプレートNeRFと，編集分に対応したNeural Fieldで実装
  - カラーと変形に分けてモデル化
  - SDFも利用するなど最適化・教師の与え方にも工夫
- 2Dでの画像編集(TextでのEditも可)で最適化

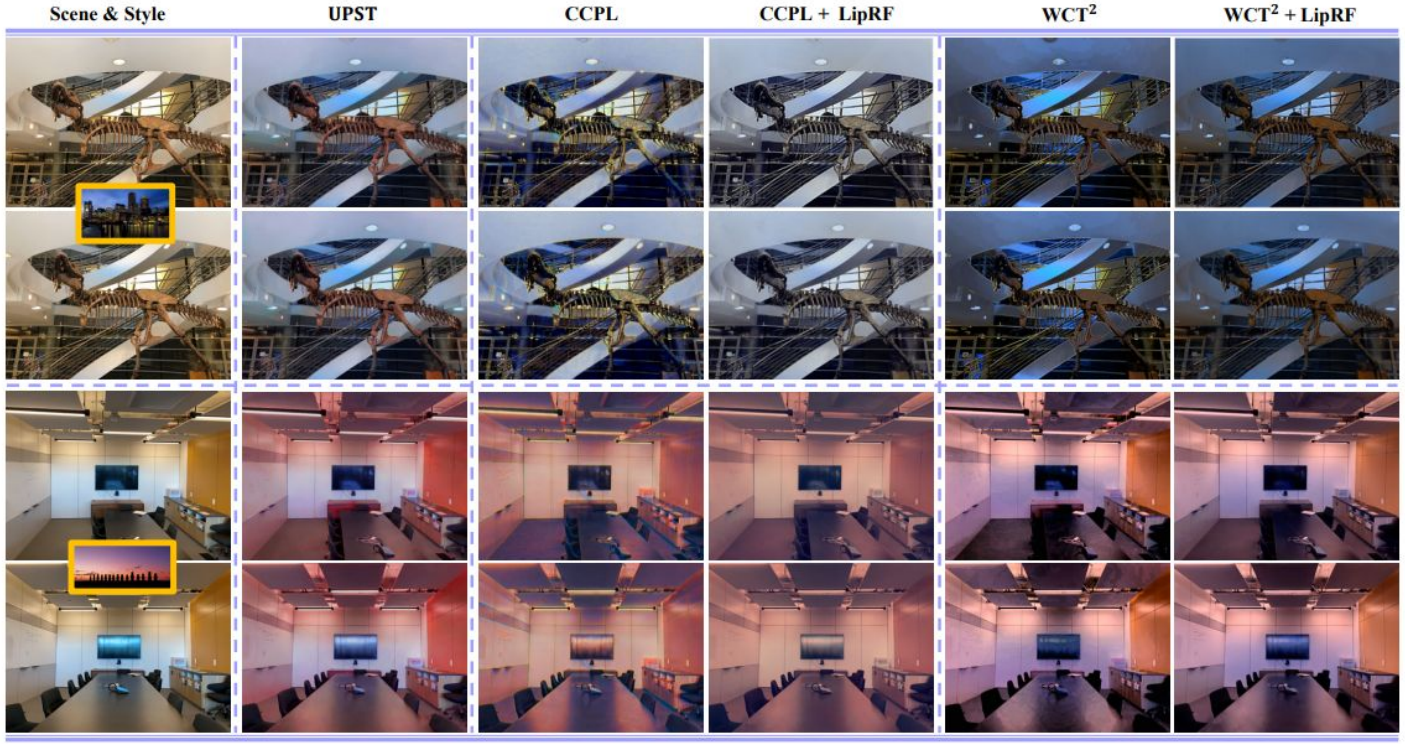




CVPR 2023 の動向・気付き(219/259)

Transforming Radiance Field with Lipschitz Network for Photorealistic 3D Scene Stylization

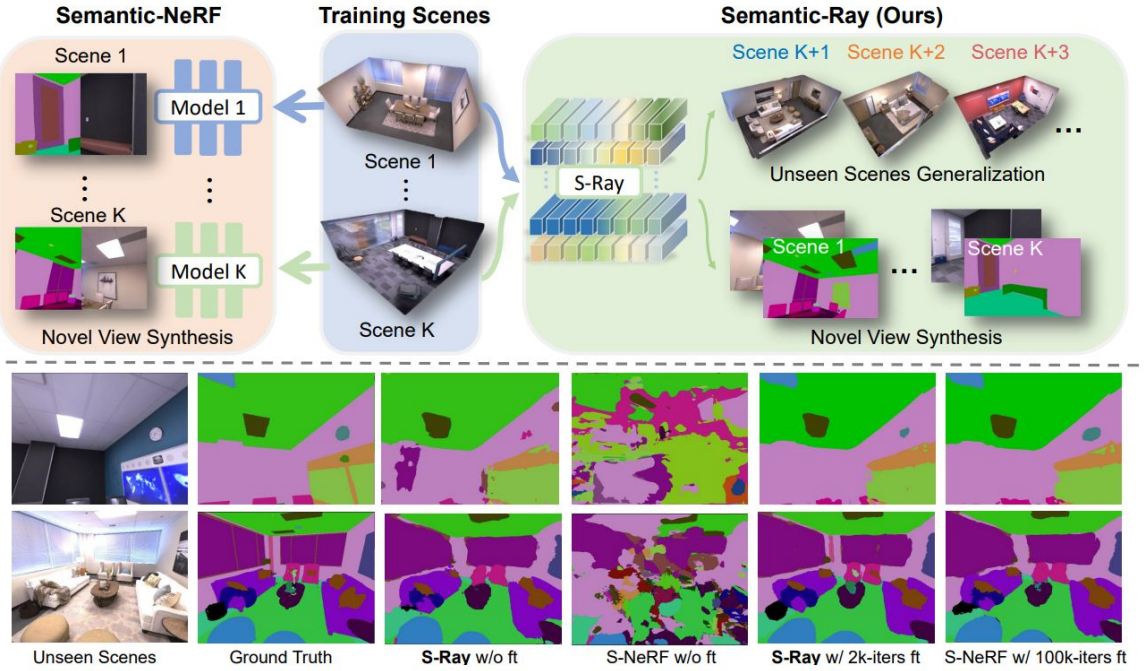
- スタイル変換可能なNeRFの派生
- スタイル変換は2Dで行う
- Lipschitz MLPを用いるというシンプルな工夫で視点間で整合するように
  - ただしメモリ使用量の問題があるため、効率のよい計算方法と合わせて提案

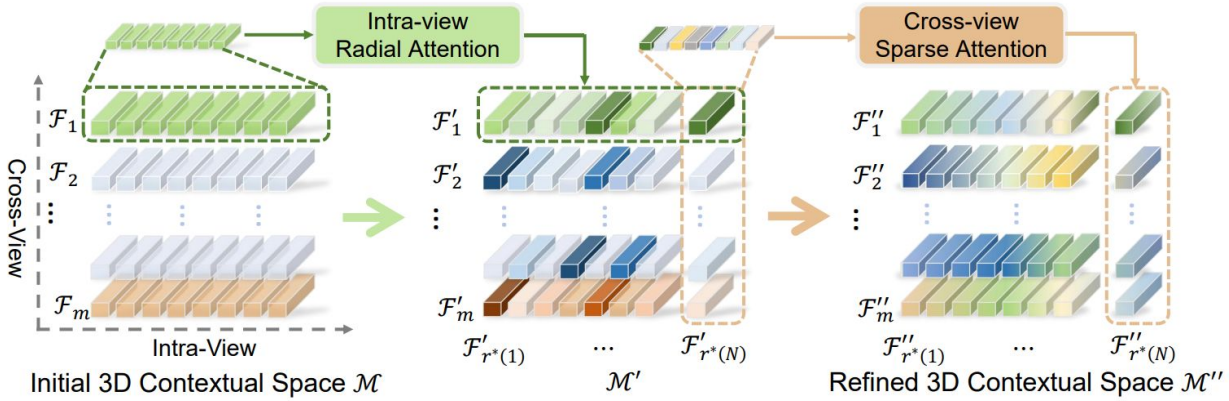


CVPR 2023 の動向・気付き(220/259)

Semantic Ray: Learning a Generalizable Semantic Field with Cross-Reprojection Attention

- Semantic NeRFの3Dは一貫性のあるラベルを提供するが，シーンごとに最適化が必要
- シーンを跨いで一般化したセマンティック推定を実現した
- 2D CNNで計算した特徴量をもとに光線単位でAttention
  - Intra-viewとCross-viewでAttentionを適用する

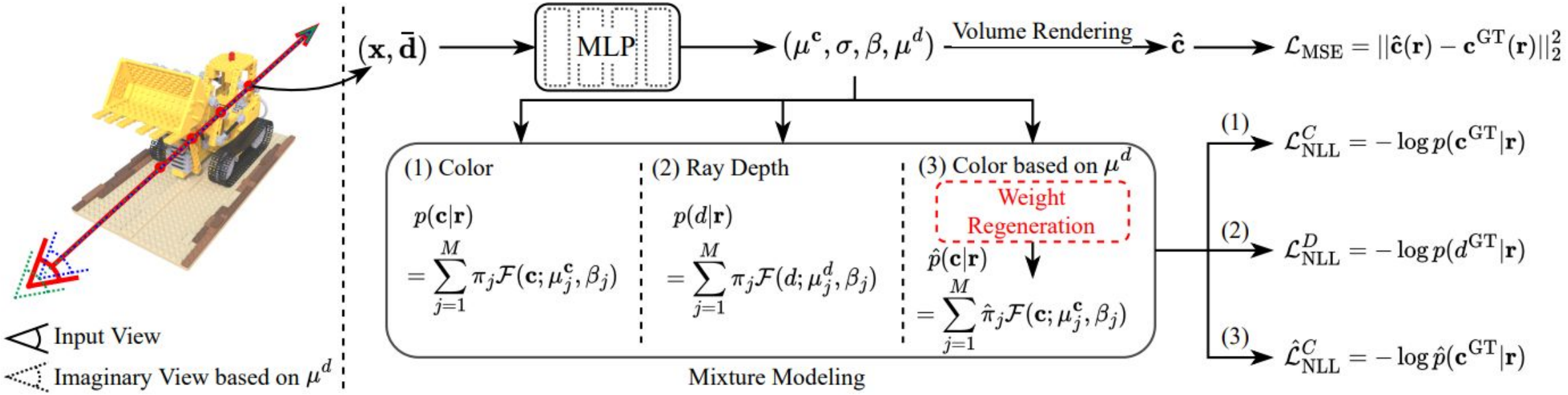




CVPR 2023 の動向・気付き(221/259)

MixNeRF: Modeling a Ray with Mixture Density for Novel View Synthesis from Sparse Inputs

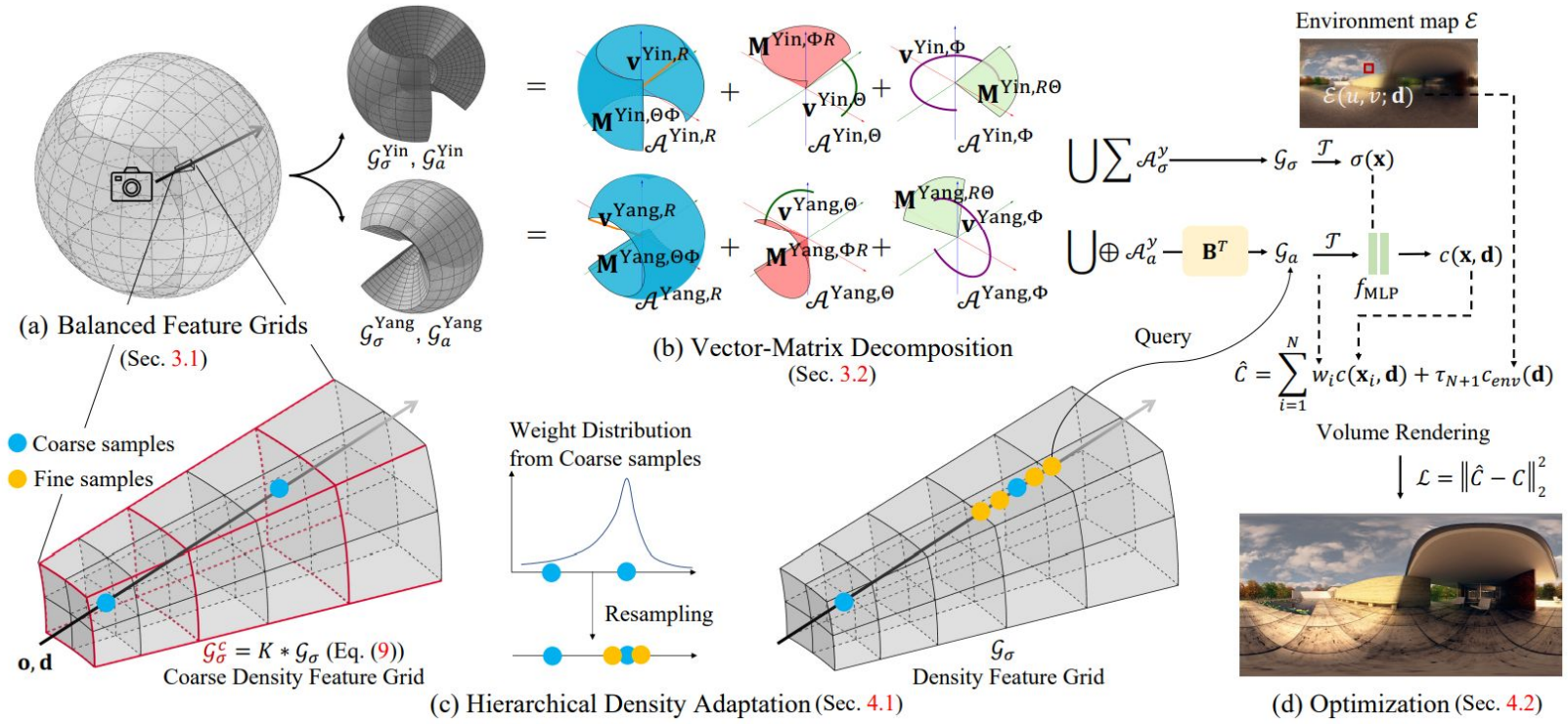
- GMMを用いてNeRFに確率モデルを導入
- 不確かさを陽に扱うことのできるNeRFを実現した
  - Negative Log-likelihoodでもロスを与える

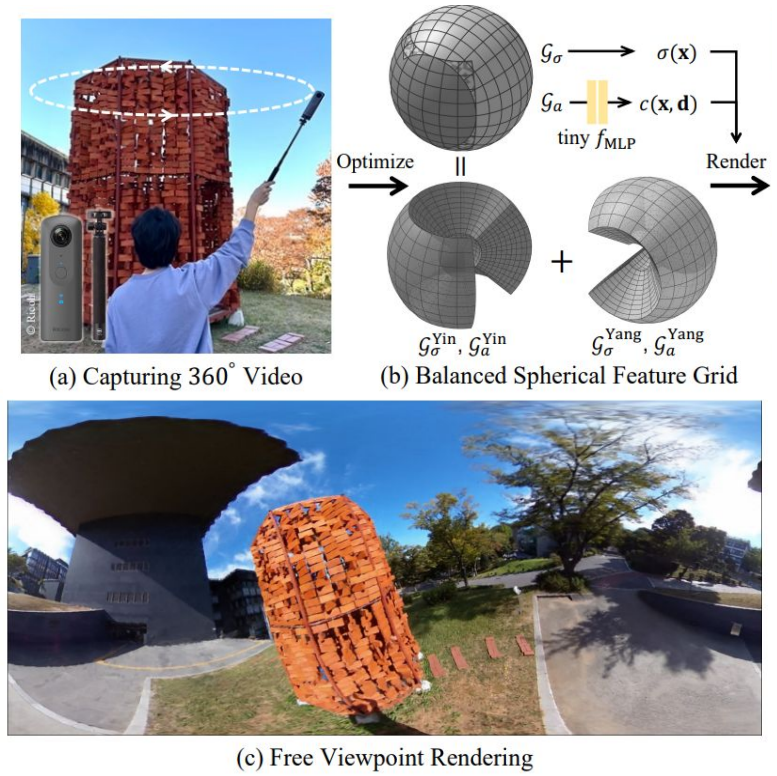


CVPR 2023 の動向・気付き(222/259)

Balanced Spherical Grid for Egocentric View Synthesis

- 全方位カメラからのNeRFを学習する
- カメラからの距離に応じてバランスするようにグリッドを設計

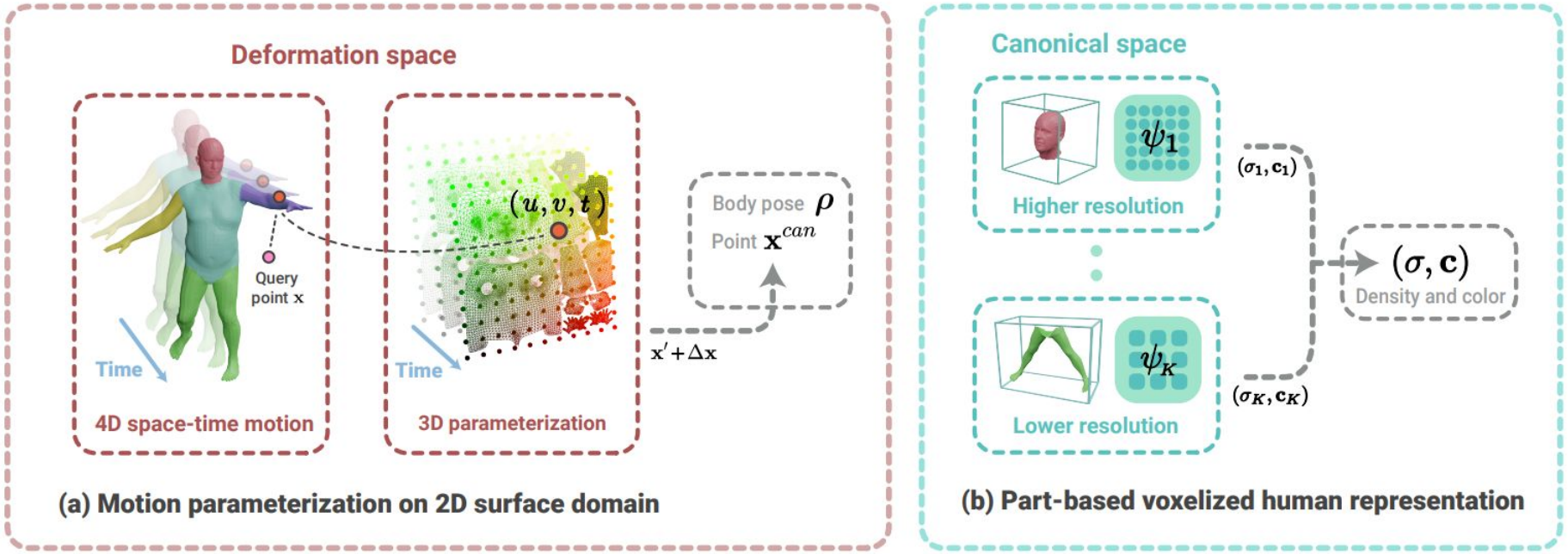




CVPR 2023 の動向・気付き(223/259)

Learning Neural Volumetric Representations of Dynamic Humans in Minutes

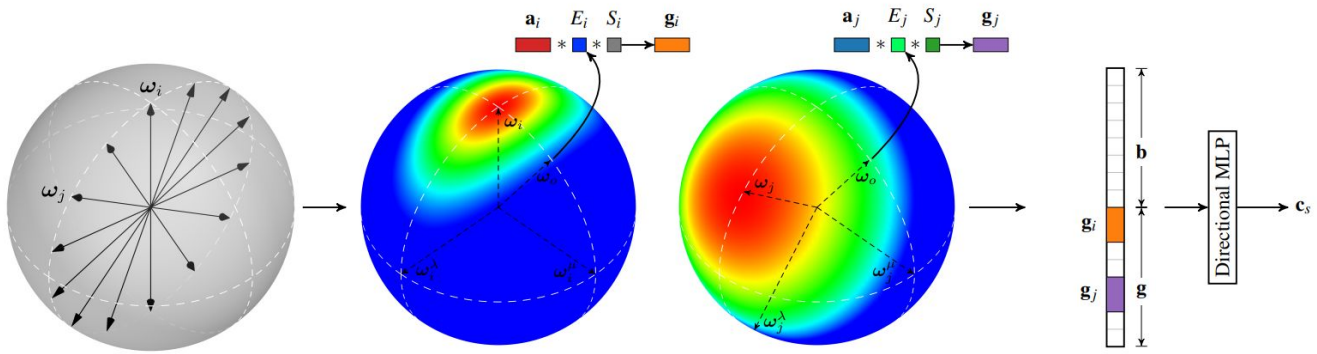
- ビデオから人体のNeRFを最適化
  - SMPLを利用して大まかな形状を記述・変形をNeural Fieldでモデル化
    (最近では一般的なアプローチ)
  - パーツ単位・ボクセル表現で分解

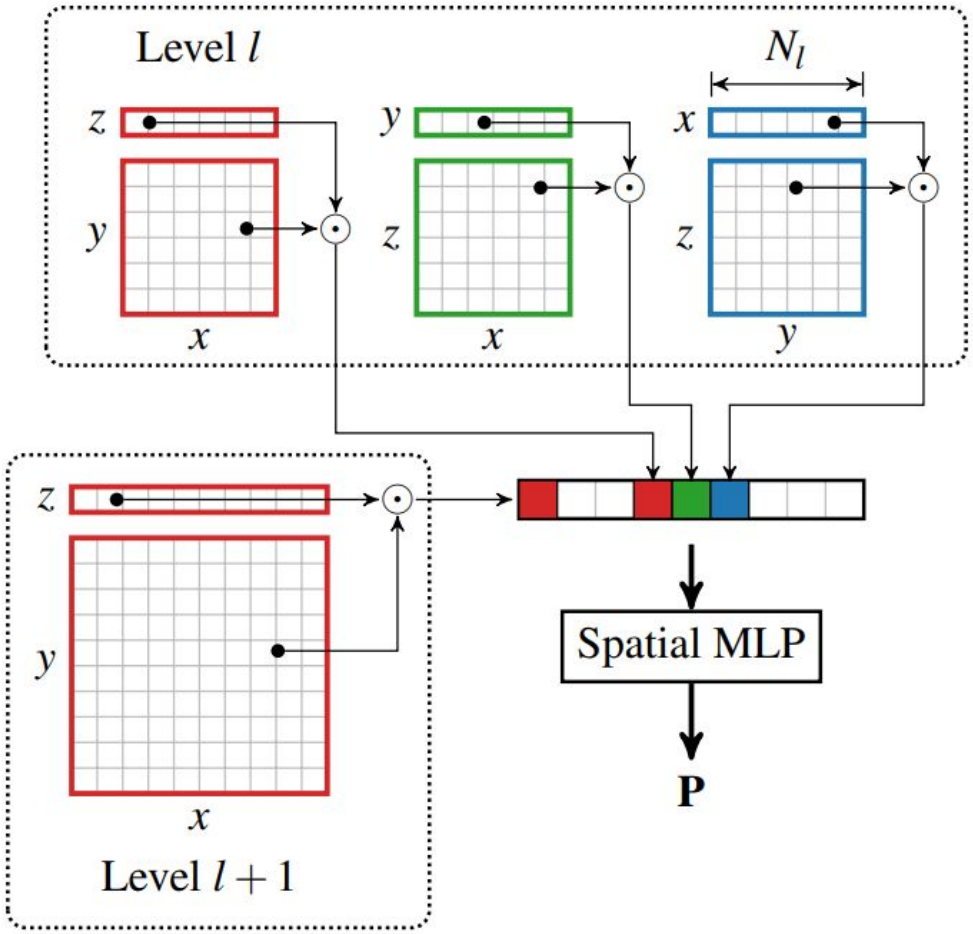


CVPR 2023 の動向・気付き(224/259)

Multiscale Tensor Decomposition and Rendering Equation Encoding for View Synthesis

- テンソル分解＋マルチスケールで
  Neural Fieldの特徴量をモデル化
  - NRFF (Neural Radiance Feature Field)
- 視線方向依存の成分を
  MLPに任せるのではなく
  特徴量空間での混合ガウスでモデル化
  - 鏡面反射などもモデル化できる

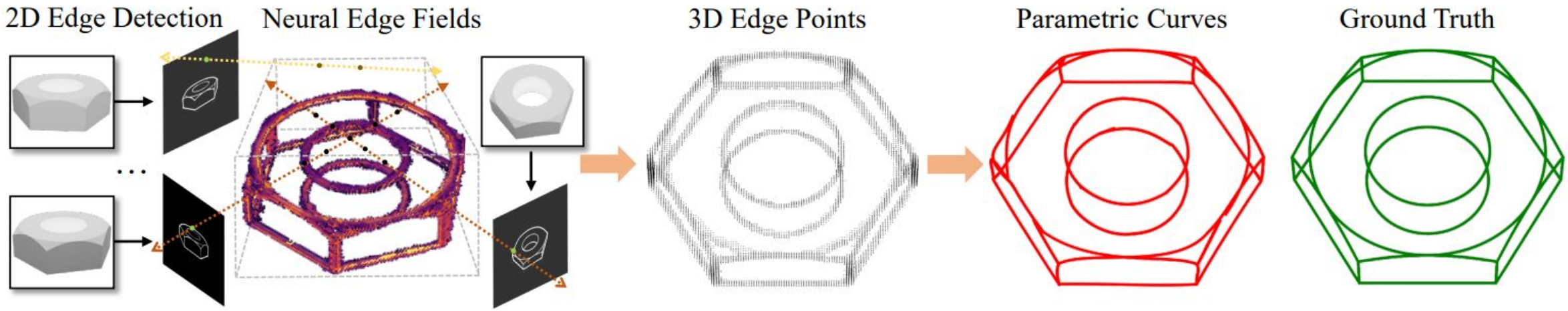




CVPR 2023 の動向・気付き(225/259)

NEF: Neural Edge Fields for 3D Parametric Curve Reconstruction from Multi-view Images

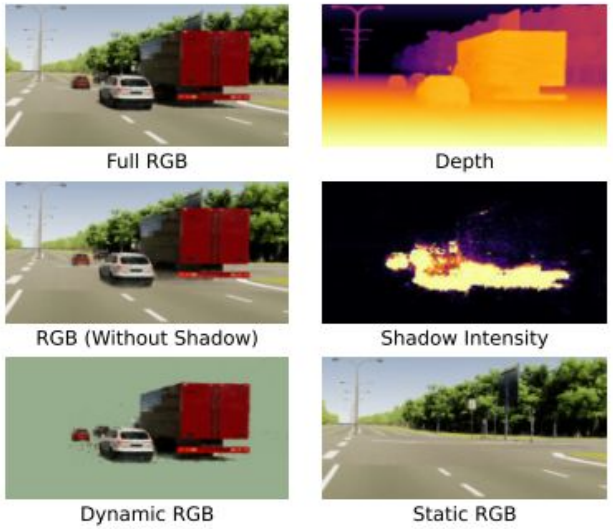
- パラメトリックな3Dエッジ検出
  - 多視点画像からエッジ検出，3Dで整合性をとるためにNeRFを利用
  - 3Dのエッジ点集合からパラメトリックなエッジ線に変換・フィッティング



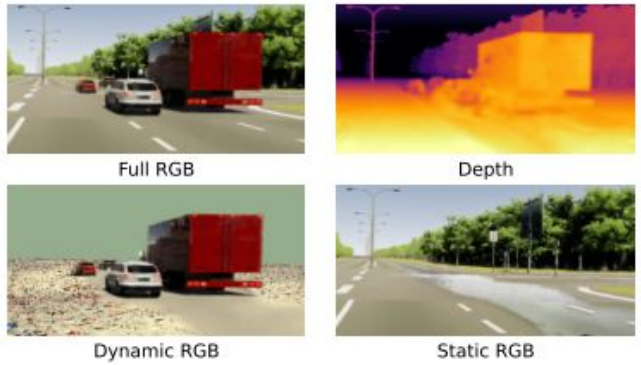
CVPR 2023 の動向・気付き(226/259)

SUDS: Scalable Urban Dynamic Scenes

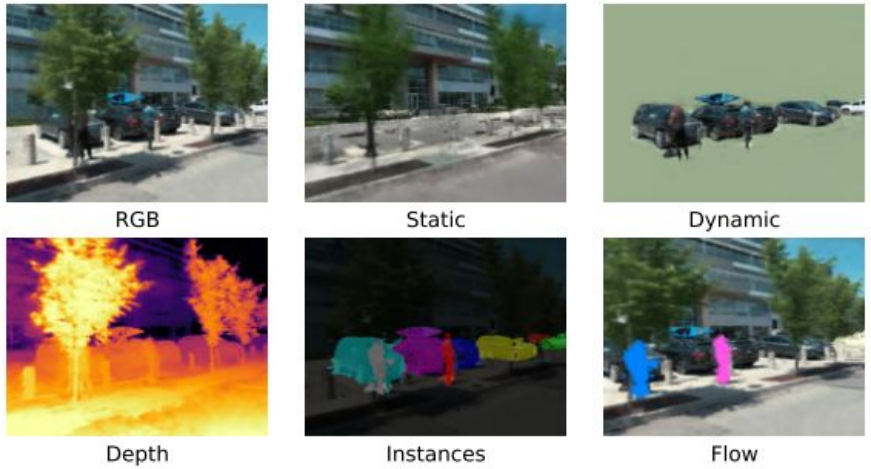
- 大規模(都市レベル)かつ動的な(移動物体のある)シーンでNeRF
- 動く物体と動かない(Static)なシーン，視点の影響の小さい遠方のビューに分けてモデル化
  - 影が問題になることを指摘し陽に記述することで対応



(a) Shadow Field



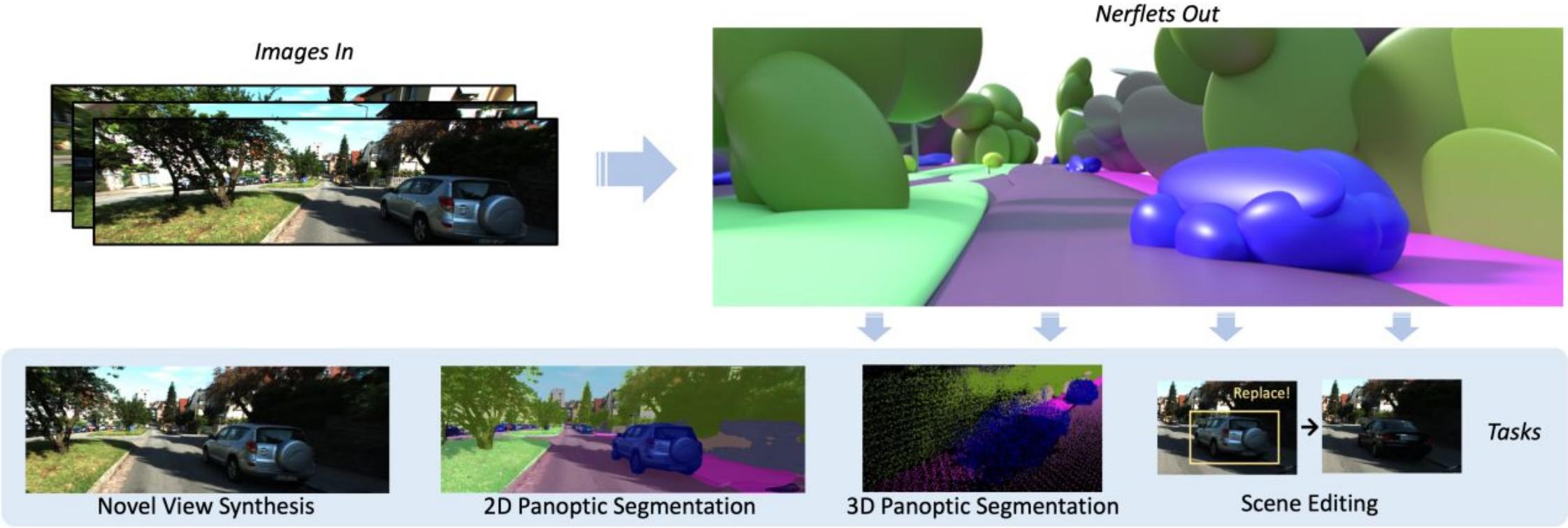
(b) No Shadow Field



CVPR 2023 の動向・気付き(227/259)

Nerflets: Local Radiance Fields for Efficient Structure-Aware 3D Scene Representation from 2D Supervision

- シーン中の各インスタンスをNeRFでモデル化し最適化
  - 2DのPanoptic Segmentationを利用



CVPR 2023 の動向・気付き(228/259)

A Data-Based Perspective on Transfer Learning

- ソースデータセットの構成に注目し，下流タスクの性能に与える影響を調べるためのフレームワークを提案
  - 転移学習において事前学習データを多く含むほど性能が向上するとされているが，ソースデータセットからある程度データを削除することも有効であると示唆
  - 本手法により転移学習の脆弱性を提示

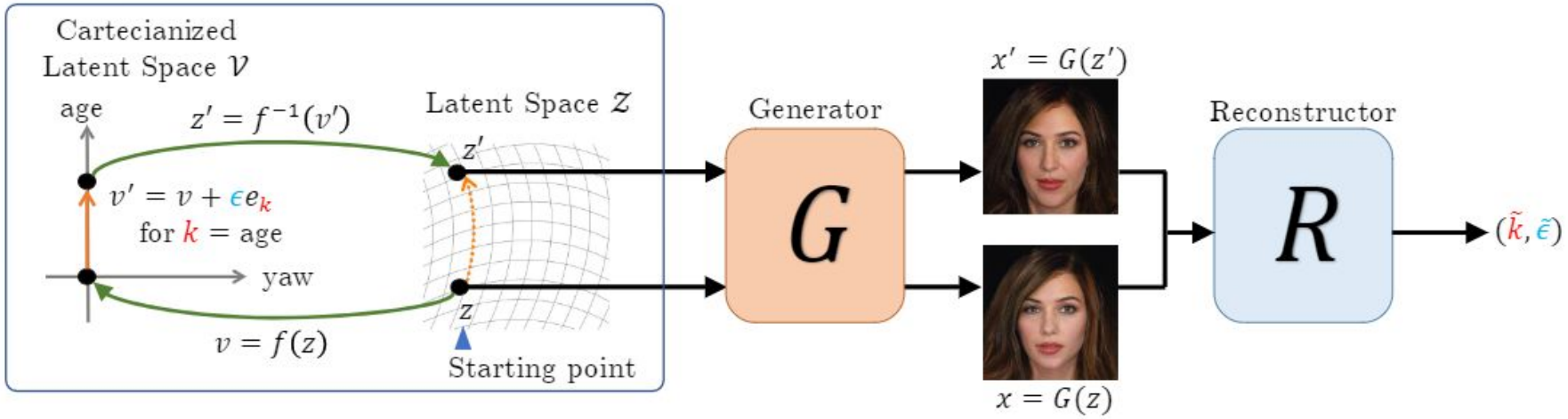
ImageNetから事前学習において悪影響を及ぼす画像データの探索/削除に成功

CVPR 2023 の動向・気付き(229/259)

Deep Curvilinear Editing: Commutative and Nonlinear Image Manipulation for Pretrained Deep Generative Model

- 潜在空間上の意味的な可換なベクトル場を決定する,「DeCurvEd」を提案
  - GANのような生成モデルは, 高品質な画像の生成が可能であるが, 意味的に画像を編集するのは困難
  - DeCurvEdの非線形かつ可換な性質が高品質な画像編集を提供

GANに限らない潜在変数を使うあらゆる生成モデルに利用可能な DeCurvEdが画像の属性を分離した, より高品質な編集を実現

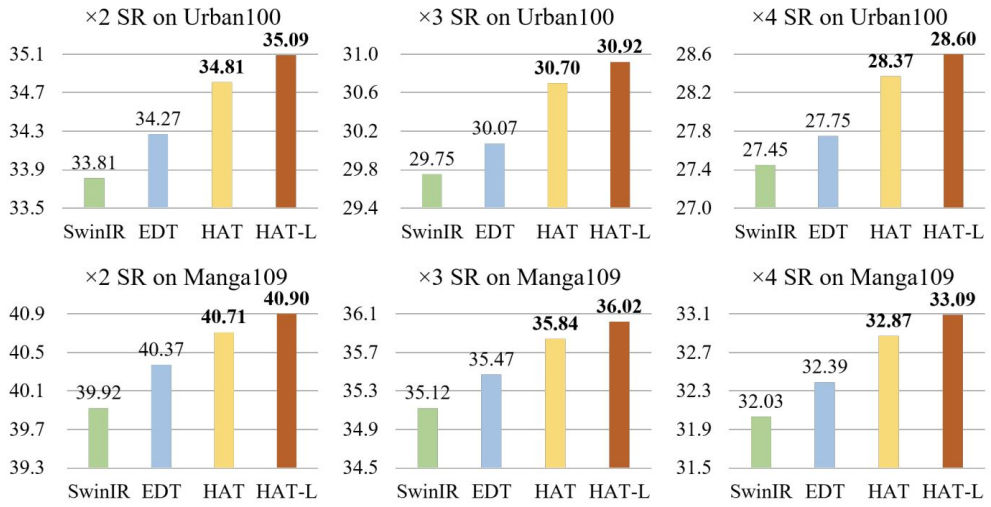


CVPR 2023 の動向・気付き(230/259)

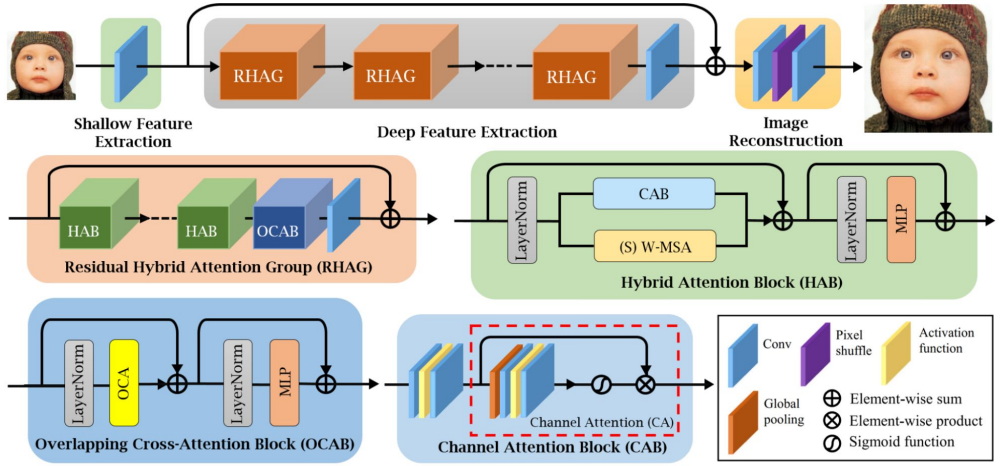
Activating More Pixels in Image Super-Resolution Transformer

Transformerの特徴を生かした超解像手法

- 超解像における既存のTransformerベース手法は大域的な情報を効果的に利用できていない問題を解決するため，attention機構と事前学習手法を見直したHybrid Attention Transformer (HAT) を提案
  - self-attentionとchannel attention，提案するoverlapping cross-attentionを組み合わせることでより大域的なピクセル情報を超解像時に利用
  - 事前学習データに大規模データセット(ImageNet)を利用，更に事前学習とファインチューニングで拡大倍率を合わせるsame-task pre-trainingを導入



SOTAを上回る性能を達成

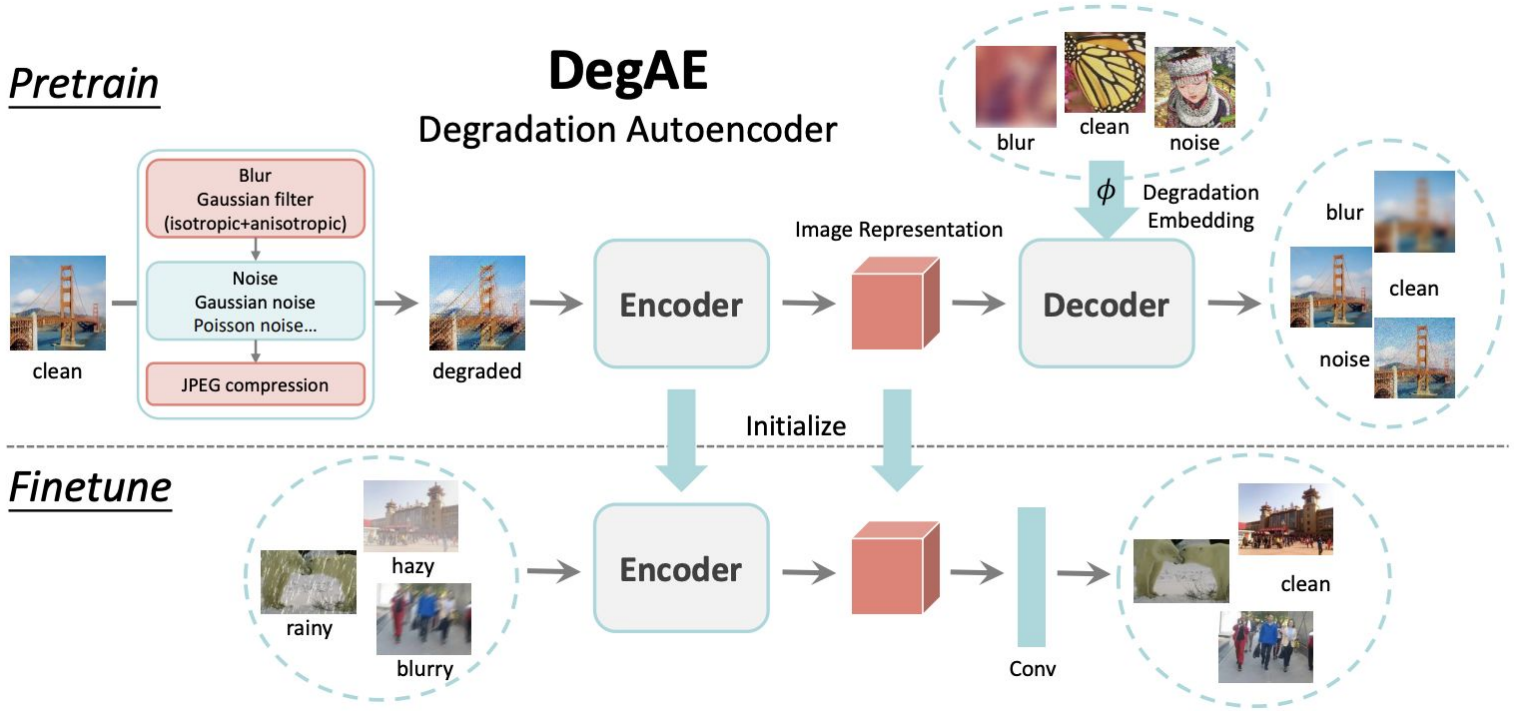


CVPR 2023 の動向・気付き(231/259)

DegAE: A New Pretraining Paradigm for Low-level Vision

low-level visionタスク向けの事前学習手法を提案

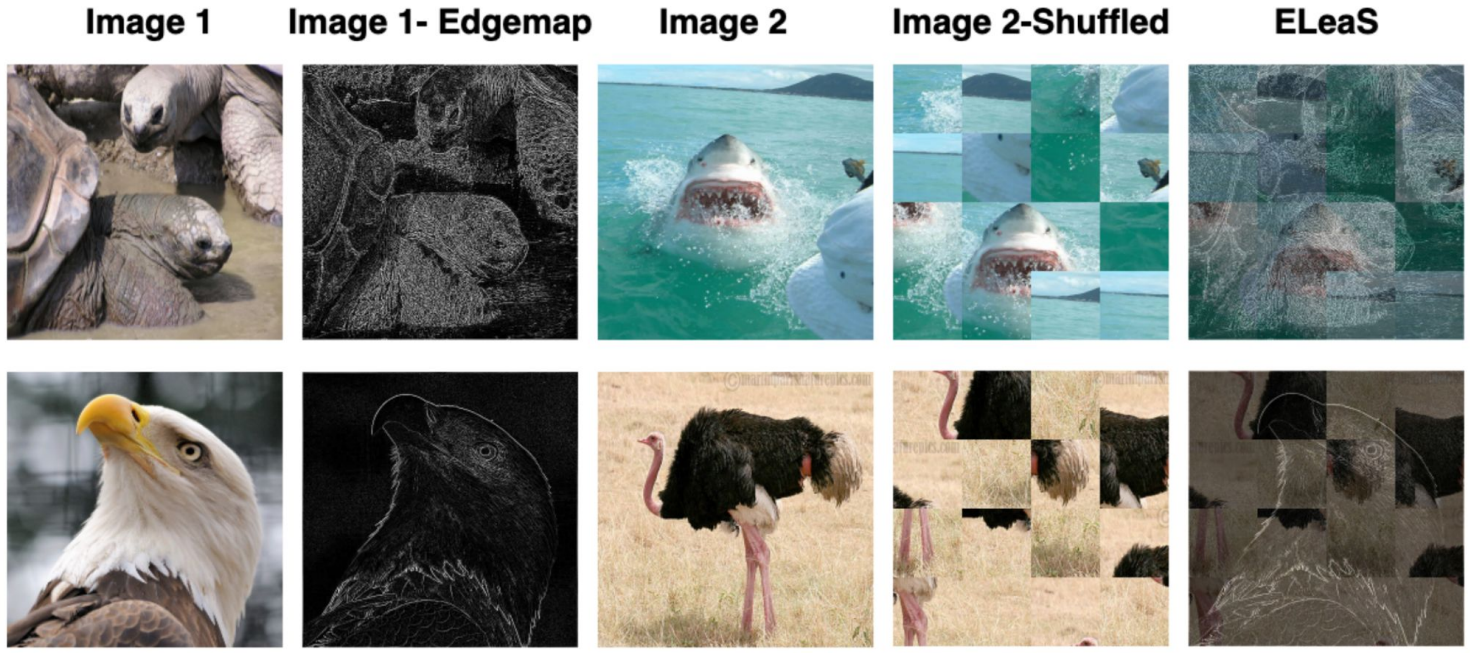
- low-level visionタスクでは事前学習手法が確立されていない
  - high-level visionタスクとlow-level visionタスクでは性質に大きな乖離がある
    - high-level vision：画像を意味レベルで扱う ex) 物体検出，セグメンテーション
    - low-level vision：画像のピクセル単位の特徴を扱う ex) ノイズ除去，超解像



CVPR 2023 の動向・気付き(232/259)

Edges to Shapes to Concepts: Adversarial Augmentation for Robust Vision

- モデルが持つ形状への感度を上昇させるaugmentation手法ELeaSを提案
  - edgeを取り出してパッチをシャッフルしたImageと合成するという軽量なaugmentation手法
  - 一例としてVIT-Sの精度向上,ResNet152の形状への偏重度の上昇,ImageNet-Cに対する広範なモデルにおいて性能上昇を確認

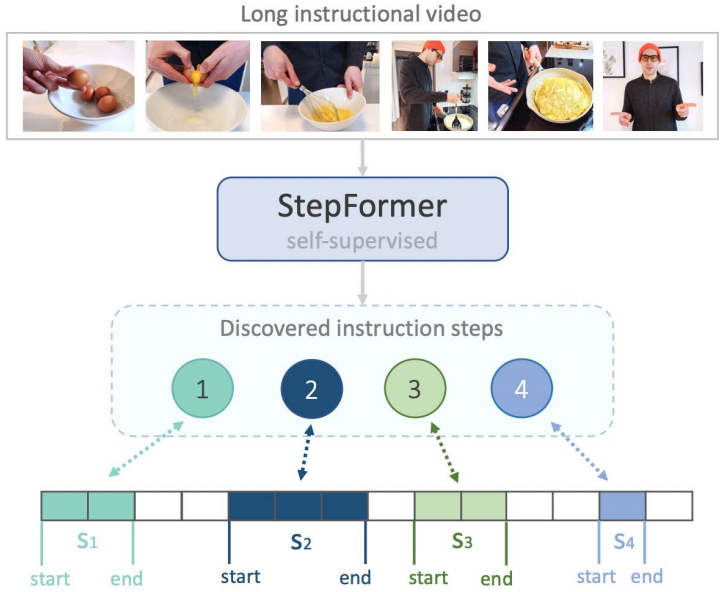


Model	Method	IN-A(↑)	IN-R(↑)	IN-C(↓)	IN-Sketch(↑)	IN-1K(↑)
Resnet50	Vanilla	2.0	36.2	75.0	23.5	76.4
	TSD [21]	3.3	40.8	67.5	28.3	76.9
	ELEAS	**5.4**	**41.7**	**58.5**	**29.7**	**77.1**
Resnet101	Vanilla	5.6	39.3	69.8	27.1	78.0
	TSD [21]	8.8	44.3	62.2	32.3	**78.8**
	ELEAS	**13.4**	**44.4**	**53.5**	**32.4**	78.6
Resnet152	Vanilla	5.9	41.3	67.2	28.4	78.6
	TSD [21]	12.5	45.5	58.9	33.3	79.7
	ELEAS	**15.4**	**45.7**	**53.0**	**34.7**	79.0
ViT-S	Vanilla	16.6	36.1	55.1	33.2	74.6
	TSD [21]	27.4	44.4	42.2	32.4	76.4
	ELEAS	**28.3**	**44.7**	**41.5**	**34.7**	**80.6**
ViT-B	Vanilla	34.6	40.7	50.8	45.8	79.5
	ELEAS	**62.9**	**56.4**	**35.9**	**46.4**	**85.5**
ViT-L	Vanilla	63.4	63.3	33.1	52.7	85.8
	ELEAS	**67.4**	**65.9**	**29.5**	**54.1**	**86.2**

CVPR 2023 の動向・気付き(233/259)

StepFormer: Self-supervised Step Discovery and Localization in Instructional Videos

- 字幕を用いた自己教師あり学習でインストラクション動画のキーステップ特定
  - **背景**
    - 長いインストラクション動画から，手順を学ぶためには動画内から手順に必要なキーステップを特定する必要がある.
  - **過去の手法背景**
    - キーステップを特定するための教師/弱教師あり学習を行っていた.
      → 教師を用意するのが大変.
  - **本研究の視点**
    - 動画の字幕から抽出したステップを教師として，動画のステップを学習する．このとき，映像系列を字幕におけるステップと対応づけるのに，Drop-DTWアルゴリズムを用いる.

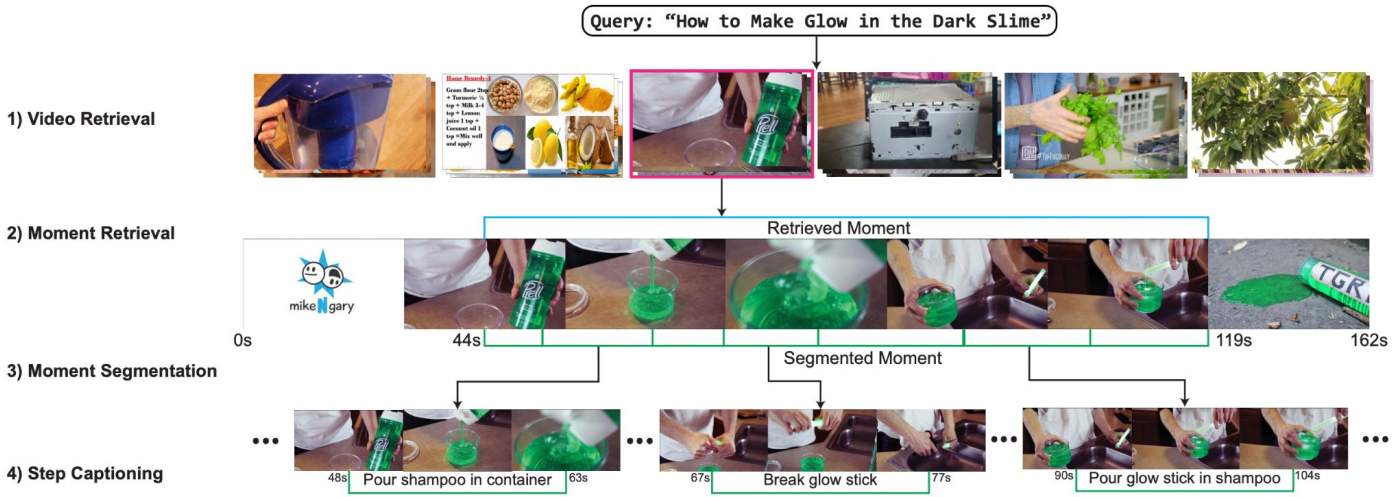


N. Dvornik et al. “StepFormer: Self-supervised Step Discovery and Localization in Instructional Videos”, in CVPR 2023.

CVPR 2023 の動向・気付き(234/259)

Hierarchical Video-Moment Retrieval and Step-Captioning

- 大規模なインストラクション動画にステップ・キャプションをアノテーション
  - **背景**
    - 大量の映像資料が蓄積されていく近年，動画内からテキストに応じた適切な部分を抽出したり要約する技術はますます重要になっている.
  - **過去の手法背景**
    - COINやCrossTaskなど，動画にステップごとのキャプションを付与したデータセット.
      → 動画の内容が限られたタスクのみ.
  - **本研究の視点**



ごとのアノテーションを付与し，retrieval 提案.

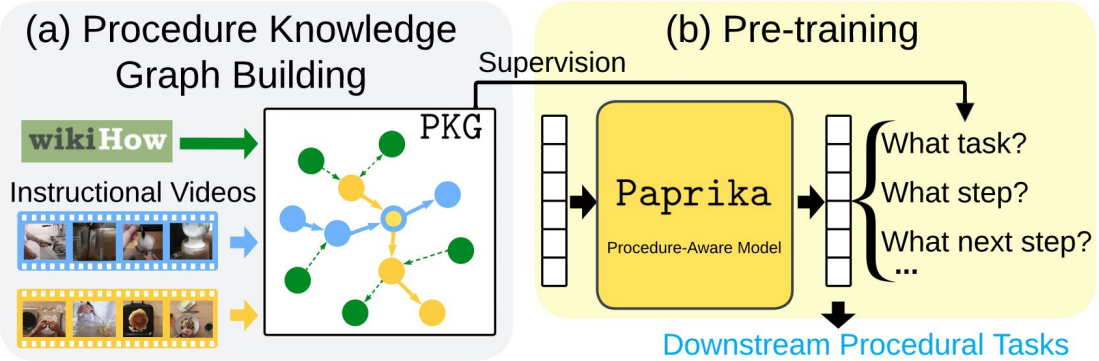
A. Zala et al. "Hierarchical Video-Moment Retrieval and Step-Captioning", in CVPR 2023.

CVPR 2023 の動向・気付き(235/259)

Procedure-Aware Pretraining for Instructional Video Understanding

○ 手順知識を利用したインストラクション動画理解のための事前学習

- **背景**
  - 人間がインストラクション動画から手順を学ぶように，人工知能もインストラクション動画から，タスクや手順を構造化して理解したい．さらには手順予測をしたい．
- **過去の手法背景**
  - HowTo100MなどWeb上の動画から得られる字幕は手順の学習に使うにはノイジーなので，字幕を参考に既存の手順知識データベース(wikiHow)における手順説明に置換．
    → 各手順について，動画表現を獲得しただけで，手順の文脈は意識していない
- **本研究の視点**
  - 手順はその**順序や関係性**などの情報が含まれるので，単純にwikiHowの手順に置換するのではなく，**wikiHowを知識グラフ化してそれを学習の教師として用いる**．

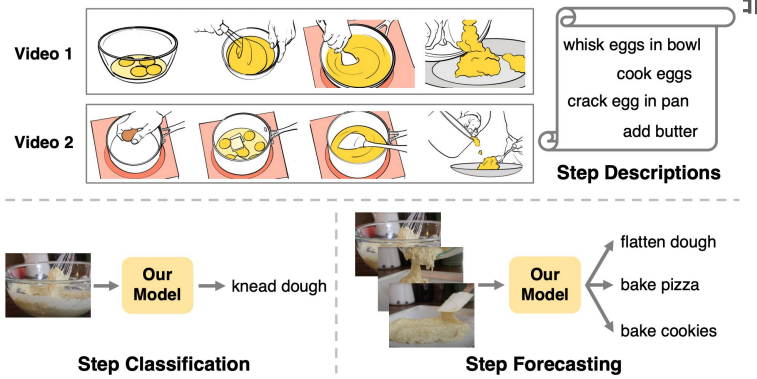


H. Zhou et al. "Procedure-Aware Pretraining for Instructional Video Understanding", in CVPR 2023.

CVPR 2023 の動向・気付き(236/259)

Learning Procedure-aware Video Representation from Instructional Videos and Their Narrations

- 手順を意識したインストラクション動画理解のための動画表現獲得
  - **背景**
    - 人工知能がインストラクション動画における手順を理解し学習できると，欠けている手順を補ったり，手順予測ができたりする.
  - **過去の手法背景**
    - HowTo100MなどWeb上の動画から得られる字幕は手順の学習に使うにはノイジーなので，字幕を参考に既存の手順知識データベース(wikiHow)における手順説明に置換.
      → 各手順について，動画表現を獲得しただけで，手順の文脈は意識していない
  - **本研究の視点**
    - 各手順の**順番や関係性**を意識して学習を行う．また，**拡散プロセスを利用した確率モデル**により，手順予測の際に可能な複数の手順を推論できる(手順の**不確実性**への対処)

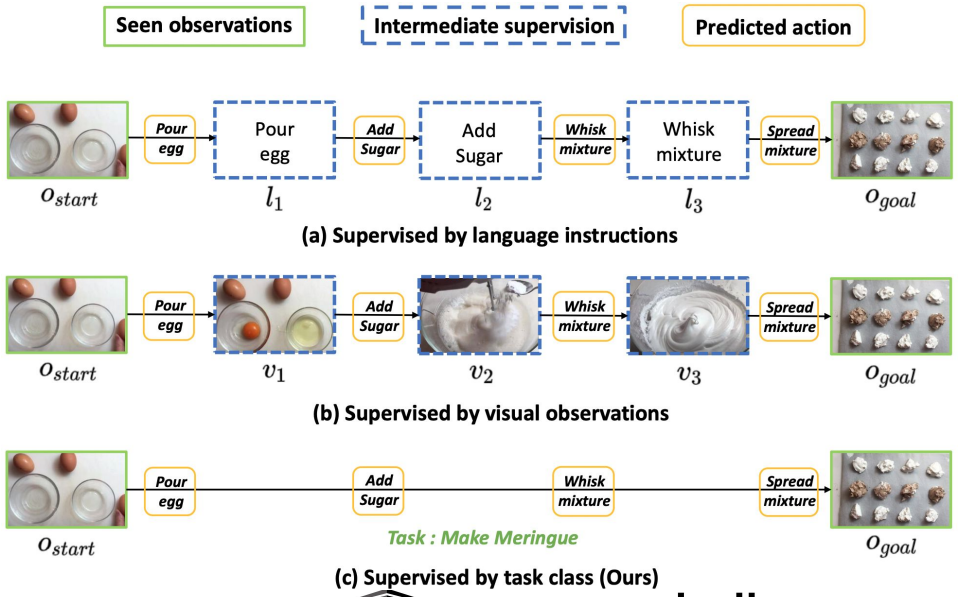


Y. Zhong et al. "Learning Procedure-aware Video Representation from Instructional Videos and Their Narrations", in CVPR 2023.

CVPR 2023 の動向・気付き(237/259)

PDPP: Projected Diffusion for Procedure Planning in Instructional Videos

- 手順計画を分布フィッティング問題として捉え拡散モデルを適用
  - **背景**
    - 人工知能がインストラクション動画から手順を理解して学習し，ゴールに向けた手順計画ができると料理をしたり修理をしたりなどの問題が解けるようになる.
  - **過去の手法背景**
    - 手順シーケンスを正確に予測する問題として捉えて，中間状態をステップ毎に予測していた.
      → 中間状態予測の誤差が募る，学習時の中間状態教師が必要.
  - **本研究の視点**
    - 手順の**不確実性**(入れ替え可能性など)に対処するため，**分布フィッティング問題**として捉え，学習プロセスにランダム性を追加する.
      中間状態の教師を必要とせず，拡散過程で一気に予測.



H. Wang et al. "PDPP: Projected Diffusion for Procedure Planning in Instructional Videos", in CVPR 2023.

CVPR 2023 の動向・気付き(238/259)

インストラクション動画解析の動向

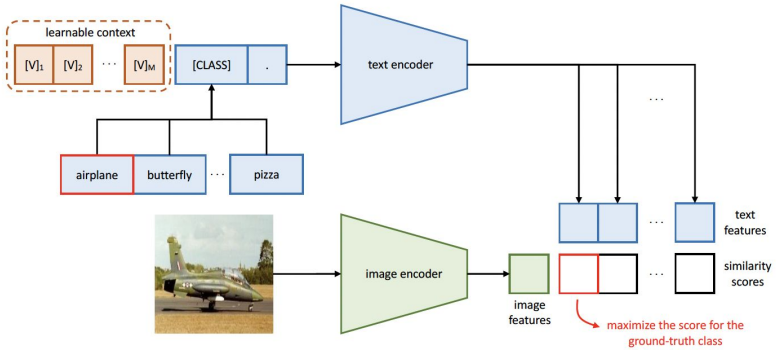
- 字幕を効果的に用いた教師情報
  - 字幕をそのまま使うのではなく，字幕をうまく利用して動画内の各ステップをうまく区分し，教師情報とする.
- 知識データベースを利用した学習
  - wikiHowなどの手順に関する知識データベースを教師として学習.
- 拡散過程の利用
  - 手順の不確実性に拡散モデルの特性を活かす.

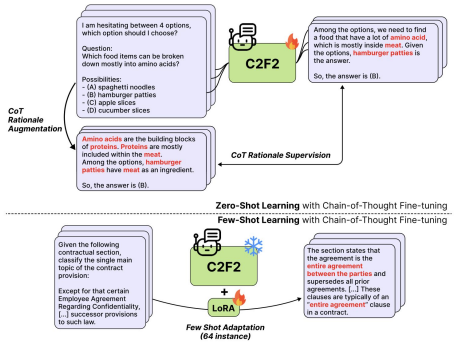
CVPR 2023 の動向・気付き(239/259)

Tutorial: Prompts in Vision

https://prompting-in-vision.github.io

- Vision and Languageにおけるプロンプティングの背景を網羅的に解説
- 手動プロンプティングは最適化が難しい
  - CoOp(Context Optimization):プロンプト内の文脈語を学習可能なベクトル表現にすることでプロンプトエンジニアリングを自動化する
  - CoOpは学習データに過適合してしまう→Conditional Context Optimization (CoCoOp)
- LLMはChain-of-Thought (CoT)で性能が上がることが知られているが，小さいLMに段階的な推論をさせるにはどうすれば？
  - 1,060タスクにわたる188万個のCoT推論を拡張したCoT Collection datasetで，Flan-T5 (3B&11B) を継続的にFTすることで性能が向上.



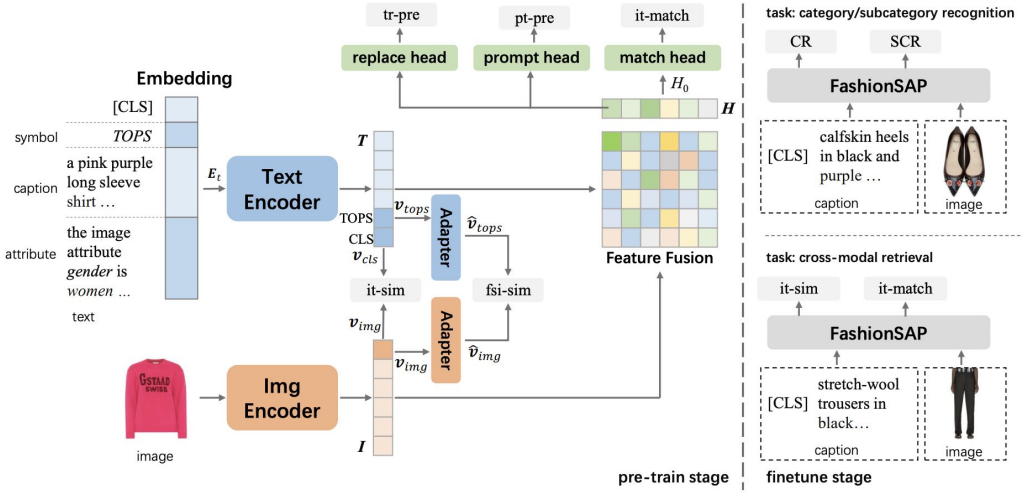


CVPR 2023 の動向・気付き(240/259)

FashionSAP: Symbols and Attributes Prompt for Fine-grained Fashion Vision-Language Pre-training

- ファッションシンボルと属性プロンプトに基づく事前学習モデルを提案
  - 一般的なVLデータにはファッションアイテム詳細が記述されないため学習には非効率
  - ファッションシンボル(左下図)と属性プロンプトの提案
    - 様々な種類のきめ細かいファッションの特徴を一般化し細かい属性をモデル化
    - ファッションデータの形式に応じたプロンプトテンプレートを設計
  - カテゴリとサブカテゴリ認識でSoTA

Fashion Symbols	Categories	Definition Rules
TOPS	tops, shirt, polo, sweater, ...	upper body
DRESSES	dress, suit, shift, ...	up-to-lower body
SKIRTS	skirt, sarong, slit, kilt, ...	lower body
COATS	jacket, parka, blazer, duffle, ...	associated with others
PANTS	jeans, shorts, breeches, ...	lower body
SHOES	boots, sneakers, pump, loafers, ...	feet
BAGS	clutches, pouches, wristlet, ...	bag & decorative
ACCESSORIES	ring, sunglasses, accessories, hat, necklace, ...	decorative & optional
OTHERS	swim-wear, lingerie, lounge-wear, ...,	-

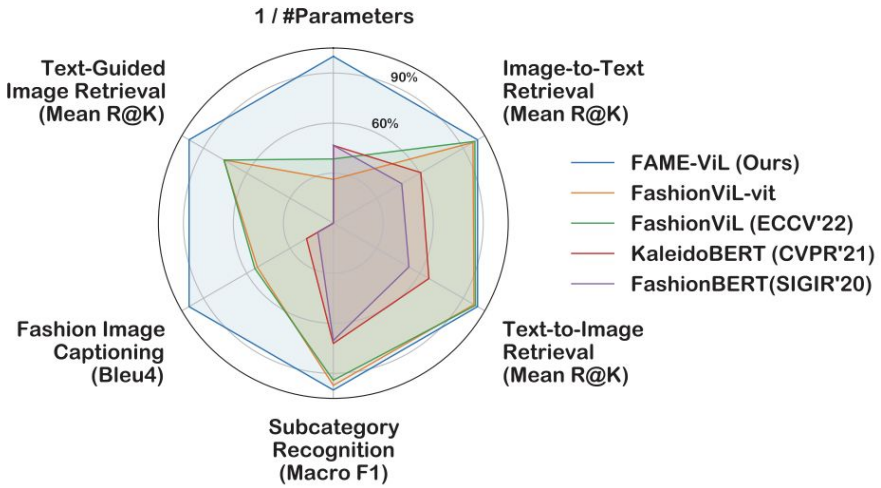


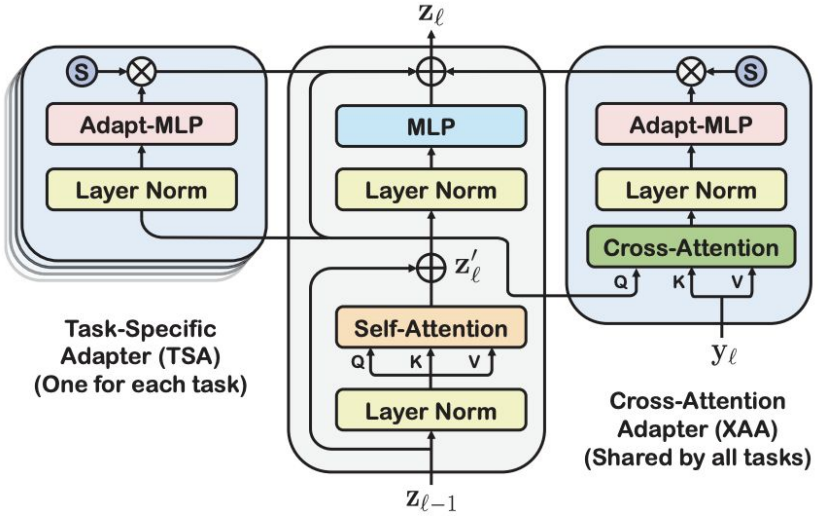
Han, Yunpeng, et al. "FashionSAP: Symbols and Attributes Prompt for Fine-grained Fashion Vision-Language Pre-training.", in CVPR2023
https://openaccess.thecvf.com/content/CVPR2023/html/Han_FashionSAP_Symbols_and_Attributes_Prompt_for_Fine-Grained_Fashion_Vision-Language_Pre-Training_CVPR_2023_paper.html

CVPR 2023 の動向・気付き(241/259)

FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks

- ファッションに焦点を当てたマルチタスクかつ効率的なモデルの提案
  - マルチタスクラーニングのために設計された2種類のアダプタを提案(右下図)
    - タスク固有のアダプタ(TSA), 異なるモダリティ間の相互作用を可能にするアダプタ(XAA)
  - 従来の単一タスクモデルと比較してパラメータを61.5%削減した
  - Cross-Modal Retrieval, Text-Guided Image Retrieval, Subcategory Recognition, Fashion Image Captioningのタスクで SoTA(左下図)



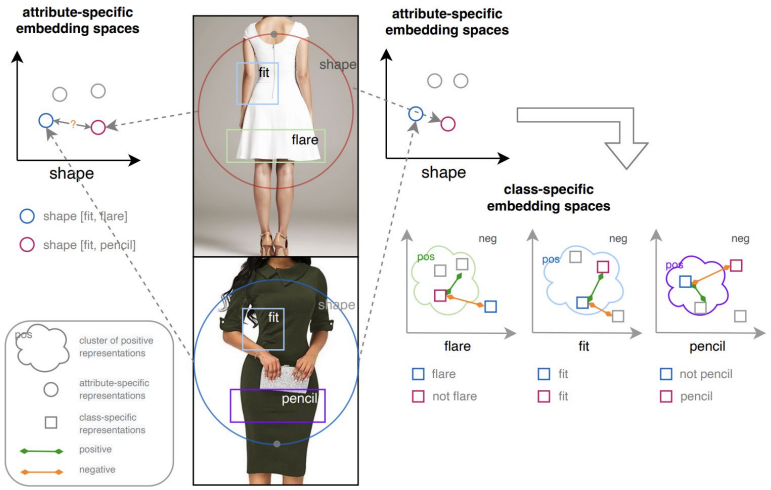


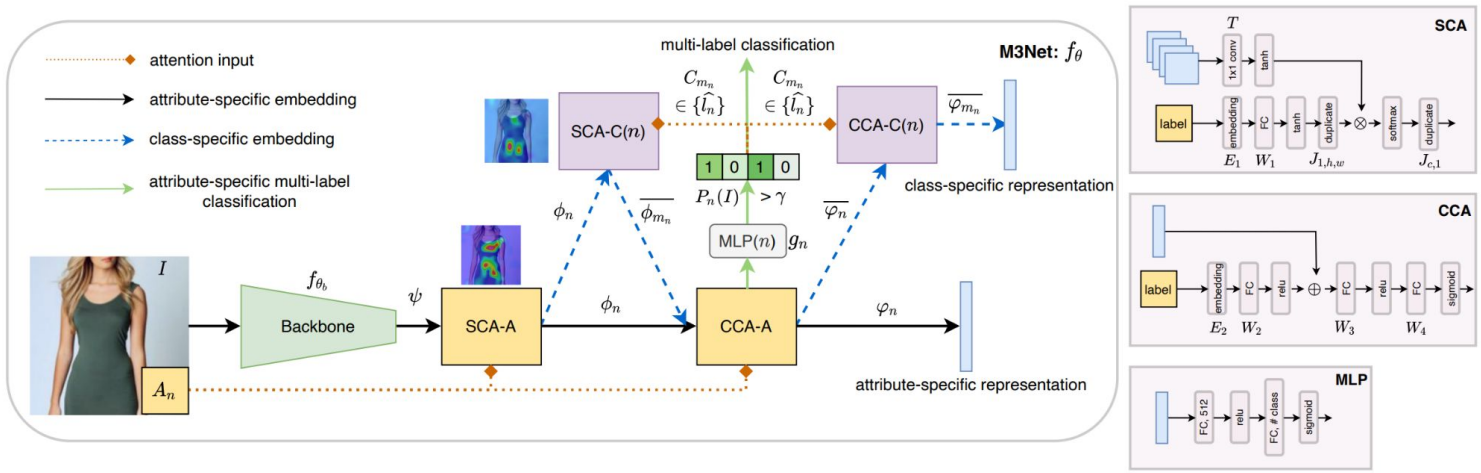
Han, Xiao, et al. "FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks," in CVPR2023
https://openaccess.thecvf.com/content/CVPR2023/html/Han_FAME-ViL_Multi-Tasking_Vision-Language_Model_for_Heterogeneous_Fashion_Tasks_CVPR_2023_paper.html

CVPR 2023 の動向・気付き(242/259)

Learning Attribute and Class-Specific Representation Duet for Fine-grained Fashion Analysis

- 詳細なファッション分析を行うためのマルチ属性・粒度・ラベルなモデル提案
  - ファッションの属性とクラスに特化した表現を共同で学習
    - クラスレベルで視覚要素の構成が異なる2つのドレスを識別可能(左下図参考)
  - 共有のバックボーンと属性とクラスの条件付きAttentionをもつサブネットワークにより属性とクラス間の相互依存性と視覚特徴量を捉える(右下図参考)





Jiao, Yang, et al. "Learning Attribute and Class-Specific Representation Duet for Fine-grained Fashion Analysis."in CVPR2023
https://openaccess.thecvf.com/content/CVPR2023/html/Jiao_Learning_Attribute_and_Class-Specific_Representation_Duet_for_Fine-Grained_Fashion_Analysis_CVPR_2023_paper.html

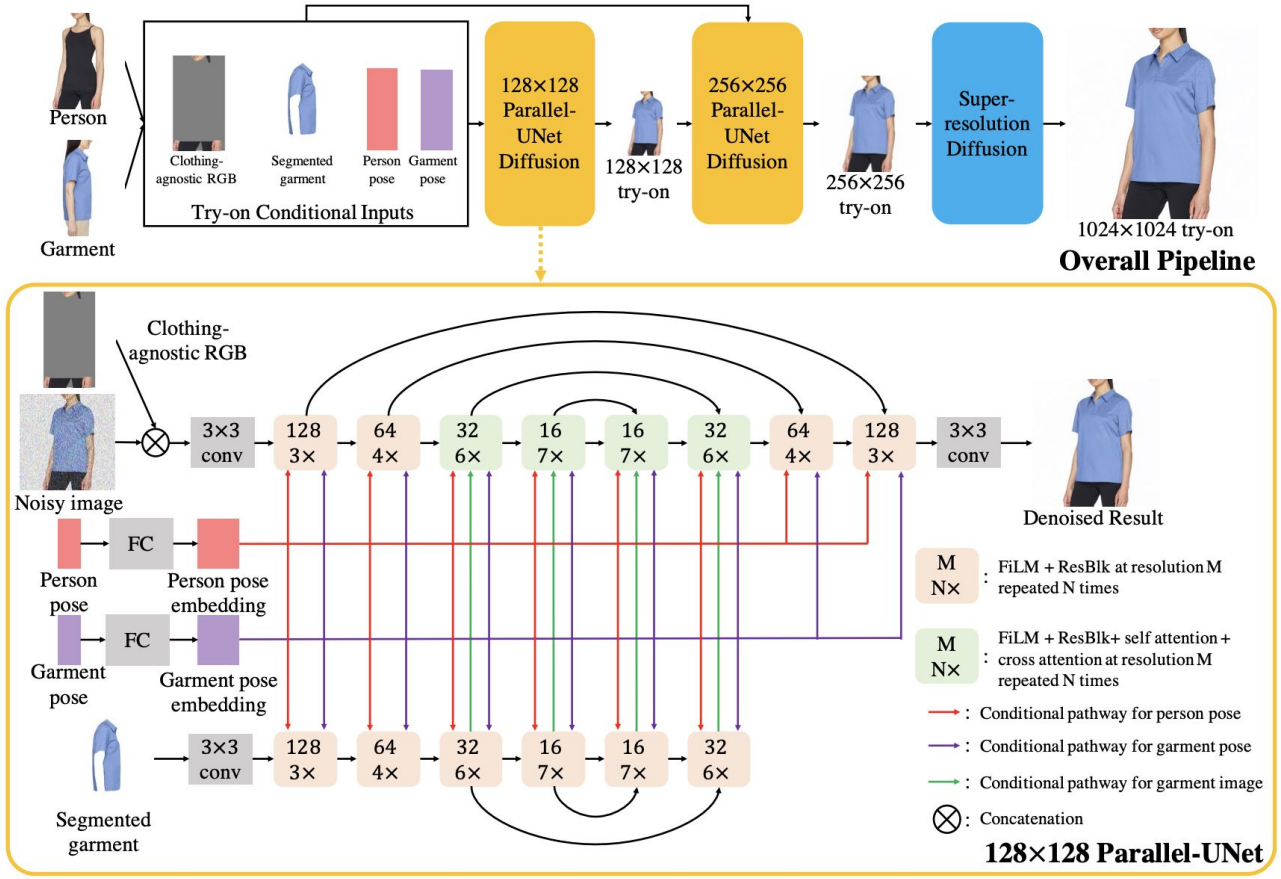
CVPR 2023 の動向・気付き(243/259)

TryOnDiffusion: A Tale of Two UNets

- ポーズや体型変化に対して衣服の詳細を保持した仮装試着が可能なモデル提案
- 2つのUNetを統合したDiffusionベースのアーキテクチャ(右下図)

左下図の上隅がターゲットの人物

左下図の下隅が別の人物の着用画像



CVPR 2023の動向・気付き(244/259)

ファッション分野への応用

- CVPR2023におけるkeywordにfashionが入る論文は3本
  - CVのトレンドにあるようにデータセットの整備がされている.
    - 主要なデータセットはFashionGenやFashionIQ
  - ファッション分野特有のタスク
    - ファッションV+Lデータ特有の記述の詳細さへの対応
    - fine-grained(詳細な)ファッションに関するタスクへの注目
  - 効率性
    - V+L事前学習モデルのパラメタの多さから下流タスクへの対応の効率性に着目
  - 仮装試着
    - これまでのSoTAモデルに比べてターゲットの人物と別の人物の着用画像の合成に違和感がない
    - ターゲット人物と大きく異なる別の人物の着用画像からの仮装試着に対応しており実用性が高い
    - 一方で衣服のセグメンテーションの誤りや仮装試着元の色が仮装試着後に残ってしまう課題あり

CVPR 2023 の動向・気付き(245/259)

医療応用:発表の動向

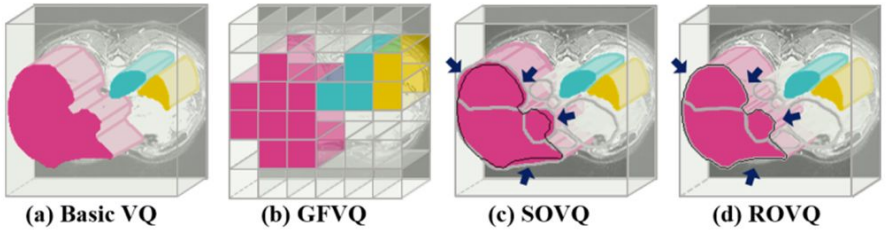
- “medical” を含む論文は12論文
- 半教師ありなどのラベル作業が減少される研究が多く，医療現場での応用が期待される
- few-shot をベクトル量子化したり，拡散モデルを用いたセグメンテーション生成など特徴的な手法

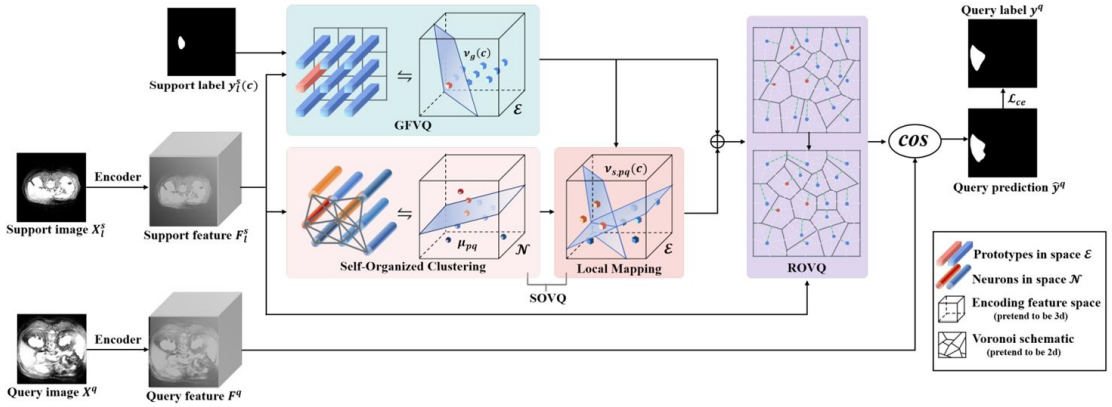
No	タイトル	領域	トピック	GitHub
1	Rethinking Few-Shot Medical Segmentation: A Vector Quantization View	教師なし学習	few-shot segmentation ベクトル量子化	×
2	SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation	セグメンテーション	2.5D	×
3	Orthogonal Annotation Benefits Barely-supervised Medical Image Segmentation	半教師あり学習	二つの直交断面を利用	https://github.com/HengCai-NJU/DeSCO
4	MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation	半教師あり学習	モデル間の認知バイアスを自動検出	×
5	Devil is in the Queries: Advancing Mask Transformers for Real-world Medical Image Segmentation and Out-of-Distribution Localization	半教師あり学習？	細長い構造 分布外領域として場所を特定	×
6	PEFAT: Boosting Semi-supervised Medical Image Classification via Pseudo-loss Estimation and Feature Adversarial Training	半教師あり学習	誤った疑似ラベルの検出	https://github.com/maxwell0027/PEFAT
7	Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation	半教師あり学習	ラベルあり、ラベルなしデータ間の分布をそろえるためのコピペ	https://github.com/DeepMed-Lab-ECNU/BCP
8	Ambiguous Medical Image Segmentation using Diffusion Models	セグメンテーション	拡散モデルを使って複数人専門家の知識を利用	https://github.com/aimansingdha/Ambiguous-Medical-Image-Segmentation-using-Diffusion-Models
9	Geometric Visual Similarity Learning in 3D Medical Image Self-supervised Pre-training	セグメンテーション	自己教師・事前学習　意味的に同じクラスタリングを統合	公開予告あり
10	Directional Connectivity-based Segmentation of Medical Images	セグメンテーション	ピクセル連結性　チャンネル方向の情報利用	https://github.com/Zyun-Y/DconnNet
11	Fair Federated Medical Image Segmentation via Client Contribution Estimation	連合学習	勾配空間とデータ空間でクライアントの貢献度を推定	https://github.com/NVIDIA/NVFlare/tree/dev/research/fed-ce
12	Pseudo-label Guided Contrastive Learning for Semi-supervised Medical Image Segmentation	半教師あり学習	対照学習　疑似ラベルの工夫	https://github.com/hritam-98/PatchCL-MedSeg

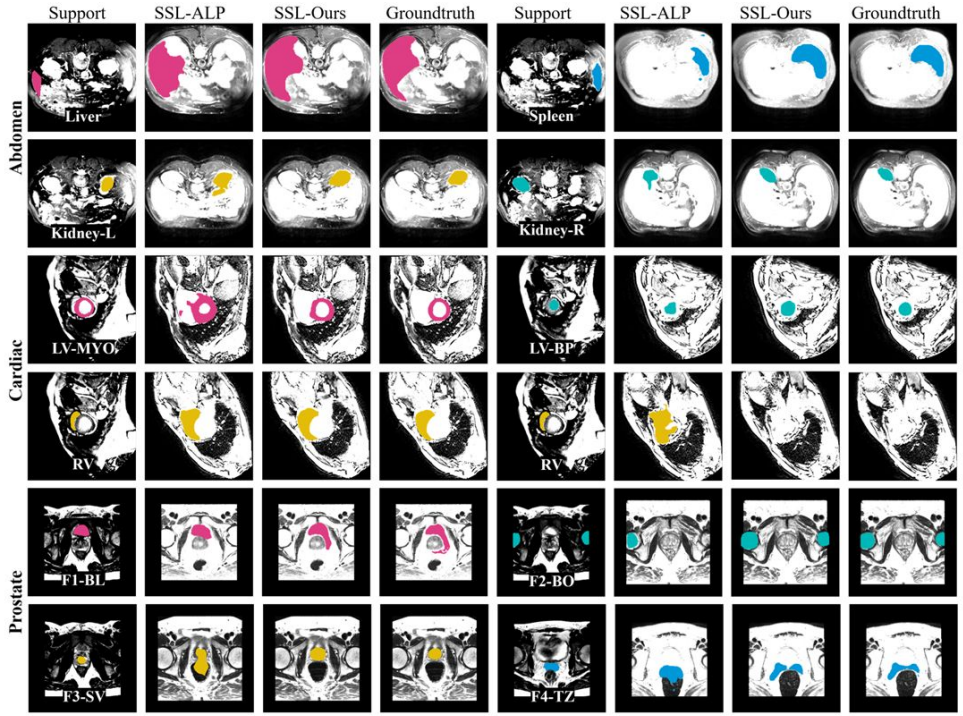
CVPR 2023 の動向・気付き(246/259)

医療応用:Rethinking Few-Shot Medical Segmentation: A Vector Quantization View

- few-shot segmentation :ベクトル量子化機能を追加して性能向上
  - ベクトル量子化ビューによりプロトタイプベクトルを抽出
  - 腹部，心臓，および前立腺MRIデータセットで良好な結果



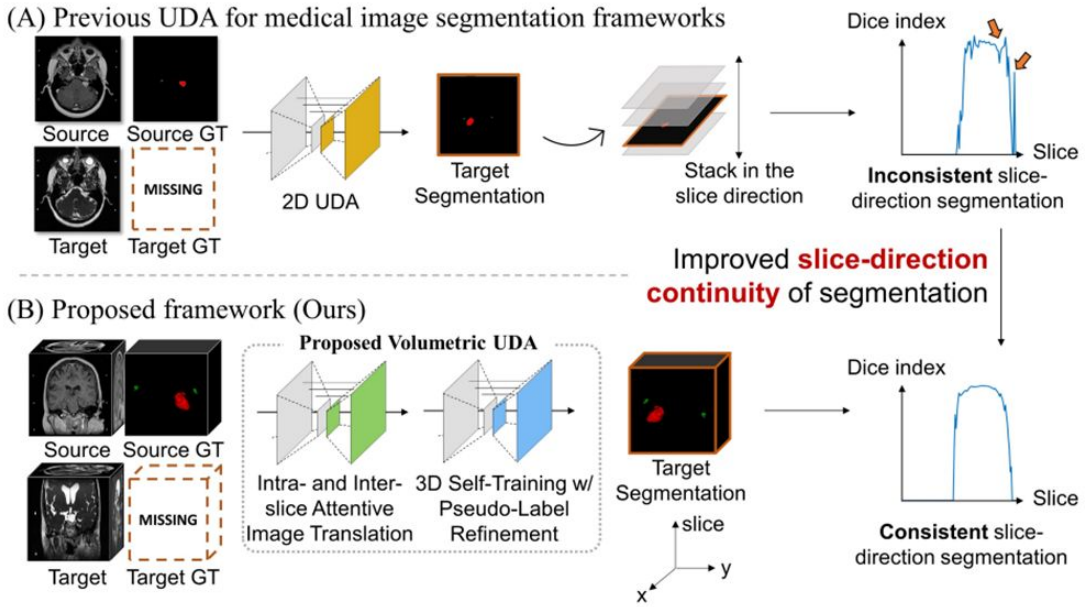


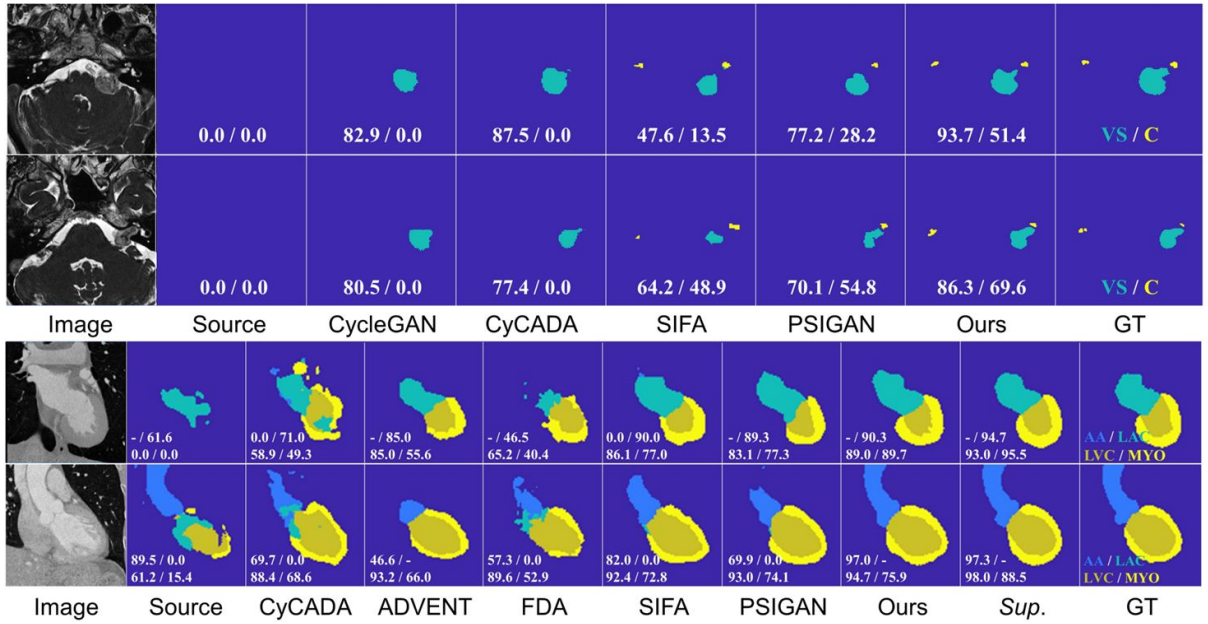


CVPR 2023 の動向・気付き(247/259)

医療応用:SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation

- スライス内およびスライス間の自己注意を介した2.5D変換
  - 合成ボリュームを次の自己学習ステップで使用(2.5D)することで，解剖学的な情報の伝達と滑らかな領域選択を可能にする.
  - 三次元的な滑らかさがポイント

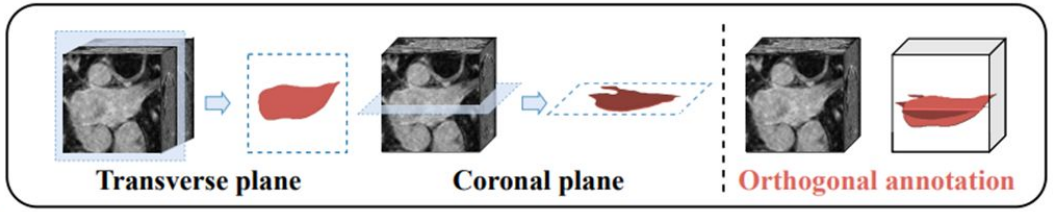


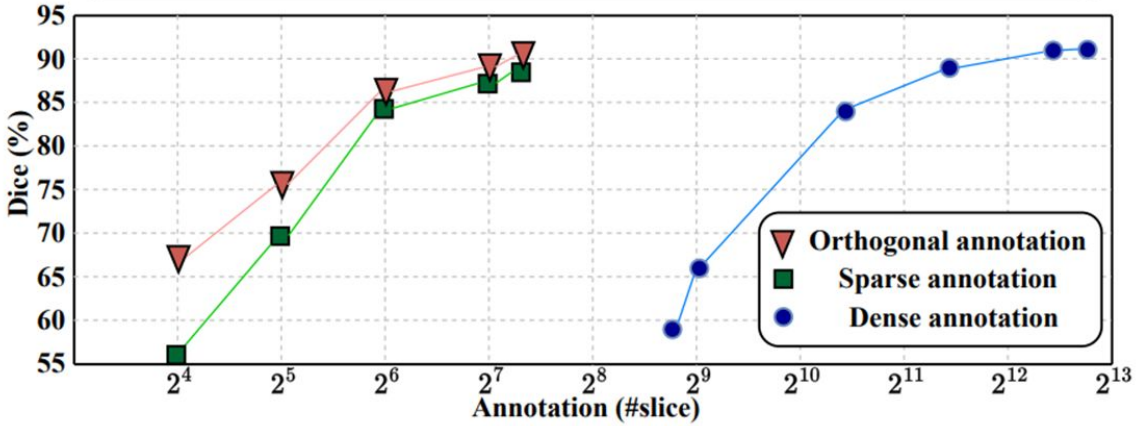


CVPR 2023 の動向・気付き(248/259)

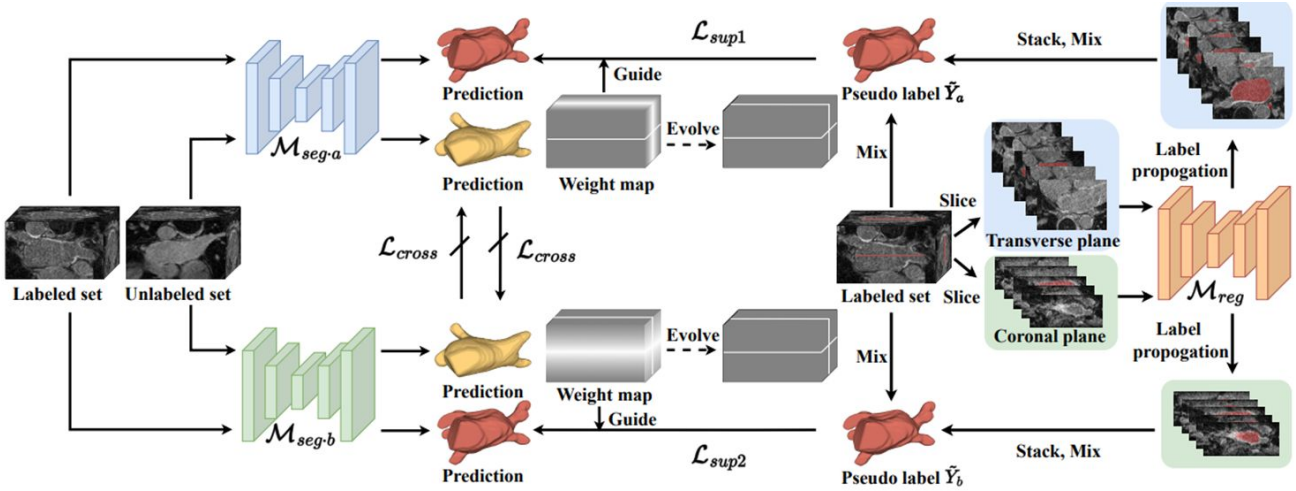
医療応用:Orthogonal Annotation Benefits Barely-supervised Medical Image Segmentation

- 半教師あり学習:直交した二つのスライスでラベル付けを行う直交アノテーションを提案
  - 3D医療画像は，水平・矢状・冠状断などの直交断面が作成可能
  - 直交アノテーションを用いてアノテーションコストを大幅に引き下げた





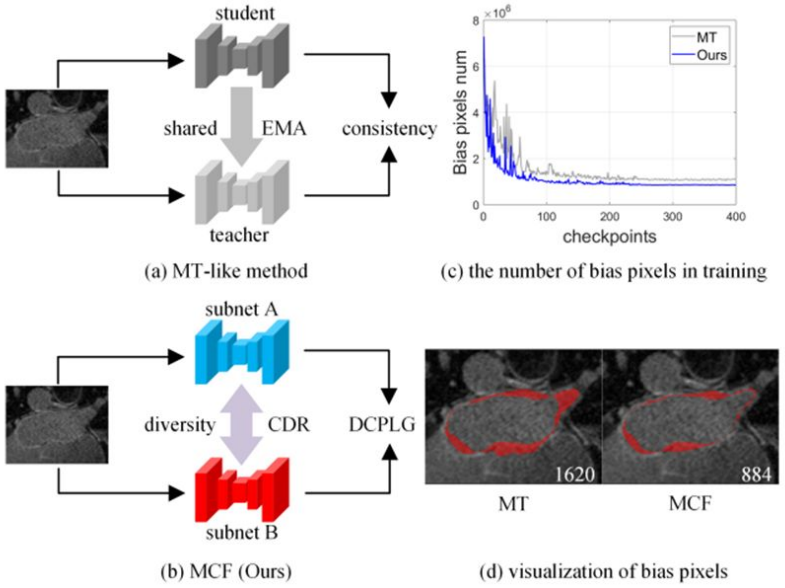
GitHub https://github.com/HengCai-NJU/DeSCO

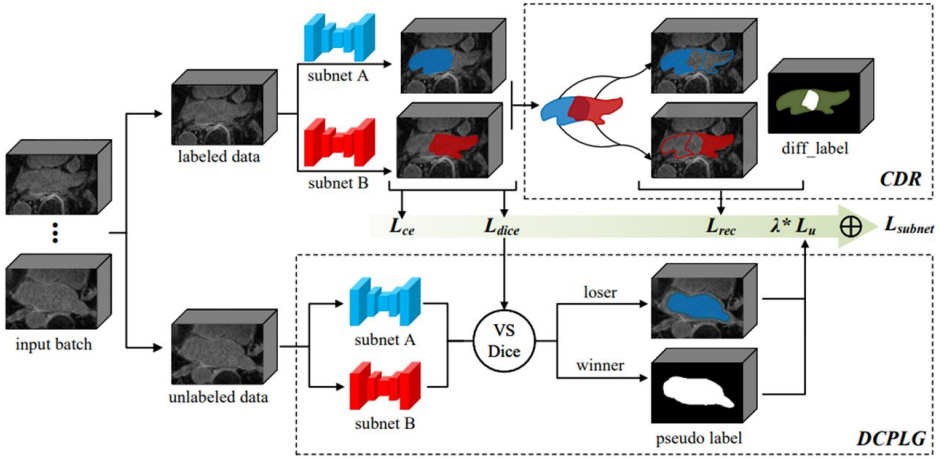


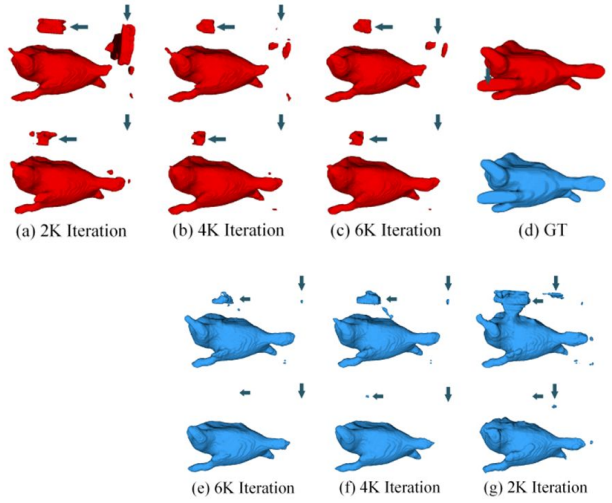
CVPR 2023 の動向・気付き(249/259)

医療応用:MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation

- 半教師あり学習:医療画像セグメンテーションモデルの認知バイアスを減らす手法を提案
  - MCFは二つの異なるサブネットを持ち，認知バイアスを修正するためにサブネット間の不一致を探索
  - 矛盾した予測領域を発見して認知バイアスの補正を行う





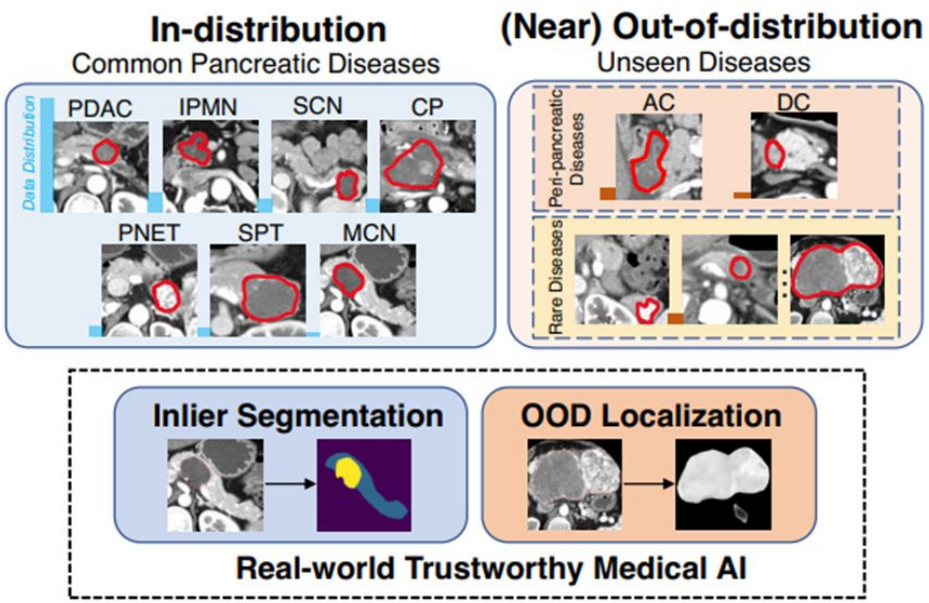


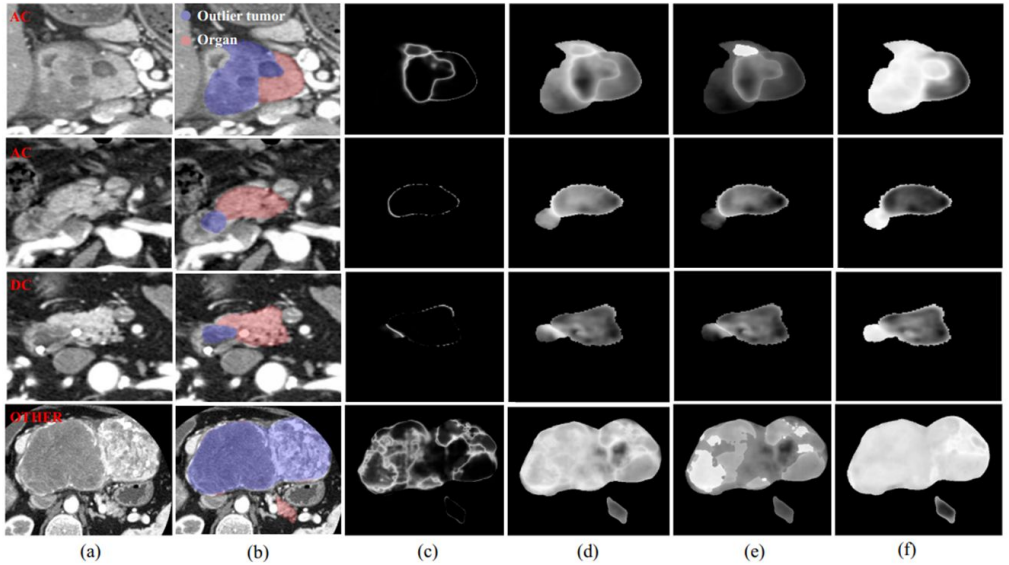
cvpaper.challenge

CVPR 2023 の動向・気付き(250/259)

医療応用:Devil is in the Queries: Advancing Mask Transformers for Real-world Medical Image Segmentation and Out-of-Distribution Localization

- セグメンテーション:膵臓のように長い尾のような構造の尾の先は，一種の分布外であり学習は困難．分布外領域を検出し極材化する手法を提案
  - MaxQuery: ピクセルとクエリの類似性によって分布外領域を検出.
  - 膵臓腫瘍と肝臓腫瘍データセットで良好な結果

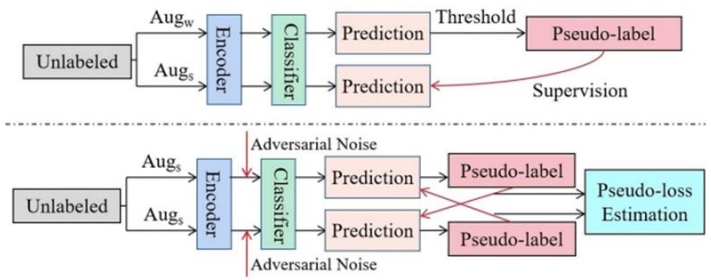




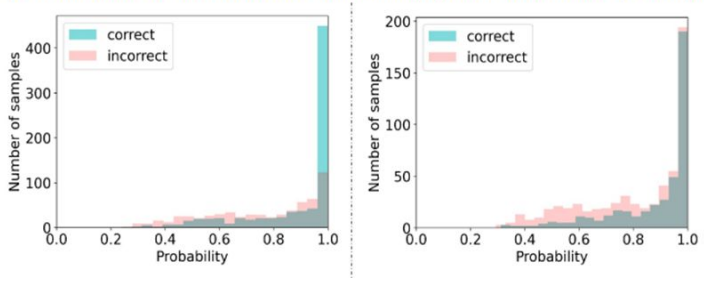
CVPR 2023 の動向・気付き(251/259)

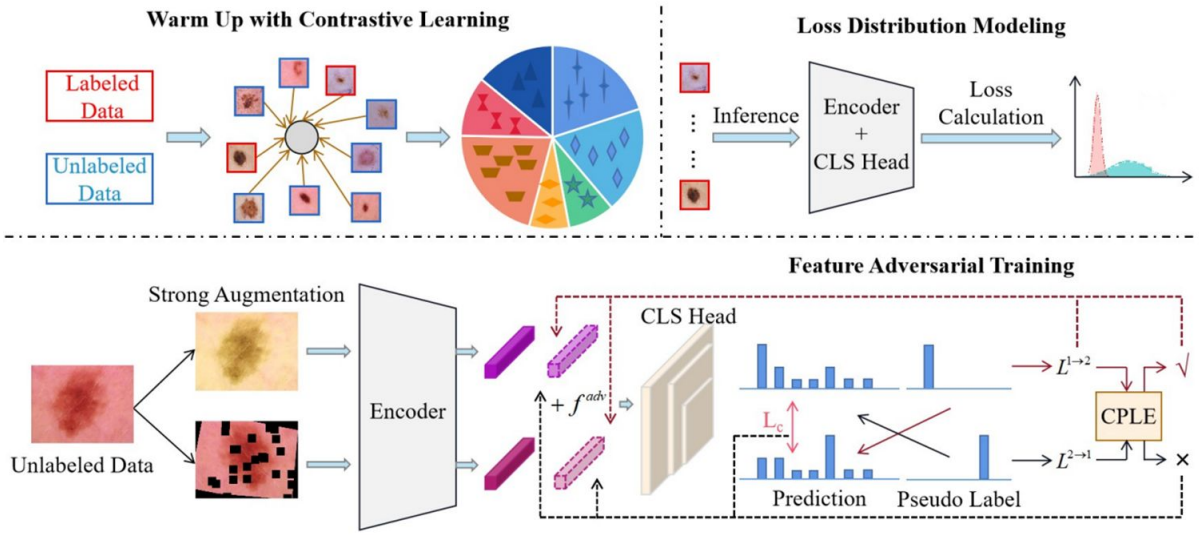
医療応用:PEFAT: Boosting Semi-supervised Medical Image Classification via Pseudo-loss Estimation and Feature Adversarial Training

- 半教師あり学習:誤った疑似ラベルデータに対する対応策
  - 綺麗なデータは損失が小さく，ノイズデータはその逆であるという疑似損失の仮定に基づき，高品質な疑似ラベル集合を抽出
  - 判定境界を平滑化するために特徴レベルで敵対ノイズを注入する新しい法を提案



(a) Illustration of traditional SSL methods (top) and our method (bottom).



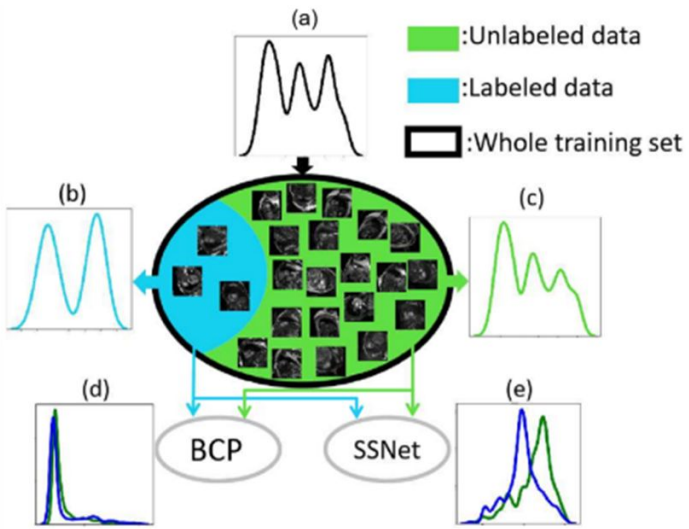


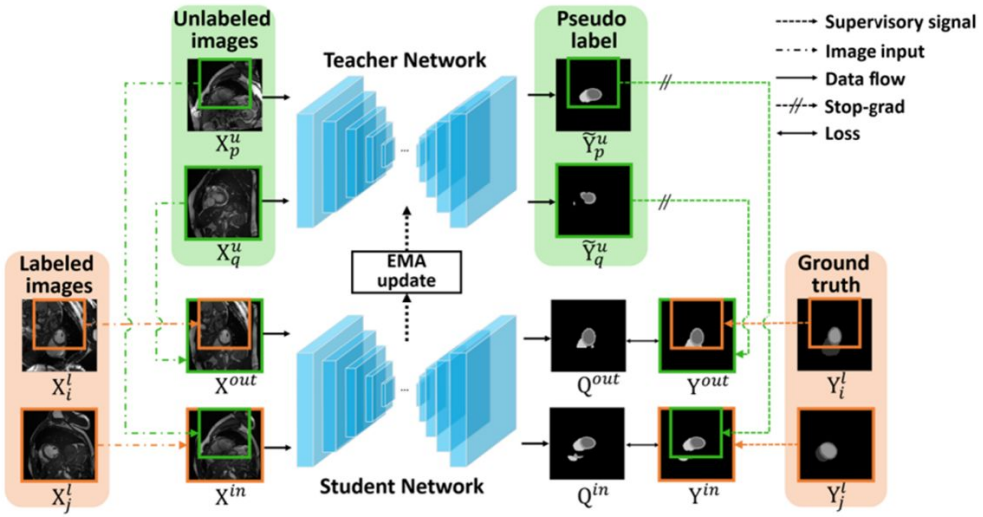
https://github.com/maxwell0027/PEFAT

CVPR 2023 の動向・気付き(252/259)

医療応用:Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation

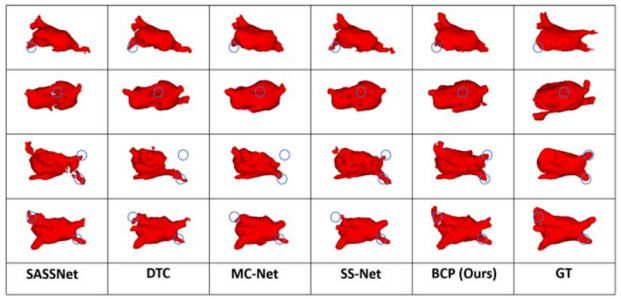
- 半教師あり学習:ラベル付けされたデータとされていないデータの分布にミスマッチがあった場合，ラベル情報が伝わらないという問題解決のために，ラベルありデータとラベル無しデータを双方向にコピー・ペースト
  - 綺麗なデータは損失が小さく，ノイズデータはその逆であるという疑似損失の仮定に基づき，高品質な疑似ラベル集合を抽出
  - 判定境界を平滑化するために敵対ノイズを注入する新しい正則化手法





GitHub

https://github.com/DeepMed-Lab-ECNU/BCP

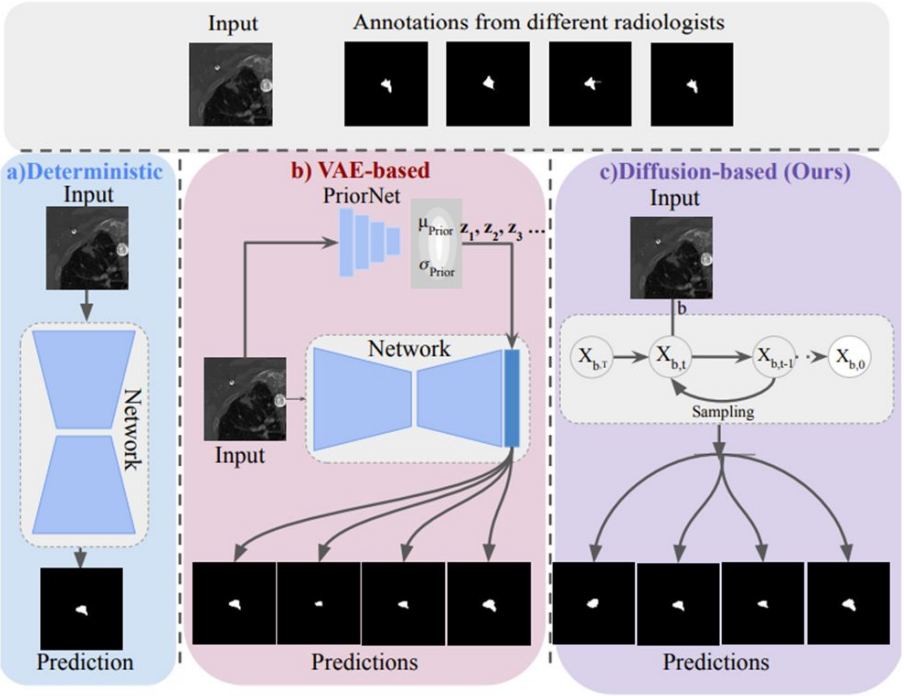


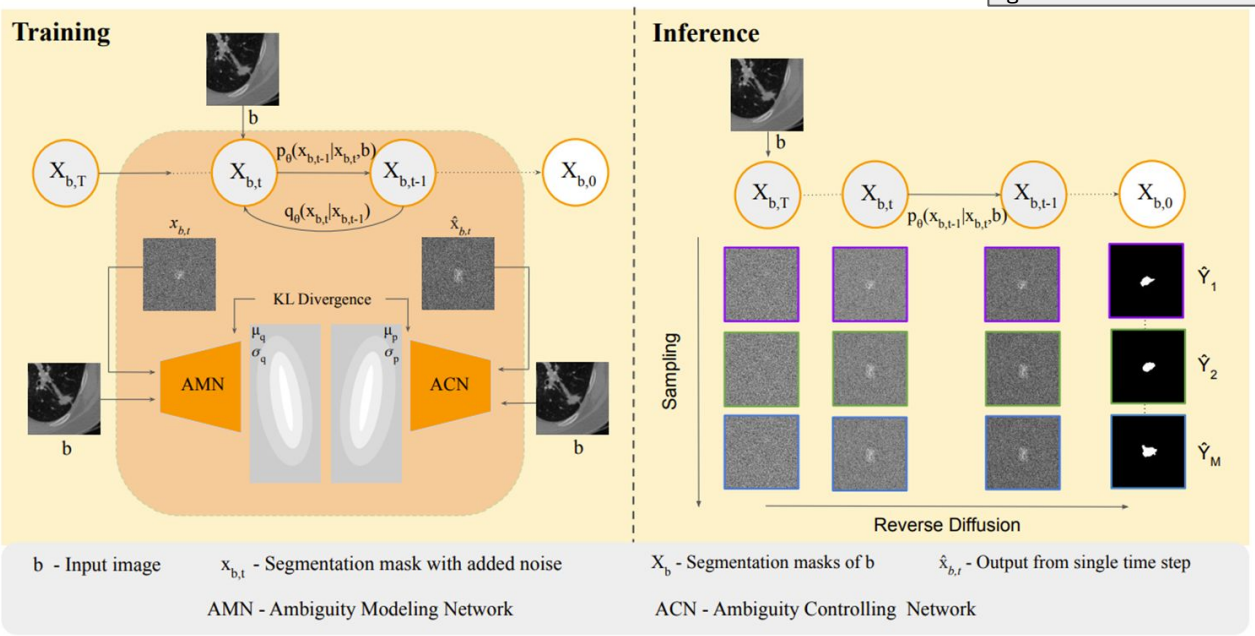
CVPR 2023 の動向・気付き(253/259)

医療応用:Ambiguous Medical Image Segmentation using Diffusion Models

- 拡散モデル:複数人の専門家のセグメンテーション知識を統合する手法
  - CT, 超音波, MRIなどの異なる3つのモダリティーで検証

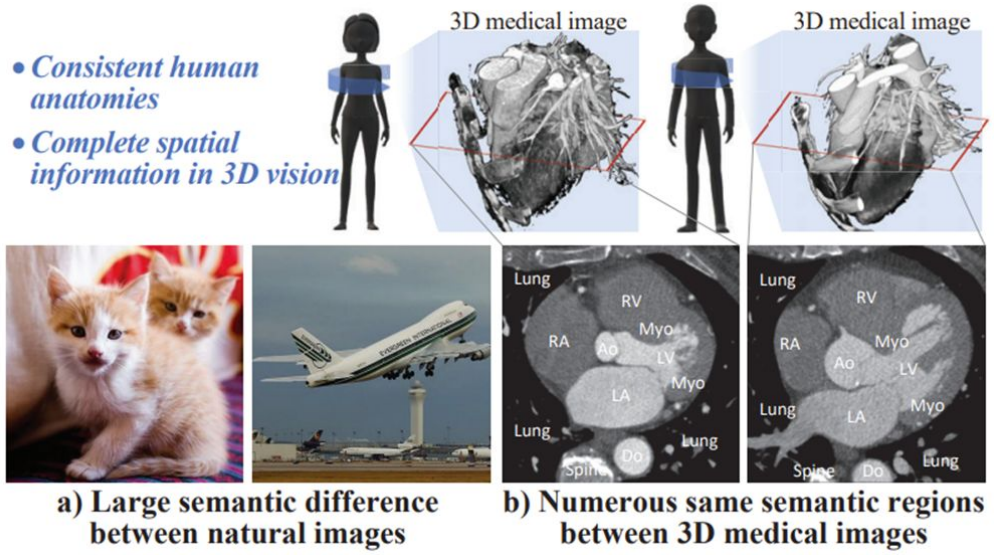
https://github.com/aimansnigdha/Ambiguous-Medical-Image-Segmentation-using-Diffusion-Models

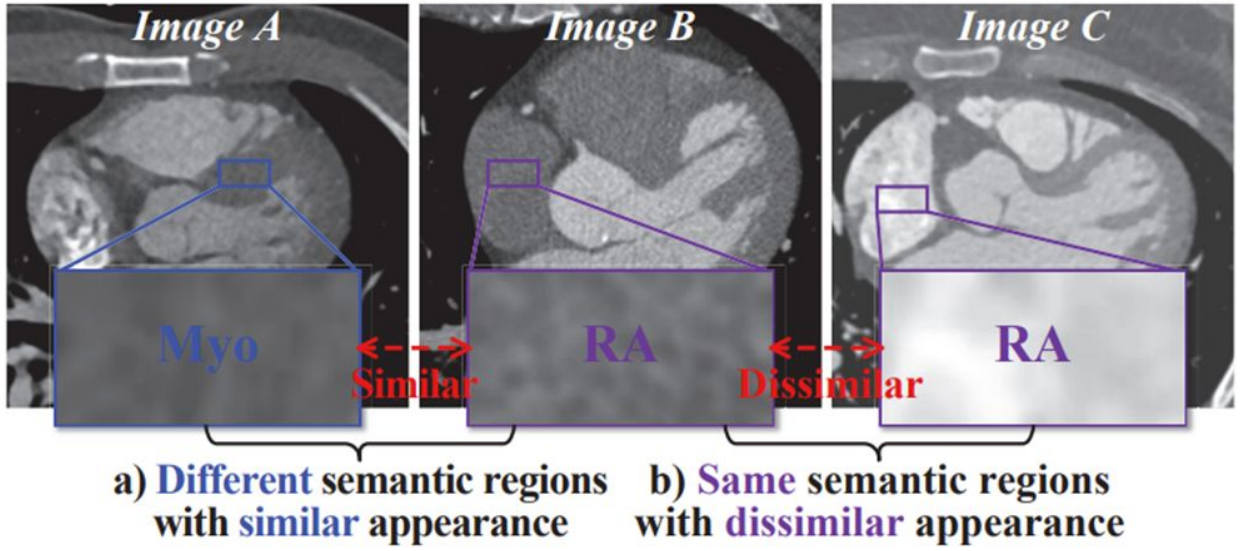




CVPR 2023 の動向・気付き(254/259)

- 医療応用:Geometric Visual Similarity Learning in 3D Medical Image Self-supervised Pre-training
  - 自己教師事前学習:3D医療画像のセグメンテーション
  - 同じ意味特徴をもつクラスタリングに対する画像間の一貫した表現を学習
    - ４つの3Dモダリティーで良好な成績

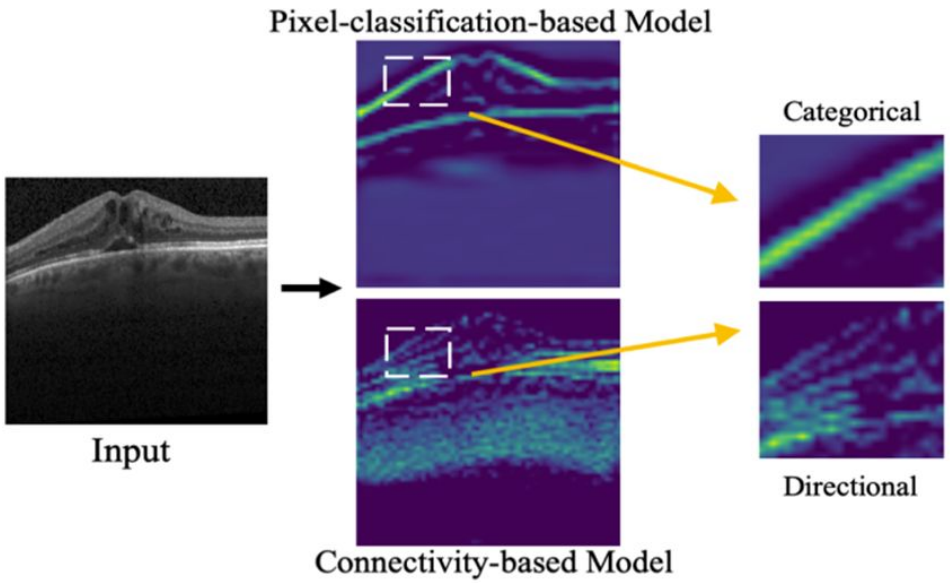


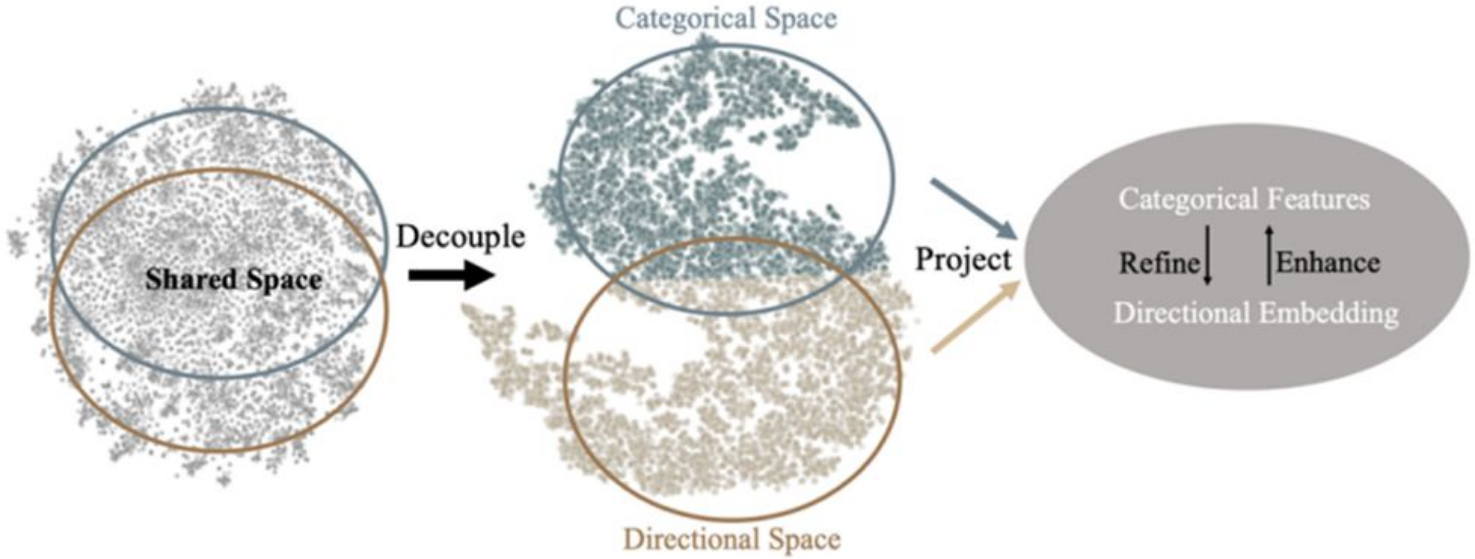


CVPR 2023 の動向・気付き(255/259)

- 医療応用:Directional Connectivity-based Segmentation of Medical Images
  - セグメンテーション:ピクセル連結性をチャンネルごとの方向性を分離することでセグメンテーション

https://github.com/Zyun-Y/DconnNet



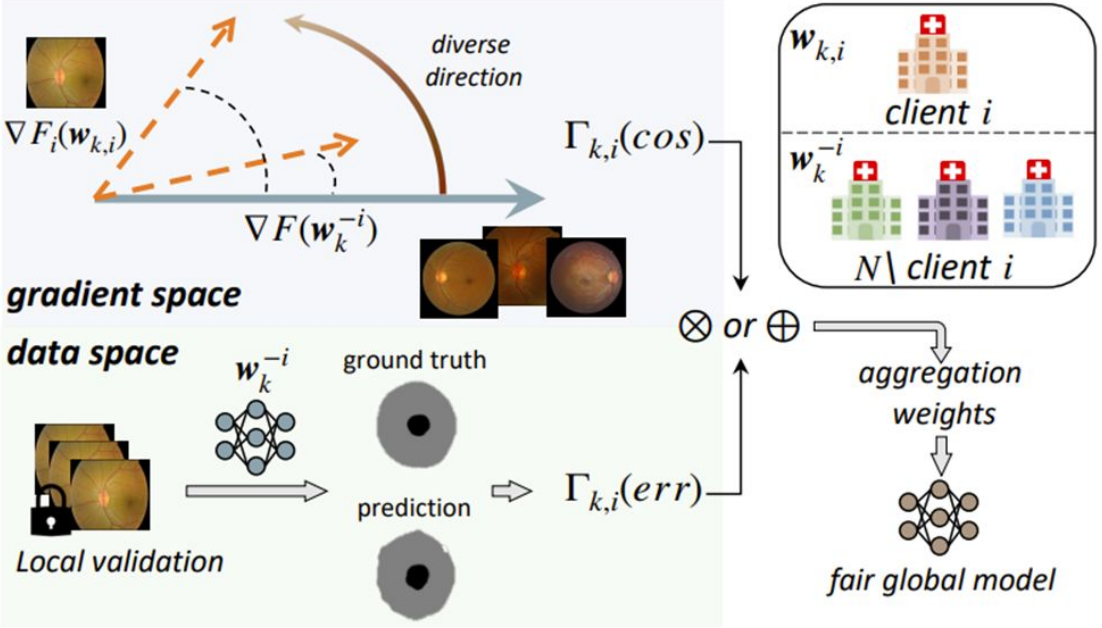


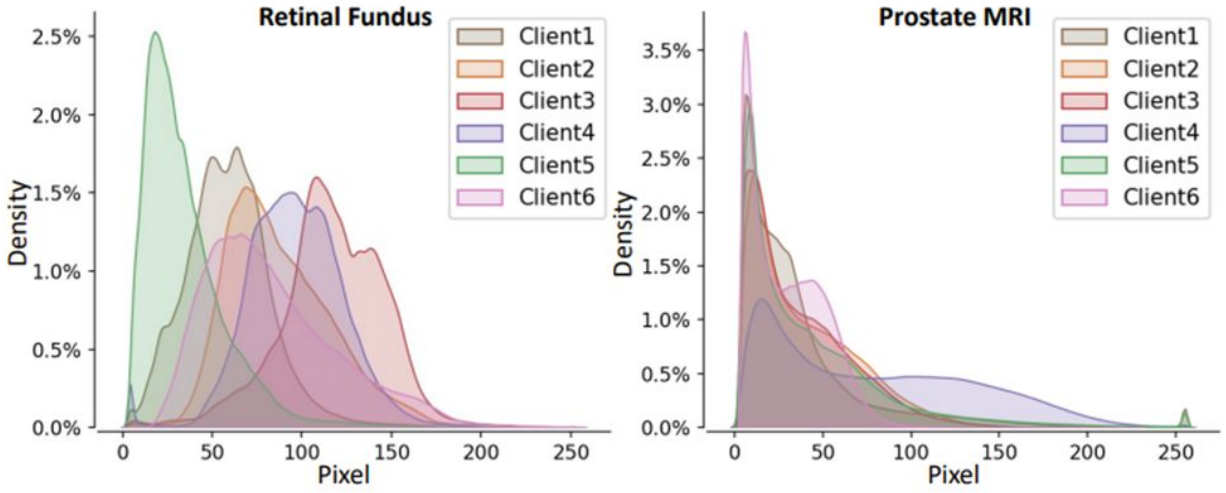
CVPR 2023 の動向・気付き(256/259)

医療応用:Fair Federated Medical Image Segmentation via Client Contribution Estimation

- 連合学習:クライアントの貢献度とパフォーマンスをそれぞれ計測する手法はあったが，同時に評価する手法はなかった
  - 勾配空間とデータ空間でクライアントの貢献度を推定を提案
  - 一つのデータセット(実際の検証)で有用性が示せた

https://github.com/NVIDIA/NVFlare/tree/dev/research/fed-ce



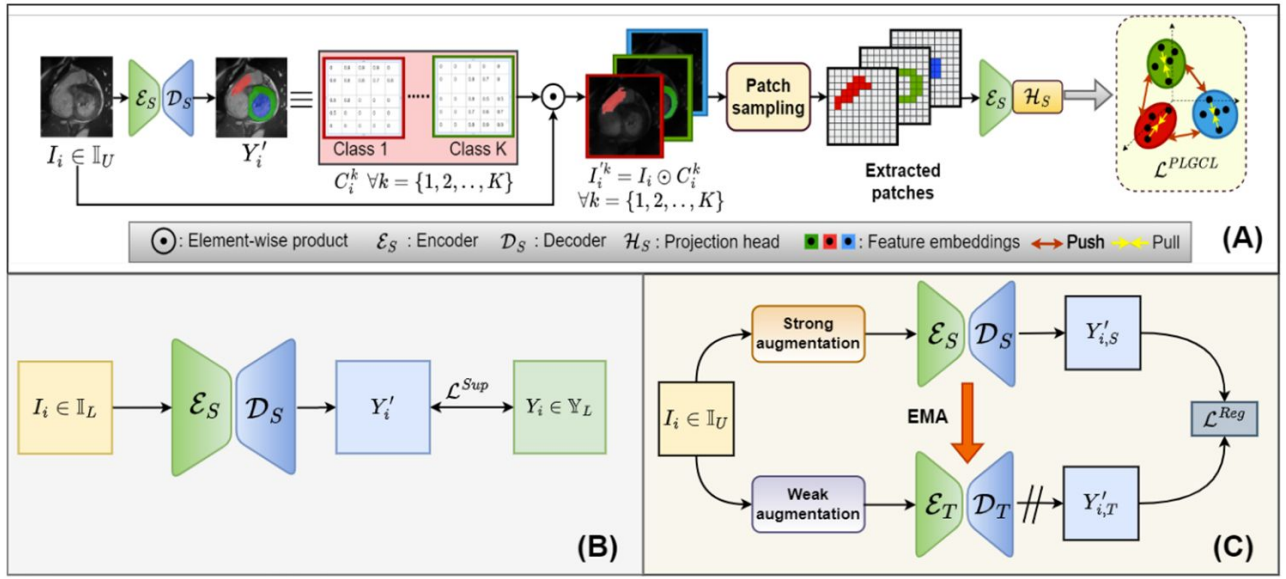


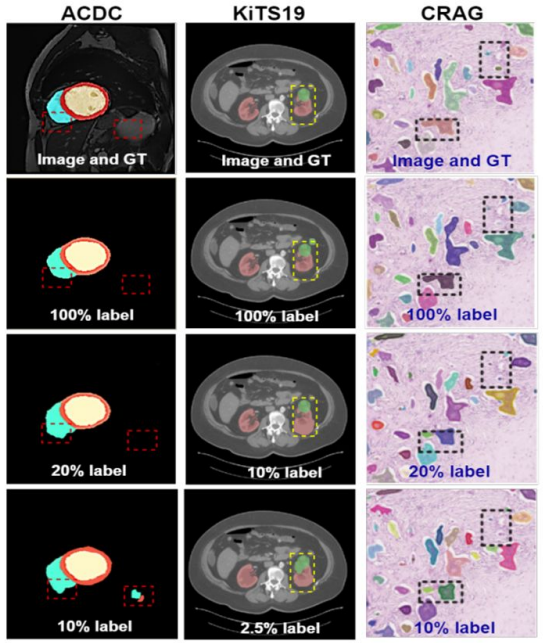
CVPR 2023 の動向・気付き(257/259)

医療応用:Pseudo-label Guided Contrastive Learning for Semi-supervised Medical Image Segmentation

- 半教師あり学習:一貫性正則化と疑似ラベルを用いて対照学習を拡張
- 3つの医療データセットで良好な結果

https://github.com/hritam-98/PatchCL-MedSeg

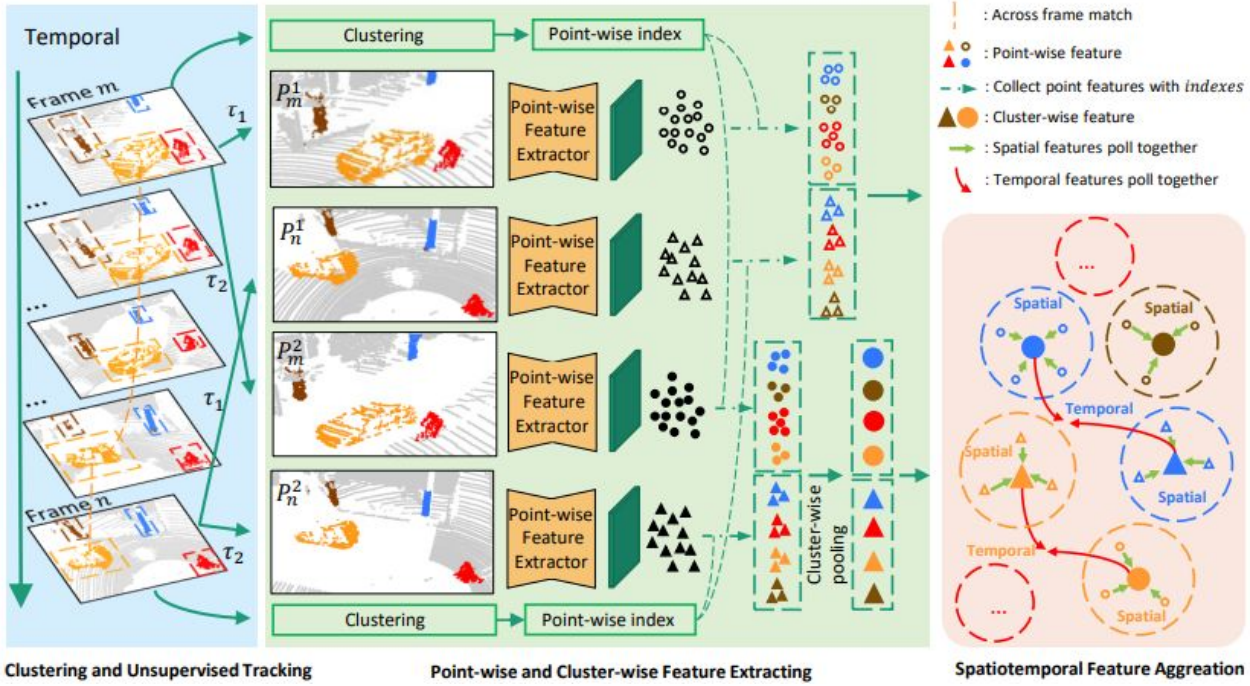




CVPR 2023 の動向・気付き(258/259)

Spatiotemporal Self-supervised Learning for Point Clouds in the Wild

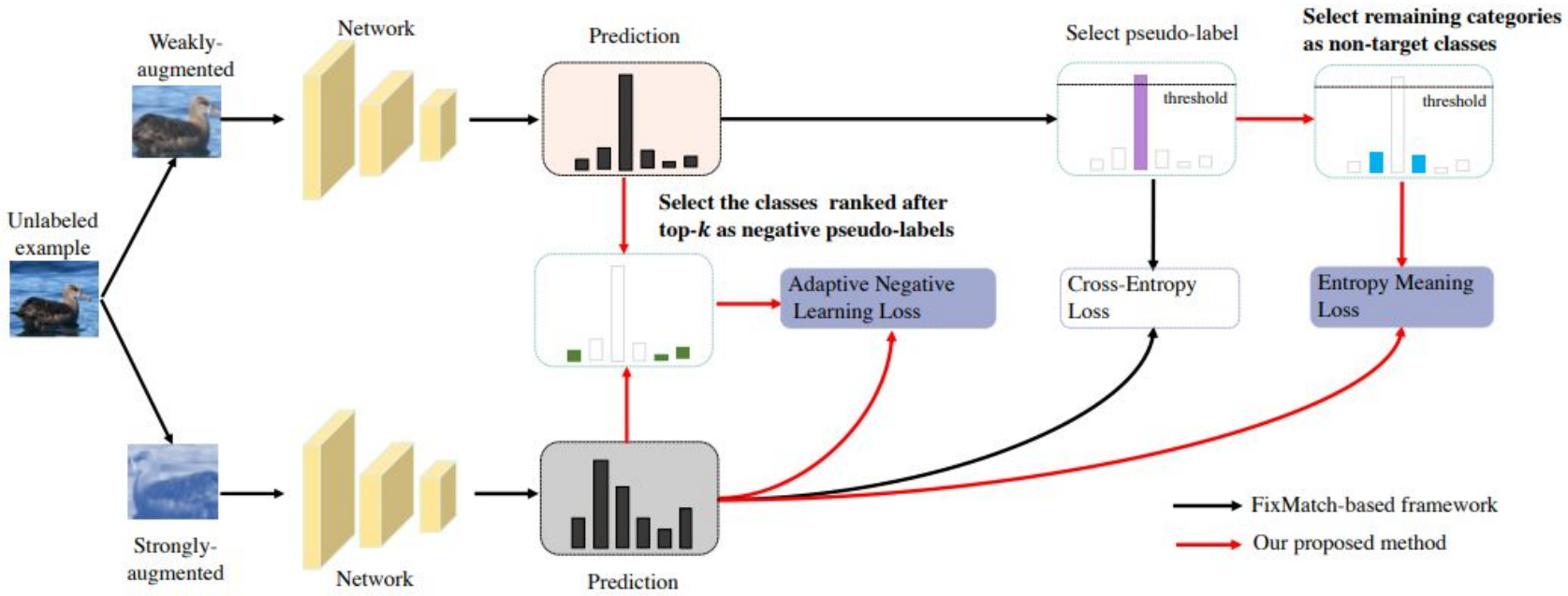
- LIDARデータが持つ空間領域に加え，時間領域も活用したSSL戦略を提案
  - 2つの大規模なLiDARデータセットで自己教師あり学習を実行
    - 空間情報を集約してオブジェクトを区別するポイント間学習
    - 時間的対応を利用する教師なしオブジェクト追跡に基づくクラスタ間学習
  - 得られたモデルを他の点群セグメンテーションベンチマークに転移
  - 最ち



CVPR 2023 の動向・気付き(259/259)

Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data

- 大規模ラベル無しデータセットをより良く活用するためのフレームワークFullMutchを開発
  - 従来のSSLフレームワークであるFixMatchでは大規模ラベル無しデータセットにおいて学習に使用されない画像が多く発生
  - 新手法であるFullMatchではより多くのラベル無しデータセットを効果的に学習で使用できるようにすることでモデルの性能向上を達成



今後の展望

- 今後，CV分野はどうなる？
- 我々としてはどうすれば良いか？

今後の展望(1/4)

大規模事前学習 + ゼロショット認識/小規模追加学習*

- Large-scale Pre-training / Vision-Language Pre-trainingの効果もあり，ゼロショット認識ができるようになってきた．今後，より少数のデータ・少量リソースの学習でチューニングできるか，についてもより多くの研究が行われる可能性が高い

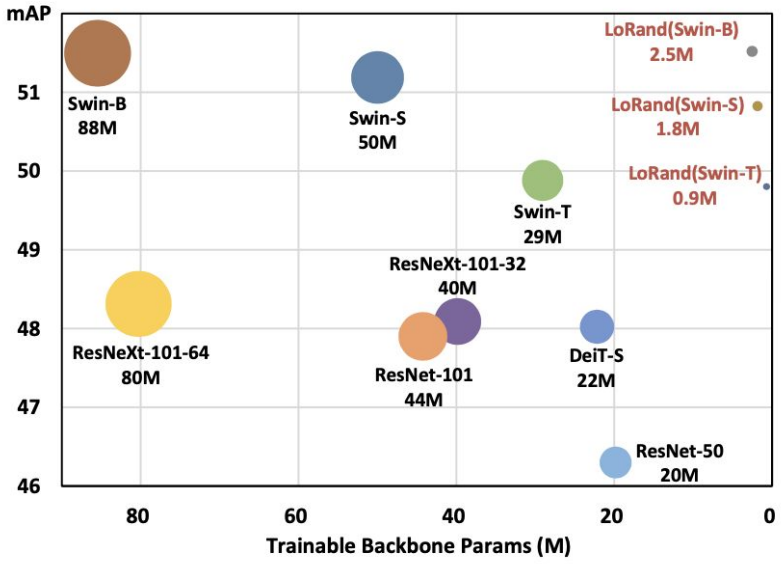
*あくまでも主要な流れということで，基盤モデル関連以外の研究が採択されないというわけではありません

今後の展望(2/4)

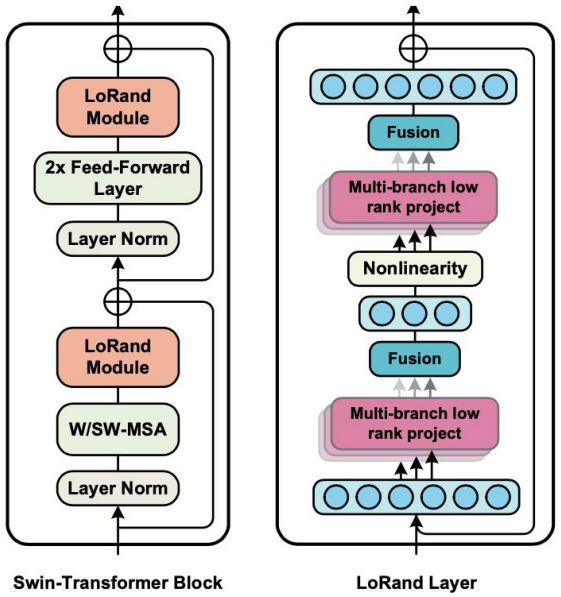
大規模事前学習 + ゼロショット認識/小規模追加学習*

- 1% VS 100% → 事前学習モデルを用いる前提. FT時に数%のみのチューニングで高い水準と同等レベルの認識ができる

基盤モデル関連以外の研究が採択されないわけではありません



大規模事前学習モデルにおける冗長性を見極め，少数パラメータのみの学習で高性能な画像認識(左図は物体検出の例)を実現. 提案手法のLoRand(Swin-B)は2.5Mのみの追加学習でSwin-Bと同等の精度まで到達.



提案手法のLoRand(Low-Rank Adapter)はSwinTransformerのブロック(図中左)に2箇所LoRand層(図中右)を追加. 複数の低ランク行列の乗算で全結合層のパラメータ数を削減.

これらの他にもNLPにて提唱されているAdapter Tuningに着目. 事前学習のbackboneのパラメータは固定，ごく少量のパラメータ追加と学習を行うことでdecoder/headを同時に最適化できる.

D. Yin et al. “1% VS 100%: Parameter-Efficient Low Rank Adapter for Dense Predictions”, in CVPR 2023.

今後の展望(3/4)

(最初の年)新規問題設定→

(次の年)精度向上→

(さらに次の年)計算コスト問題解消→

の3点セットは黄金ルートで，以降繰り返す

- 例としてCLIPは2021年に提唱
- 2022年はCLIPを応用して各トピックで精度向上を確認
- 2023年はCLIPにより実施されるVision-Language Pre-trainingの計算コストを削減

2023後半～2024に大きな流れを作る問題設定が登場するかも(常に登場している感はある)

今後の展望(4/4)

今回CVPR 2023は何か大きなことが起こる前夜的な立ち位置か

- CVPR 2023 の投稿直後にChatGPT(GPT-3.5/4)などLLMの衝撃
- LLM(e.g., GPT, LLaMA, Bard), VLM(e.g., ViT, CLIP, SAM)など部品は揃ってきて，事前学習・追加学習の整備が進行
- ICCV 2023(3月前半投稿)ではある程度出てきそう，CVPR 2024(2023/11月投稿予定)で爆発するか

(基盤モデル関連に限らず)Brave Newな研究をしよう！

トレンドを創るために挑戦し続けよう！！

来年のCVPR2024は発表者側で参加しよう！！！