cvpaper.challenge

ICCV 2023 速報

片岡裕雄, Qiu Yue, 岩田健司, 中村凌, 篠田理沙, 松澤郁哉, 中原龍一,
柴田優斗, 千葉直也, 井口悠司, 大島遼祐, 山田亮佑, 堀田大地, 速水亮, 荒井
智貴, 森江梨花, 舘野将寿, 松尾雄斗, 中條亨一, 牧原昂志,
品川政太朗, 上田樹, 佐々木馨, 武田, 佐藤和仁, 江藤謙, 武藤良

ICCV 2023 の動向・気付き

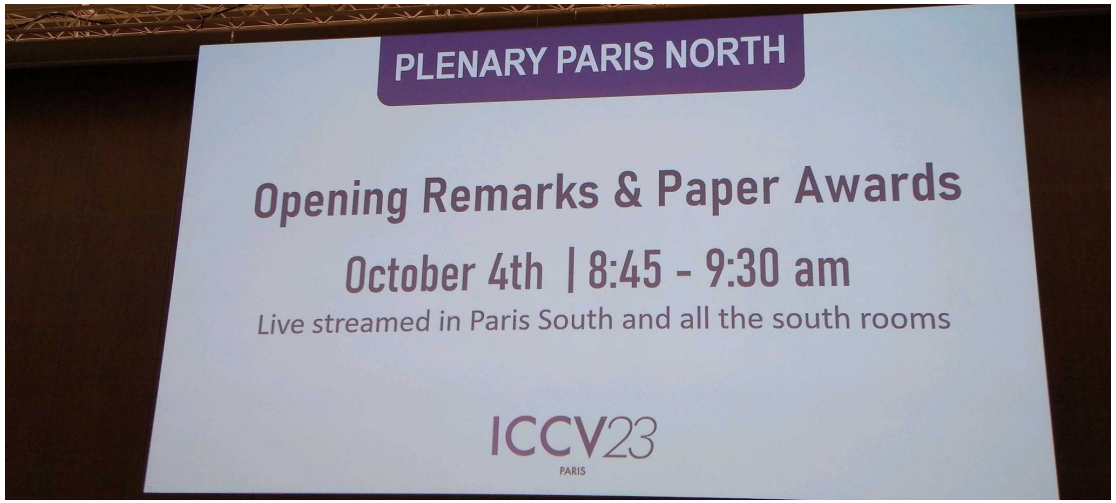
- 今回どんな研究が流行っていた？
- 海外の研究者は何をしている？
- 「動向」や「気付き」をまとめました

– https://openaccess.thecvf.com/ICCV2023
– https://openaccess.thecvf.com/ICCV2023 workshops/menu

ICCV 2023 の動向・気付き(1/165)

ICCV 2023

- プログラムの概観
  - Day 1-2 Workshop
  - Day 3-5 Main conference
  - Day 3 Reception
  - Day 4 PAMI Meeting(国際会議の取り決め)



	Day 1 Monday 2nd	Day 2 Tuesday 3rd	Day 3 Wednesday 4th	Day 4 Thursday 5th	Day 5 Friday 6th
08:00am	Welcome coffee 8:00 am-8:30am				
08:30 am	Workshops & Tutorials	Workshops & Tutorials	Opening remarks & awards 8:45am-9:30am	Welcome coffee 8:30am-9:00am	
09:00 am				Orals THU-AM 1A / 1B	Orals FRI-AM 1A / 1B
09:30 am			Orals WED-AM 1A / 1B	Orals THU-AM 2A / 2B	Orals FRI-AM 2A / 2B
10:00 am			Orals WED-AM 2A / 2B	Orals THU-AM 3A / 3B	Orals FRI-AM 3A / 3B
10:30 am			Demos-WED / Posters WED-AM	Demos-THU / Posters THU-AM	Demos-FRI / Posters FRI-AM
11:00 am					
11:30 am	Workshops & Tutorials	Workshops & Tutorials			
12:00 pm					
12:30 pm	Lunch 12:30pm-1:30 pm	Lunch 12:30pm-1:30 pm	Lunch 12:30pm-1:30 pm	Lunch 12:30pm-1:30 pm	Lunch 12:30pm-1:30 pm
01:00 pm					
01:30 pm	Workshops & Tutorials	Workshops & Tutorials	Orals WED-PM 3A / 3B	Keynote 1	Keynote 2
02:00 pm			Orals WED-PM 4A / 4B		
02:30 pm			Posters WED-PM	Posters THU-PM	Posters FRI-PM
03:00 pm					
03:30 pm					
04:00 pm	Workshops & Tutorials	Workshops & Tutorials	Orals WED-PM 5A / 5B	Orals THU-PM 4A / 4B	Orals FRI-PM 4A / 4B
04:30 pm					
05:00 pm			Orals WED-PM 6A / 6B	Orals THU-PM 5A / 5B	Orals FRI-PM 5A / 5B
05:30 pm			Orals WED-PM 7A / 7B	Orals THU-PM 6A / 6B	Orals FRI-PM 6A / 6B
06:00 pm				PAMI Meeting	
06:30 pm			ICCV23 Reception at the convention center		
07:00 pm					
07:30 pm					
08:00 pm					

ICCV 2023 の動向・気付き(2/165)

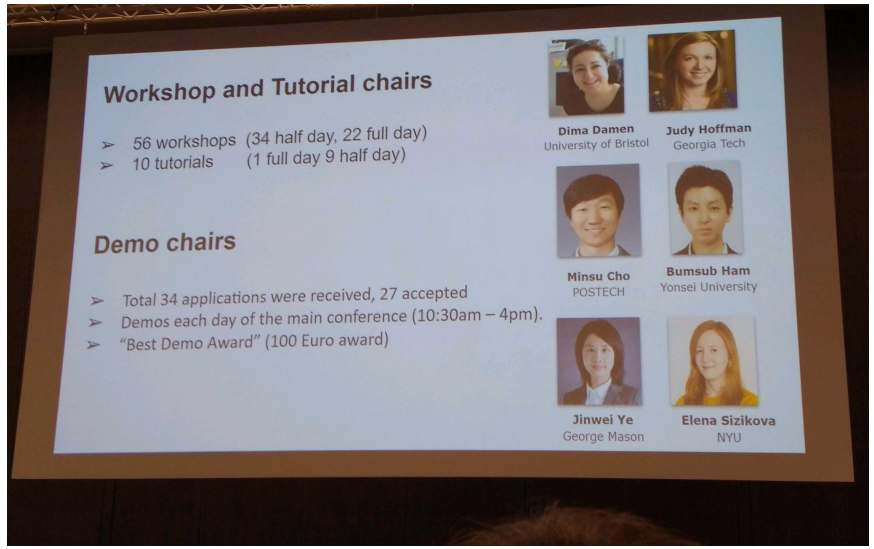
Opening Remarks & Paper Awards

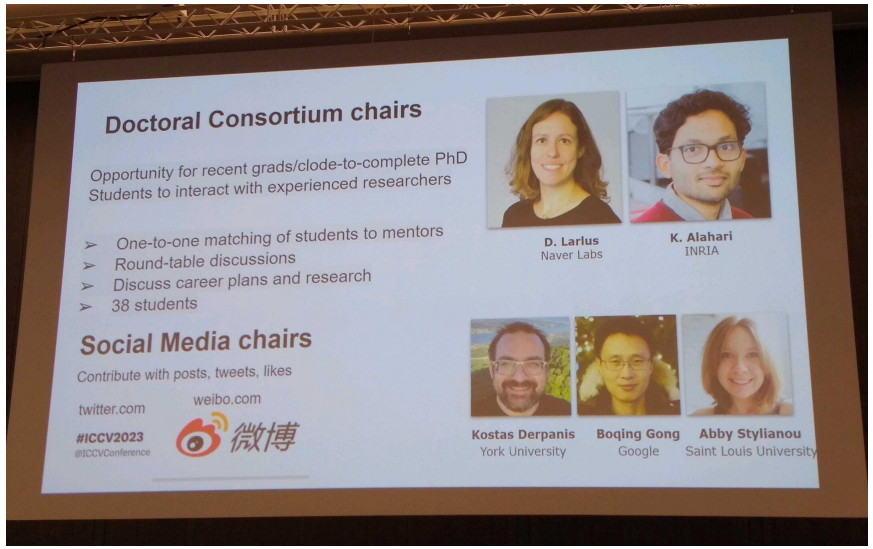
- General Chairs / Program Chairs
  - General Chairs は EU勢が多い
  - GCは全体調整役，PCは査読やプログラムを管理
  - 開催にとても貢献している(この人たちがいないと開催できない！)

ICCV 2023 の動向・気付き(3/165)

Opening Remarks & Paper Awards

- Workshop/Tutorial Chairs: 主にDay 1-2のWS/Tの運営
  - Proposal submission/reviewもしている
- Demo Chairs
  - 同様にDemoのsubmission/reviewをする上に，Best Demo Awardの審査もやる
- Doctoral Consortium Chairs
  - 博士課程学生の交流・議論・支援を検討する
- Social Media Chairs
  - SNS担当，X/Twitterやってるとよく見かける





ICCV 2023 の動向・気付き（4/165）

Opening Remarks & Paper Awards

- Local Organization
  - 現地担当，会場とのやりとりをしてくれた
  - 今回は，Registrationも担当？
  - 160名の現地学生がボランティア（ありがとう！）

ICCV 2023 の動向・気付き（5/165）

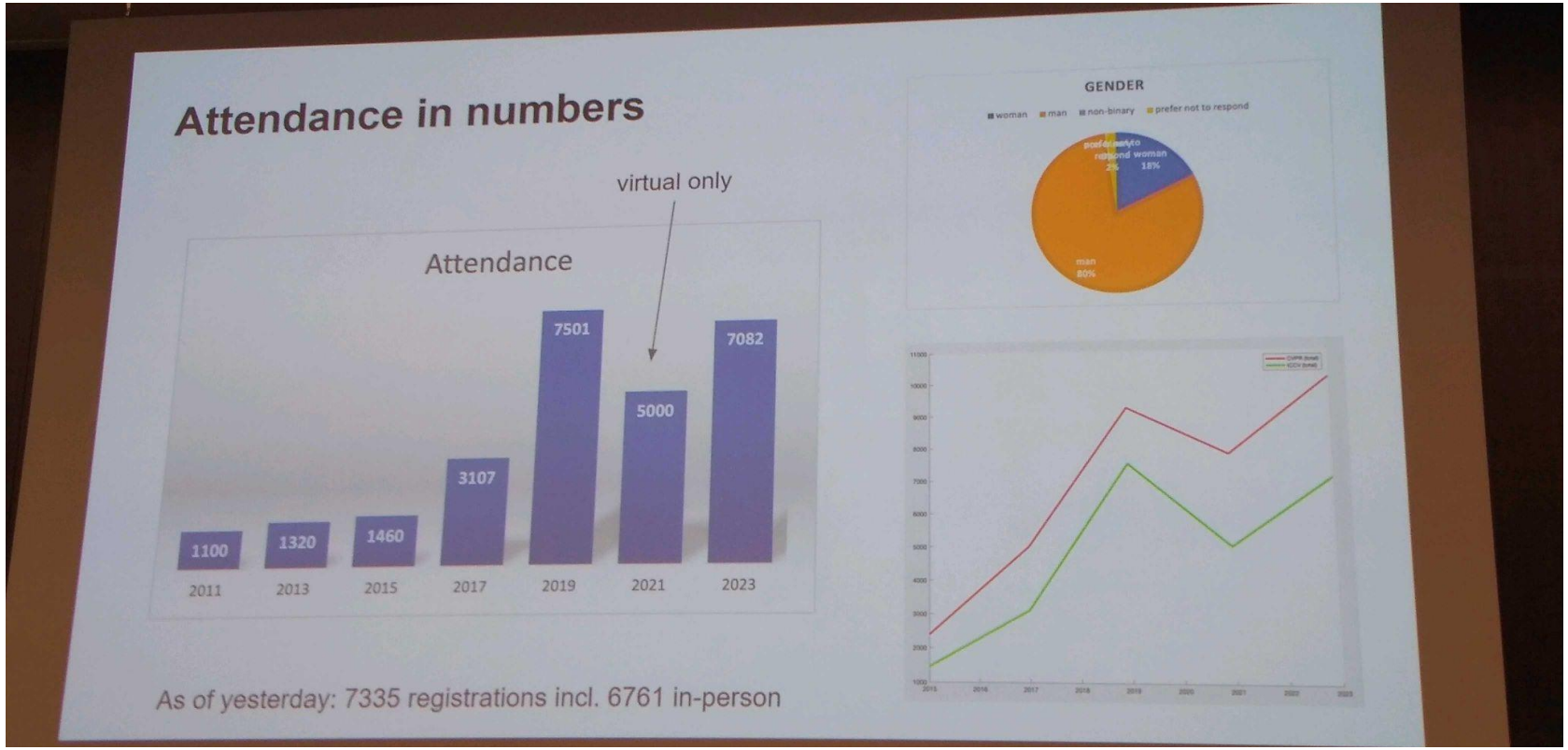
Opening Remarks & Paper Awards

- Diversity Chairs
  - ダイバーシティ確保のためのチェア
  - 今回は（今回も？）高校生イベントを開催と書いてある

ICCV 2023 の動向・気付き(6/165)

Opening Remarks & Paper Awards

- ICCV 参加者数の推移
  - 2017/2019あたりで激増
  - COVIDの影響で減少(2019: 7,501 -> 2021: 5,000)
  - ピーク時を超えなかったが戻っている(2019: 7,501 -> 2023: 7082)



ICCV 2023 の動向・気付き（7/165）

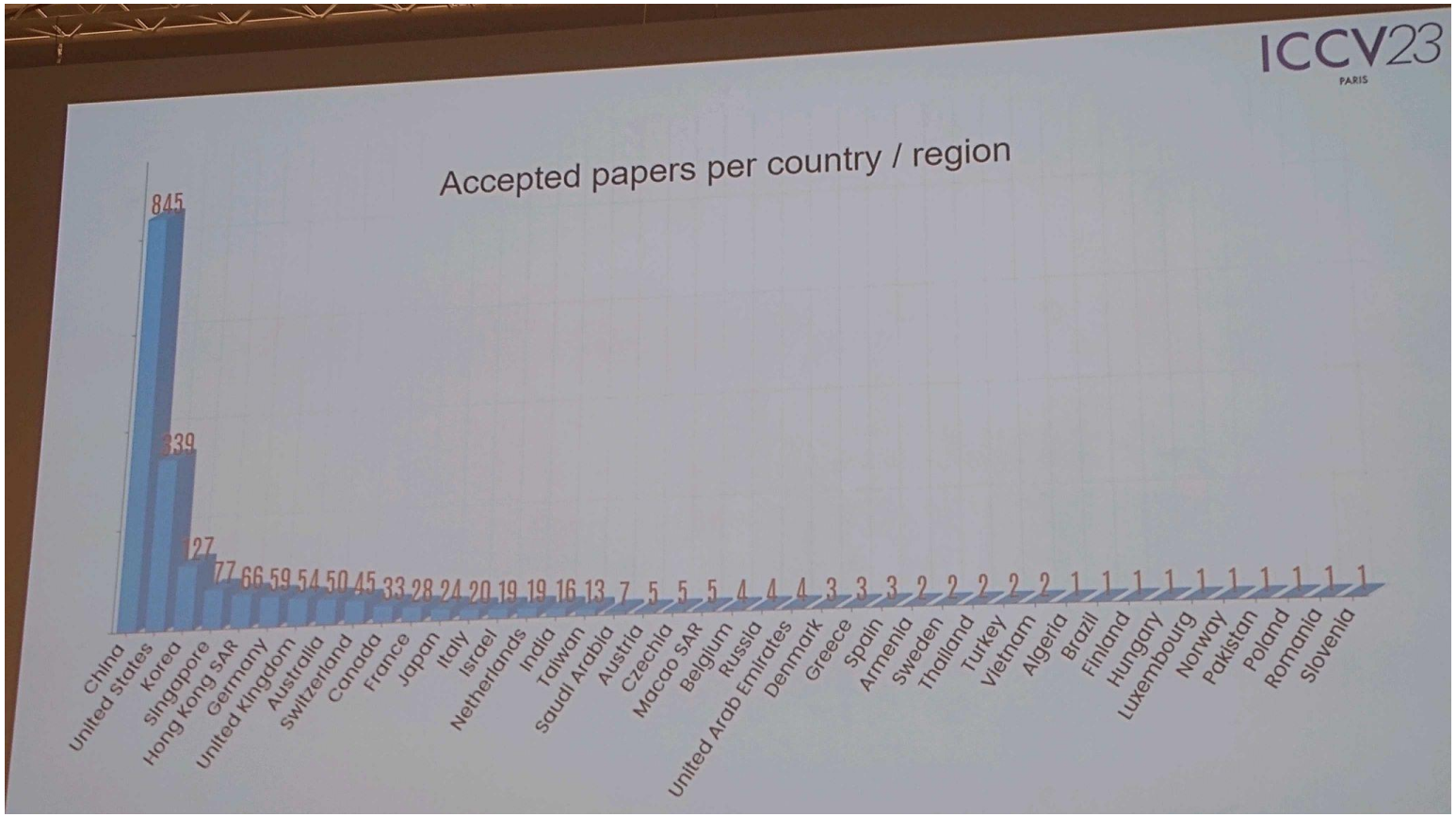
Opening Remarks & Paper Awards

- Sponsors & Exhibit
  - やはりピーク時からは少し減った？
  - 参加者と同期して，2019年くらいが最も数が多かった印象

ICCV 2023 の動向・気付き（8/165）

Opening Remarks & Paper Awards

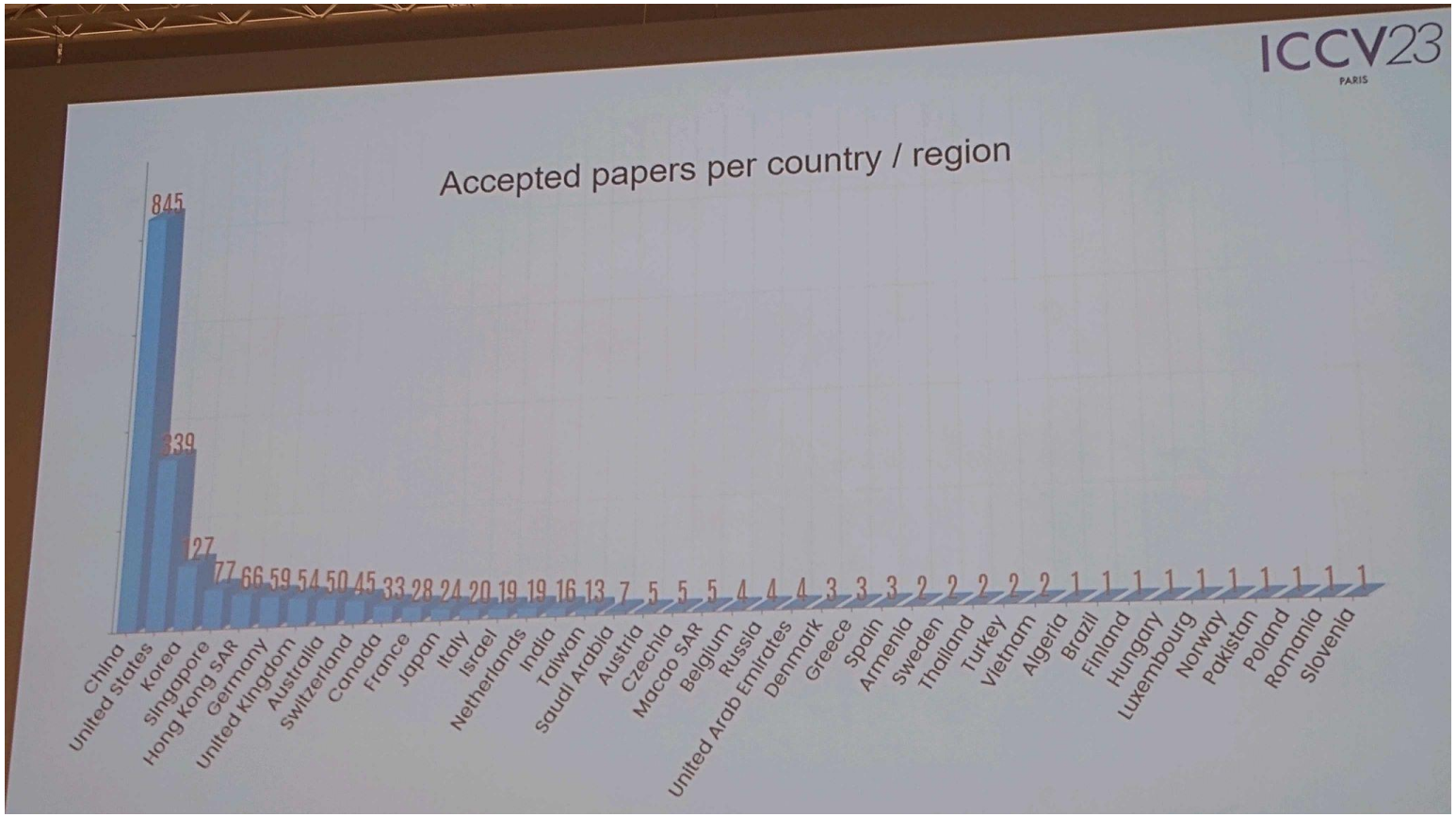
- 国・地域別の論文採択数（1 / 2）
  - 中国が圧倒（845論文で米国の339にDouble Score以上）
  - 韓国・シンガポール・香港が健闘（それぞれ127, 77, 66論文）



ICCV 2023 の動向・気付き(9/165)

Opening Remarks & Paper Awards

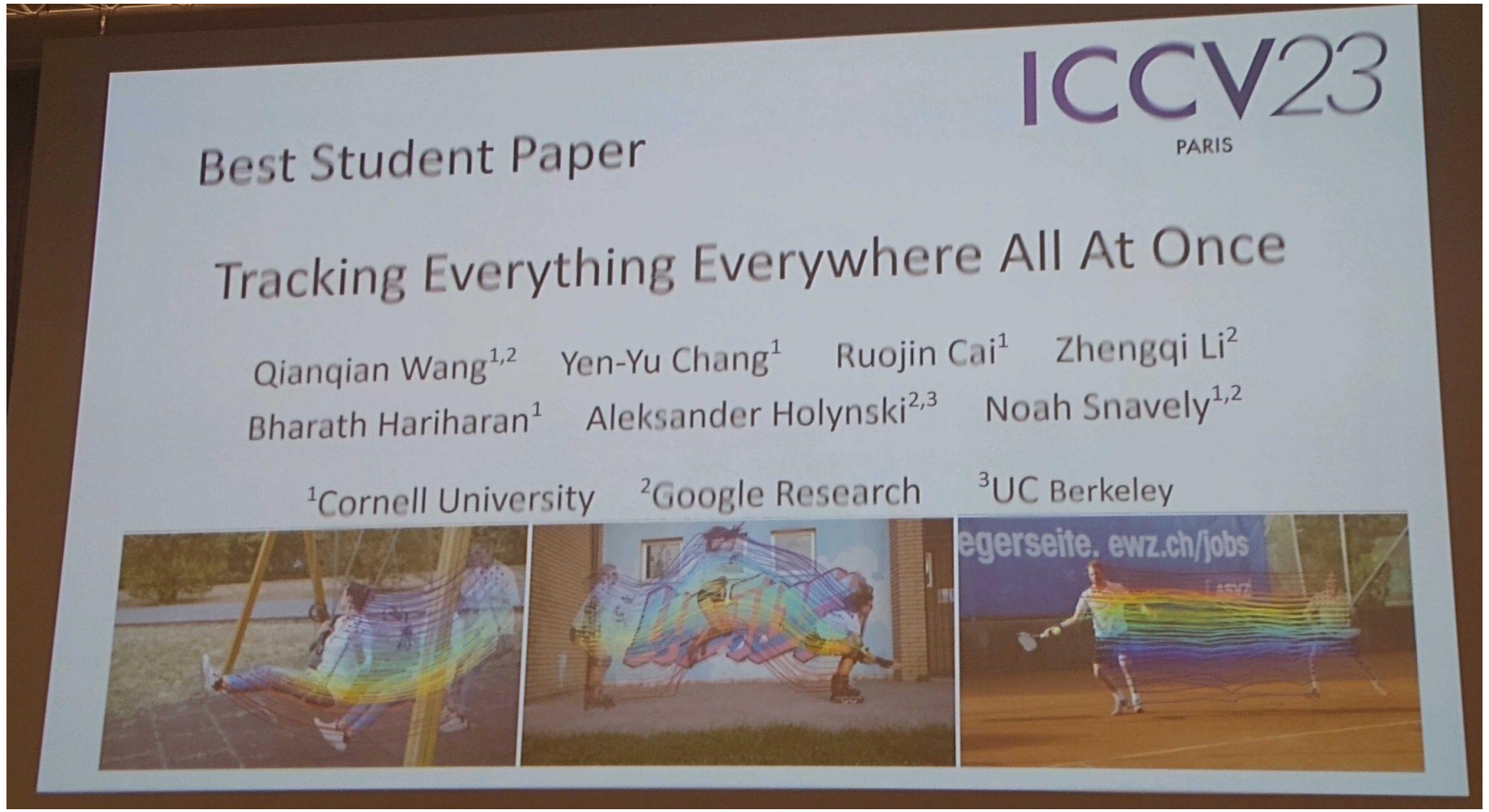
- 国・地域別の論文採択数(2 / 2)
  - EU勢は苦戦気味？(独国:59論文，英国:54論文，スイス:45論文，フランス:28論文)
  - 日本は24論文で世界第12位



ICCV 2023 の動向・気付き（10/165）

Opening Remarks & Paper Awards

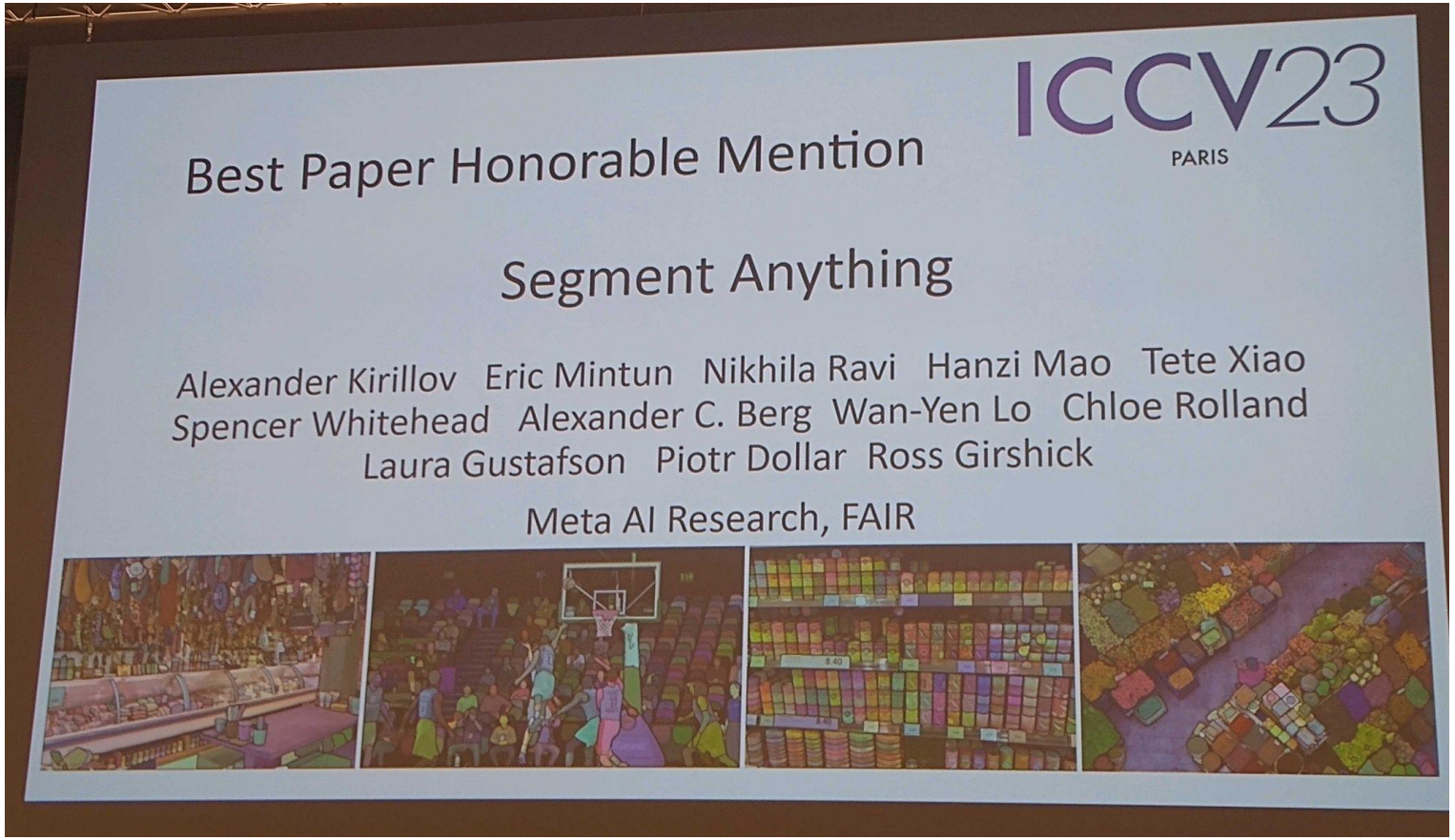
- Best Student Paper: Tracking Everything Everywhere All At Once
  https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Tracking_Everything_Everywhere_All_at_Once_ICCV_2023_paper.html
  - 物体の時空間・詳細・高精度追跡手法が “Best Student Paper Award” 受賞



ICCV 2023 の動向・気付き(11/165)

Opening Remarks & Paper Awards

- Best Paper Honorable Mention: Segment Anything
  https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html
  - なんでもセグメンテーション手法が “Best Paper Honorable Mention Award” 受賞



ICCV 2023 の動向・気付き（12/165）

Opening Remarks & Paper Awards

- Best Paper: Passive Ultra-Wideband Single-Photon Imaging
  https://openaccess.thecvf.com/content/ICCV2023/html/Wei_Passive_Ultra-Wideband_Single-Photon_Imaging_ICCV_2023_paper.html
  - 極端な幅（1～1ピコ秒）・微量光源・動的シーンの撮像技術が “Best Paper Award” 受賞

ICCV 2023 の動向・気付き（13/165）

Opening Remarks & Paper Awards

- Best Paper: Adding Conditional Control to Text-to-Image Diffusion Models
  https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html
  - 微量画像情報（bone/edgeなど）からの画像生成技術（ControlNet論文）が “Best Paper Award” 受賞

ICCV 2023 の動向・気付き（14/165）

タイトルに”language”を含む論文は89件、”dialog”は0件！

- “vocabulary”は17件、“CLIP”は19件、”Foundation”は6件、”text”は97件
- Vision and Language関係では、CLIP応用が相変わらず多い印象
- 今度のCVPR投稿ではLLM系がどっと増えるか？
  - 今回LLMと銘打っているのは2件のみ：
    - LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models
    - Enhancing NeRF akin to Enhancing LLMs: Generalizable NeRF Transformer with Mixture-of-View-Experts

ICCV 2023 の動向・気付き（15/165）

Scene Graph生成もまだまだ現役？（全7件）

- 4件はデータの偏りに注目
  - Vision Relation Transformer for Unbiased Scene Graph Generation
  - Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation
  - うち二件はデータ拡張
    - Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an Open World
    - Scene Graph Contrastive Learning for Embodied Navigation

- 挑戦的なPanoptic Scene Graphも2件（性能はまだまだ・・・・）
  - TextPSG: Panoptic Scene Graph Generation from Textual Descriptions
  - HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation

- Faster R-CNNが未だに現役・・・？
  - Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation
  - Compositional Feature Augmentation for Unbiased Scene Graph Generation

ロボットナビゲーション向けのシーングラフが登場

- Bird's-Eye-View Scene Graph for Vision-Language Navigation
- Scene Graph Contrastive Learning for Embodied Navigation

ICCV 2023 の動向・気付き（16/165）

拡散モデル

- タイトル中に”Diffusion”と入った論文数: 141 / 2,161
- 画像生成問題だけでなく，幅広い問題への応用
  - 表現学習: DiffMAE
  - 物体検出: DiffusionDet,
  - 領域分割: Open-vocabulary Object Seg. with Diffusion Models, DiffuMask, LD-ZNet
  - Text-to-image生成のカスタマイズ: SVDiff
  - 画像編集: Prompt Tuning Inversion, Fashion Image Editing, AIDI
  - 画像生成: MDT
  - 3D-aware画像生成: IVID,
  - スタイル変換: Diffusion in Style, ZeCon
  - 動画編集: StableVideo, Pix2Video, Tune-A-Video
  - モーション生成: ReMoDiffuse
  - Low-level vision: ExposureDiffusion, Diff-Retinex
  - テクスチャ生成: Point-UV diffusion
  - 3Dモデル生成: Diffusion-SDF, Make-It-3D
  - NeRFへの応用: ReferenceGuided3D
  - 概念除去: Erasing Concepts from Diffusion Models
  - 定式化改善: PSLD
    - （他にも沢山あります. ICCV2023 Proceedings）
- 事前学習モデルの上手な利用方法もあれば，タスク固有の問題を扱うために導入した研究もある
  - これからも開拓が進められ，面白い使い方が更に出てくるのではないか？

（スペースの都合上，著者名は省略）

ICCV 2023 の動向・気付き（17/165）

ICCV 2023 LIMIT Workshop

- “日本発”でICCV 2023 Workshopを開催！
- 限られたデータで深層学習の特徴表現（モデル・特徴量など広義）を獲得しよう，という趣旨
- モチベーションなどは次のページ以降に記載

WSのWebpageより https://lsfsl.net/limit23/

ICCV 2023 の動向・気付き（18/165）

ICCV 2023 LIMIT Workshop

- どのようにWSオーガナイズが始まった？
  - 理由：数式ドリブン教師あり学習（FDSL）や類似した問題設定で研究している研究者にスポットライトをあてたワークショップを企画
  - 問題設定：任意の生成モデルで教師あり画像を生成して，深層学習する手法
    - AISTはFormula-Driven Supervised Learning
      https://hirokatsukataoka16.github.io/Pretraining-without-Natural-Images/
    - MITはGenerative Models as Data++（by Phillip Isola）https://www.youtube.com/watch?v=YuRAeQsTSo8
  - WSの取扱範囲：画像・ラベル・計算リソースなど，限られたリソースの中で認識・生成モデルなどを学習

ICCV 2023 の動向・気付き（19/165）

ICCV 2023 LIMIT Workshop

- 投稿までの流れ（〆切：2023/02/10）
  - 〆切4年前：ICCV 2019 Workshopを運営 http://lsfsl.net/ws/
    - データ・ラベル・モダリティなど特殊な学習方法を扱うWorkshopを韓国で企画
  - 〆切2年前：数式ドリブン教師あり学習（FDSL）がACCV Award
  - 〆切1年前：「FDSL関連のワークショップやりたいですね」程度だった
  - 〆切2~3ヶ月前：「ICCVでワークショップやりたいけどまだ早いかも？」
  - **〆切1週間前（！？）：「今しかない，やろう！」**
  - 〆切6.5日前：テンプレをダウンロードして書き始める
  - 〆切6日前：日本側オーガナイザ結成，招待講演2人を選定，FDSL論文引用の研究者にメールして追加オーガナイザ・査読者を招集
  - 〆切1~5日前：返信されたメール（人選）を元にプロポーザルを修正
  - 〆切当日：ICCV 2023 Workshop Chairに投稿！

ICCV 2023 の動向・気付き（20/165）

ICCV 2023 LIMIT Workshop

- 採択発表（〆切：2023/03/27 -> 03/31）
  - 予想以上に投稿が多く，査読は難航した
  - 結果通知（Notification）が延期
  - 会場の制限で，130件中52件（40%）のみが採択

We are delighted to inform you that your workshop proposal #30: Representation learning with very limited images: the potential of self-, synthetic- and formula-supervision has been accepted to the ICCV 2023 workshop. Congratulations!

採択時のメールより

ICCV 2023 の動向・気付き（21/165）

ICCV 2023 LIMIT Workshop

- Workshop あとがき（1 / 3）
  - 〆切1週間前に思い立った時は「どうせRejectなら自分の好きなようにやろう！」だった
  - FDSL論文を引用してくれた研究者への感謝のつもりでメールを送り続けた
  - 査読依頼もシステム上でなく,直接メールしたかったので合計数百通は送っている,直接のやりとりも大事（手間をかけてでも，継続する研究コミュニティを作ることが重要）
  - 「FDSLを提案してくれてありがとう！」「おかげで論文が通った！」など連絡が届いて，投稿の時点でもう良いことをした気持ち
  - 査読者は70名程度集めたがそれでも40本の論文を捌くのは困難，次回は倍の査読者が必要
  - 博士課程の産総研RAが運営に活躍してくれた（感謝！→中村・篠田・山田），研究コミュニティ内では博士課程学生レベルでも運営できる人材が育っている

ICCV 2023 の動向・気付き(22/165)

ICCV 2023 LIMIT Workshop

- Workshop あとがき(2 / 3)
  - Workshop運営は世界進出にとても寄与する
    - ICCV 公式ページに名前が残る
    - 自分たちでプログラムをアレンジできる
    - トピックを同じく研究する仲間が増える
  - 「研究トレンドを創る」という研究コミュニティの大目標にマッチする
  - コツが必要！数年後を見据えた研究を考えてそのテーマをベースに企画など

ICCV 2023 の動向・気付き(23/165)

ICCV 2023 LIMIT Workshop

- Workshop あとがき(3 / 3)
  - 日本国内でCVPR/ICCV/ECCV勉強会(のみを)してる場合ではない！
  - 今すぐ連携チーム構成してトップ会議併設のワークショップに投稿しよう
    - 1. 日本からのWorkshop (e.g., CVPR/ICCV/ECCV)が増える
    - →2. 国内外の人的交流が劇的に加速
    - →3. 情報も増えて研究が加速
    - →4. トップ国際会議に論文数が増加(今回ICCVは日本ドメインから24本のみ...)
    - →5. 新規研究問題設定・手法が増えて産業も活性化！

Workshopを増やすことが日本のプレゼンスを上げる！

ICCV 2023 の動向・気付き(24/165)

ICCV 2023 LIMIT Workshop

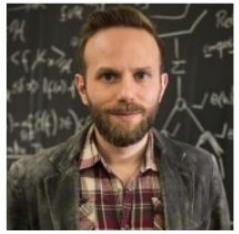
- Organizing Team
  - 日本・米国の研究者により構成
  - Dan: GeLU, OOD benchmark, ImageNet-Cなど
  - Xavier: SALICONなど
  - Connor: Improving Fractal Pre-training
  - 片岡/横田/井上/中村/山田/篠田: FDSL
  - Qiu: VQA, Change CaptioningなどV&L


Hirokatsu Kataoka
AIST/LINE


Rio Yokota
Tokyo Tech/AIST


Nakamasa Inoue
Tokyo Tech/AIST


Dan Hendrycks
Center for AI Safety


Xavier Boix
MIT


Yue Qiu
AIST


Connor Anderson
BYU


Ryo Nakamura
Fukuoka Univ./AIST


Ryosuke Yamada
Univ. of Tsukuba/AIST


Risa Shinoda
Kyoto Univ./AIST

ICCV 2023 の動向・気付き（25/165）

ICCV 2023 LIMIT Workshop

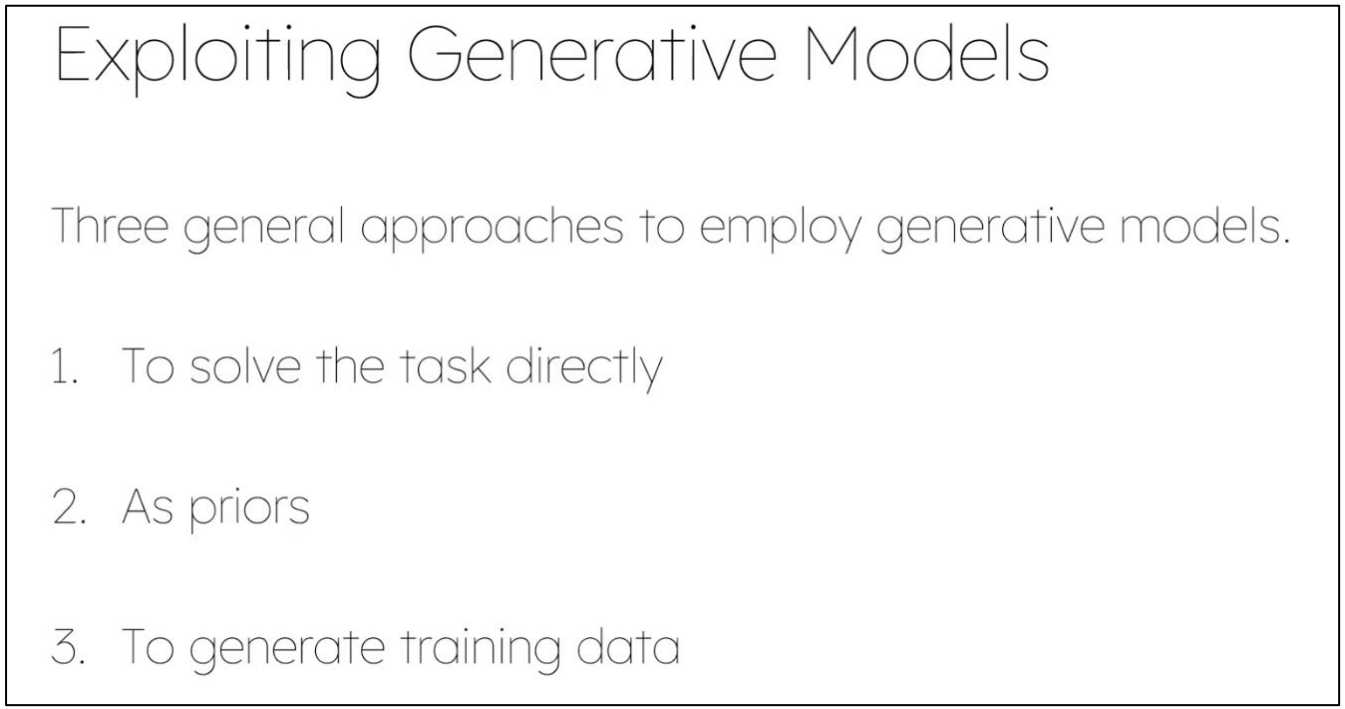
- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - 教師なし学習（Unsupervised Learning）x 限られたデータ（Limited Data）という非常に厳しい環境での特徴表現学習により視覚的タスクを解決するという講演

画像は動画より
https://youtu.be/IJBIFCiV_WI

ICCV 2023 の動向・気付き（26/165）

ICCV 2023 LIMIT Workshop

- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - 現在のレベルでは生成データでも深層学習の訓練ができる！
  - CVPR 2023 Workshop（Generative Models for Computer Vision）でも下記のように，生成モデルは (i) 直接タスクを解決，(ii) 事前知識として，(iii) 学習データ生成，の側面で使えると主張（下画像参照）



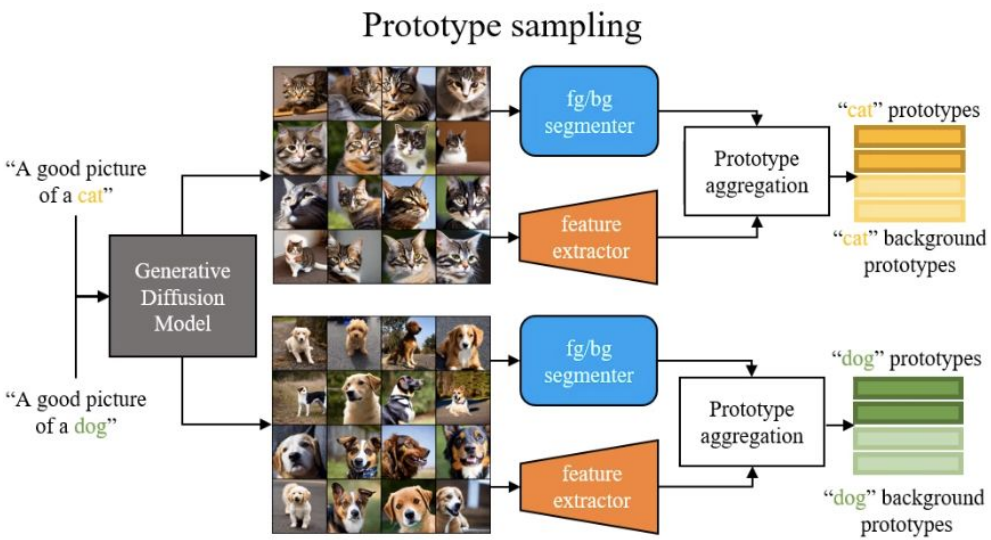
画像は動画より
https://www.youtube.com/watch?v=HUyP2C2rYto

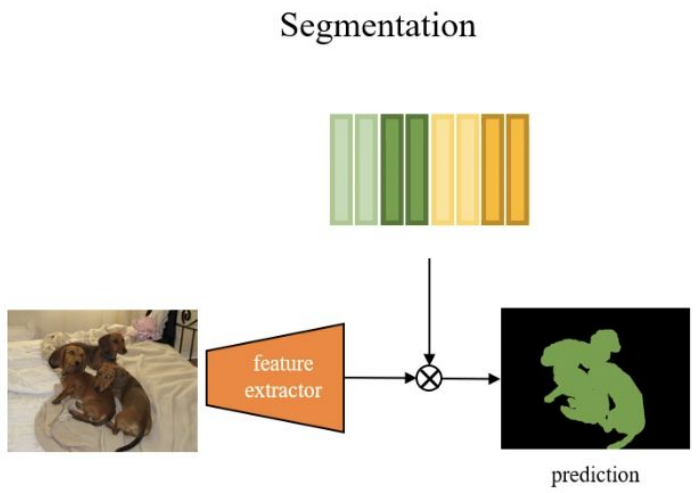
ICCV 2023 の動向・気付き（27/165）

ICCV 2023 LIMIT Workshop

- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - データを生成してセグメンテーションの学習
  - Diffusion Model（text-to-image） + セグメンテーションモデルでUnsupervised・Open Vocaburary・Zero-shot Recognitionのセグメンテーションを実現している

学習データは生成により作れる！



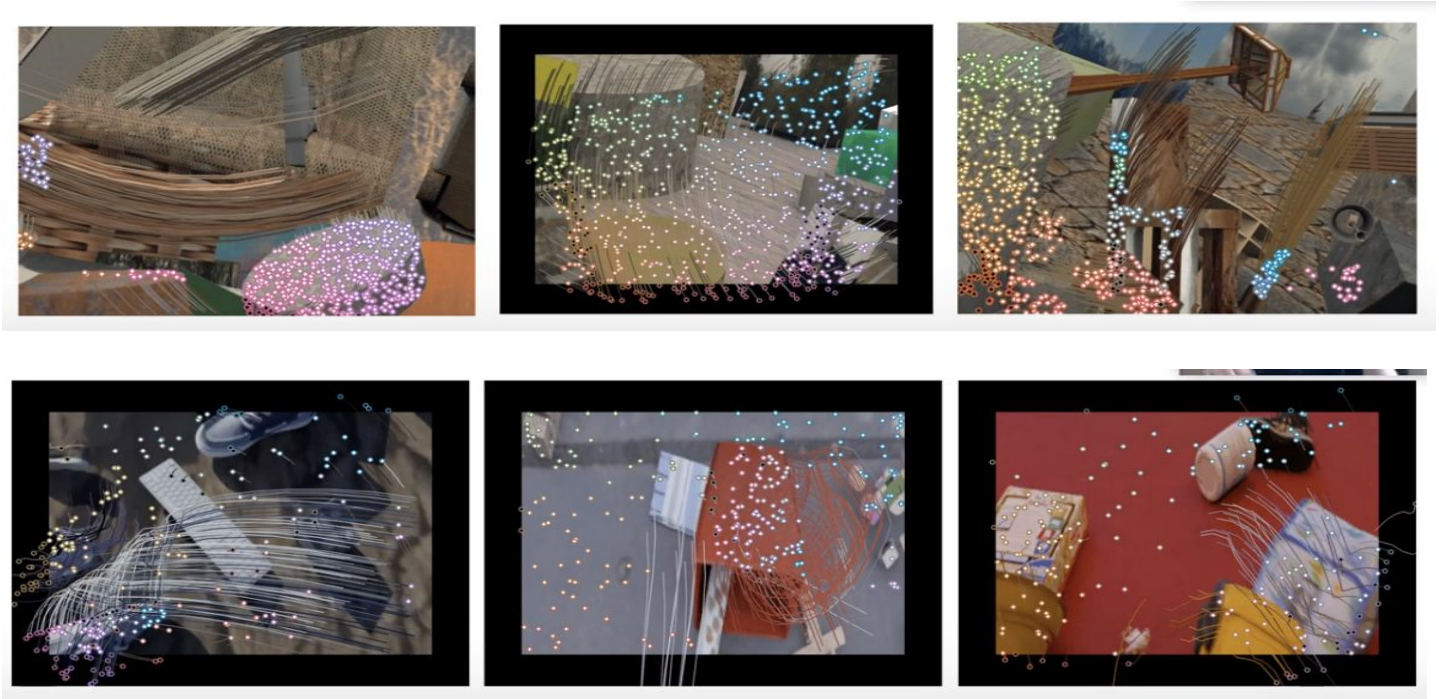


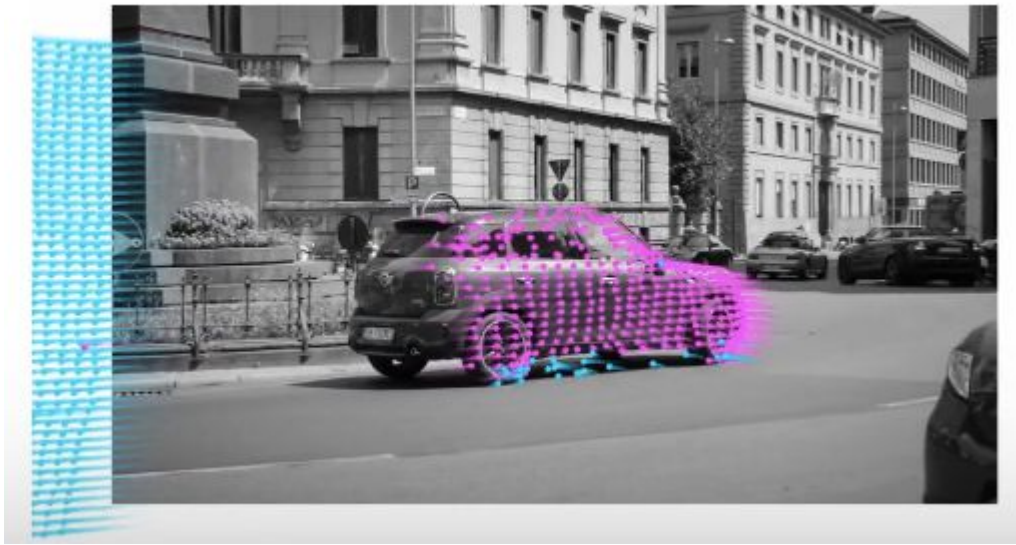
画像は動画より
https://youtu.be/IJBIFCiV_WI

ICCV 2023 の動向・気付き（28/165）

ICCV 2023 LIMIT Workshop

- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - データを生成して物体追跡（Object Tracking）の学習
  - 合成動画データにより物体の時空間特徴点を生成（左図）
  - 初期位置とその動向を把握して追跡（Co-Tracking）
  - 背景と切り分け，物体領域全体を追跡可能（右図; 影が一部青色の背景と判断される）



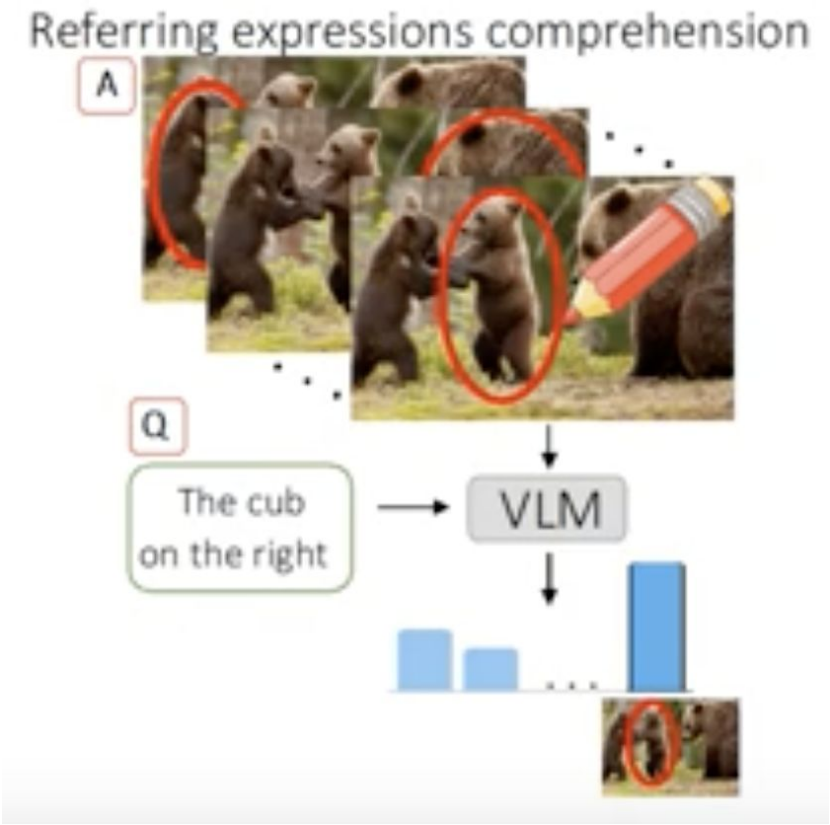


画像は動画より
https://youtu.be/IJBIFCiV_WI

ICCV 2023 の動向・気付き（29/165）

ICCV 2023 LIMIT Workshop

- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - 簡易的な人手によるデータによりVision&Language Model（VLM）を学習
  - 画像中に付けた赤丸によりVLM（CLIPを使用）の出力が変化する，という話



画像は動画より
https://youtu.be/IJBIFCiV_WI

ICCV 2023 の動向・気付き（30/165）

ICCV 2023 LIMIT Workshop

- Invited speaker 1: Christian Rupprecht (University of Oxford)
  - 今後，より汎用的な目的のモデルが増える
  - 人間による労力を減らしつつ，汎用タスクを増やす
  - 合成データ・簡易作成データ・生成モデル・強力なモデルを利用しよう！

画像は動画より
https://youtu.be/IJBIFCiV_WI

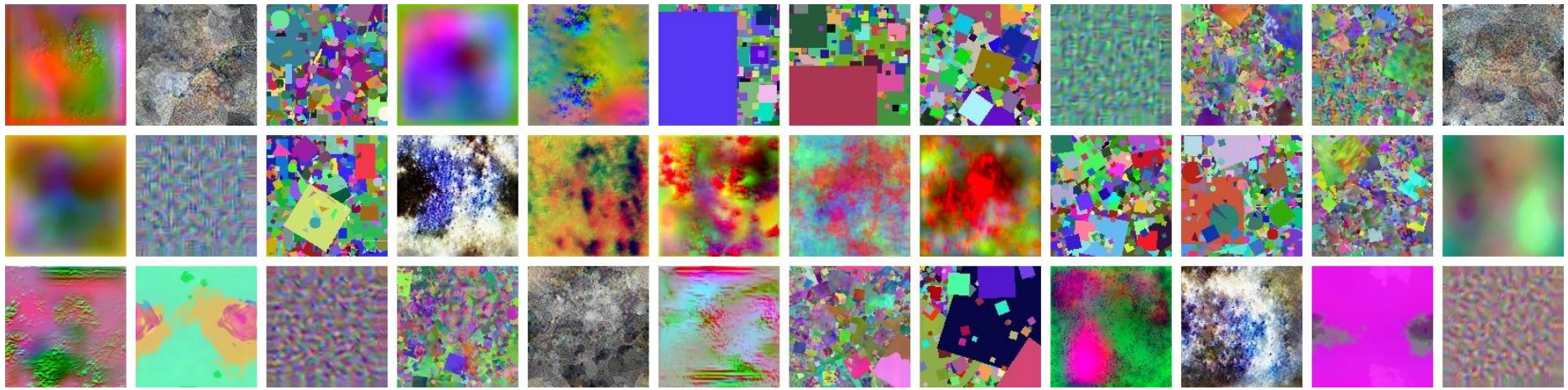
ICCV 2023 の動向・気付き（31/165）

ICCV 2023 LIMIT Workshop

画像はプロジェクトページより
https://mbaradad.github.io/learning_with_noise/

- Invited speaker 2: Manel Baradad (MIT)
  - 一画像でもない，ゼロ枚の実画像で視覚機能は学習できる
  - 簡易形状の組み合わせ（DeadLeaves），生成モデルの初期値（StyleGAN）等を利用
  - 対照学習（MoCov2）により効果的に特徴を獲得
  - まだImageNet/Placesなど実画像による学習とはギャップがあるが「できる」ことは明らかとなった

ノイズを見ることで視覚機能を学習！

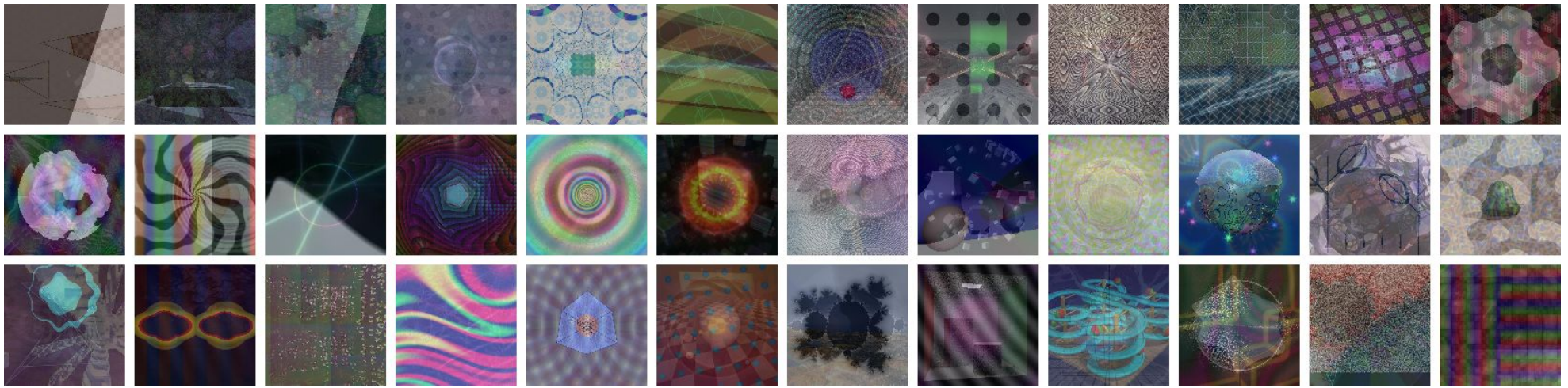


ICCV 2023 の動向・気付き(32/165)

ICCV 2023 LIMIT Workshop

画像はプロジェクトページより
https://mbaradad.github.io/learning_with_noise/

- Invited speaker 2: Manel Baradad (MIT)
  - Generative Art(プログラミングによるグラフィクス)を活用
  - 1プログラム1カテゴリとして認識モデルが学習
  - ImageNetを想定して1,000/21,000カテゴリを準備して学習



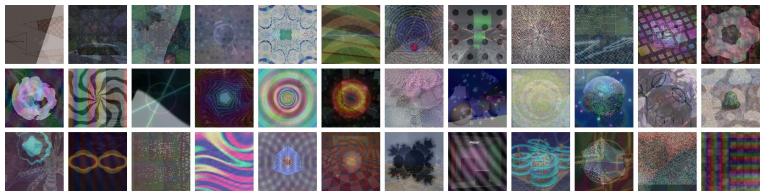
ICCV 2023 の動向・気付き（33/165）

ICCV 2023 LIMIT Workshop

画像はプロジェクトページより
https://mbaradad.github.io/learning_with_noise/

- Invited speaker 2: Manel Baradad (MIT)
  - Generative Art（プログラミングによるグラフィクス）を活用
  - ImageNet-1kは参考値として，Places365と比較
  - 場合によりとても近い値を示す（例：VTAB Nat. 59.72 vs. 57.18）

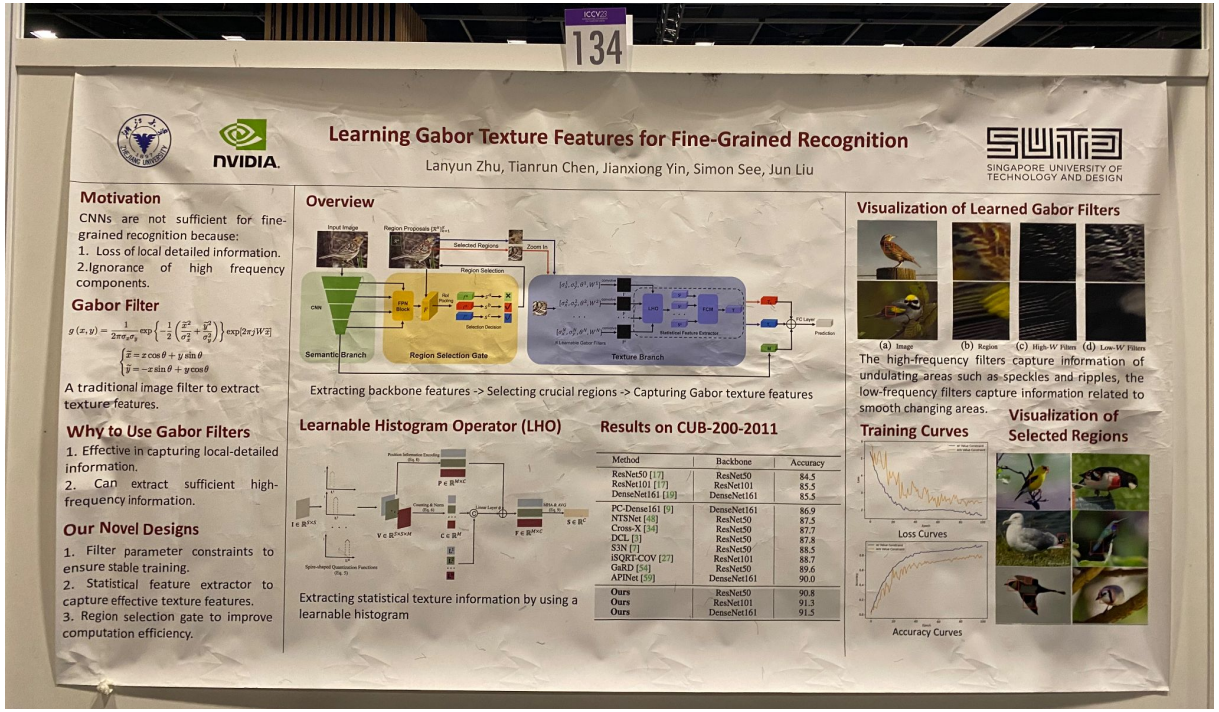
Pre-train Dataset	I-1k	I-100	VTAB Nat.	VTAB Spec.	VTAB Struct.
Random init	4.36	10.84	10.98	54.30	22.64
Places365 [6]	55.59	76.00	59.72	84.19	33.58
ImageNet-1k	67.50	86.12	65.90	85.02	35.59
StyleGAN O. [2]	38.12	58.70	54.19	81.70	**35.03**
StyleGAN O. [2] (MixUp)	31.73	53.44	51.26	81.39	33.21
FractalDB-1k [1]	23.86	44.06	38.80	76.93	31.01
Dead-leaves Mixed [2]	20.00	38.34	35.87	74.22	30.81
S-1k	16.67	34.56	32.39	75.28	28.23
S-1k MixUp	38.42	60.04	53.24	82.08	30.32
S-21k	30.25	51.52	45.23	80.75	32.85
S-21k MixUp	**44.83**	**66.36**	**57.18**	**84.08**	31.84



ICCV 2023 の動向・気付き（34/165）

Poster 4th am:Learning Gabor texture Features for Fine-Grained Recognition

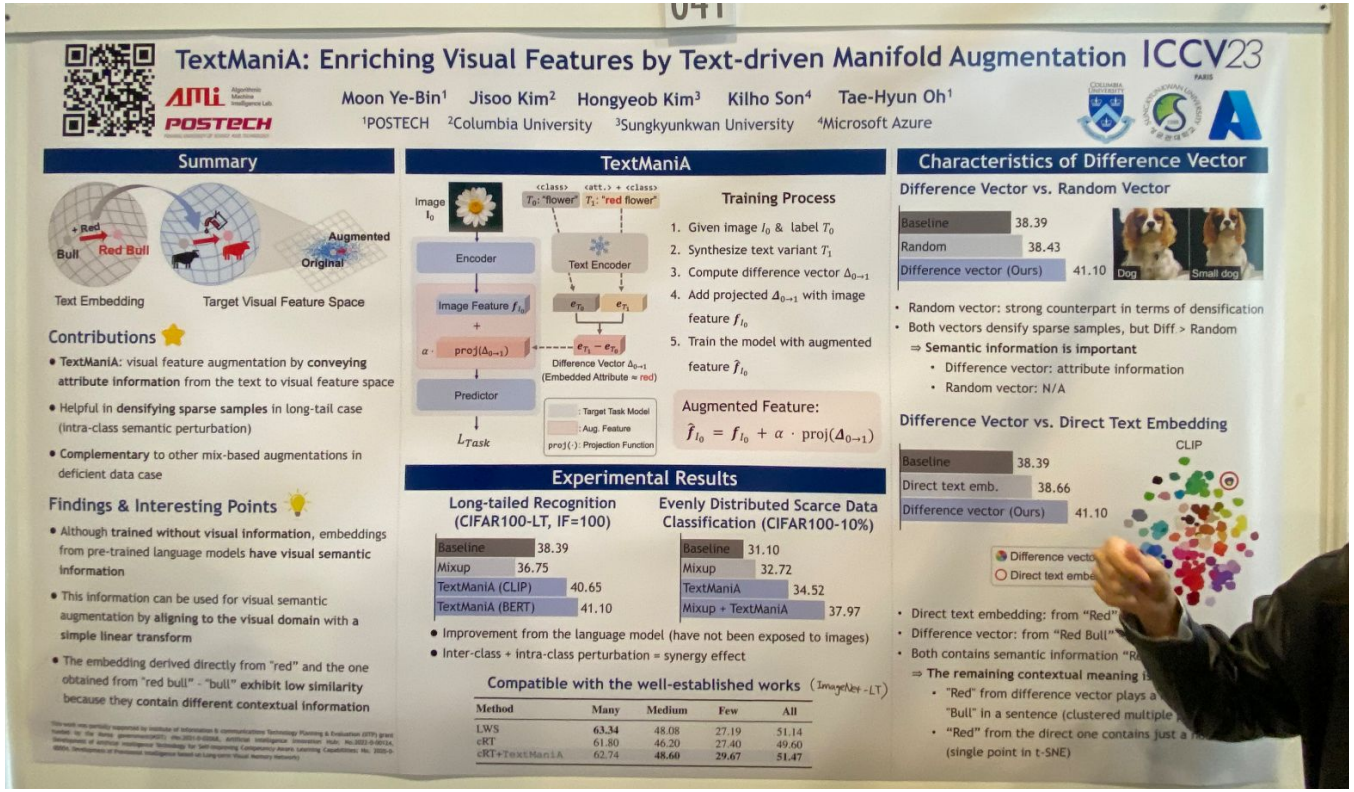
- Fine-grained な認識においてGabor Fileterで抽出したtexture が有効であることを示した論文
  - CNN Encoderから抽出した特徴とRegion Proposalで切り出したパッチをGabor filterに通した特徴を使って分類学習を行う

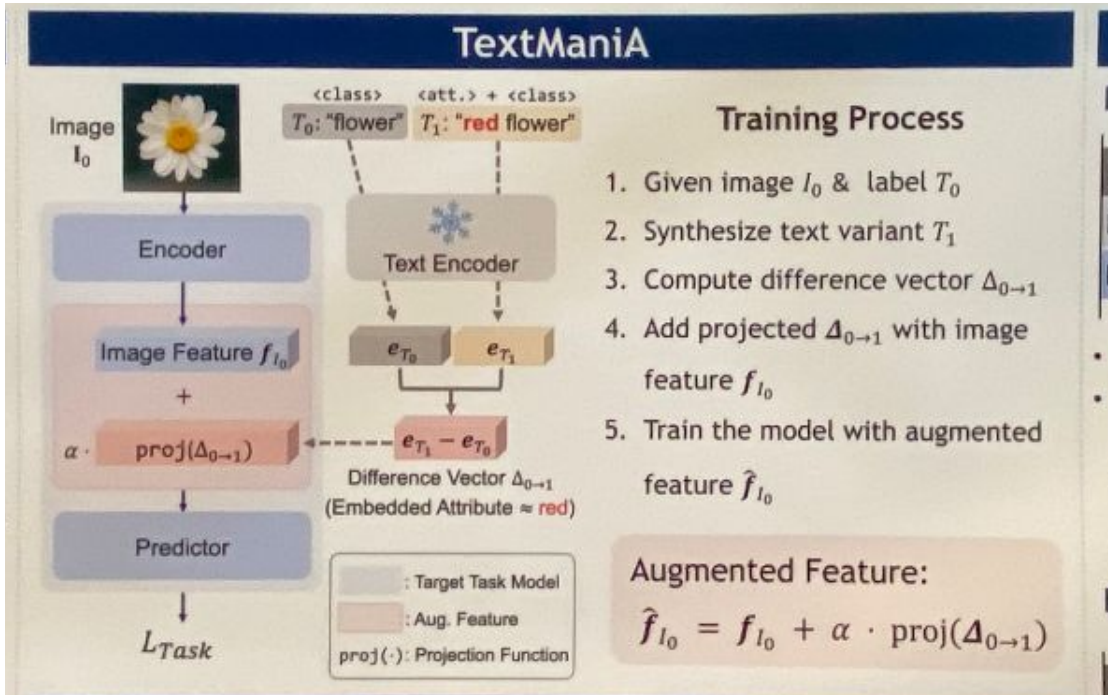


ICCV 2023 の動向・気付き（35/165）

Poster 4th am:TextManiA:Enhancing Visual Features by Text-driven Manifold Augmentation

- Text embedding空間のセマンティクスを用いて色の違いなどの特徴を画像特徴との足すことでData augmentationをする論文
  - Text embeddingにはCLIPやBERTを用いている

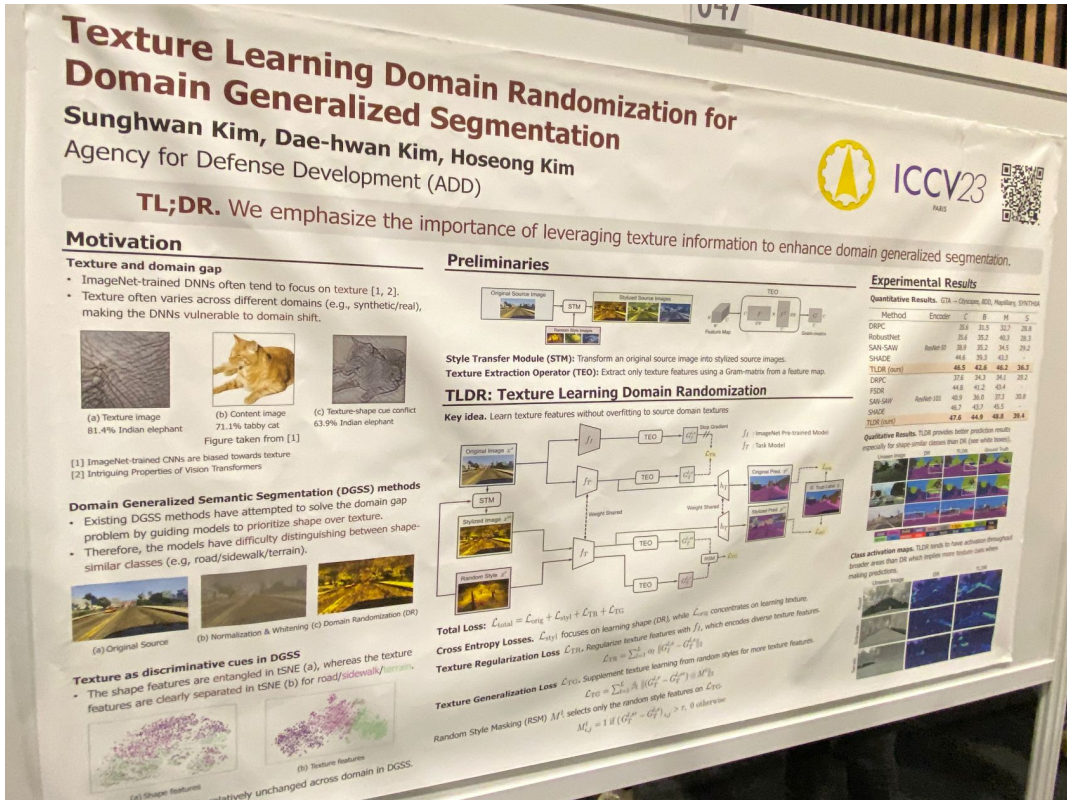




ICCV 2023 の動向・気付き（36/165）

Poster 4th am:Texture Learning Domain Rndomization for Domain Generalized Segmentation

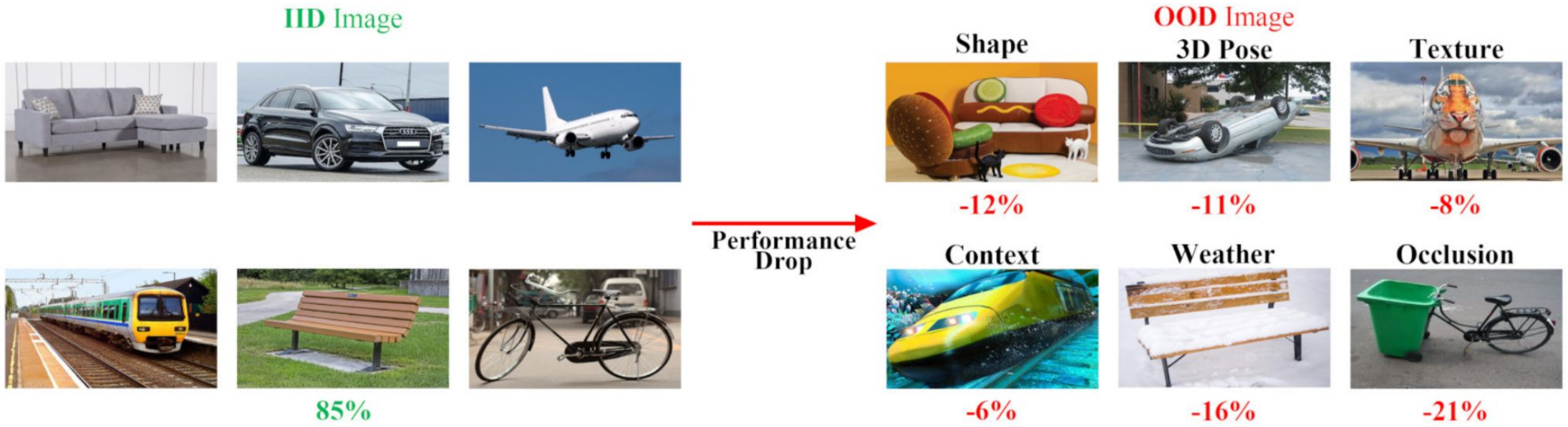
- 多様なTextureに依らない認識を行うために，多様なStlye変換を用いてスタイルにロバストなsegmentation手法の提案.
  - Encoderの出力にImageNetの出力やstyle変換画像の出力に近づけるような学習を行っているのが特徴的



ICCV 2023 の動向・気付き（37/165）

The 2nd Workshop and Challenges for Out-of-Distribution Generalization in Computer Vision

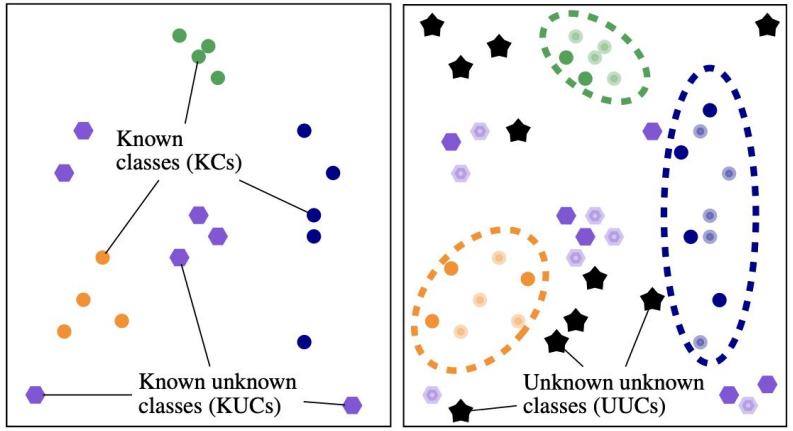
- ディープラーニングモデルは，訓練とテストデータが同じ分布から来ていると仮定されているが、実際のシナリオでは分布が異なる場合が多く、性能低下の原因となる可能性がある.
- このワークショップでは，学習と異なる分布のOOD画像に対するモデルの対応力を議論し，三つの競技を通じてその性能を評価.
- 特に，ウェブスケールで事前学習されたモデルのOOD性能に焦点を当てている.



ICCV 2023 の動向・気付き（38/165）

Oral発表: LORD: Leveraging Open-Set Recognition with Unknown Data

- Unknow classを用いることで，既知の領域と未知の領域を
- 分類モデルは通常，あらかじめ定義されたデータセットで訓練され，未知の特徴空間については考慮されていないため，分布外のデータの推論時に問題が生じる.
- この論文では，LORDというフレームワークを提案し，**訓練中に未知の空間を明示的にモデリング**することで，**オープンセット認識(OSR)の性能を向上**させる方法を検討する。
- また，背景データの依存を軽減するためのデータ生成技術としてmixupの効果を探ると、mixupが背景データの代替として効果的であることが実験で示される　.



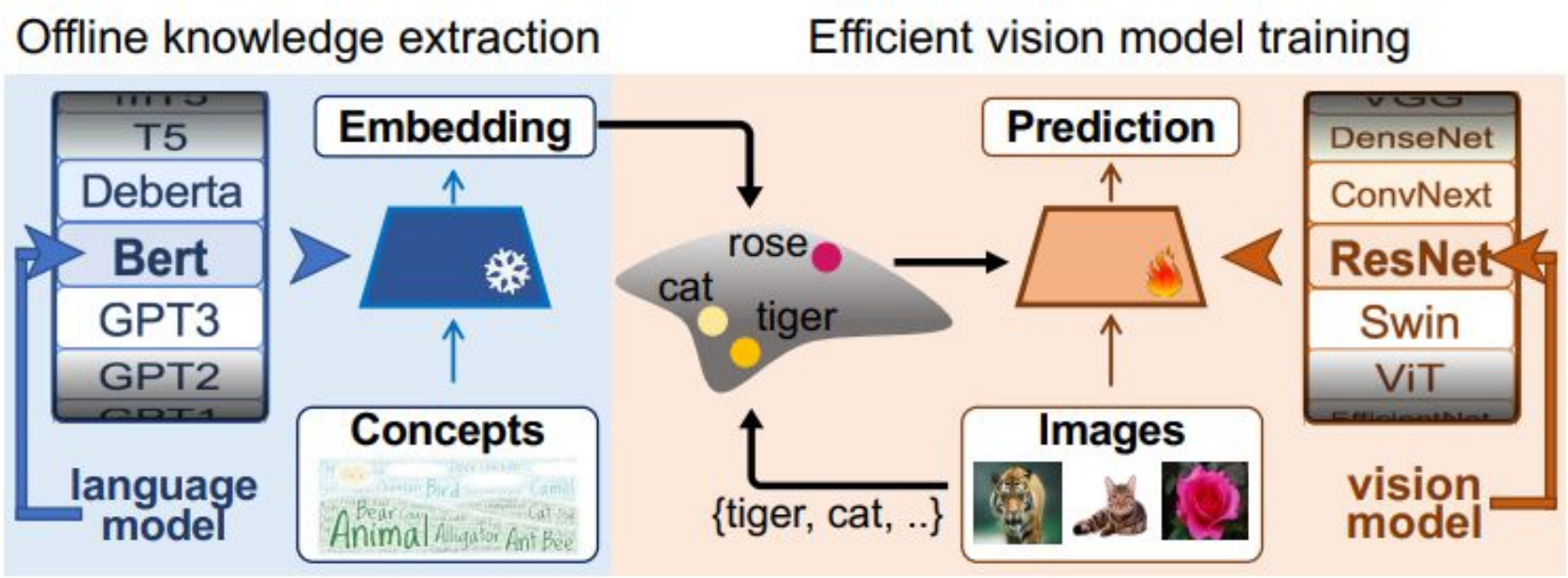
(a) Training data (opaque). (b) Test data (opaque) with training data (shaded).

Figure 1. Overview of data types in open-set recognition. The training set in (a) includes known classes (KCs) and known unknown classes (KUCs) (⬢). The trained classifiers in (b) model decision boundaries for KCs as dashed ellipses. KUCs correlate with the training set's KUCs, exhibiting higher identifiability in comparison to the unknown unknown classes (UUCs) (⬟), which can exist anywhere in the feature space.

Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm

- 事前学習済みの言語モデル(Language model)を利用した視覚モデル(Vision model)の効率的な追加学習手法BorLanの提案
  - 視覚モデルが失いがちな意味的概念を補完可能
  - 複数のベンチマークで最高の性能を達成
    - FGVC Aircraft
    - Stanford Cars
    - CUB-200

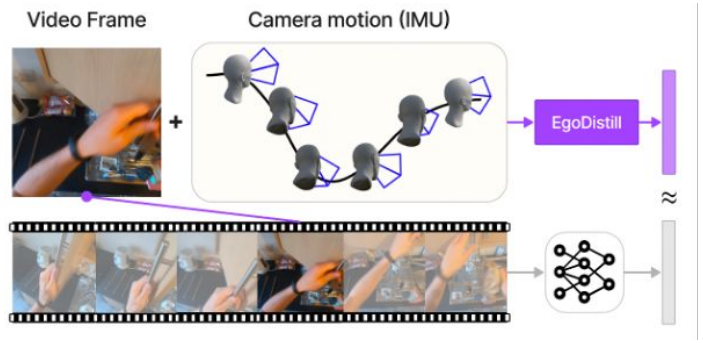
ICCV Open access:
[Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm (thecvf.com)]
GitHub:
[GitHub - BIT-DA/BorLan: [ICCV2023] Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm]



ICCV 2023 の動向・気付き（40/165）

Workshop: What is Next in Multimodal Foundation Models (1/2)

- タイトル： Goals, Memories, and Summaries from Large-Scale Narrated Video
- 発表者：Kristen Grauman
- 発表趣旨：Learning from large-scale narrated videos -> 下記の4つのトピック
  - Hierarchical video-language embeddings
    - Action descriptions (what) と summaries (why) の階層関係を学習-> HierVL [CVPR, 2023]
  - Episodic memory language queries
    - Augment NLQ(natural language queries) training by learning to temporally localize narrations -> NaQ [CVPR, 2023]
    - Denseビデオクリップの特徴の代わりにクリップを選択することでコスト削減-> SpotEM [ICML, 2023]
  - Visual Narrations in how-to’s
    - Visual related / non-visual narrationsがVideo narrationsデータセットに遍在
    - アプローチ：視覚と密に関係するnarrationsをdetectionすることで良い特徴表現
  - Fast video features with IMU
    - Denseビデオクリップの特徴のコストが高い
    - アプローチ：IMUから、Headのモーションなどのデータを得てsemantics と関連付け-> EgoDistil [NeurIPS, 2023]
    - Low-level ego dataでSemanticsを得るのが流行ってきている？



EgoDistil [NeurIPS, 2023]

ICCV 2023 の動向・気付き(41/165)

Workshop: What is Next in Multimodal Foundation Models (2/2)

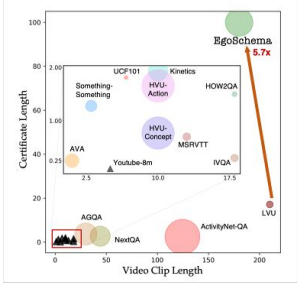
- タイトル: Visual Commonsense Reasoning with Large Language ModelsTowards 3D Representation Learning at Scale
- 発表者:Chuang Gan
- 現在のMultimodal Foundationモデルの問題点:
  - 15ヶ月の赤ちゃんと比べると劣るCommonsense 3D Scene Reasoning能力(使用モデル:ChatGPT＋OpenFlamingo)
- 人間の3D Reasoning:
  - コア能力:空間探索, 3次元特徴表現, 様々な3次元ベースタスクを解く
- 人間レベルの 3D Reasoning を実現する試み:
  - 大規模3次元・言語データセットを提案(box-demonstration-instruction-based prompting)
  - 手法提案1:多種類の3次元特徴表現、フレームワークを結合:
  - 3D-LLM framework: 3D scenes -> Multiview -> CLIP feature
  - Direct construction, nerf, slam
  - 手法提案2:Communicative decoding＋3次元特徴表現、対話的・能動的なモデルをしよう
- 未来の方向性:3次元Visual Reasoning、CompoistionalReasoningなど(こういった領域ではLLMの強みを発揮するべき)

ICCV 2023 の動向・気付き(42/165)

Workshop: Vision and Language Algorithmic Reasoning (1/5)

- Invited talk内容:unsolved problems in video understanding
- 発表者:Jitendra Malik
- Unsolved problem 1: 同時にreconstruction, recognition, trackingを高精度で行う
  - 論文紹介:Human in 4D: Reconstructing and Tracking Humans with Transformers [ICCV, 2023]
  - 内容:ビデオから、同時に3次元reconstructionとTrackingを行う手法を提案。
- Unsolved problem 2: long-form video recognition
  - 論文紹介: EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding
  - 内容:1. Long-formビデオを評価する指標temporal certificates (人間が判断するに必要な最小サブクリップの集合)を提案;2. 大規模Long-formビデオデータセットを提案。EgoSchemaは既存の15ビデオベンチマークと比べて、最も高いtemporal certificatesを得た。合計250時間以上の3分ビデオから構成する。
  - EgoSchemaでの結果:既存の最も良い手法は30%程度の精度となる(ランダムは20%)。Scaling token-based LLMはessence of the 4D worldを理解できないため、EgoSchemaは今後のLong-form動画認識の良いベンチマークとなる。
- 未来の方向性:映画の認識、YouTubeなどの巨大なビデオソースから学習など

Human in 4Dの結果例



EgoSchemaと既存データセットの比較

ICCV 2023 の動向・気付き(43/165)

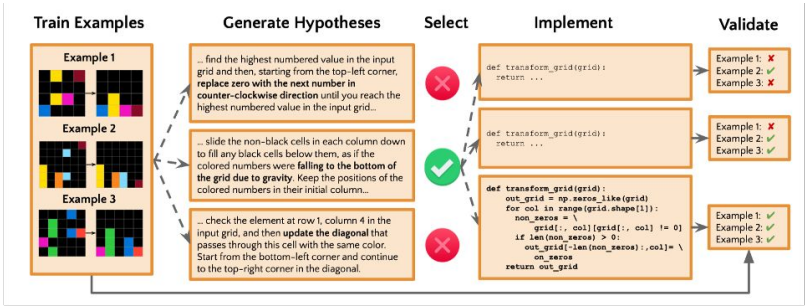
Workshop: Vision and Language Algorithmic Reasoning (2/5)

- Invited talk内容:The Missing Rungs on the Ladder to General AI
- 発表者:François Chollet
- LLMの制限(e.g., hallucinations, …)
  - 記憶されたデータの小さい変動に対してロバスト性不足
  - Rephraseに関してロバスト性が不足
  - 学習データに含まれていないタスクに弱い
  - 汎化性能が不足
- Conceptualizing intelligent systemsのコア能力:
  - fluidity -> on the flyで新しいプログラムを生成
  - 学習データより広ければ広いほどよいoperational area
  - information-efficiency
- Generalization はAIの中心的な問題
  - uncertainty, novelty, autonomy, unknown unknownsなどへの対応が必要。
- Generalizationの評価
  - experience と priors をコントロール-> 提案ARCデータセット (abstract and reasoning corpus)
  - LLMs are not good at ARC (5%to10%)

ICCV 2023 の動向・気付き（44/165）

Workshop: Vision and Language Algorithmic Reasoning (3/5)

- Invited talk内容：The Missing Rungs on the Ladder to General AI
- 発表者：François Chollet
- Abstraction はGeneralizationのコア：
  - Abstraction：representation, patternsを発見し、再利用
  - Abstractionのレベル：Pointwise factoids -> organized knowledge (LLMs はここくらい) -> generalizable models ->on-the-fly model synthesis -> AGI
  - Abstractionの２本のPole: prototype-centric (value-centric), program-centric（LLMsは前者が強く、後者が弱い）
  - How LLMs do abstraction (word2vec analogy): good at type 1 but not good for type 2
- 未来の方向性：
  - prototype-centric と program-centricを結合する。例：DLとプログラム合成を組み合わせる
  - 関連研究紹介：Hypothesis search: inductive reasoning with language models (2x performances on ARC)



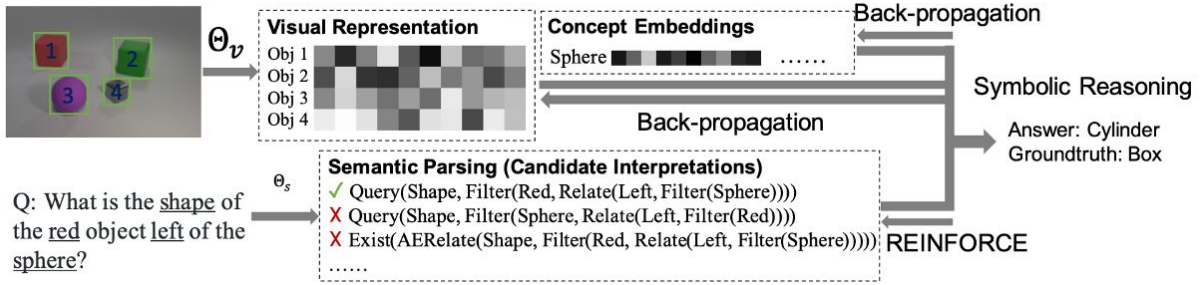
Hypothesis search:
LLMを利用し、階層的にExplicit hypothesisを生成し、問題を解いていく

cvpaper.challenge

ICCV 2023 の動向・気付き(45/165)

Workshop: Vision and Language Algorithmic Reasoning (4/5)

- Invited talk内容:Concept Learning Across Domains and Modalities
- 発表者:Jiajun Wu
- コンセプト理解とSymbolic reasoningを結合したモデルの試み:
  - 視覚の理解とSymbolic reasoningを分けたモデル:
    - Scene parsing + semantic parsing + symbolic reasoning [NS-VQA, NeurIPS'2018]
    - 最近の手法:ViperGPT、Visual Programming [CVPR'2023]
  - 視覚の理解とSymbolic reasoninを結合させたモデル
    - Visual representation + concept embeddings + semantic parsing + Neural-Symbolic learning [NS-CL, ICLR 2019] (2次元画像)
    - temporal reasoning, causal reasoning [CLEVRER, ICLR'2020](NS-CLの動画像への応用)
    - Learn to execute neural programs in 3D point clouds [NS-3D, CVPR'2023] (NS-CLの3Dへの応用)
    - BABEL-QA dataset [NS-Pose, ICML'2023] (NS-CLの3D Humansへの応用)
    - program-based modular paradigm [ProgramPort, ICLR'2023] (NS-CLのrobotic manipulationへの応用)
  - 視覚の理解とSymbolic reasoninを結合させた、ドメイン任意のFoundationモデル
    - LEFT (logic-enhanced foundation models) [NeurIPS, 2023]
    - LEFTの構造:domain-independent reasoning (LLM), domain-specific grounding;LLM interpreter + first-order logic executer
    - LEFT が最近のMultimodal LLMs(例:OpenFlamingo)と比べ、unseen complex reasoning tasksに対して性能が圧倒的に高い

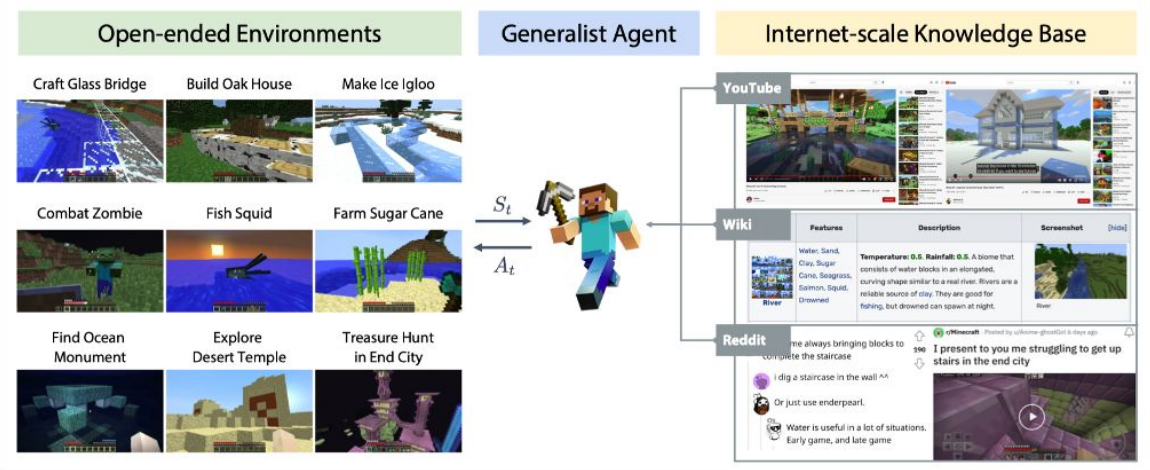


NS-CLモデル構造

ICCV 2023 の動向・気付き（46/165）

Workshop: Vision and Language Algorithmic Reasoning (5/5)

- Invited talk内容：An Open-Ended Embodied Agent with Large Language Models
- 発表者：Anima Anandkumar & Christopher Choy
- MINEDOJO: [NVIDIA, NeurIPS’2022]
  - generalist agent（open-ended環境＋インターネット知識：YouTube、Wiki、Reddit）
  - 3000+タスク（programmatic タスク, creative タスクなど含める）
  - MineCLIPモデルも提案（contrastive video-language foundation モデル）
- Voyager: an open-ended embodied agent with large language models
  - 初めてのLLM-powered lifelong learning agent
  - コアアイディア: コードでEmbodied Agentをコントロール
  - 構造: automatic curriculum maximizing exploration; 拡張し続けるSkill library; interactive prompting mechanism;



MINEDOJO

ICCV 2023 の動向・気付き（47/165）

Workshop: 5th Workshop on Closing the Loop Between Vision and Language

- 採択論文（抜粋）
  - Instruction-tuned Self-Questioning Framework for Multimodal Reasoning
    - LLMで盛んに研究されている、self-refineをMLLMでもVQAに関して行った論文
  - Zero-Shot and Few-Shot Video Question Answering with Multi-Modal Prompts
    - 動画のエンコーダーの出力をLLMの入力に合わせる研究（BLIP2, InstructBLIPなどの大規模画像言語モデルが行っていることを、動画でも行った論文。LLMとしては、RoBeRTaを使用していた）
  - Simple Token-Level Confidence Improves Caption Correctness
    - Image captioningのモデルの出力に関して、トークン一つずつにモデルの出力の自信どあい（confidence）を算出し、このconfidenceが一定基準を満たすまで、モデルにビームサーチをさせる研究
      - 単純なconfidenceの平均と、confidence予測モデルの二つでキャプションのconfidenceを考えている
- 所感
  - 本会議に比べて投稿時期が遅い（7月ごろ）ので、MLLMに関する研究が多い印象を受けた。
  - 他にも、より推論が難しい問題設定とデータセットの提案という従来の通りの流れを汲む論文も見受けられた。

https://www.aclweb.org/portal/content/5th-workshop-closing-loop-between-vision-and-language-clvl-iccv-2023

ICCV 2023 の動向・気付き（48/165）

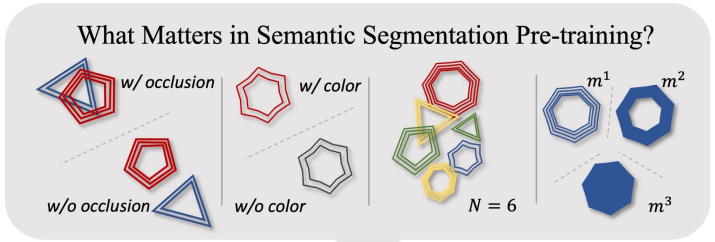
Workshop : 1st Workshop on Open-Vocabulary 3D Scene Understanding (OpenSUN3D)

- WS概要：三次元シーン認識をオープンボキャブラリーに拡張。そのためのタスク定義や評価指標、データセットに関して議論するワークショップ。
- 所感：採択論文はCLIPなどの大規模な画像ーテキストペアで学習された特徴量を三次元の認識に組み込む研究がほとんどであった。
- Keynote（抜粋）
  - CLIP goes 3D: Leveraging Prompt Tuning for Language Grounded 3D Recognitio
    - 発表者：Vishal Patel
    - https://arxiv.org/abs/2303.11313
    - 論文の紹介。三次元点群、レンダリング画像、キャプションのトリプレットを用いて3Dエンコーダを学習。
    - Zero -shotで分類、image-3d検索、物体認識などのタスクが可能となった。
  - 3D Simulation for Embodied AI: Emerging Challenges and Opportunities
    - 発表者：Manolis Savva
    - これまでのEmbodied AI 環境の紹介や、今後シミュレーション環境がどうあるべきかについての議論。
    - シミュレーションは有益だが、ブラックボックスとして扱うのは賢明ではない。
    - インタラクティブシーンであることが大切
    - これまでのデータセットはオープンボキャブラリーと言える？
    - ⇨さらに大きくインタラクション可能なデータセットが必要

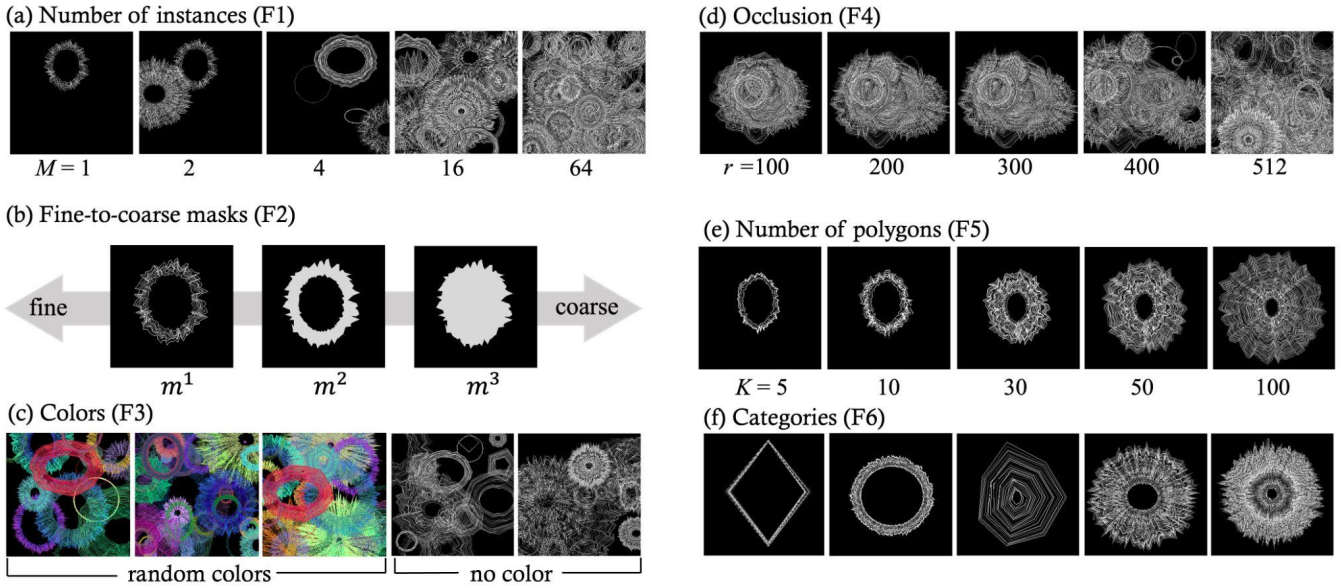
ICCV 2023 の動向・気付き（49/165）

SegRCDB: Semantic Segmentation via Formula-Driven Supervised Learning

- 実画像を用いずセマンティックセグメンテーションの事前学習を可能にするデータセットを提案
  - どのような要素がセマンティックセグメンテーションに効くのかを調査
  - 商用利用も可能
  - COCO-Stuffと同枚数でより高い事前学習効果を達成



Fine-tuning	@ ADE-20k	@ Cityscapes	
COCO-Stuff-164k	43.39	GTA5	71.00
RCDB	41.07	RCDB	69.66
SegRCDB (Ours)	**43.85**	SegRCDB (Ours)	**73.06**

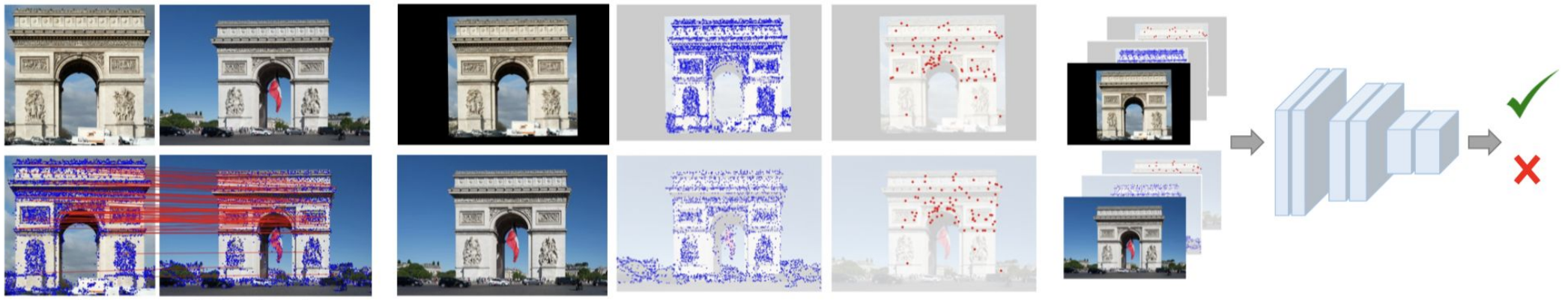


Project page: https://dahlian00.github.io/SegRCDBPage/

ICCV 2023 の動向・気付き（50/165）

Doppelgangers: Learning to Disambiguate Images of Similar Structures

- 2枚の画像から、画像の中の物体（例：建物）が同一表面なのかどうかを判断するタスク設定・データセット・手法を提案。
  - データセット提案：222シーン（ランドマーク等）の76kインターネット画像から構成される。
  - ベース実験：画像ペアからキーポイント検出し、キーポイントの対応関係で2枚の画像が同一表面であるかどうかのバイナル推定を行う。
  - 提案タスクの有効性検証：提案手法をフィルターとしてノイズ画像を排除した後既存手法より綺麗な3次元再構成ができた。



(a) Image pair (top) and keypoints and matches (bottom)

(b) Aligned image pair, keypoint mask, and match mask

(c) Binary classifier

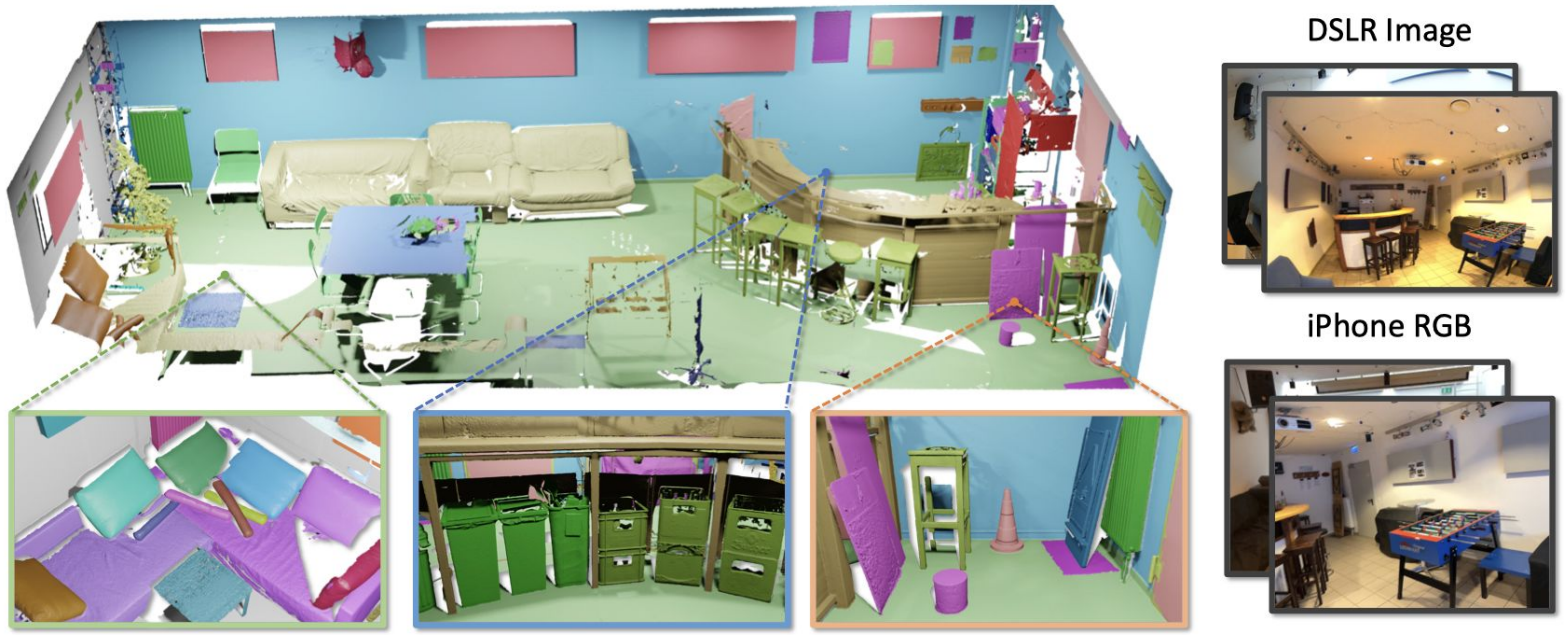
提案手法のパイプライン

https://arxiv.org/abs/2309.02420

ICCV 2023 の動向・気付き(51/165)

ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes

- 大規模high-fidelity室内3次元シーンデータセットScanNet++の提案
  - データセットの構成：460high-fidelity室内シーンとセグメンテーション付き、シーンごとのDSLR、iPhone RGB画像も含まれる。
  - データセットの有用性：Scene recognitionタスクの他、既存の3次元シーンデータセットで精度のため対応しづらいNovel View Synthesisタスクの学習・テストに活用できる。



ScanNet++の
インスタンスの例

ICCV 2023 の動向・気付き(52/165)

Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images

- HumanのCommonsenseを反するベンチマークデータセットWHOOPSの提案。
  - WHOOPSデータセット:画像中に人間のCommonsenseと反する要素(例:ビルゲーツがMacを使うなど)が含まれる画像集(合成方法で作成)とその反Commonsense要素を人間によりアノテーションするテキストデータで構築される。
  - WHOOPSでの実験:BLIP+GPT3、Ground truth caption+GPT4などの強いモデルでもWHOOPSでの性能が改善する余地があることを示した。また、画像をキャプションをテキストに変更し、それをベースにreasoningする性能が人間レベルと距離があり、画像側からのWHOOPSの認識が必要であることを示した。

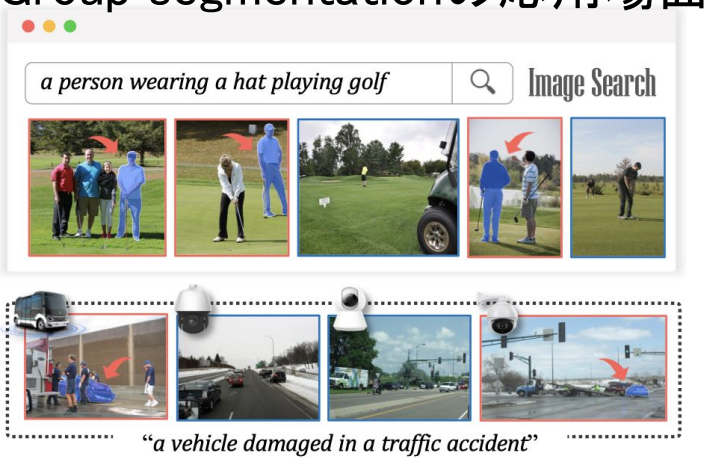
WHOOPSデータセットで行える4つのタスク

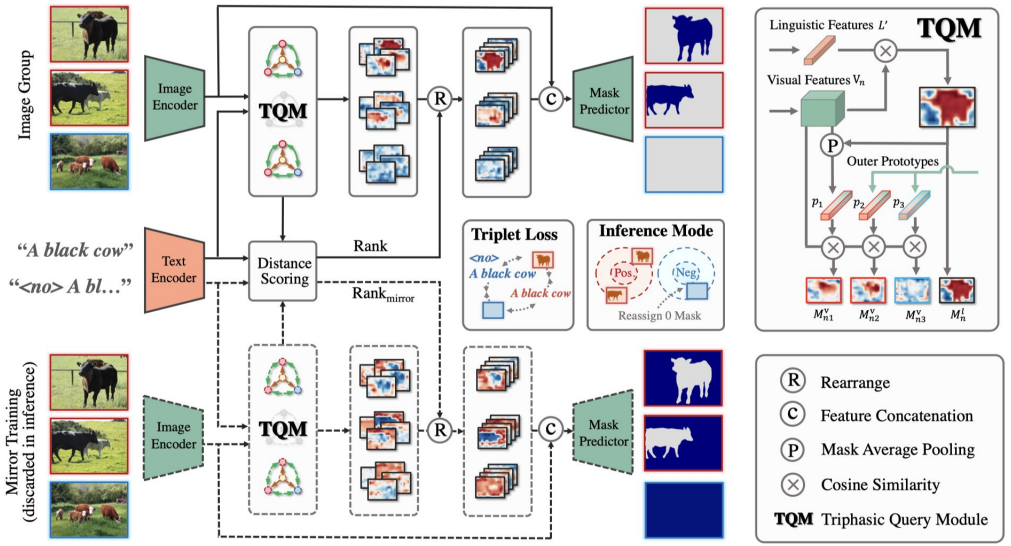
ICCV 2023 の動向・気付き（53/165）

Advancing Referring Expression Segmentation Beyond Single Image

- グループ画像と入力テキストから、入力と合う領域をセグメントするタスク・データセット・手法の提案。
  - Group Referring Expression Segmentationのメリット：実環境設定とより近い。指定の文章がない画像の対応や、同時に複数領域の対応を可能にした。Group image searchやRetrievalなどのタスクを可能にした。
  - 提案手法：基本のSOTAなセグメンテーション手法を提案タスクに適応したシンプルなベースライを構築。

Group segmentationの応用場面





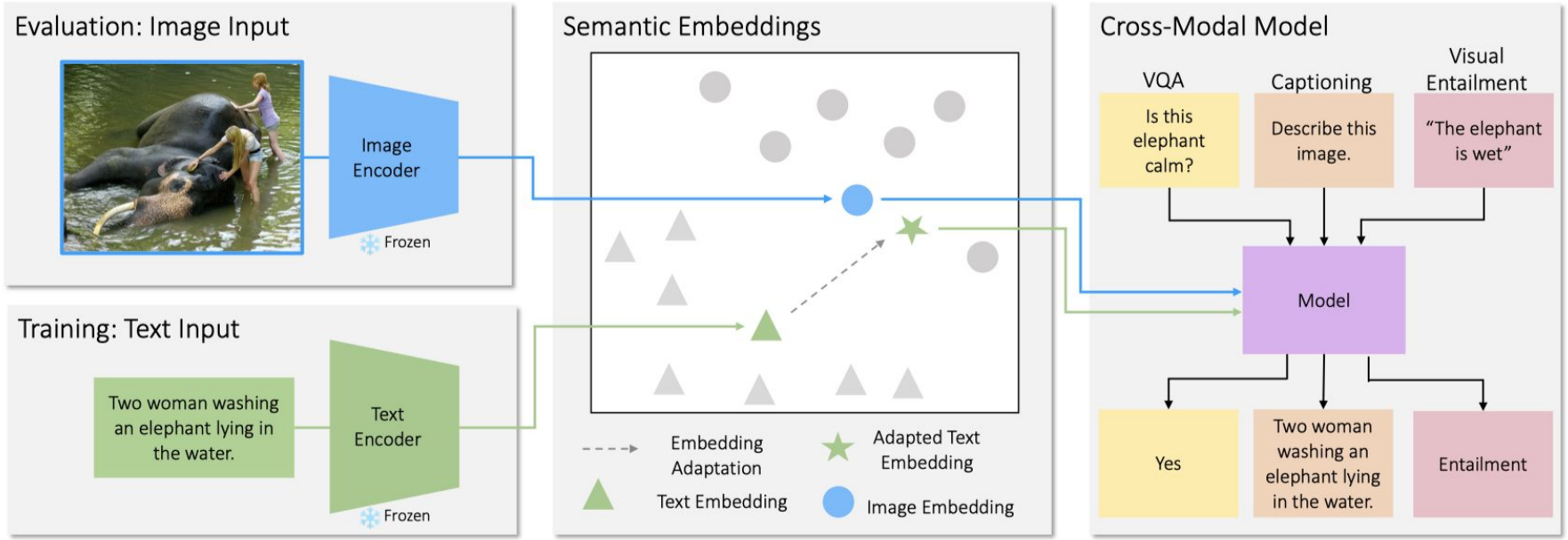
提案手法

cvpaper.challenge

ICCV 2023 の動向・気付き（54/165）

I can’t believe there’s no images! Learning Visual Tasks Using Only Language Supervision

- テキストのみの教師信号からVisual Tasksに使える画像・テキスト特徴量の提案。
  - 提案手法：CLIPの画像・テキスト特徴量をベースとする。画像とテキストの特徴量のSemantic Embeddingをリファインするため、タスクごとへのAdaptationするモデルを提案。（下図）
  - 実験結果：4つVisual タスクで高い精度を達成。



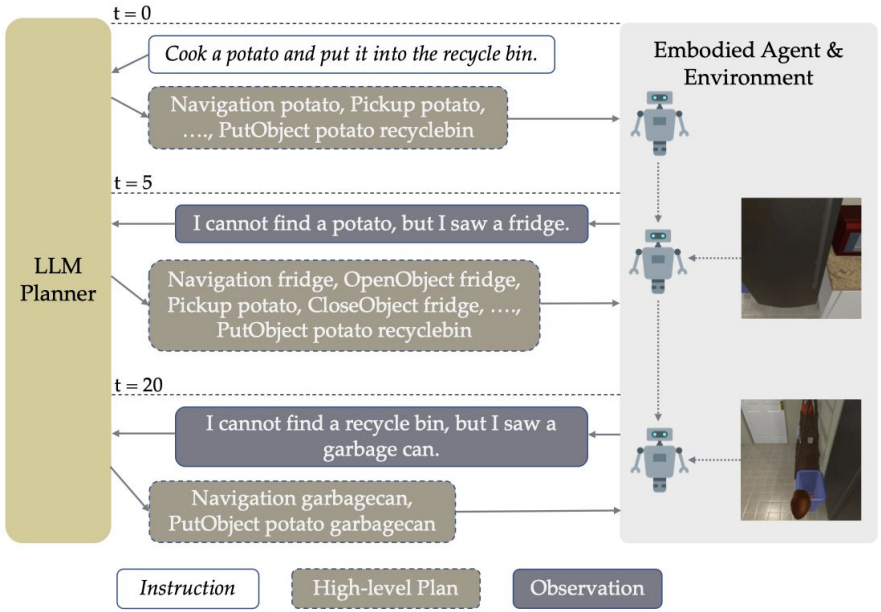
提案手法

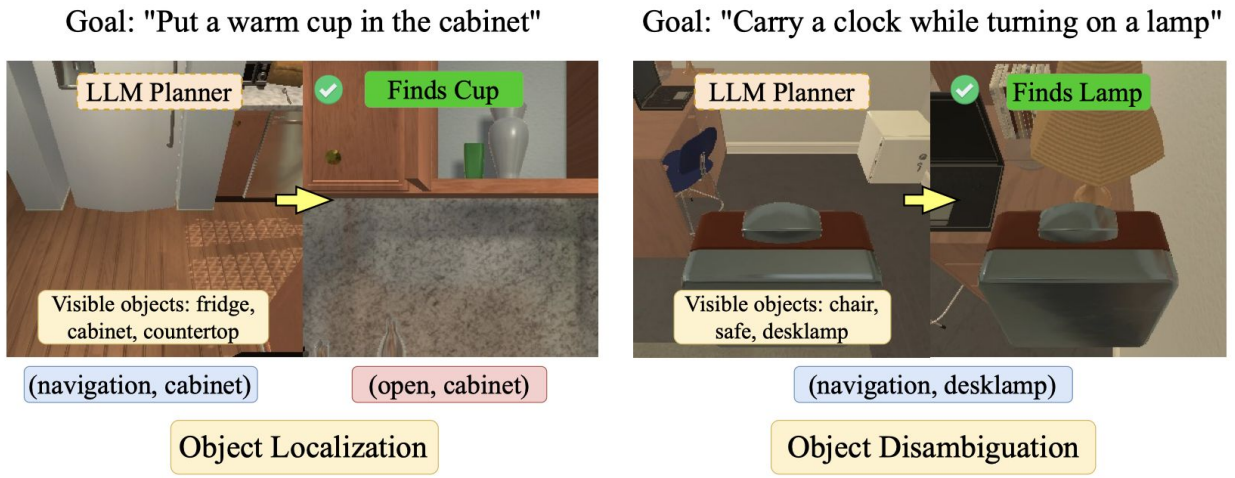
ICCV 2023 の動向・気付き（55/165）

LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models

- Embodied Agentのタスク実行でLLMで行動のPlanningを行う手法の提案。
  - 提案手法：提案手法はLLMを用いて、Embodied Agentのhigh-level planningを行う。具体的に、Agentsの行動プランを言語指示として提示する。同時にLow-level plannerも用いることで、high-level planningを実行する。Low-level plannerはLLMに依存せずに設定可能となる。更に、LLM-plannerはAgentの観測によりRe-planなどが実行可能となる。
  - 実験結果：少量な学習データでSOTAな結果を達成。

提案手法



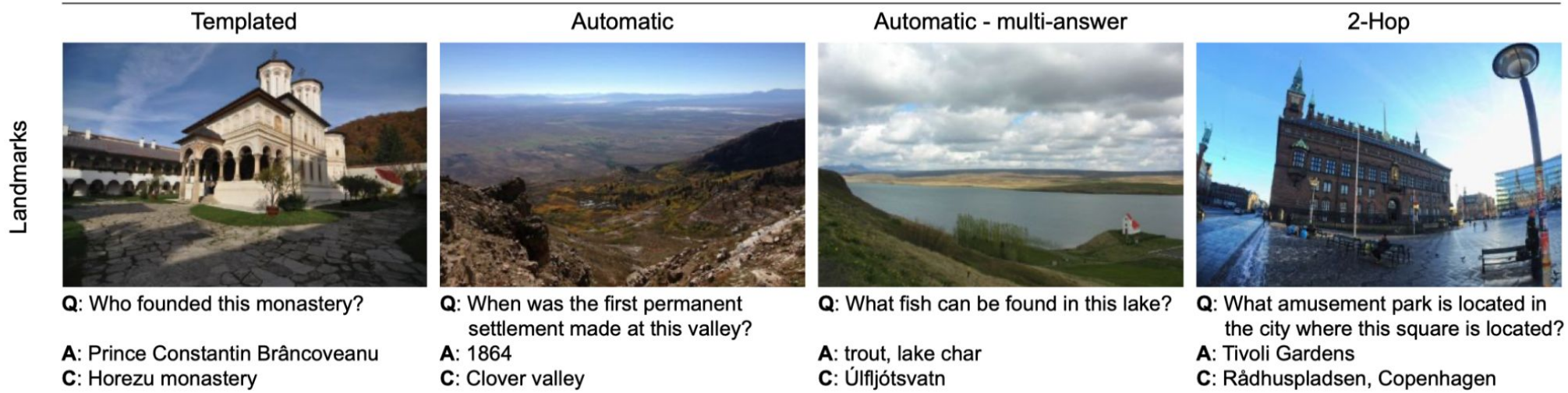


実行例

ICCV 2023 の動向・気付き(56/165)

Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories

- 地理情報、Speciesの詳細情報・他の詳細的な知識が必要となる新しいVQAデータセットEncyclopedic VQAの提案。
  - Encyclopedic VQAデータセット:データセット作成が主に自動となり、Google LandmarkとNatural Worldデータセットから画像を収集し、Wikipediaから知識情報を得る。Wikipediaのテキスト情報からQAを自動生成する。One-stepや複数の推理stepが必要なQA問題を集めた。
  - ベンチマーク実験:SOTAなLLMsと知識Retrievalモデルでも精度が50%以下となる。



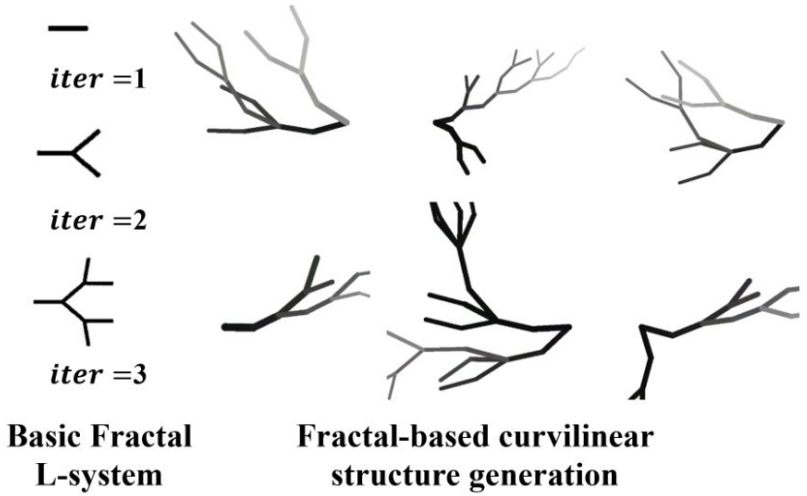
https://arxiv.org/abs/2306.09224

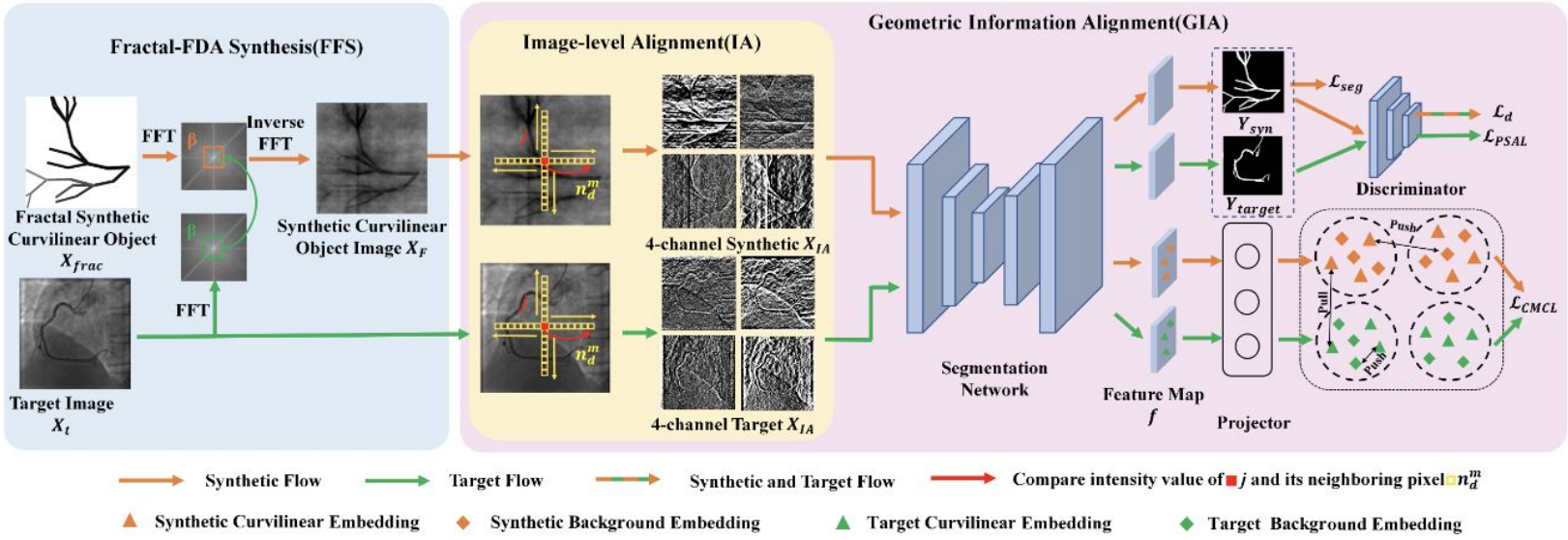
Encyclopedic VQAデータセット例

ICCV 2023 の動向・気付き（57/165）

FreeCOS (Curvilinear Object Segmentation)

- フラクタルにより枝分かれの線分を描画して医用画像の学習
  - フラクタルによる線分の合成画像生成（左図）
  - FFT/InverseFFTによる変換，学習（右図）

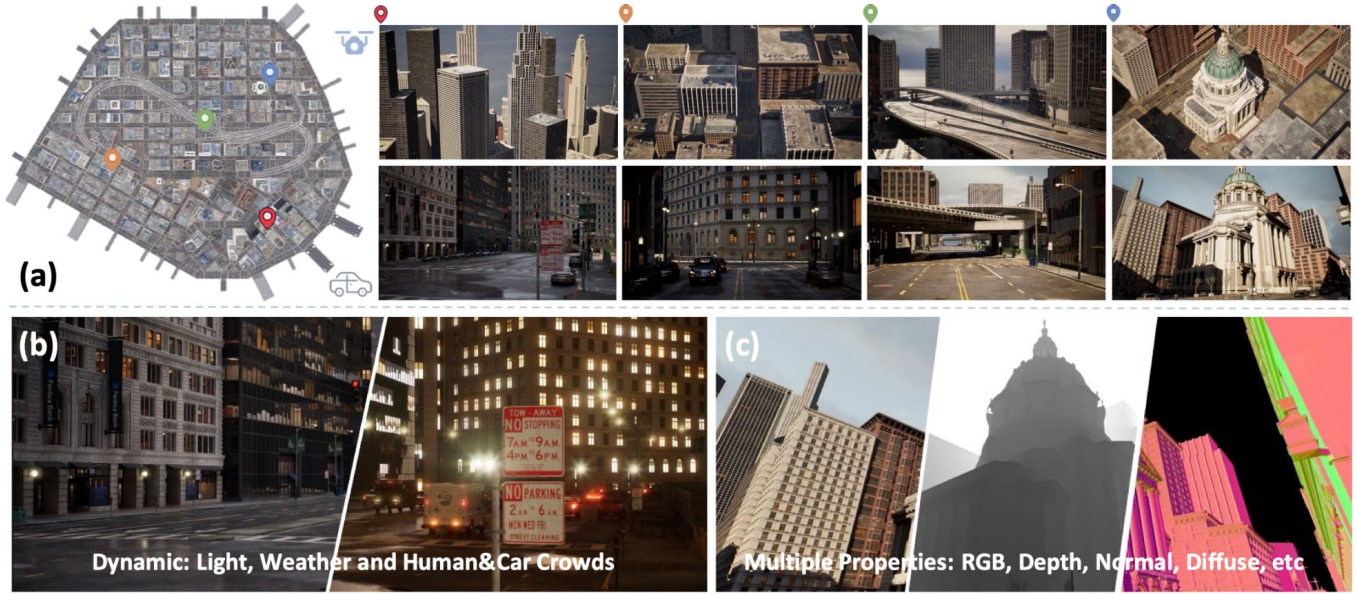




ICCV 2023 の動向・気付き（58/165）

MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond

- 屋外環境のSceneのNeural renderingタスクの学習・テストのための大規模データセットMatrixCityの提案。
  - MatrixCityデータセット：屋外環境のNeural Renderingを可能にするため、大規模の室外Synthetic sceneデータセット（2つのScene、28平方キロの範囲）を提案。MatrixCityデータセットはMultimodalityであり、天気・車・人などの変化をFlexibledで対応可能となる。
  - MatrixCityでの実験：Aerial-viewとStreet-viewデータをベースとしたNeural Renderingのベンチマーク実験により提案データセットがまだまだ難しいと示した。

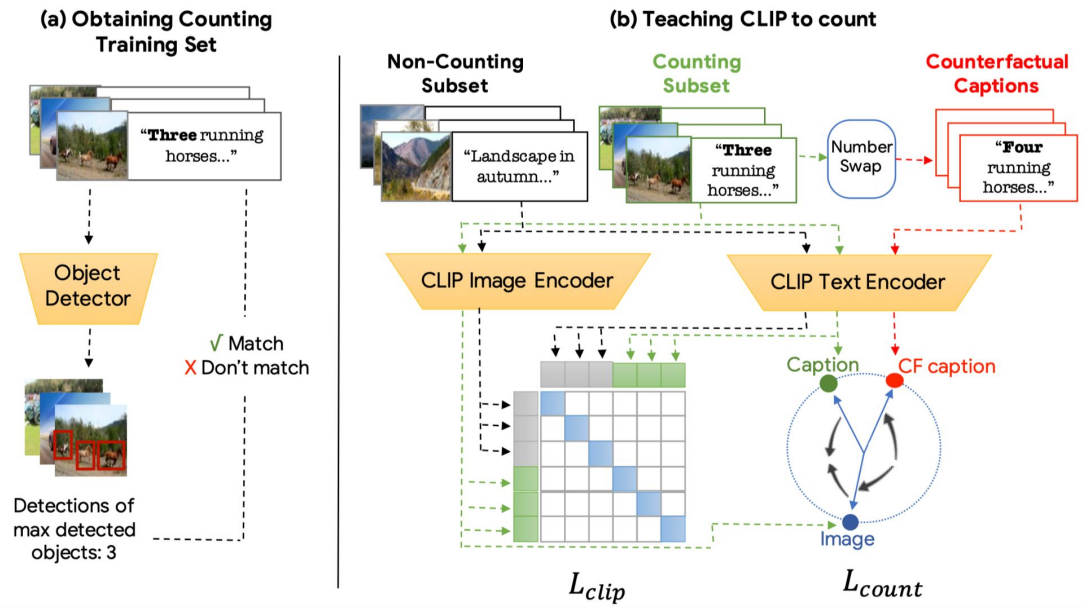


MatrixCityデータセット

ICCV 2023 の動向・気付き（59/165）

Teaching CLIP to Count to Ten

- CLIPなどのPre-trained Vision and Languageモデルの Counting能力を向上する提案。
  - 提案手法：CLIPやBASICなどのモデルの上、新しいCountingに関するContrastive Lossを追加し、Counting能力を向上させた。
  - ベース実験：まずObject counting能力を検証するための新しいBenchmarkデータセットCountBenchを提案。オリジナルCLIPやBASICなどと比べCounting能力を大幅に向上。
  - Downstreamタスクでの実験：Countingが入ったImage Retrievalや、text-to-image generationなどのタスクで提案手法の有効性を示せた。

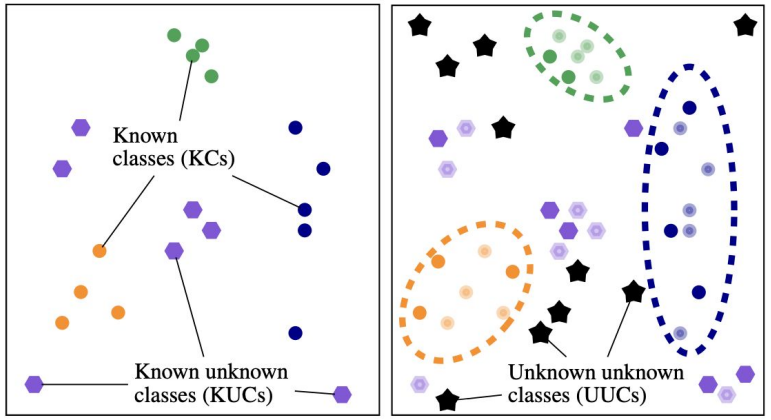


提案手法構造

ICCV 2023 の動向・気付き（60/165）

Oral発表: LORD: Leveraging Open-Set Recognition with Unknown Data

- Unknow classを用いることで，既知の領域と未知の領域を
- 分類モデルは通常，あらかじめ定義されたデータセットで訓練され，未知の特徴空間については考慮されていないため，分布外のデータの推論時に問題が生じる.
- この論文では，LORDというフレームワークを提案し，**訓練中に未知の空間を明示的にモデリング**することで，**オープンセット認識(OSR)の性能を向上**させる方法を検討する。
- また，背景データの依存を軽減するためのデータ生成技術としてmixupの効果を探ると、mixupが背景データの代替として効果的であることが実験で示される　.



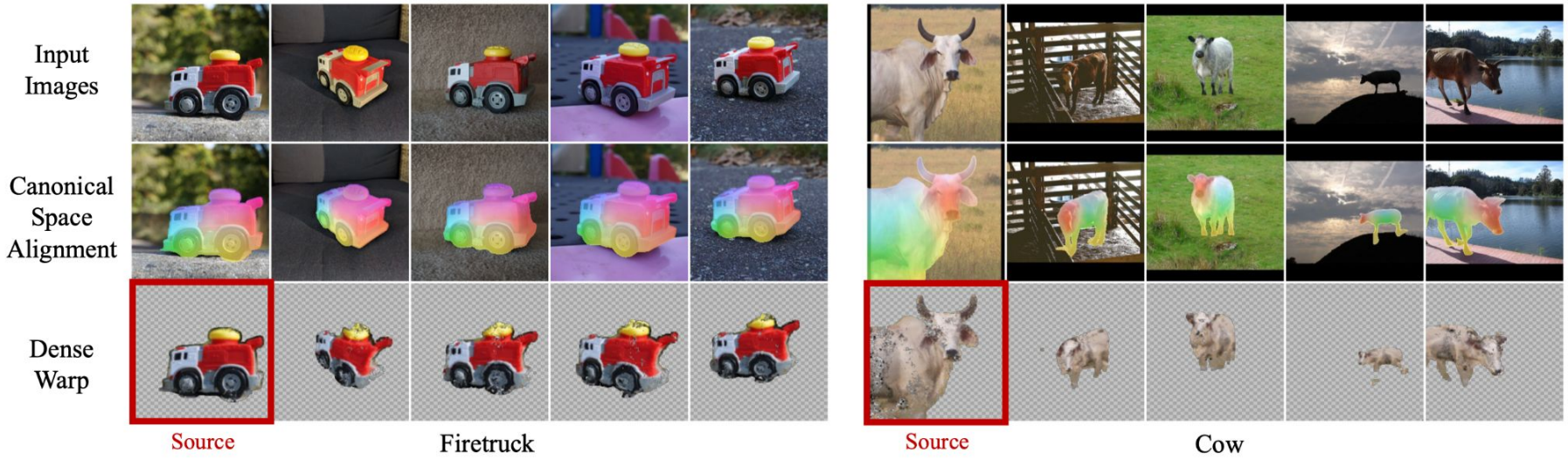
(a) Training data (opaque). (b) Test data (opaque) with training data (shaded).

Figure 1. Overview of data types in open-set recognition. The training set in (a) includes known classes (KCs) and known unknown classes (KUCs) (⬢). The trained classifiers in (b) model decision boundaries for KCs as dashed ellipses. KUCs correlate with the training set's KUCs, exhibiting higher identifiability in comparison to the unknown unknown classes (UUCs) (⬟), which can exist anywhere in the feature space.

ICCV 2023 の動向・気付き(61/165)

ASIC: Aligning Sparse in-the-wild Image Collections

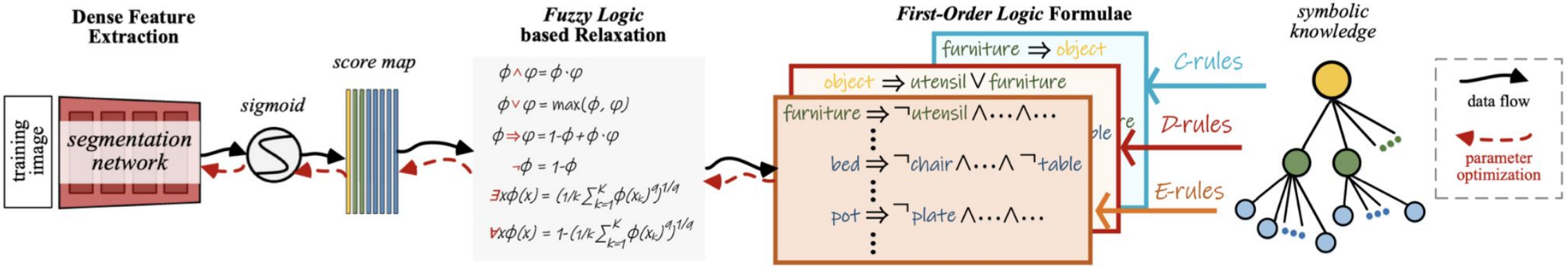
- 指定物体のグループ画像から物体のCanonical Space Alignmentを行う手法の提案。
  - 提案手法:まずViTなどのPretrainedエンコーダーで疎なKeypoint Alignmentを行う。また、グループの画像にJointlyでAlignmentを行う。次に、Canonical Gridを用いて、疎なAlgimentを統一する。Geometry equivariantロスを使用した。
  - 既存手法と比べる際のメリット:Keypointなどのアノテーションを用いずにself-supervisedで行える;大規模学習データを用いずにSparseな学習データから学習可能
  - 実験:CUB と SPair-71kデータセットで既存のself-supervisedより高質な結果を得た。



ASICの結果例

LOGICSEG: Parsing Visual Semantics with Neural Logic Learning and Reasoning

- Symbolic reasoningが可能な新たなVisual Semantic Parsing手法の提案。
  - 既存のSemantic Parsingでは、物体カテゴリの階層関係の理解がある程度できるが、その階層関係をベースとしたSymbolic reasoningの検討がこれまでにあまりされてこなかった。
  - 提案手法：First-Order Logic FormulateなどをSemantic parsingモデルの上に設定し、厳格的なDisentangled Semanticの認識や、それをベースとしたSymbolic Reasoningを可能にした。
  - 実験：提案のLOGICSEGを既存のSOTAな手法に追加することで、4つの既存Semantic Segmentationデータセットで最も高い精度を達成。その上、segmentation以外に、Logical操作なども可能にした。

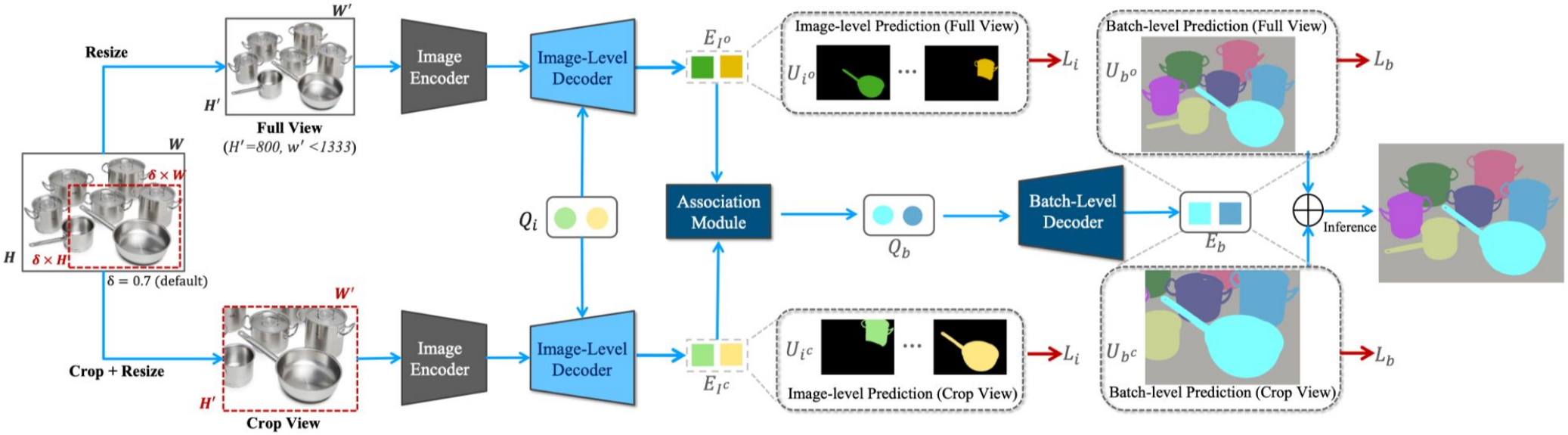


提案手法

ICCV 2023 の動向・気付き（63/165）

High-Quality Entity Segmentation

- 高解像度・精度を持ったsegmentationデータセットEntitySegと手法の提案。
  - EntitySegデータセット：33k高解像度（avg. 2700.7）で物体の細かいエッジや物体間の境目を高精度のアノテーションが寄与されている。
  - 提案手法：解像度を下げた視点と元の画像からCropされた画像のセグメンテーション及びBatchレベルの処理（物体ごとに処理）を結合し、計算コストを抑えながら高精度実現。
  - 実験結果：既存の様々なsegmentationデータセットより画像が少ないが、提案データセットでsegmentationの学習で既存の他の大規模データセット以上の精度向上を示せた。

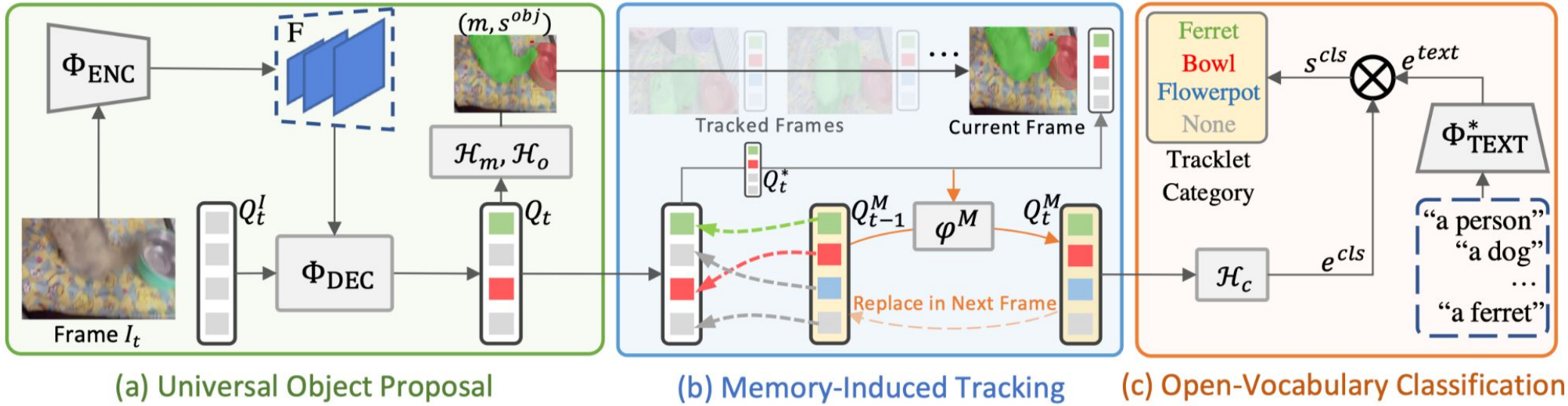


提案手法

ICCV 2023 の動向・気付き（64/165）

Towards Open-Vocabulary Video Instance Segmentation

- Open-Vocab Video Instance Segmentationの問題定義・データセット・手法の提案。
  - データセット：4,828ビデオ(6.8 hours)、1,196物体カテゴリ（レアなカテゴリを含め）から構成される。テストセットには学習データに含まれていないカテゴリも使用。
  - 手法：Open-vocab VISタスクの初めてのベースライン手法を提案。提案手法がまず単独のフレームに対してセグメンテーションを行い、Memory Induced Trackingで各フレーム間の関連付けをする。
  - 実験：提案手法OV2Segがテスト時にZero-shotで良い精度で物体を追跡・セグメンテーションできる結果を示した。

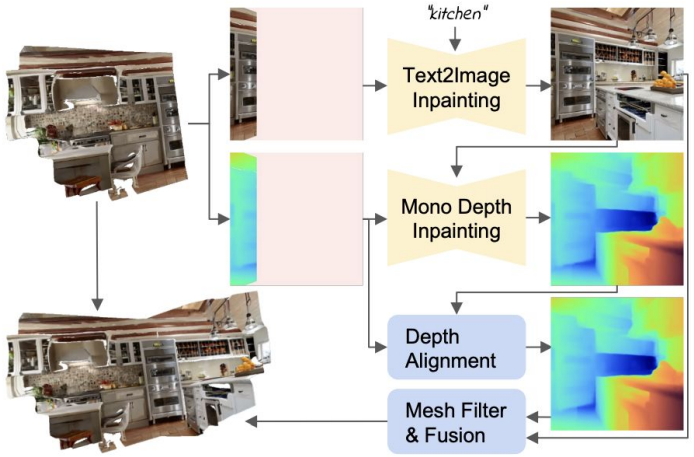


提案手法

ICCV 2023 の動向・気付き(65/165)

Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models

- Sceneを記述するテキストから3次元メッシュを生成する手法の提案。
  - 提案手法:手法全体はテキストから画像生成とデプスマップ推定をコンバインとなる。また、生成プロセスはマルチステップとなる。まずテキストから画像とデプスを別々で推定し、その後Inpaintingを利用し連続の画像・デプスを推定(コア部分、複数画像・デプス画像間の幾何的一致性を考慮している);最後に、全体的なシーンの結果を最適化する。
  - 既存手法との比較:既存手法はシーンレベルの生成を適応できるものがほぼなかった。Text2Roomはテキストのみから、良い精度で3次元シーンを生成できる手法となる。今後の大規模3次元コンテンツ生成に良い方向性を示せた。



Iterative scene generation

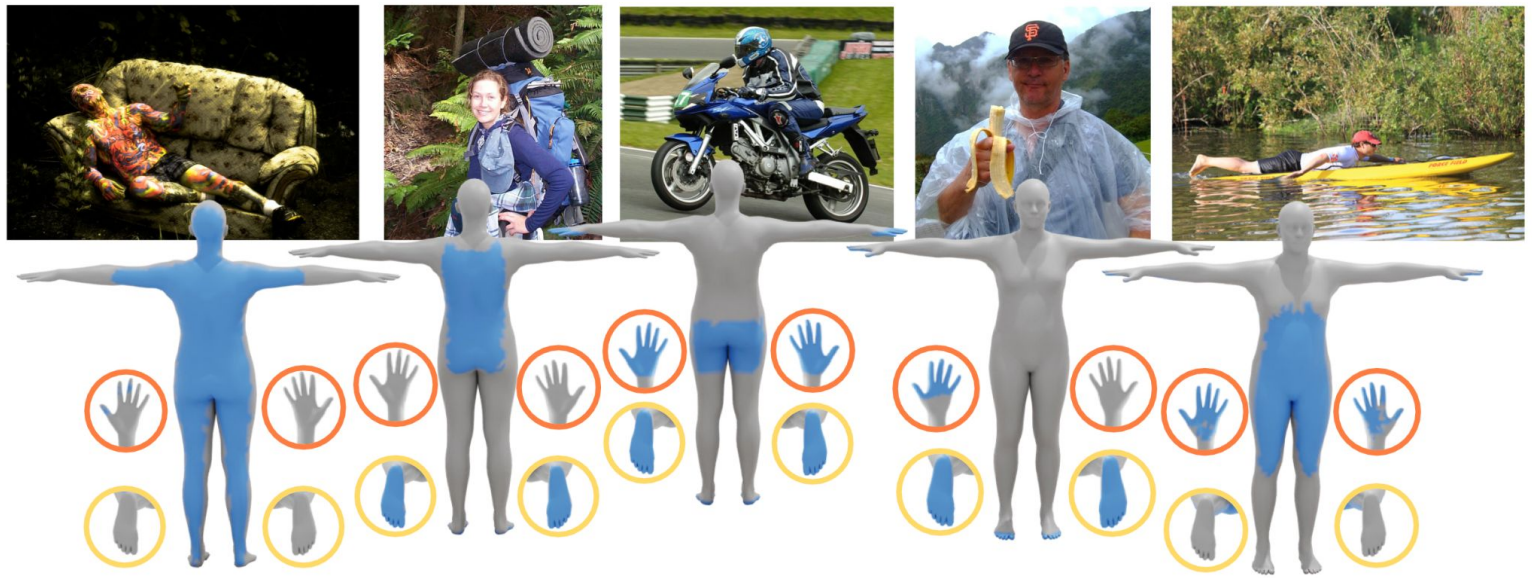
Editorial Style Photo, Coastal Bathroom, Clawfoot Tub, Seashell, Wicker, Mosaic Tile, Blue and White

結果例

ICCV 2023 の動向・気付き(66/165)

DECO: Dense Estimation of 3D Human-Scene Contact In The Wild

- In-the-wildでhuman-object/sceneの3次元接触情報を推定する手法・データの提案。
  - データセット提案:AMTで手動でアノテーションした大規模データセットDAMONを提案。リアル画像とその画像中のHuman Sceneの3次元コンタクト情報が記録されている。データセットには5,522枚画像とそれらのアノテーションから構成される。
  - 提案手法:提案手法がbody partとシーンコンテキストの二つの情報を融合し人間の身体構造と画像中の情報から推定を行う。提案のDAMONデータセットとRICH、BEHAVEなどの既存データセットで既存手法より大幅に精度を向上した。

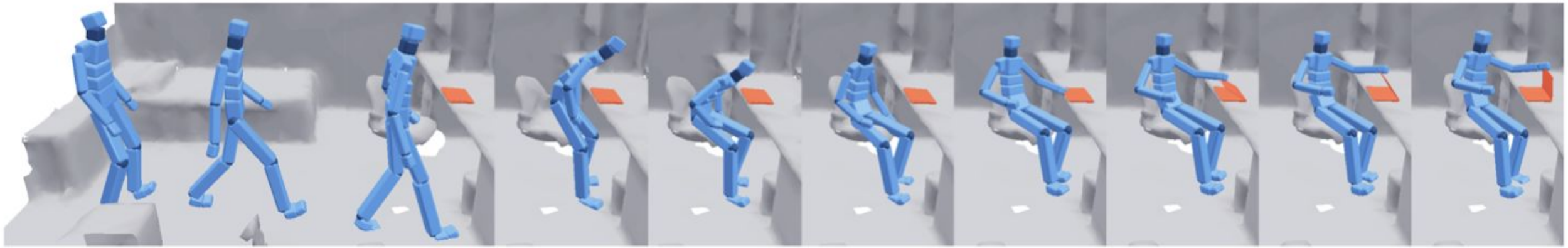


DECOの出力例

ICCV 2023 の動向・気付き（67/165）

Locomotion-Action-Manipulation: Synthesizing Human-Scene Interactions in Complex 3D Environments

- 画像、最初のポーズ、3次元シーンからLong-term行動のモーションを自動生成する手法・タスクを提案。
  - 提案手法の特徴：学習する際にMotion Datasetのみ用意し、テストする際に様々な3次元環境で動作するモーションの生成が可能。短いモーションデータを組み合わせて最適化する方法となり、Long-term motionの学習データが必要ない。
  - 提案手法LAMAの構成：最適化するためにモーションマッチングを提案した。また、強化学習により最適化プロセスの学習を行う。また、manifold learningをベースとしたモーション編集フレームワークも導入した。

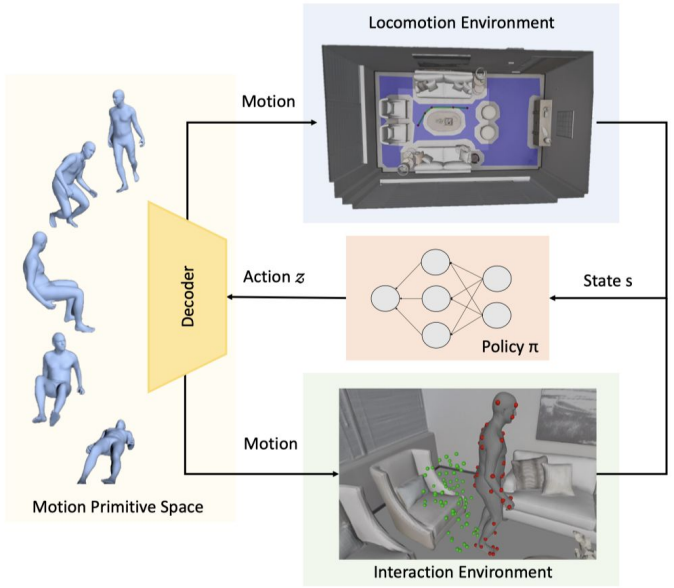


提案手法で生成したモーションの例

ICCV 2023 の動向・気付き(68/165)

Synthesizing Diverse Human Motions in 3D Indoor Scenes

- 3次元環境から、高精度のlong-form human motion (移動、シーン・物体とInteractionなど)を生成する手法の提案。
  - 提案手法:全ページのLAMAと類似するように、ハイコストのlong-form motionの学習データを頼らずに、Motionのコントロールをベースとした強化学習フレームワークにより、3次元空間との関係性を考慮した複数動作の序列が生成可能である。
  - 感想:3次元シーンの生成、テキスト生成とコンバインして、3次元シーンでの日常活動生成が可能になりそう。



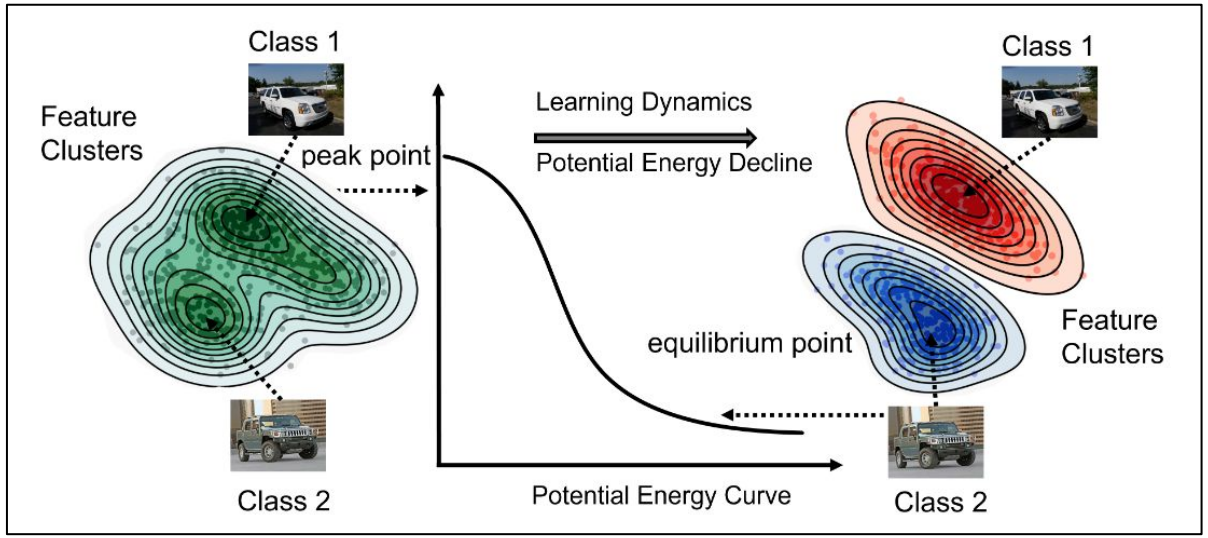
提案手法

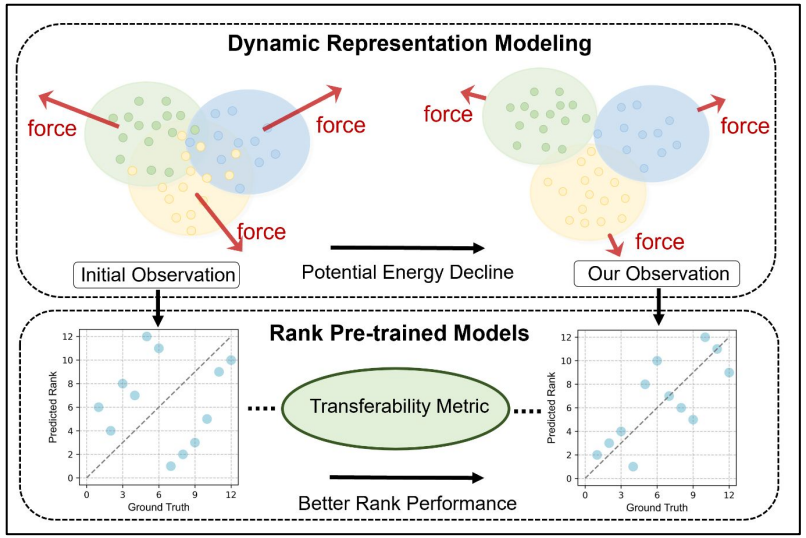
結果例

ICCV 2023 の動向・気付き（69/165）

Exploring Model Transferability through the Lens of Potential Energy

- 物理学のポテンシャルエネルギーに着想を得た、新たなTransferabilty計測アプローチである**PED(Potential Energy Decline)**を提案
- ダウンストリームタスクのデータから得られる**特徴量を、物理学のアナロジーを用いて更新**する
- PEDを用いて得られる特徴量と既存手法と組み合わせることで、多くのダウンストリームタスクで**高い予測性能**を発揮



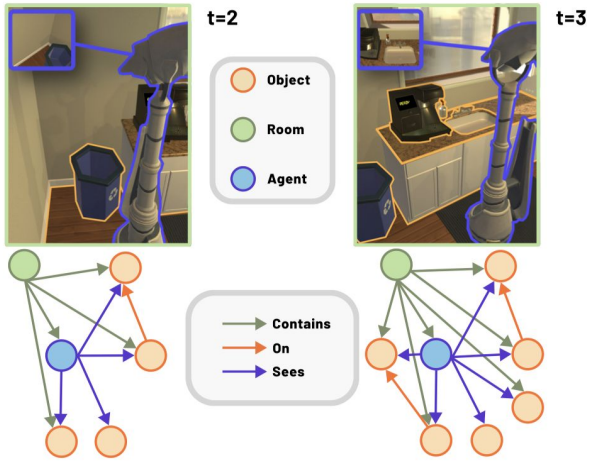


https://openaccess.thecvf.com/content/ICCV2023/html/Li_Exploring_Model_Transferability_through_the_Lens_of_Potential_Energy_ICCV_2023_paper.html

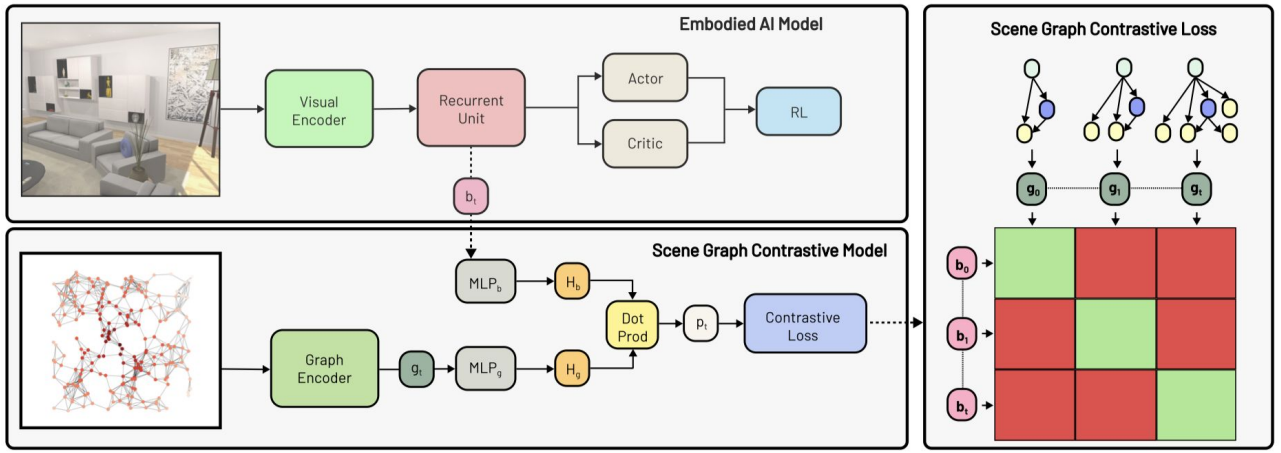
ICCV 2023 の動向・気付き（70/165）

Scene Graph Contrastive Learning for Embodied Navigation

- 様々なEmbodied AIタスクに容易に適応できるScene Graph Contrastive Lossの提案。
  - 提案手法：提案手法が、学習時のみScene GraphのContrastive Lossで学習させる。Embodied Navigationタスクなどを実行しながら、Agentが環境に対しての詳細理解（部屋、物体、部屋物体関係、物体物体関係、Agentと物体/部屋の関係（すでに見たかどうか））をSGC Lossで学習させる。テスト時に、Scene Graphの推定をせずにタスクを実行する。
  - 実験：Object Navigation, Multi-Object Navigation, と Arm Point Navigationの3つのEmbodied Navigationタスクで提案のSGCロスを導入することでSOTAを達成。



Iterative Graph Building



提案手法

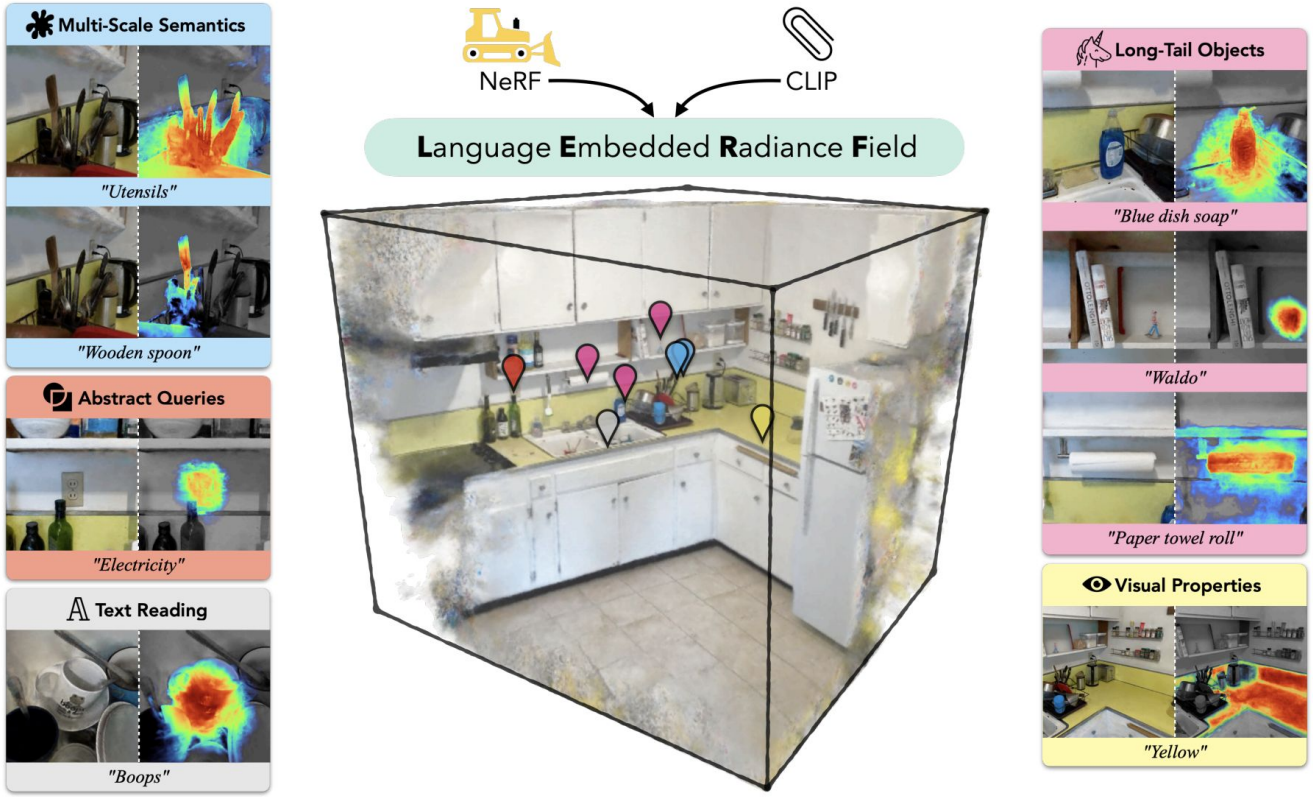
ICCV 2023 の動向・気付き(71/165)

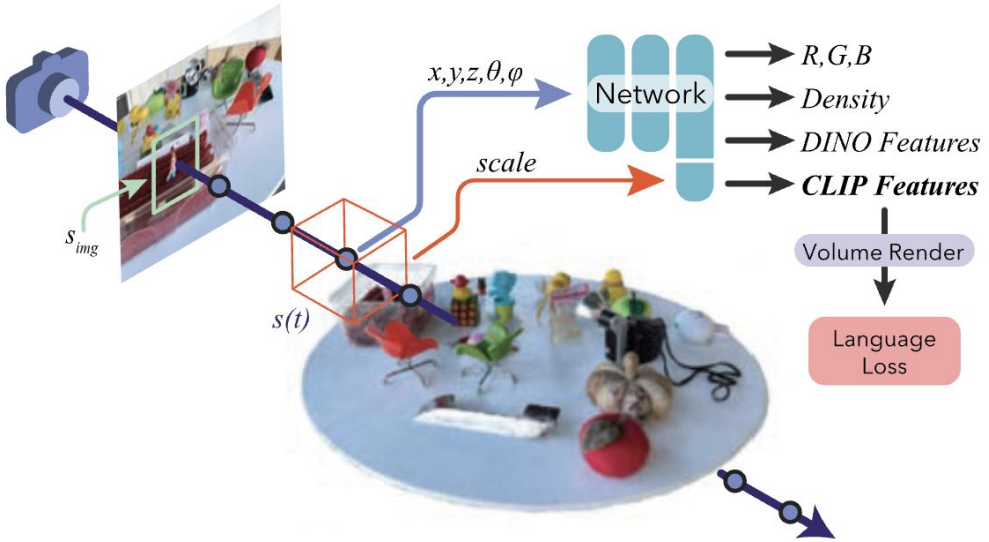
LERF: Language Embedded Radiance Fields

- 3D(NeRF)×CLIP

  NeRFにCLIPの言語特徴量を埋め込む

  - Radiance Field との自然言語によるインタラクションが可能
  - CLIPを使用することで、幅広いクエリに対応可能(抽象的クエリ,視覚的クエリ,珍しい単語)



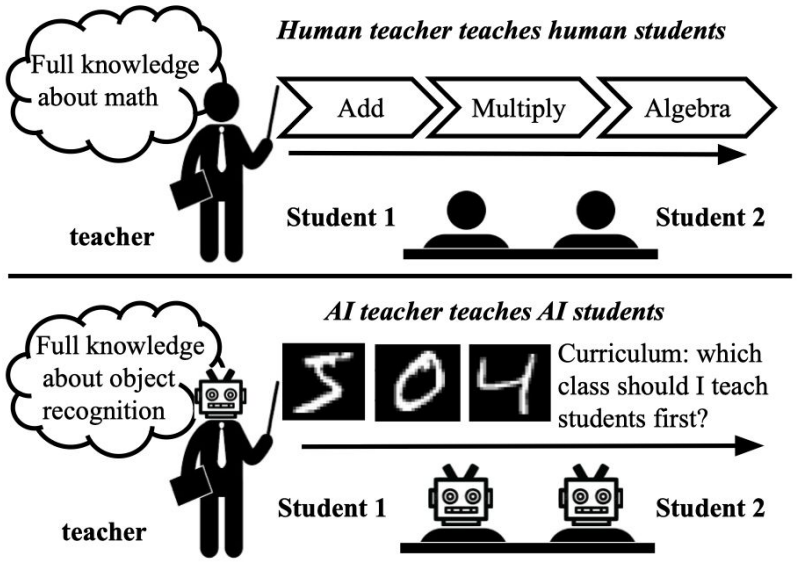


LERF: Language Embedded Radiance Fields, Justin Kerr et al.,arXiv:2303.09553, 2023

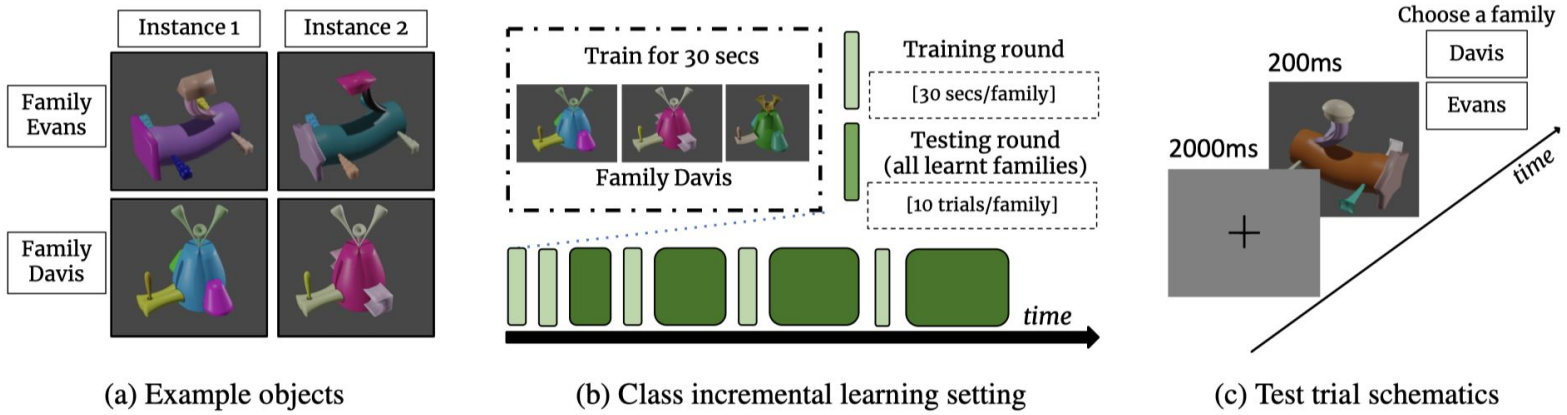
ICCV 2023 の動向・気付き（72/165）

Learning to Learn: How to Continuously Teach Humans and Machines

- Curricula（タスクを学習する順番）が人間とAIのOnline Class Incremental Learningにおける影響の分析及び自動的に最適Curricula生成の実現に向ける手法の提案。
  - 実験分析：人間とAI両方に関してOnline Class Incremental Learning（毎回一つのみのタスクを学習）において、各クラス／タスクが学習効果（精度と忘却の程度）に大きな影響を与えることを実験で示した。
  - 手法：クラス間の類似性によりクラスの学習順番を決めるCurriculum Designerを提案した。



curricula



Human実験者におけるClass incremental Learning設定

ICCV 2023 の動向・気付き（73/165）

Equivariant Similarity for Vision-Language Foundation Models

- Vision-language modelsのequivarianceを検討する提案。また、既存のVLMsはequivarianceへの対応が弱い、equivarianceに対応可能な新たなロスも提案。
- Regularizationロス：画像ペア間の相互への類似性スコア（下図1）とそのペアのテキストの相互への類似性スコア（下図2）が同一になるように学習する。

$$s_{11} - s_{12} = \underbrace{\sum_{T_1}^{T_2} \mu(T)}, \quad s_{22} - s_{21} = \underbrace{\sum_{T_2}^{T_1} \mu(T)}, \qquad (1)$$

Semantic Change Measured by Text Change

$$s_{11} - s_{21} = \underbrace{\sum_{I_1}^{I_2} \mu(I)}, \quad s_{22} - s_{12} = \underbrace{\sum_{I_2}^{I_1} \mu(I)}, \qquad (2)$$

Semantic Change Measured by Image Change

ICCV 2023 の動向・気付き（74/165）

OmniLabel: A Challenging Benchmark for Language-Based Object Detection

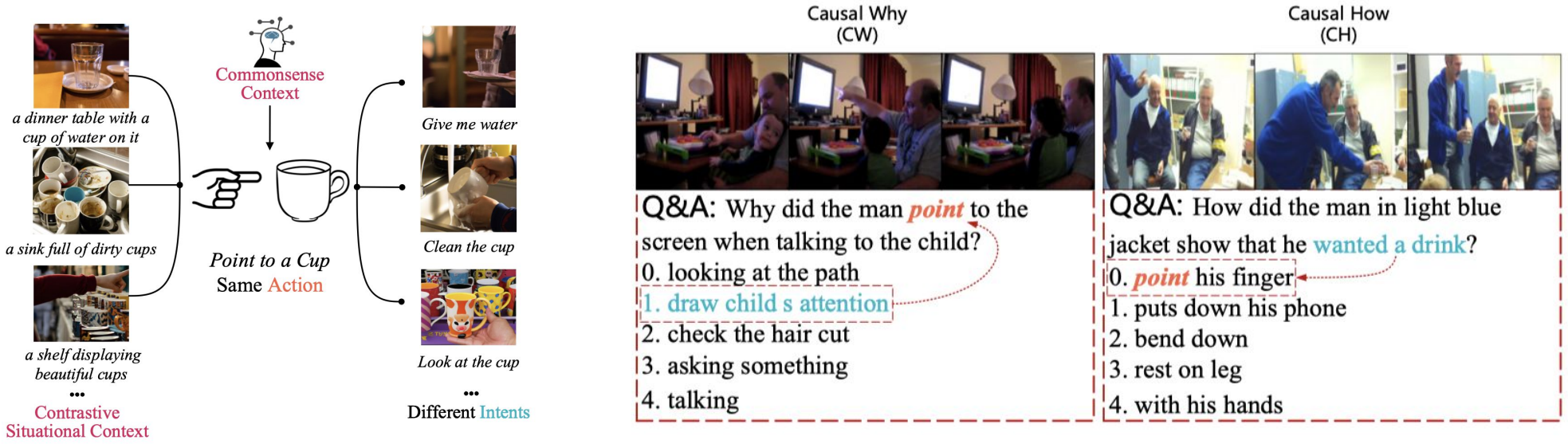
- Referring Expression、Open-vocabulary detectionなどの言語ベースDetectionタスクを網羅した新しいタスク設定、ベンチマークデータセットOmniLabelを提案。
- OmniLabelのインスタンスは、一つのセンテンスと画像から構成され、そのセンテンスと画像中の0、1、1以上の複数の物体領域と対応する。データセットは人間よりアノテーションし、合計28Kの説明文と25K画像から構成される。
- 既存手法がOmniLabelにおいて既存のデータセットより劣る性能となり、OmniLabelが今後Object detectionの重要なベンチマークとして活用できる。

OmniLabelの概要図、既存データセットとの比較、データセット例

ICCV 2023 の動向・気付き（75/165）

IntentQA: Context-aware Video Intent Reasoning

- ビデオから、Intentを推定するタスクとベンチマークデータセットを提案。
- データセットには4つのタイプ（Causal Why、Causal How、Temporal Next、Temporal Previous）の質問が人間のアノテーターによりラベリングされる。
- また、3つのモジュール（Situational Reasoning、Contrastive Reasoning、Commonsense Reasoning）から構成するベンチマーク手法を提案し、既存手法よりIntentQAで高い精度を実現。

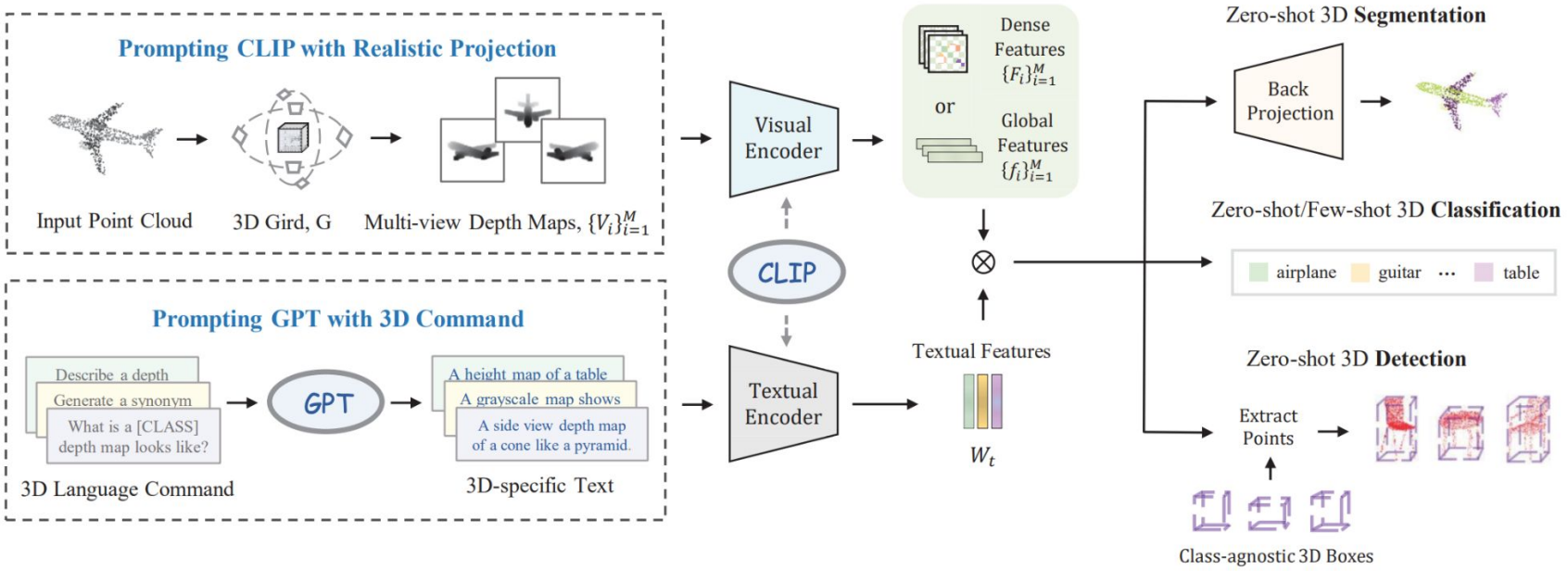


IntentQAデータセットのデザイン、データセット例

ICCV 2023 の動向・気付き（76/165）

PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-world Learning

- 3D（点群）×CLIP
  - 点群からよりリアルな”2D Depth Map”を生成、CLIPを活用することで、3D点群分類を行う
    - 点群を一度ボクセルにしてからDepth Mapを作成することで、よりリアルなDepth Mapを作成
    - 分類の際CLIPのテキストエンコーダに入力するテキストをGPTで工夫することによって、豊かな3Dセマンティクスを持つテキストを生成

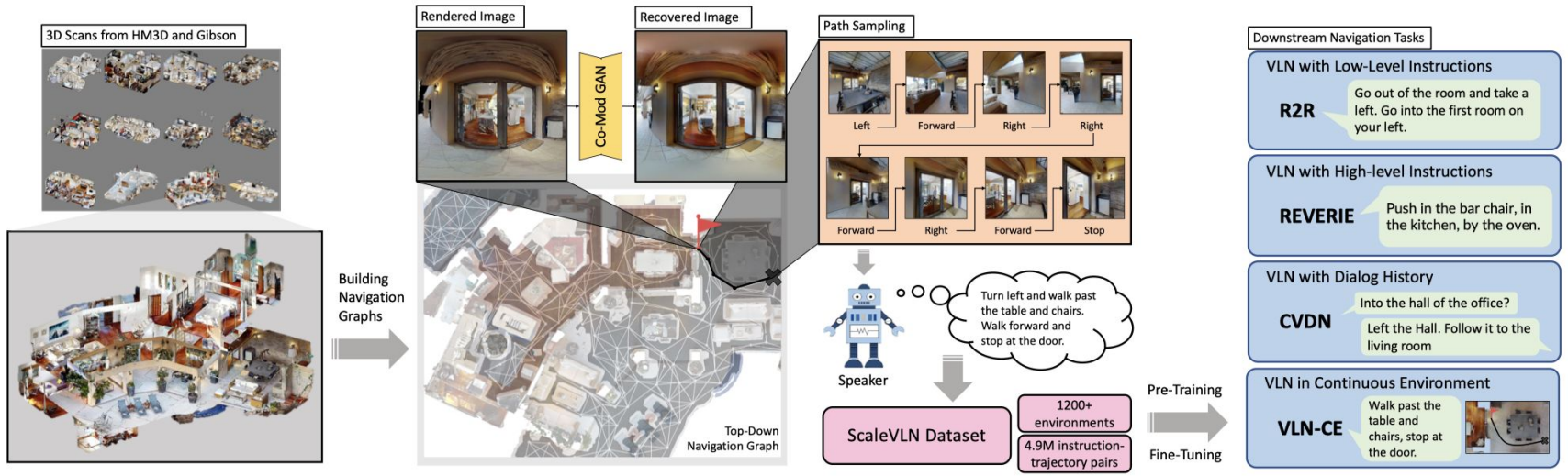


PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-world Learning, Xiangyang Zhu et al.,arXiv:2211.11682,2023

ICCV 2023 の動向・気付き（77/165）

Scaling Data Generation in Vision-and-Language Navigation

- Embodied VLNの学習・評価のための新たなベンチマークデータセットScaleVLNを提案。ScaleVLNは既存のデータセットより遥かにデータセットの規模が大きい（1200＋のシーンと4.9MのInstructions）。
- ScaleVLNは既存のデータセットをベースに自動生成され、同じ仕組みで更にデータセットの規模を向上することが可能。
- ScaleVLNデータセットの学習し他のVLNベンチマークで最大11%のSuccess Rateを向上した。また、実験でシーンとInstructionの数両方の重要度をそれぞれ示した。

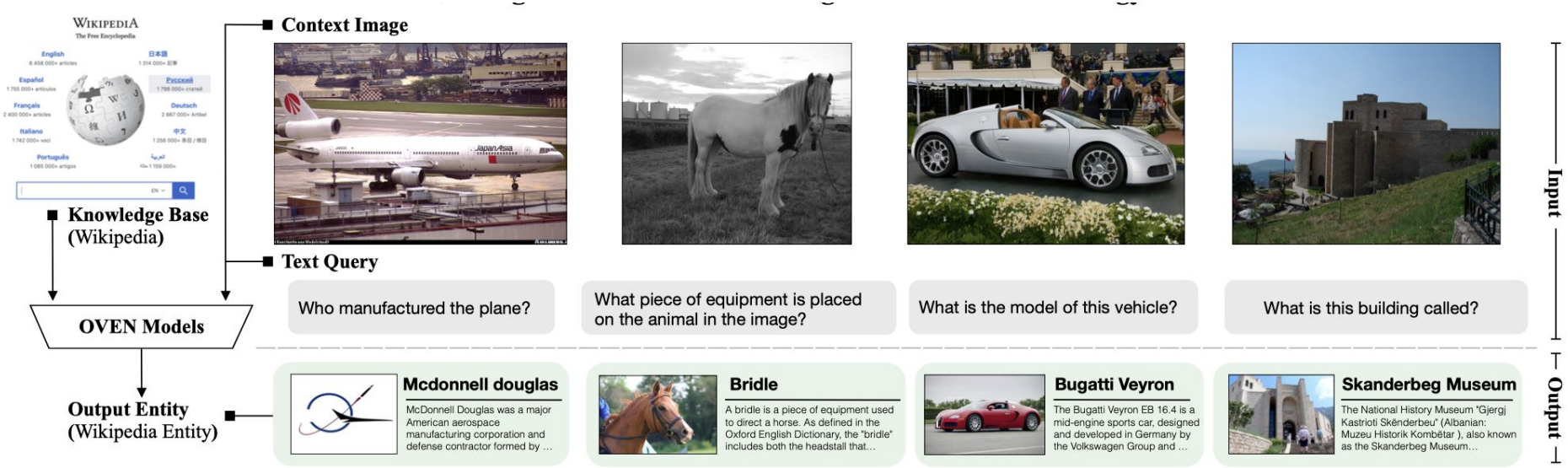


ScaleVLNの説明図

ICCV 2023 の動向・気付き（78/165）

Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities

- 画像とWikipedia Entityを関連し、Wikipediaで載せている詳細的なEntity情報をベースとした画像理解を行うタスクOVENを提案。
- 14種類の既存データセットとWikipedia Entityを関連させたデータセットOVEN-Wikiも提案。OVEN-Wikiでは6MのWikipedia Entityと関連し、認識モデルがそれらを区別できるようなFine-grained entity知識を認識する必要がある。

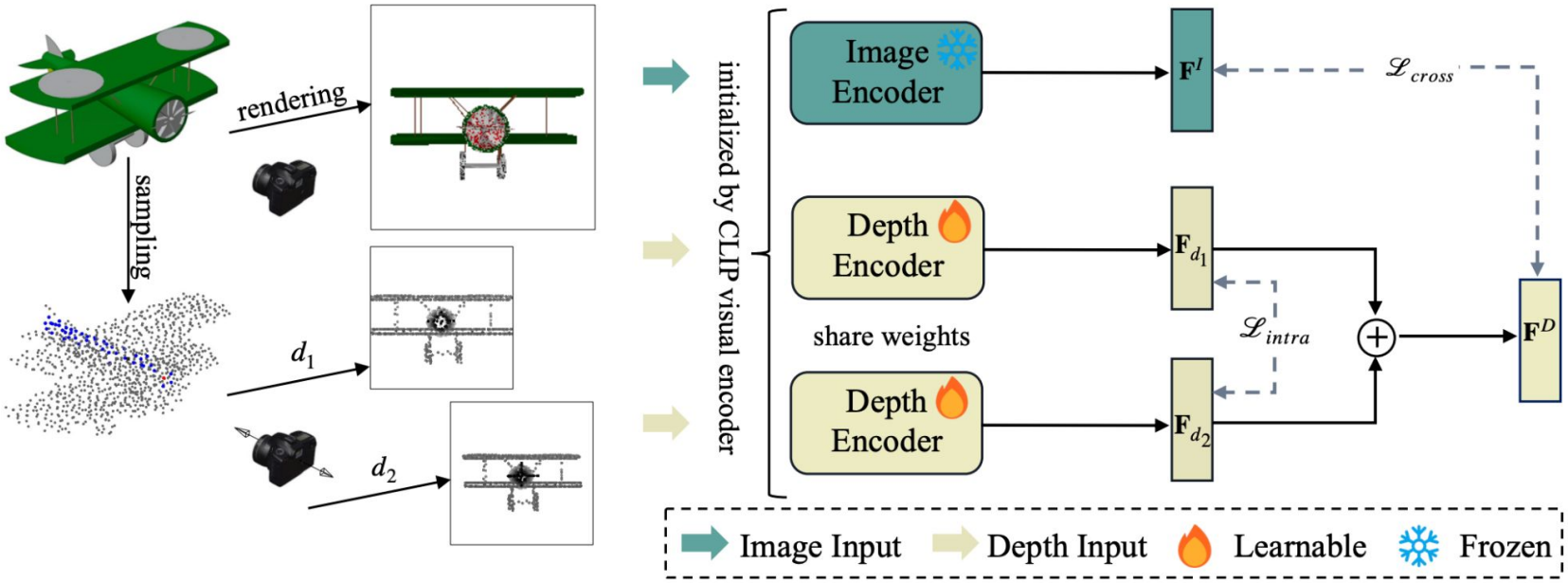


OVENタスクの説明図

ICCV 2023 の動向・気付き（79/165）

CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training

- 3D（点群）×CLIP
  - 点群を”2D Depth Map”に投影し、CLIPを活用することで、3D点群分類を行う
    - 既存研究は投影されたDepth Mapに直接CLIP画像エンコーダーを適用するもドメインギャップが存在していた
    - Depth Mapと単純な2D画像とのドメインギャップを埋めるためcontrastive learning
    - ゼロショット点群点群分類の精度が向上

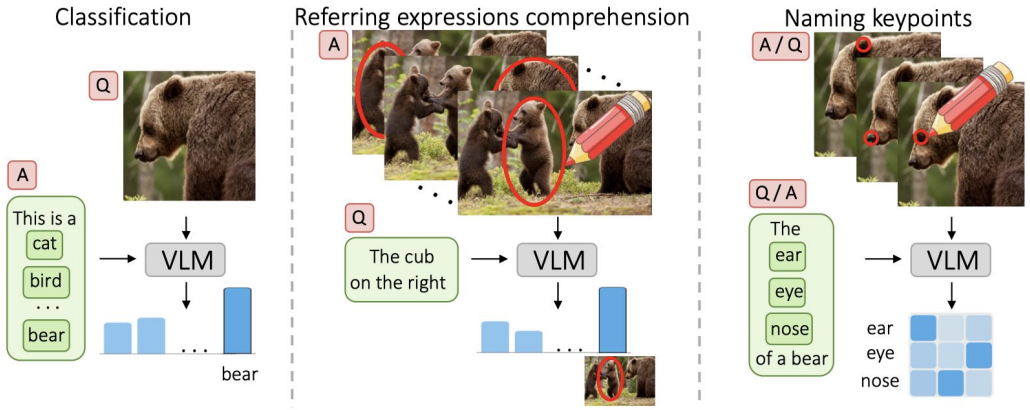


CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training, Tianyu Huang et al.,arXiv:2210.01055,2023

ICCV 2023 の動向・気付き（80/165）

What does CLIP know about a red circle? Visual pr mpt engineering for VLMs

- CLIPの性能を向上させる（特にAttentionの改善）の新しいPrompt手法の提案。
- 提案手法が画像中の赤い円で指定されている領域と指定のテキスト（Referring Expressionや物体のパーツなど）と関連させるようにPromptingを行う。提案のPromptを使用して、既存のCLIPの性能を向上させた。
- また、データセットの分析から、赤い円が特別に有用であることの原因を明らかにした。（YFCC15Mなどのデータセットでは一部赤い円で領域指定されている）

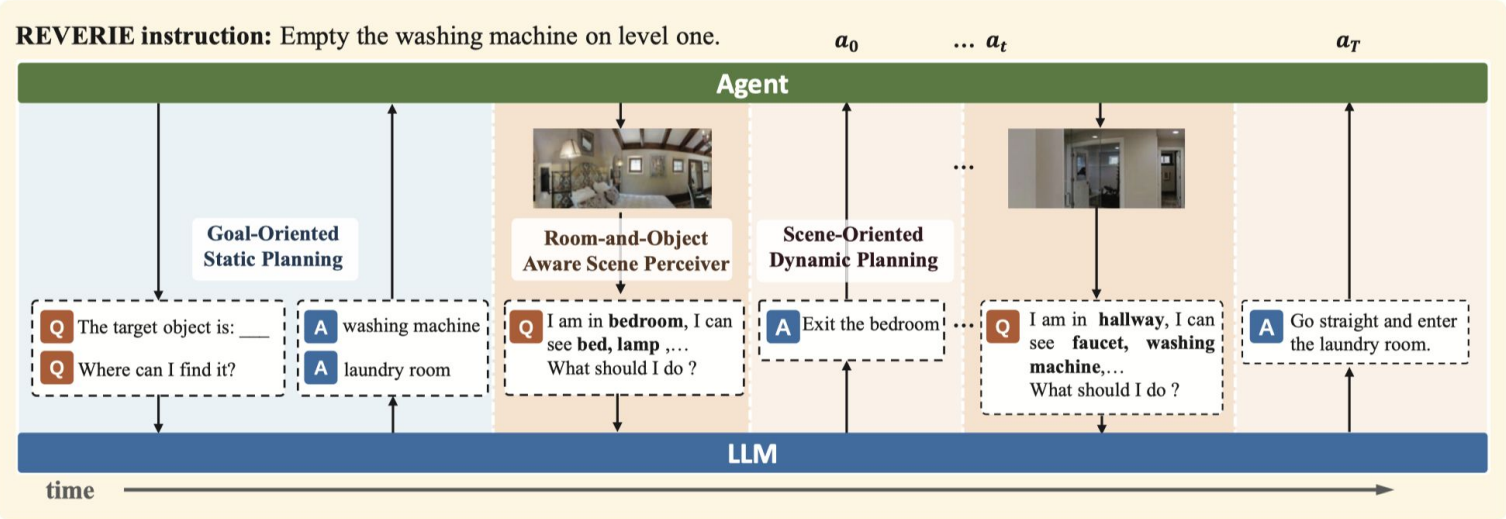


提案のPrompt Engineering

ICCV 2023 の動向・気付き（81/165）

March in Chat: Interactive Prompting for Remote Embodied Referring Expression

- Object oriented Navigationタスクのための LLM Plannerモデルを提案。LLM Plannerが Agentの状況から、ターゲットへ導くSub goalを生成する（例：部屋を出る）。
- 提案のLLM Plannerの入力がタスクのターゲット（寝室の画像をみたい）とAgentの状況（Agent現在の位置、周りの物体など）から、詳細的な言語指示を提示。

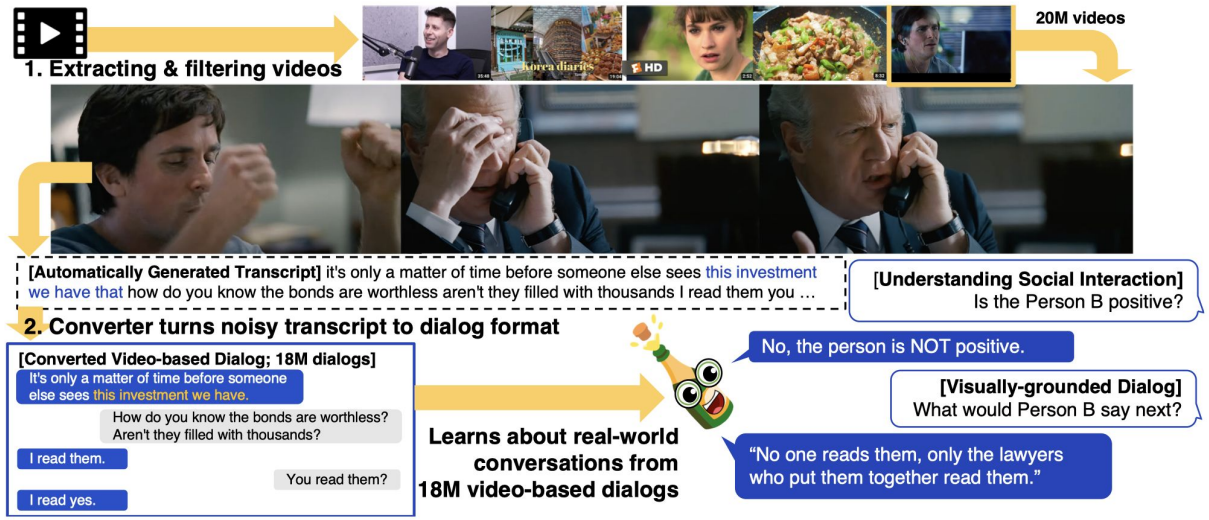


提案手法の構造

ICCV 2023 の動向・気付き（82/165）

CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos

- 大規模ビデオ対話データセットChampagneの提案。ChampagneはYouTubeの人間対話動画、そして対話の言語テキストから構成する（合計18Mビデオ対話）。
- Champagneで学習することで、対話から次の話者の内容の予測や、Social Intelligenceの認識、Visual Commonsense Reasoning認識などでSOTAな精度を達成。

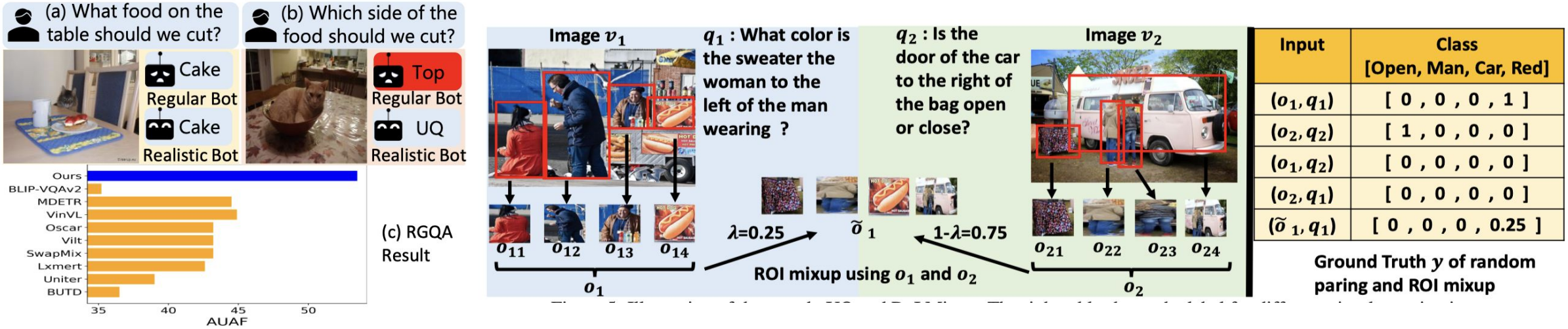


Champagneの説明図

ICCV 2023 の動向・気付き（83/165）

Toward Unsupervised Realistic Visual Question Answering

- 既存のVQAデータセットが回答可能と仮定している。それと比べてこの研究では回答可能と回答不可のQAが含まれるVQAデータセット RealisticVQAを提案。
- Questionと対応する画像と対応しない画像ペアの構築、局所的な画像パッチを変更する工夫の二つの設計を持ったSelf-supervised手法を提案。
- 提案のRealisticVQAデータセットで既存手法と提案手法を評価した結果、既存のVQA手法が回答不可QAの正解率が低い傾向となり、提案手法が大幅に既存手法より精度が高かった。

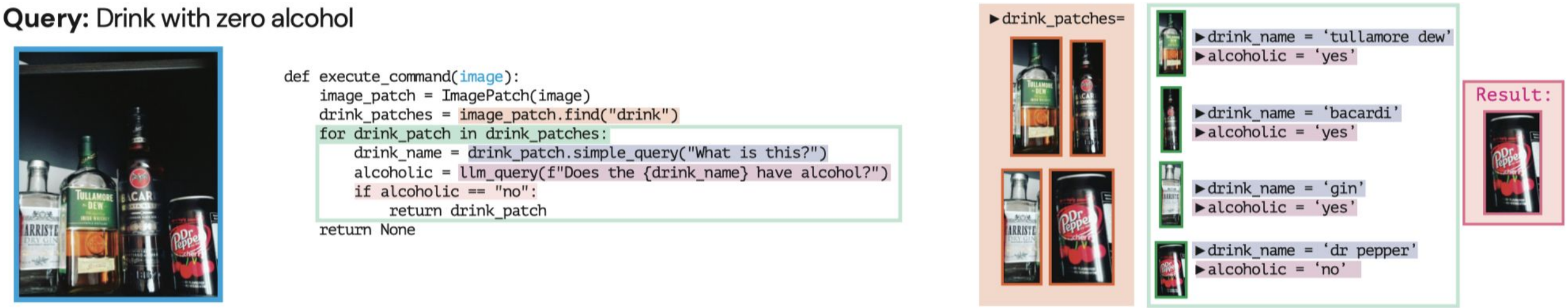


RealisticQAタスクと提案手法

ICCV 2023 の動向・気付き（84/165）

ViperGPT: Visual Inference via Python Execution for Reasoning

- 画像と画像に関する質問から自動的に回答のコードを生成し、Python API＋Function APIで直接質問をStep-by-stepで解く新たなSymbolic手法の提案。
- コード生成の部分はプログラミングLLMsを利用し、Promptで操作可能となる。学習が必要としない。また、基本的な画像操作（Find、Attention、Countingなど）をPython APIにより実装するAPIも提案した。
- 回答のプロセスがよりTransparentの他、いくつか既存のVQAベンチマークデータセットでSOTAな精度を実現した。

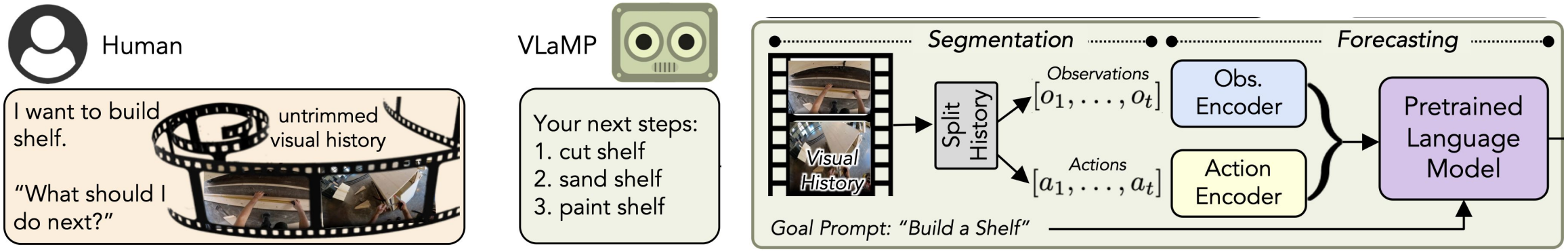


ViperGPTの結果例

ICCV 2023 の動向・気付き（85/165）

Pretrained Language Models as Visual Planners for Human Assistance

- Episodic video（3次元環境で行動の歴史のEgocentric動画）と言語ターゲット（例：タンスを組み立てる）から、タスクを実行するためのStep-by-step言語指示を推定するタスクを提案。
- 提案手法がLLMsを利用して行動の指示を生成する。また、具体的に提案手法がビデオ認識をSegmentationとForcastingの二つのステップから構成する。Segmentationではビデオのなかの人間行動を細かくセグメントする。Forcastでは行動のSequenceをモデリングする。



提案タスクと手法

ICCV 2023 の動向・気付き(86/165)

Kick Back & Relax: Learning to Reconstruct the World by Watching SlowTV

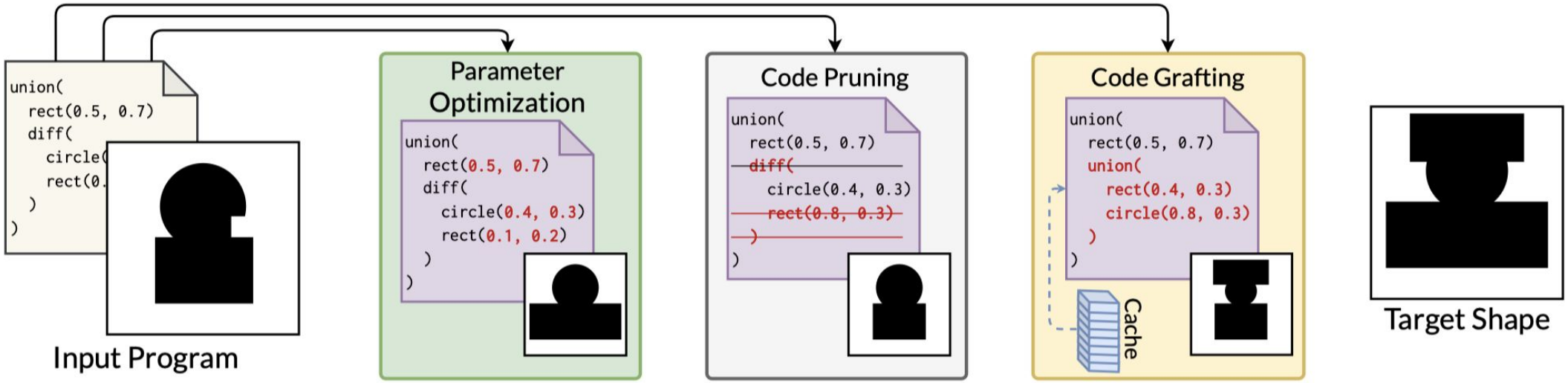
- 新しい大規模動画像データセットSlowTVを提案。SlowTVはYouTube動画から屋外のシーン(ハイキング、ドライビング、ダイビングなど)をピックアップし、合計1.7Mの画像が含まれる。
- SlowTVデータセットでself-supervised monocular depth estimation (SS-MDE)モデルを学習し、SS-MDEが複数のOutdoorとIndoorデータセットでZero-shotで既存の教師学習手法と同レベル/もっと高い精度を実現。

SlowTVデータセット例

ICCV 2023 の動向・気付き（87/165）

Improving Unsupervised Visual Program Inference with Code Rewriting Families

- Visual Program Inference（3次元物体の構造推定）のための新しい手法を提案。
- 提案手法がBootstrapped LearningとRewriting仕組みを結合し（下図）、複数の2D/3Dベンチマークでの SOTAな精度を実現。
- 論文では複数のCode rewriter仕組みを提案（Parameter Optimization、Code Pruning、Code Graftingなど）し、他のVPI手法へも適応しやすい。

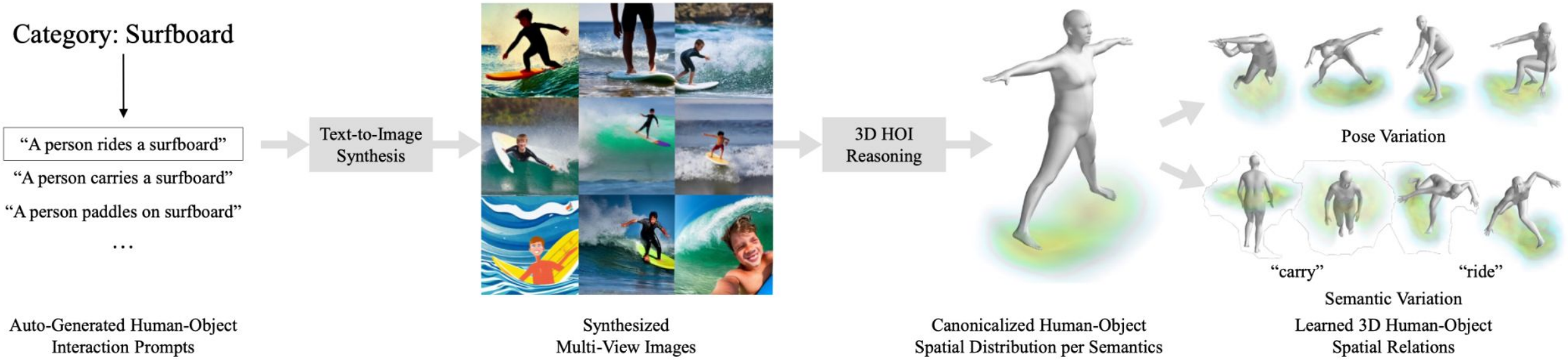


提案のRewritingの仕組み説明

ICCV 2023 の動向・気付き（88/165）

CHORUS: Learning Canonicalized 3D Human–Object Spatial Relations from Unbounded Synthesized Images

- 物体カテゴリ/（カテゴリ＋画像）からHuman–Objectの3次元空間配置を推定するタスク・Self–supervisedな手法の提案。
- 提案手法のプロセス（下図）は、まずLLMによりカテゴリからHumanとその物体カテゴリのInteractionの文章を生成。次にDiffusionモデルでテキストから複数の画像（多視点）を生成し、生成した複数画像からCanonicalizedな空間配置を計算。

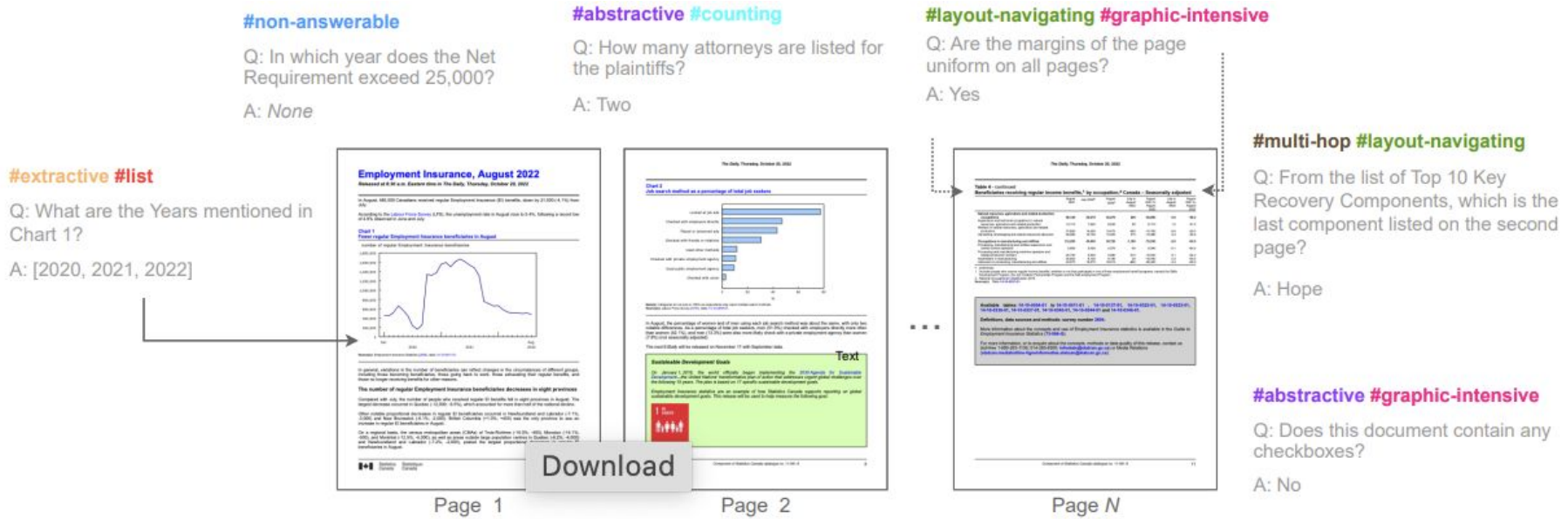


CHOURSの仕組み

ICCV 2023 の動向・気付き（89/165）

Document Understanding Dataset and Evaluation

- Document（テキスト、表、図表、図、リスト、チェックボックス、スタンプなど）理解のための新しいDocument QAデータセットを提案。データセットにはmulti-page, multi-domainのデータが含まれる。また、既存ベンチマークと比べて実環境で使われる図表データとの類似度が高い。
- DUDEデータセットではSOTA手法が人間の精度と差があり、今後実環境での図表理解のための学習・評価に活用できる。

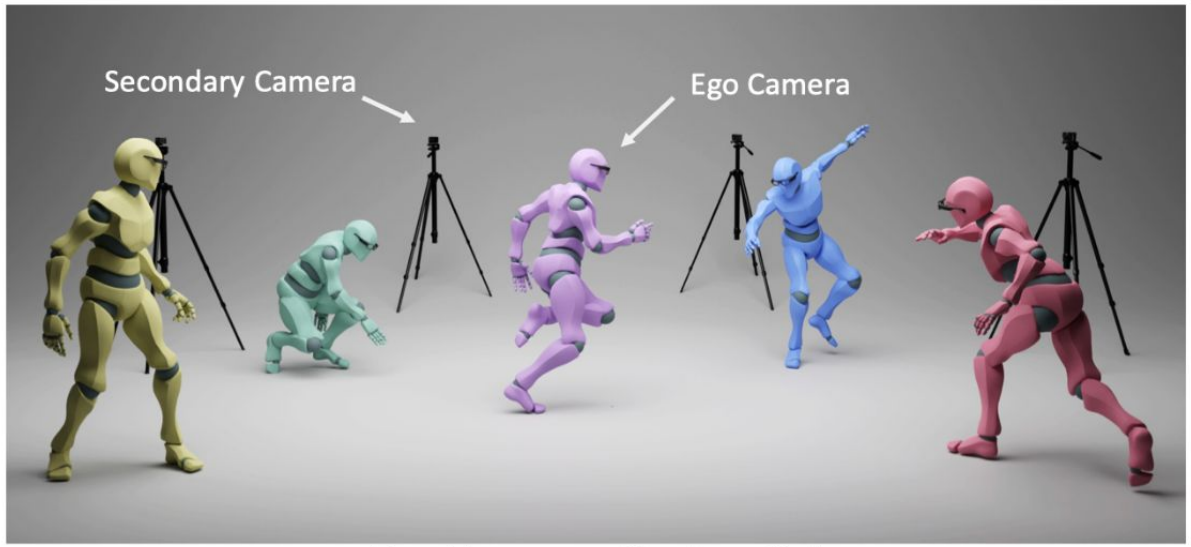


DUDEデータセット例

ICCV 2023 の動向・気付き（90/165）

EgoHumans: An Egocentric 3D Multi-Human Benchmark

- Multi-view multi-humanビデオデータセットEgoHumansの提案。EgoHumansは複数の人間にカメラ付きGlassを着装させ、また他のSecondary Cameraも設定しデータ撮影を行った。複数の人間のより複雑な動作（例：テニスなど）も高精度で記録可能。EgoHumansデータセットで3次元姿勢推定・トラッキング、Mesh復元などに活用できる。
- Multi-stream transformer+3次元Reasoningの手法を提案し、EgoHumansでSOTA。



Ego-Humans - Capture Setup


Ego-Views

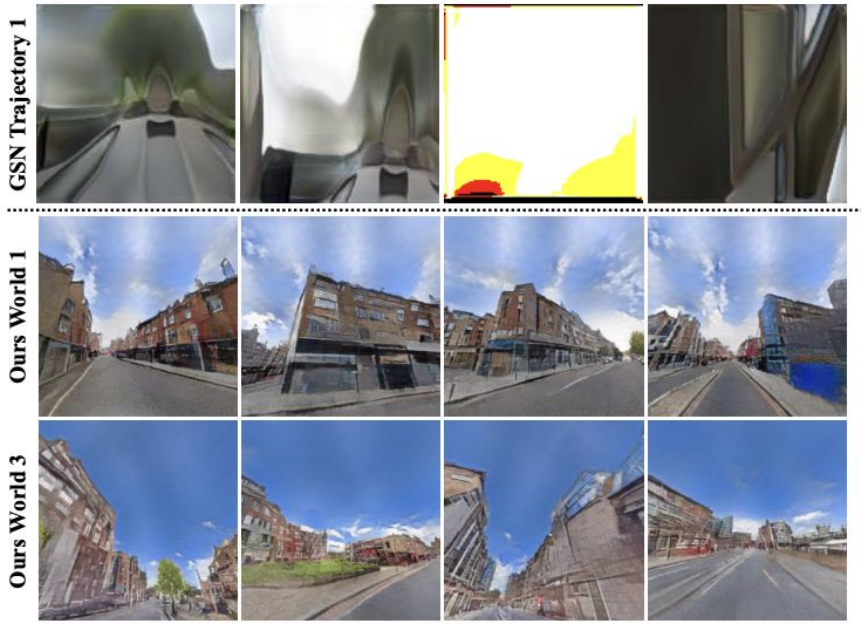

Secondary-Views

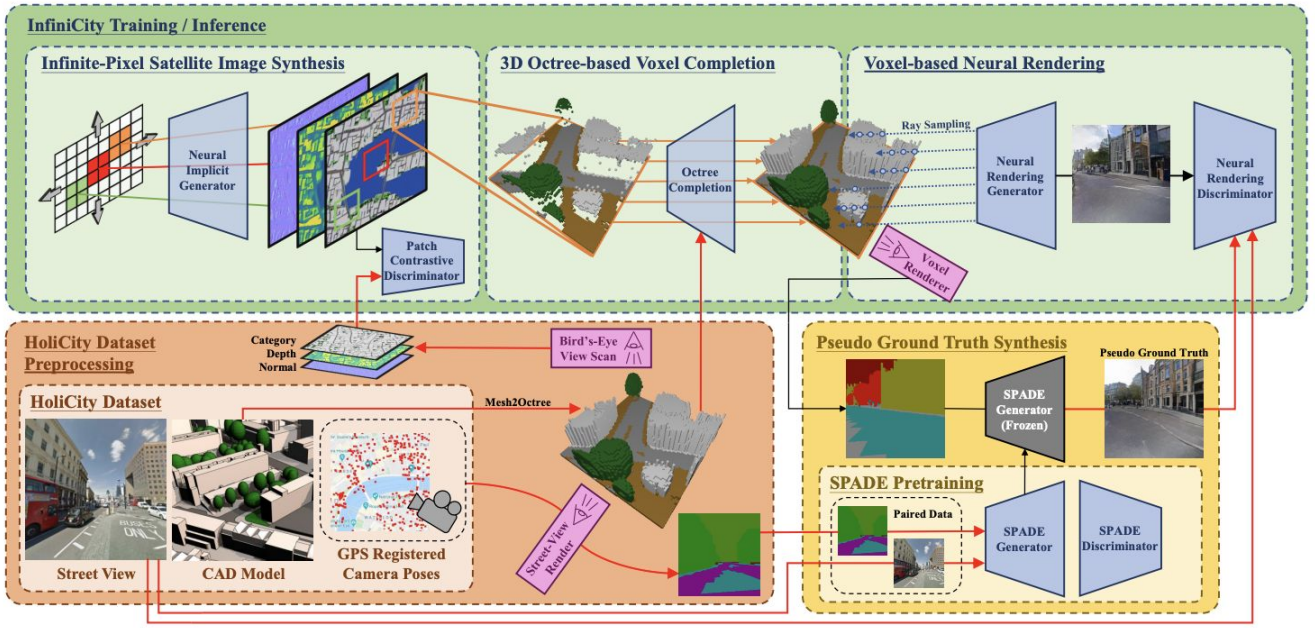
EgoHumansカメラ設定、Ego-View画像、合成されたSecondary-Views画像

ICCV 2023 の動向・気付き（91/165）

InfiniCity: Infinite-Scale City Synthesis

- 任意規模の都市3次元データを合成する手法の提案。
- 提案手法が3stepで合成を行っている。
  - Step1：マルチモーダル（デプス、法線、カテゴリ）の上空画像を合成。
  - Step2：Octree-based voxel補完を利用し、Voxelデータを作成。
  - Step3：Voxel neural renderingでテクスチャを生成。（既存の手法を使用）



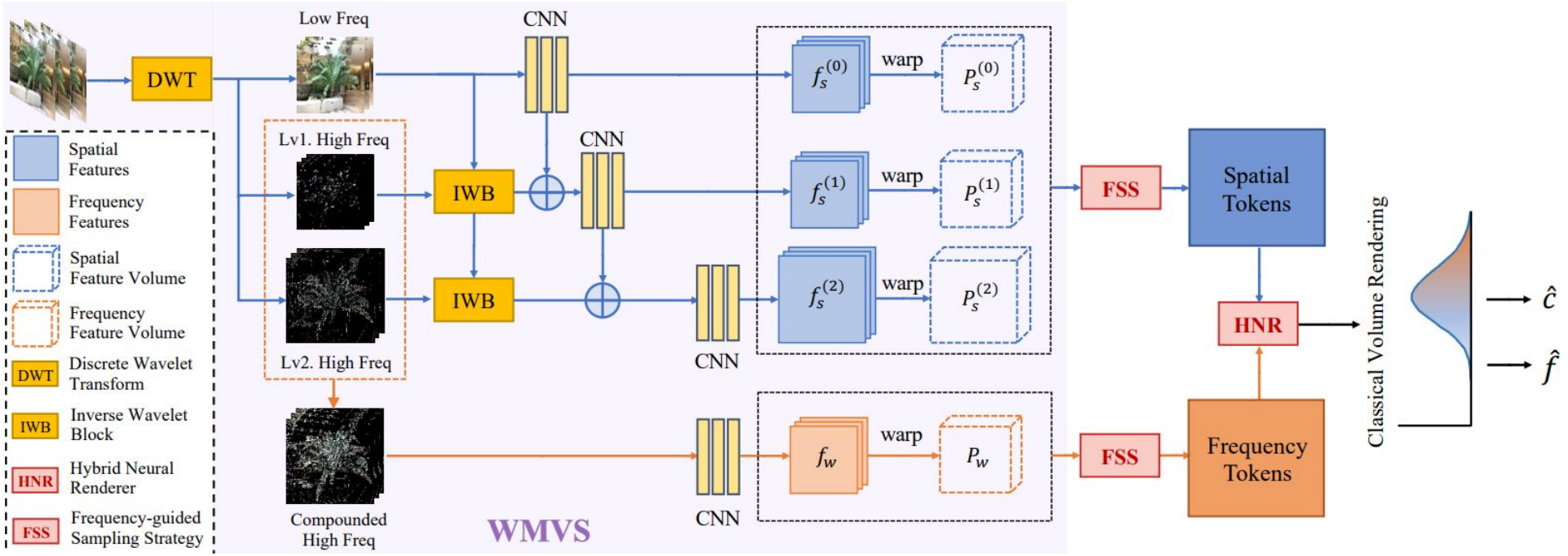


結果例、提案手法

ICCV 2023 の動向・気付き（92/165）

WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields

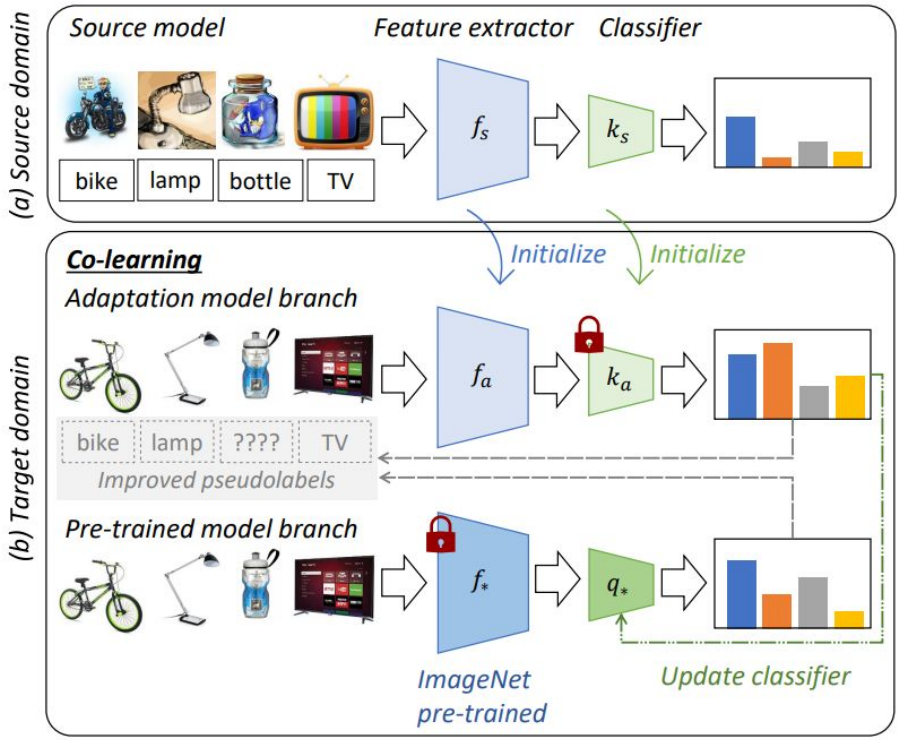
- Muyu Xu, Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Xiaoqin Zhang, Christian Theobalt, Ling Shao, Shijian Lu
  - 空間周波数を活用したNeRFの拡張
  - 入力画像をウェーブレット変換して空間周波数に応じた適応的なサンプリングを行う
  - ボリュームレンダリングに対応する計算でも周波数と実空間の両方から集約

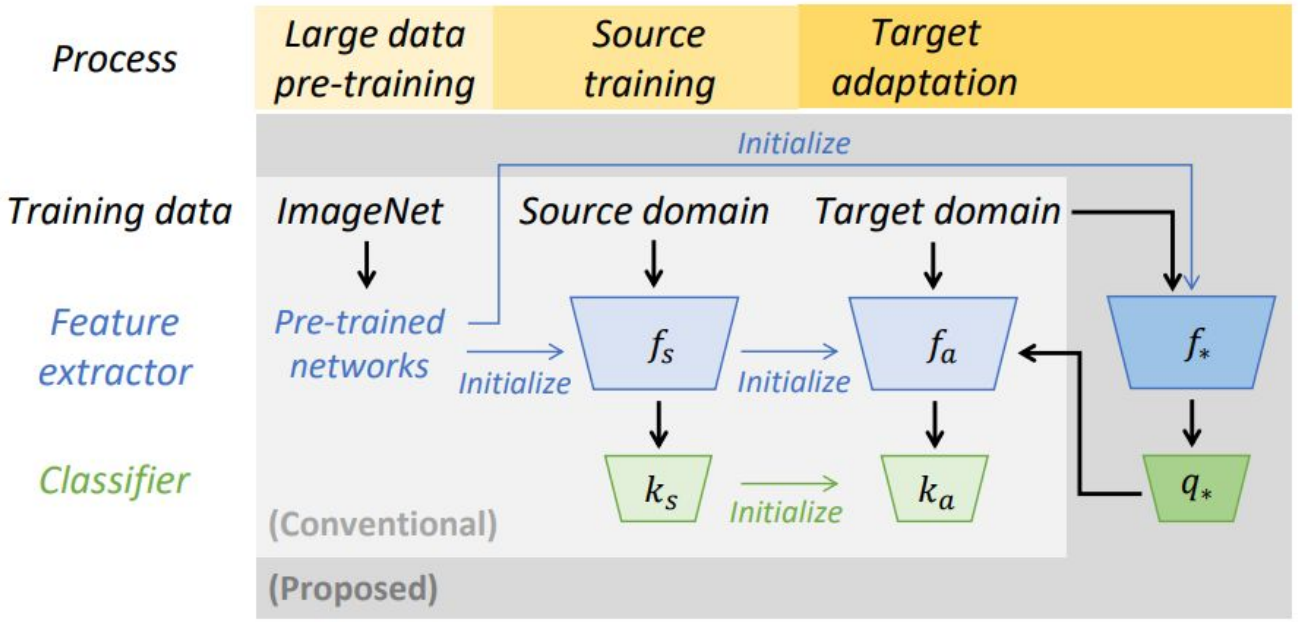


ICCV 2023 の動向・気付き（93/165）

Rethinking the Role of Pre-Trained Networks in Source-Free Domain Adaptation

- target adaptationのプロセスにPre-trained networkを統合
  - 学習済みモデルは汎化に重要で多様な特徴を持つ
  - Fine-tuning時に共学習することで下流タスクの有用なドメイン情報を抽出
  - ロバスト性と汎化性の改善を確認





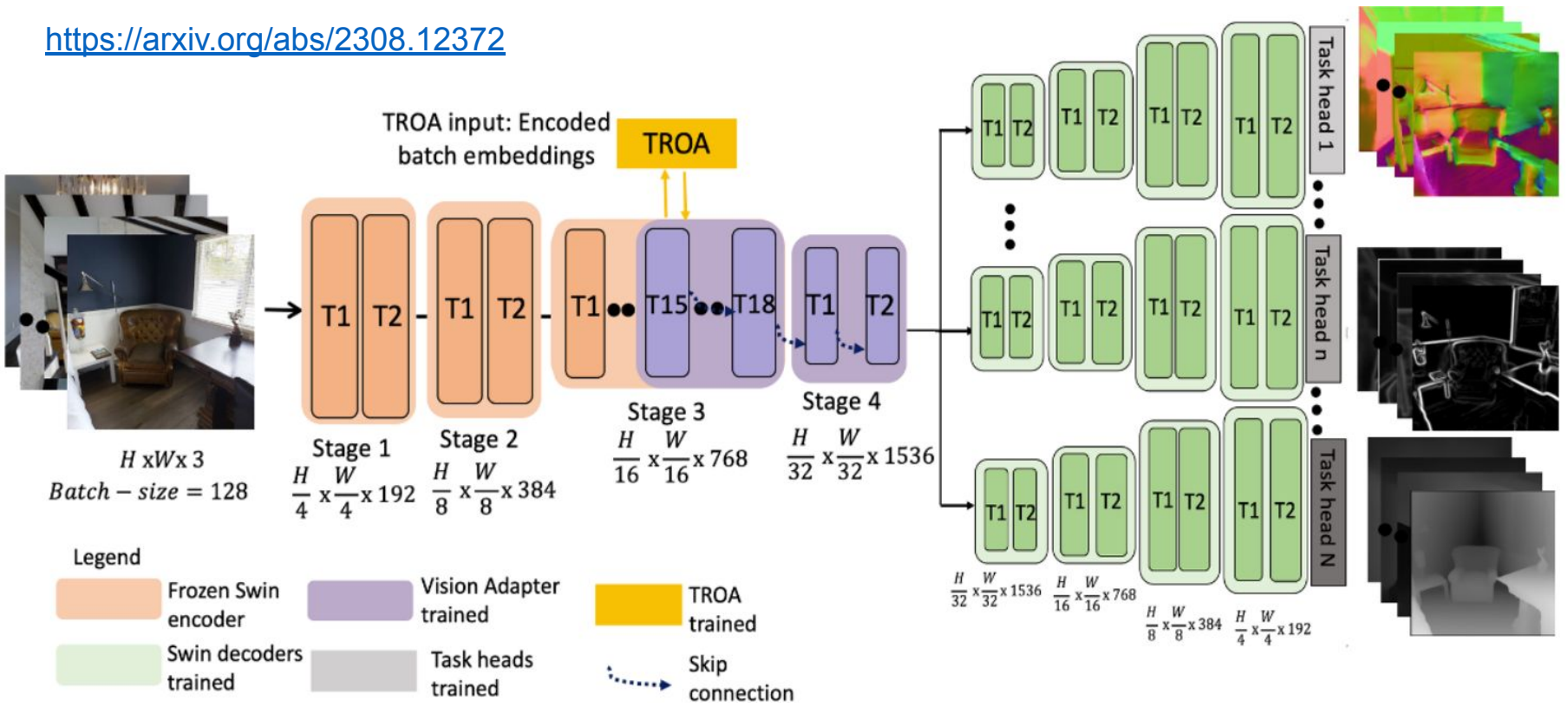
https://arxiv.org/abs/2212.07585

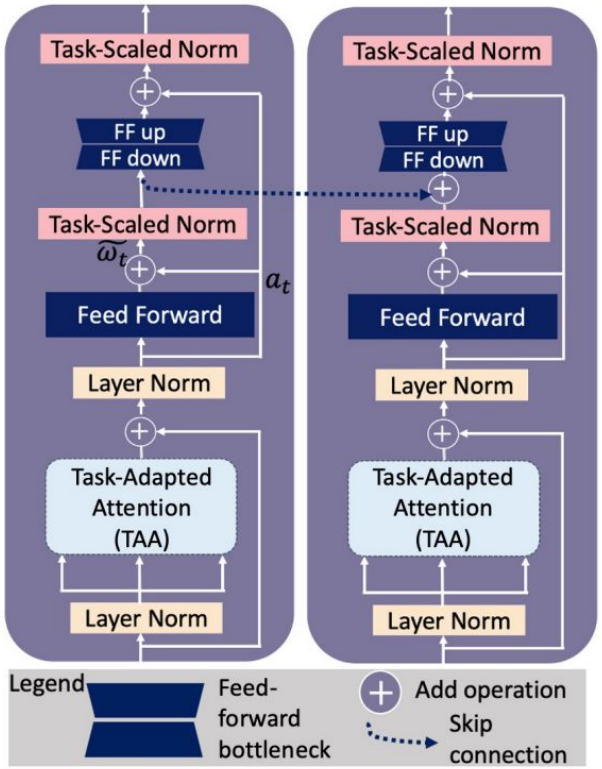
ICCV 2023 の動向・気付き（94/165）

Vision Transformer Adapters for Generalizable Multitask Learning

- タスクやドメインに適用可能なタスクの親和性を学習する
  マルチタスクViTアダプタを提案
  - アダプタに勾配ベースと注意ベースの類似性を組み合わせたタスク適応型注意メカニズム
  - ゼロショットタスク転送，教師なしドメイン適応において適応可能
  - Fine-tuningせずにターゲットタスクのドメインに汎化可能

https://arxiv.org/abs/2308.12372

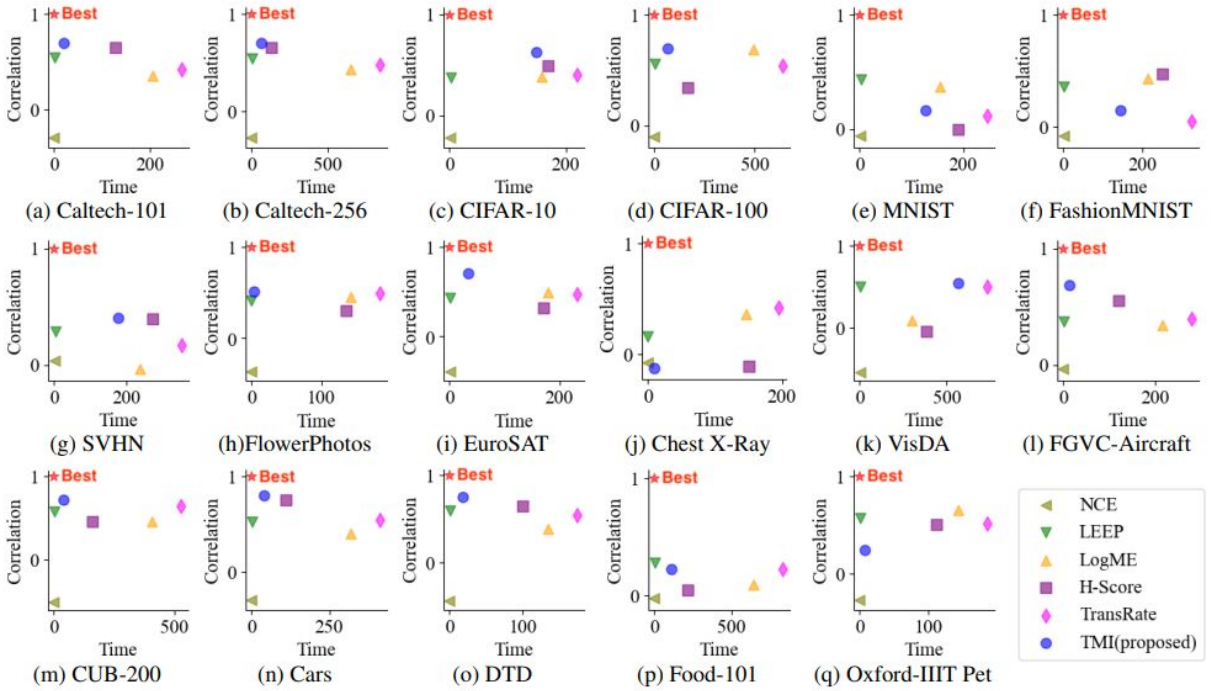




ICCV 2023 の動向・気付き(95/165)

Fast and Accurate Transferability Measurement by Evaluating Intra-class Feature Variance

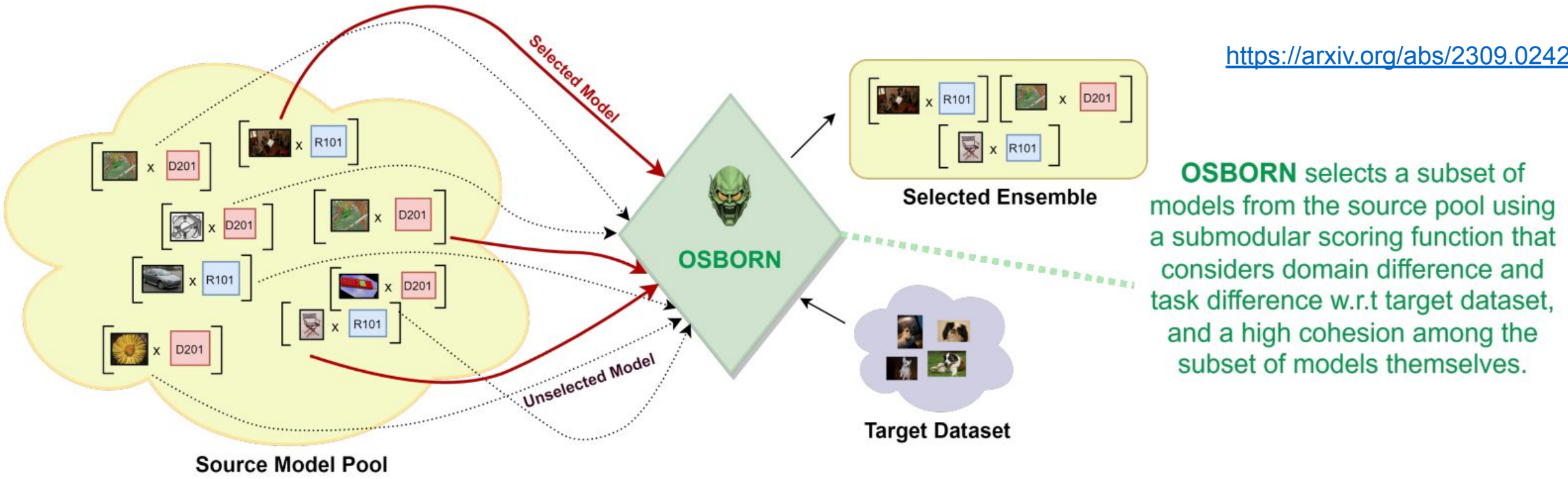
- 高速かつ高精度な転移性測定アルゴリズムTMI(TRANSFERABILITY MEASUREMENT WITH INTRA-CLASS FEATURE VARIANCE)を提案
  - クラス内特徴分散を測定することで学習済みモデルの転移性を評価
  - タスクに対するモデルの適応性を評価し，モデルがどの程度移譲可能であるか測定
  - 最適な特徴抽出器と分類器を必要としないアルゴリズム



https://arxiv.org/abs/2308.05986

Building a Winning Team: Selecting Source Model Ensembles using a Submodular Transferability Estimation Approach

- Optimal Transport-based Submodular Transferability metric (OSBORN)を提案
  - アンサンブル選択に領域類似性，タスク類似性，モデル間凝集性を組み込んだTransferability推定指標を導入



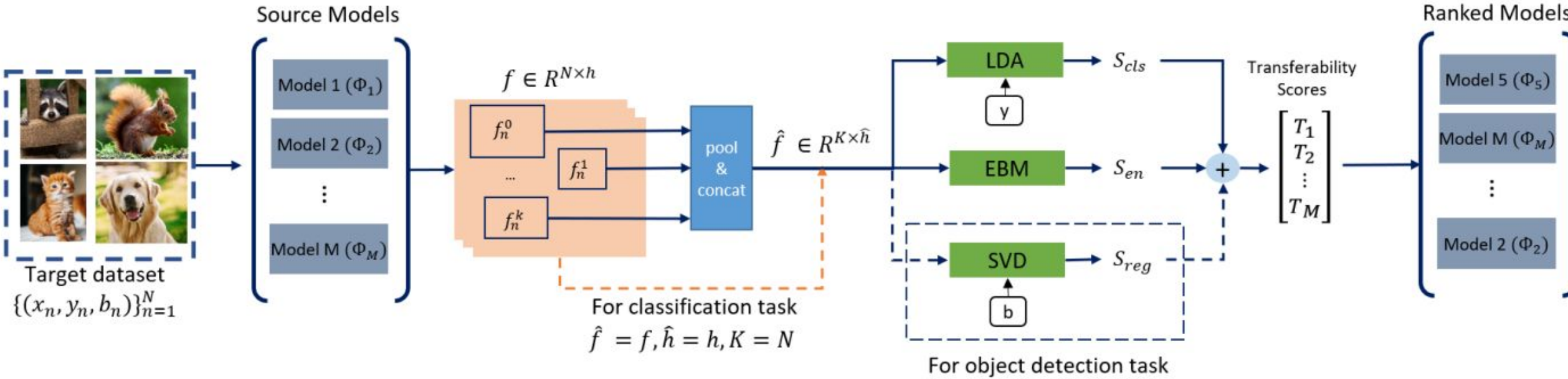
https://arxiv.org/abs/2309.02429

OSBORN selects a subset of models from the source pool using a submodular scoring function that considers domain difference and task difference w.r.t target dataset, and a high cohesion among the subset of models themselves.

ICCV 2023 の動向・気付き(97/165)

ETran: Energy-Based Transferability Estimation

- エネルギースコア，分類スコア，回帰スコアから推定するETranを提案
  - ターゲットデータセットがモデルの学習データ分布に従わない場合，抽出された特徴量は信頼できない
  - EBMを活用し，Pre-trained modelが分布内(IND)か分布外(OOD)を測るETranを提案
  - エネルギー・スコアが高いモデルであるほど識別精度が向上する傾向を示唆

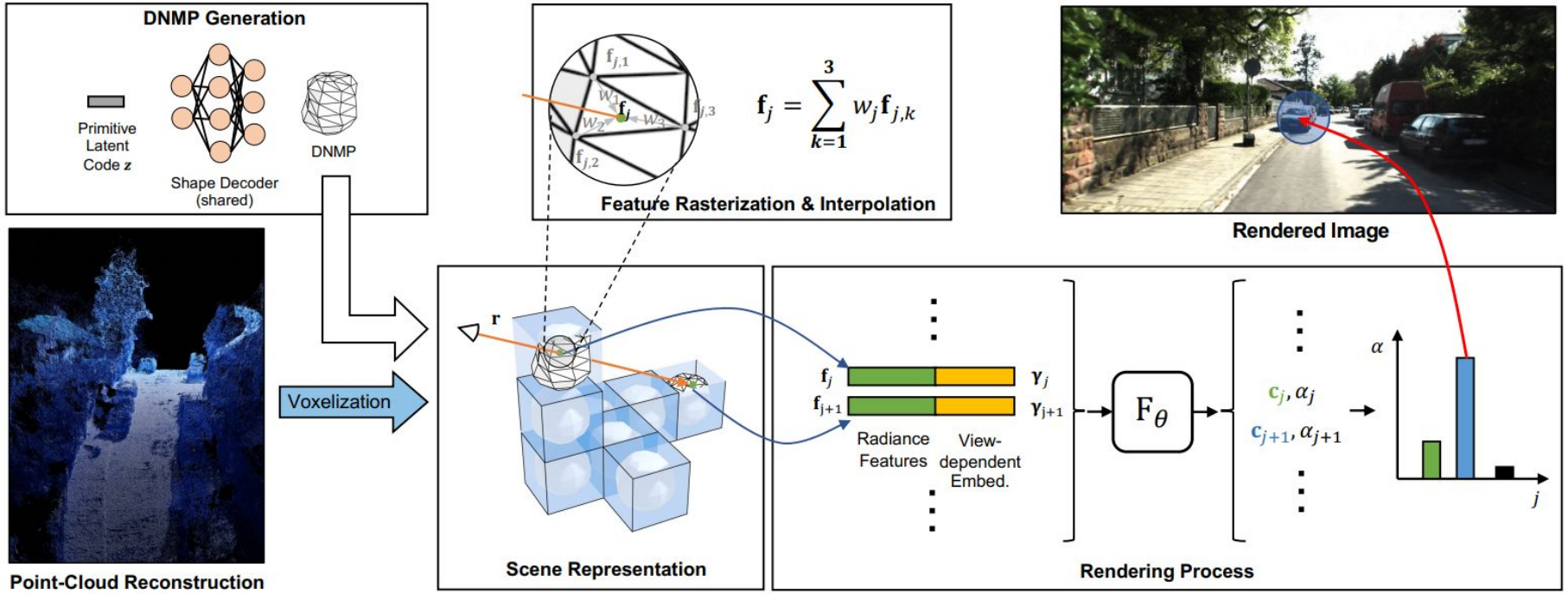


https://arxiv.org/abs/2308.02027

ICCV 2023 の動向・気付き(98/165)

Urban Radiance Field Representation with Deformable Neural Mesh Primitives

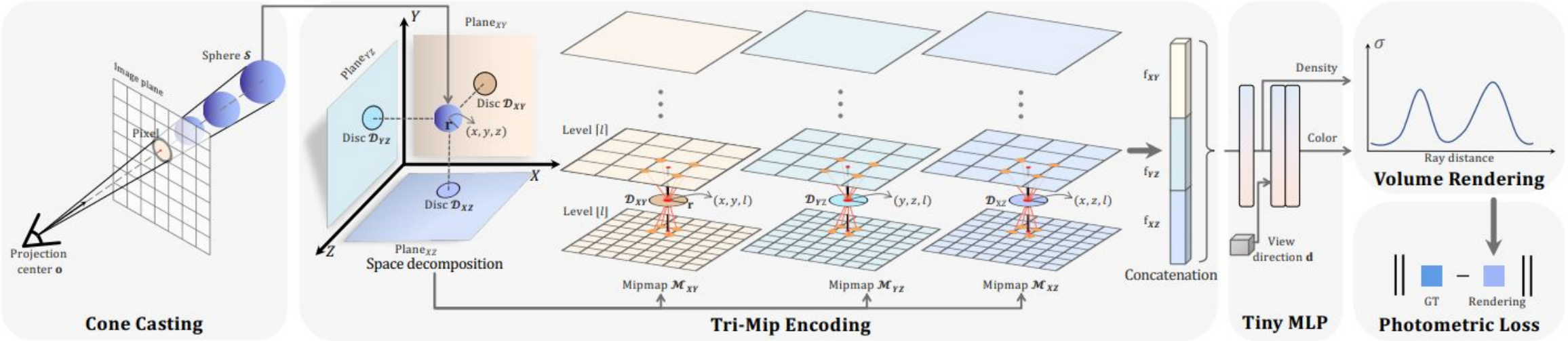
- Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, Changjun Jiang
  - 点群からボクセルごとにメッシュを設定・線形補間することで特徴量を設定し，特徴量を視線方向と合わせてMLPで変換することで表面モデルとしてレンダリングする
  - 既存のレンダリングパイプラインと相性がよく，高速で省メモリ



ICCV 2023 の動向・気付き（99/165）

Tri-MipRF: Tri-Mip Representation for Efficient Anti-Aliasing Neural Radiance Fields

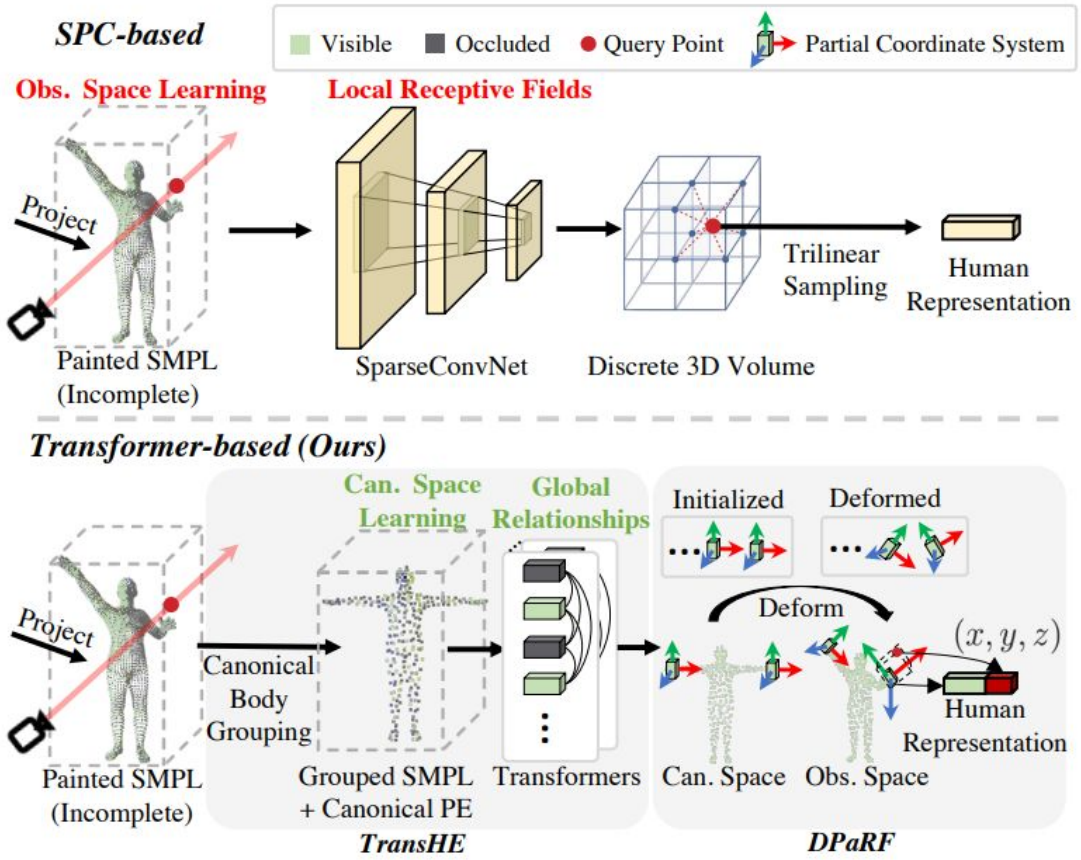
- Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, Yuewen Ma
  - Mip-NeRFとTri-planeによる特徴量表現を組み合わせた手法
  - 空間中でのサンプル点列について，距離に応じた半径の球を想定，これがTri-planeに投影すると円盤になることを利用
  - 各Planeについて階層的に解像度を設定し，結合してその点での特徴量を記述
  - 特徴量と視線方向から軽量なMLPでRFを出力



ICCV 2023 の動向・気付き（100/165）

TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering

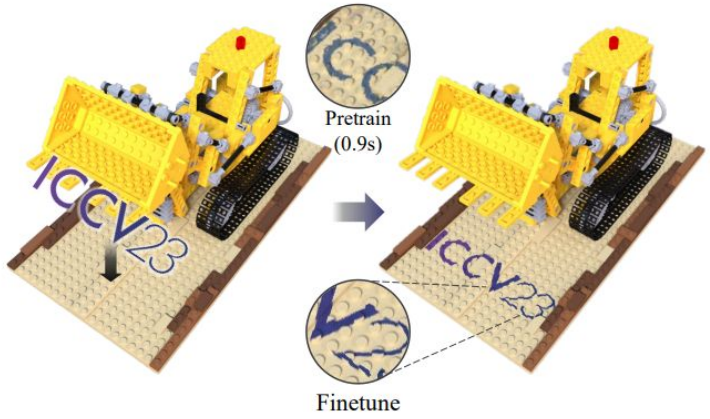
- Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, Yi Yang
  - NeRFとTransformerを用いた人体モデル
  - 変形にはSMPLを用いる
  - 人体上のPositional Encodingを与える
  - 変形に対応したパーツごとの局所座標系を設定
  - クエリ点近傍の特徴量を集約，これをクエリとして入力画像からの特徴量にAttentionを張る
  - これをConditionとしてNeRFをモデル化，ボリュームレンダリング

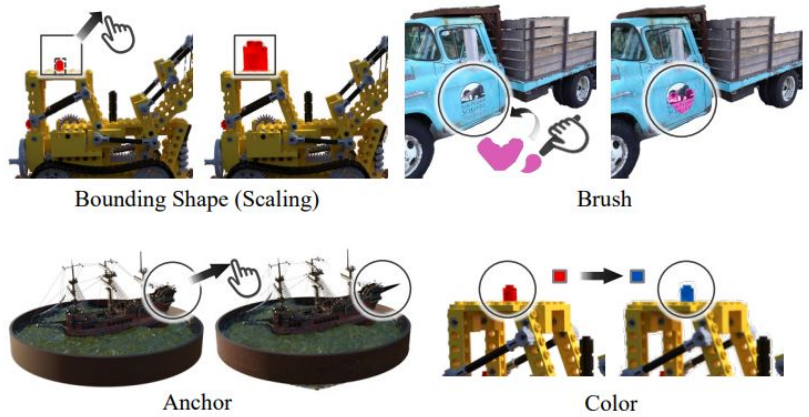


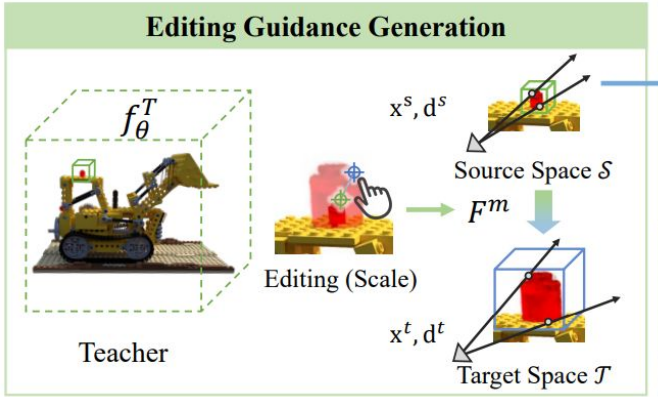
ICCV 2023 の動向・気付き（101/165）

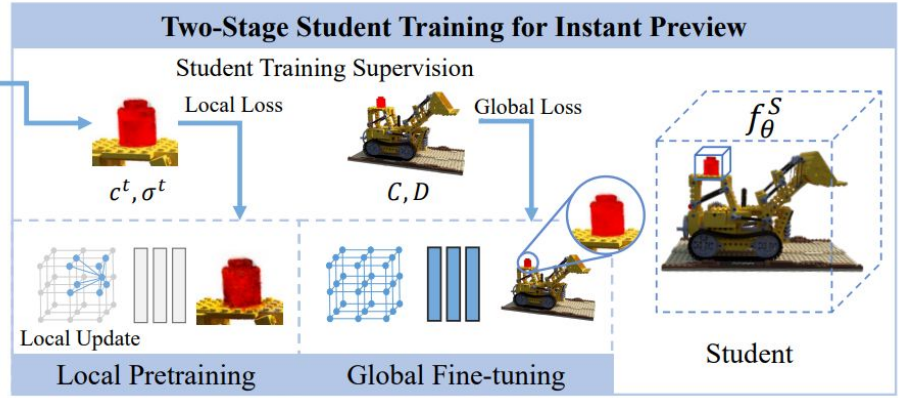
Seal-3D: Interactive Pixel-Level Editing for Neural Radiance Fields

- Xiangyu Wang, Jingsen Zhu, Qi Ye, Yuchi Huo, Yunlong Ran, Zhihua Zhong, Jiming Chen
  - 学習済みのNeRFを編集
  - カラー・形状の
    いずれもわかりやすい
    インターフェースで操作
  - 数秒で粗く編集内容を
    可視化できるように
    局所的にNeRFを上書き
  - 数分で全体のNeRFを
    最適化し，編集を適用した
    NeRFによる精緻な
    任意視点画像生成を実現





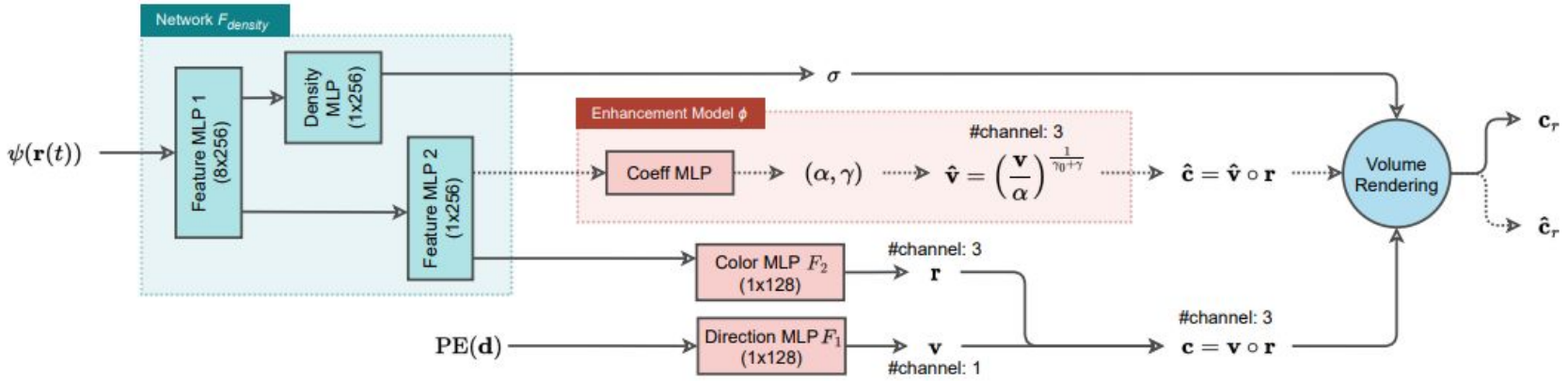


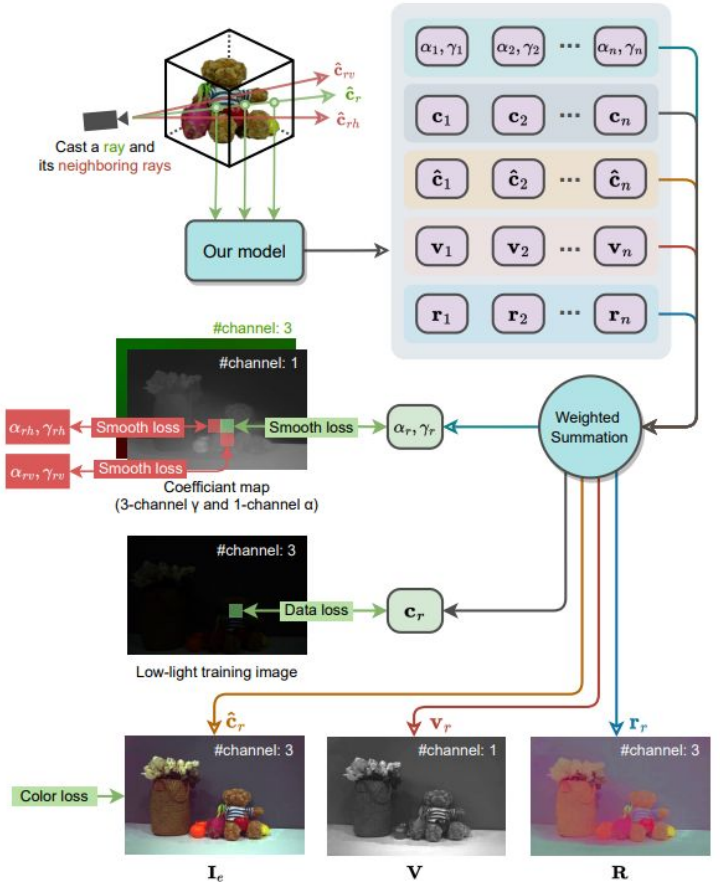


ICCV 2023 の動向・気付き（102/165）

Lighting up NeRF via Unsupervised Decomposition and Enhancement

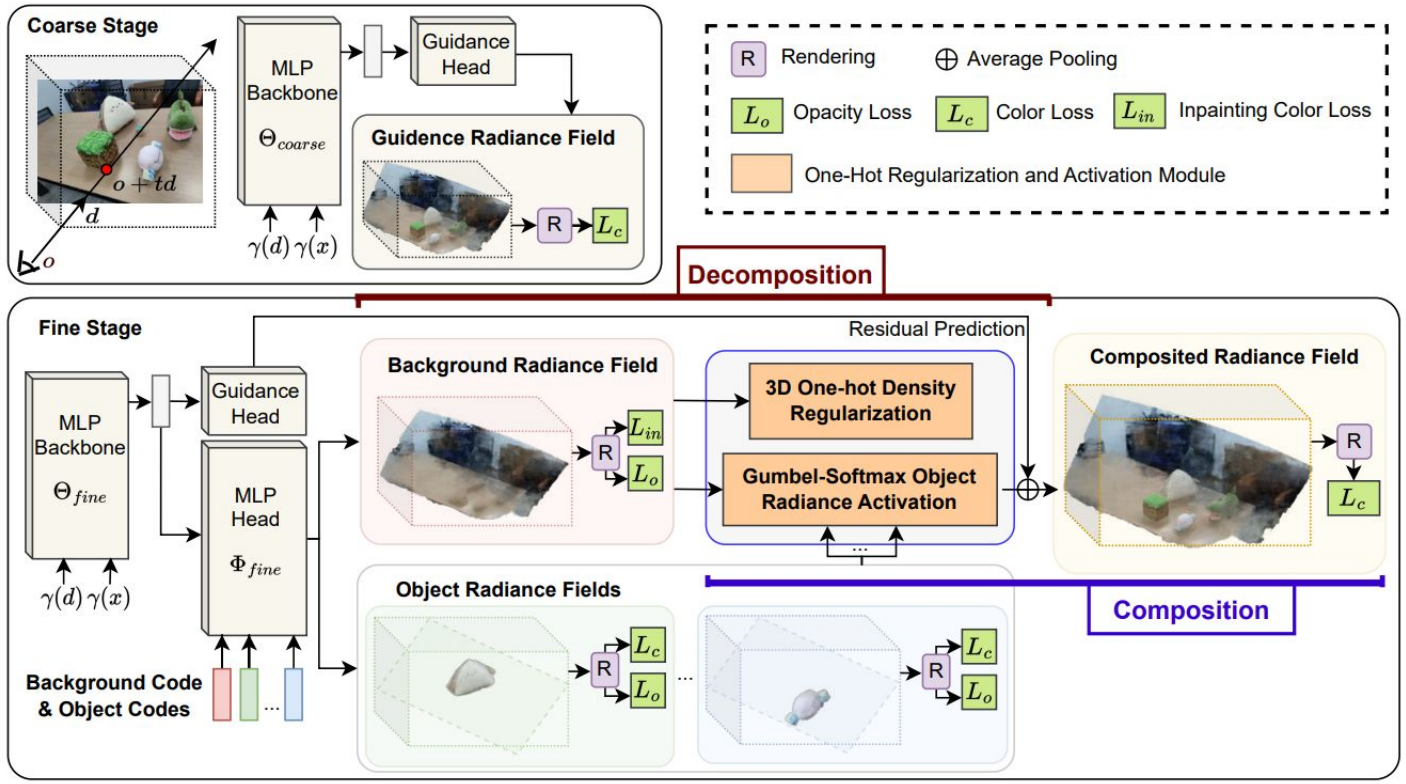
- Haoyuan Wang, Xiaogang Xu, Ke Xu, Rynson WH. Lau
  - 低照度環境でNeRFに相当する任意視点画像を実現
  - カラーについて，ガンマ補正の係数を同時に推定するMLPを導入
  - 視線非依存な成分についてのみガンマ補正
  - 近傍画素でガンマ補正の係数が滑らかになるような正則化を導入





Learning Unified Decompositional and Compositional NeRF for Editable Novel View Synthesis

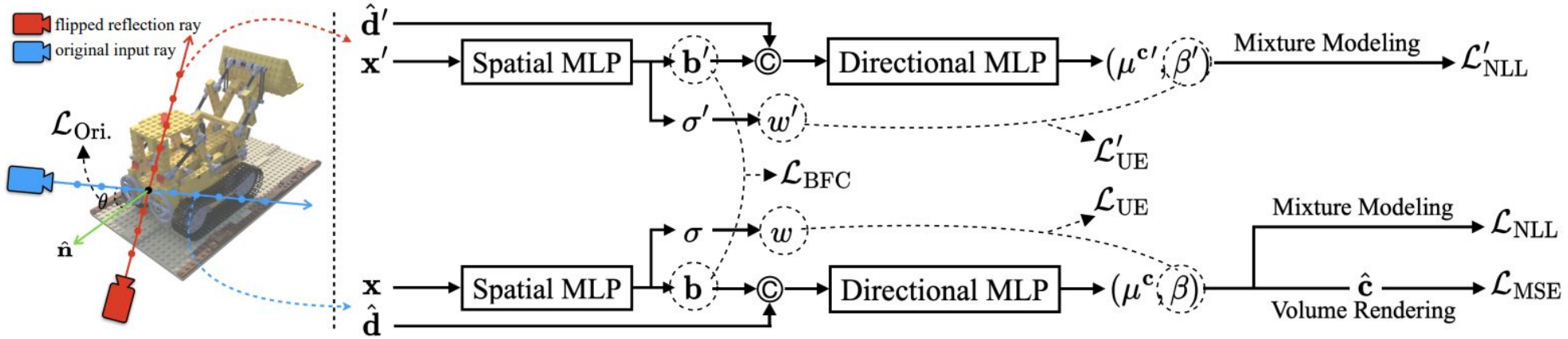
- Yuxin Wang, Wayne Wu, Dan Xu
  - NeRFを前景と背景に分離して最適化
  - 粗いNeRFを最適化してから，各点について前景と背景を重ね合わせてレンダリング
  - 前景が分離するように正則化



ICCV 2023 の動向・気付き（104/165）

FlipNeRF: Flipped Reflection Rays for Few-shot Novel View Synthesis

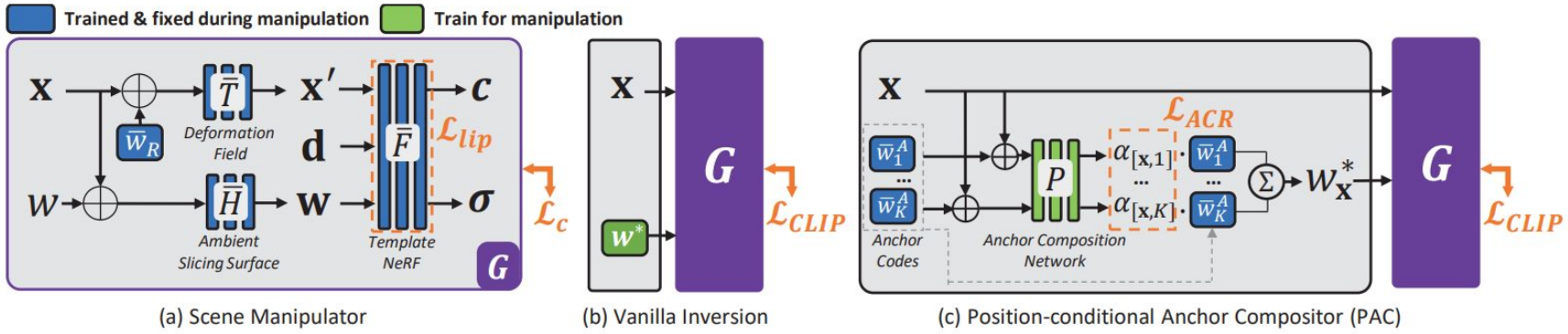
- Seunghyeon Seo, Yeonjin Chang, Nojun Kwak
  - 少数視点の画像からNeRFを最適化するための正則化手法
  - 光線が反射した点の法線を推定，反転させた方向の光線について一致するようにロス関数を導入する
  - 放射輝度だけでなく，反転させた光線について特徴量も近くなるように正則化
  - Emptiness Lossを少数視点での不確実性に対応させたUE Lossを導入
  - その他の正則化も導入（Orientation Loss

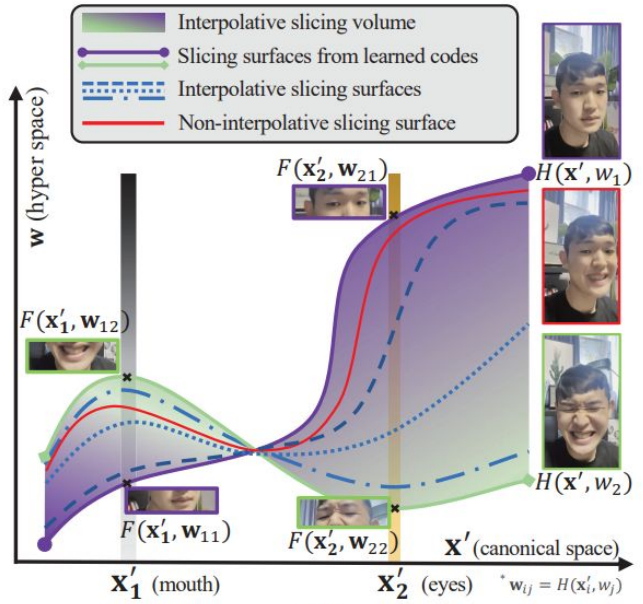


ICCV 2023 の動向・気付き（105/165）

FaceCLIPNeRF: Text-driven 3D Face Manipulation using Deformable Neural Radiance Fields

- Sungwon Hwang, Junha Hyung, Daejin Kim, Min-Jung Kim, Jaegul Choo
  - テキストで顔のNeRFを編集
  - HyperNeRFをベースに，単一の潜在ベクトルでは局所的な変形を分離して扱えないことに着目
  - 領域ごとにアンカーを設定し，クエリに合わせて領域ごとの潜在ベクトルを足し合わせてNeRFを学習
  - CLIPでテキストでの指示と比較し最適化して変形

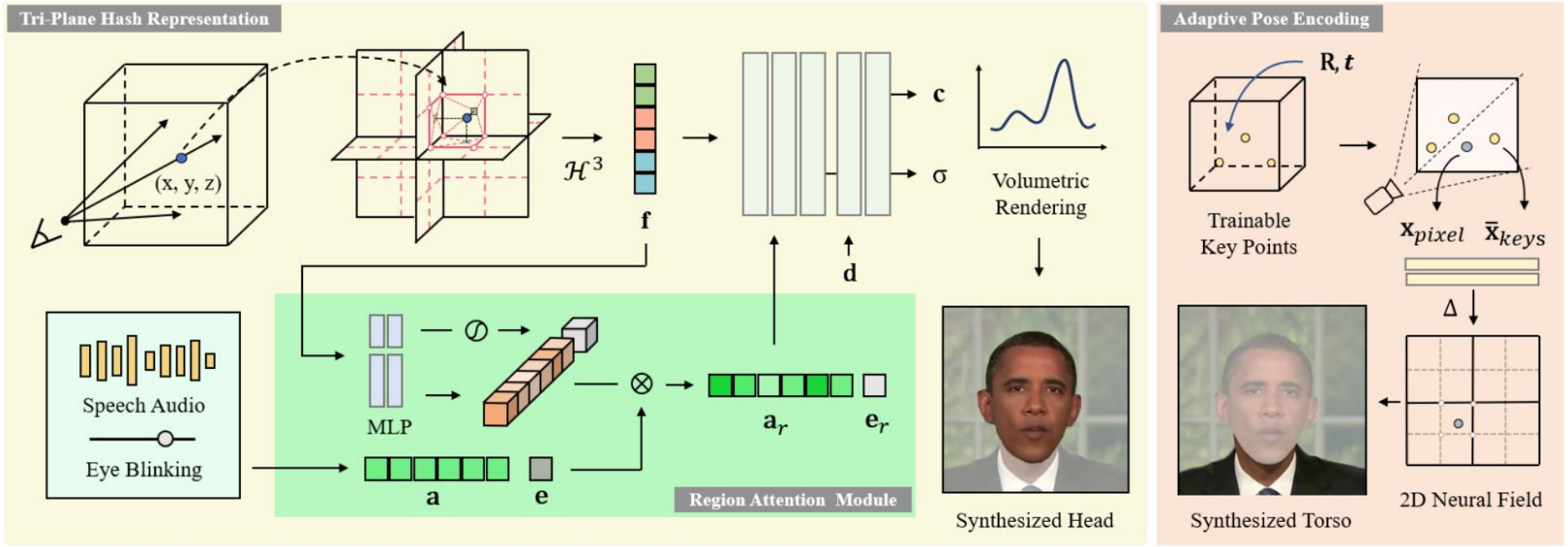




ICCV 2023 の動向・気付き（106/165）

Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis

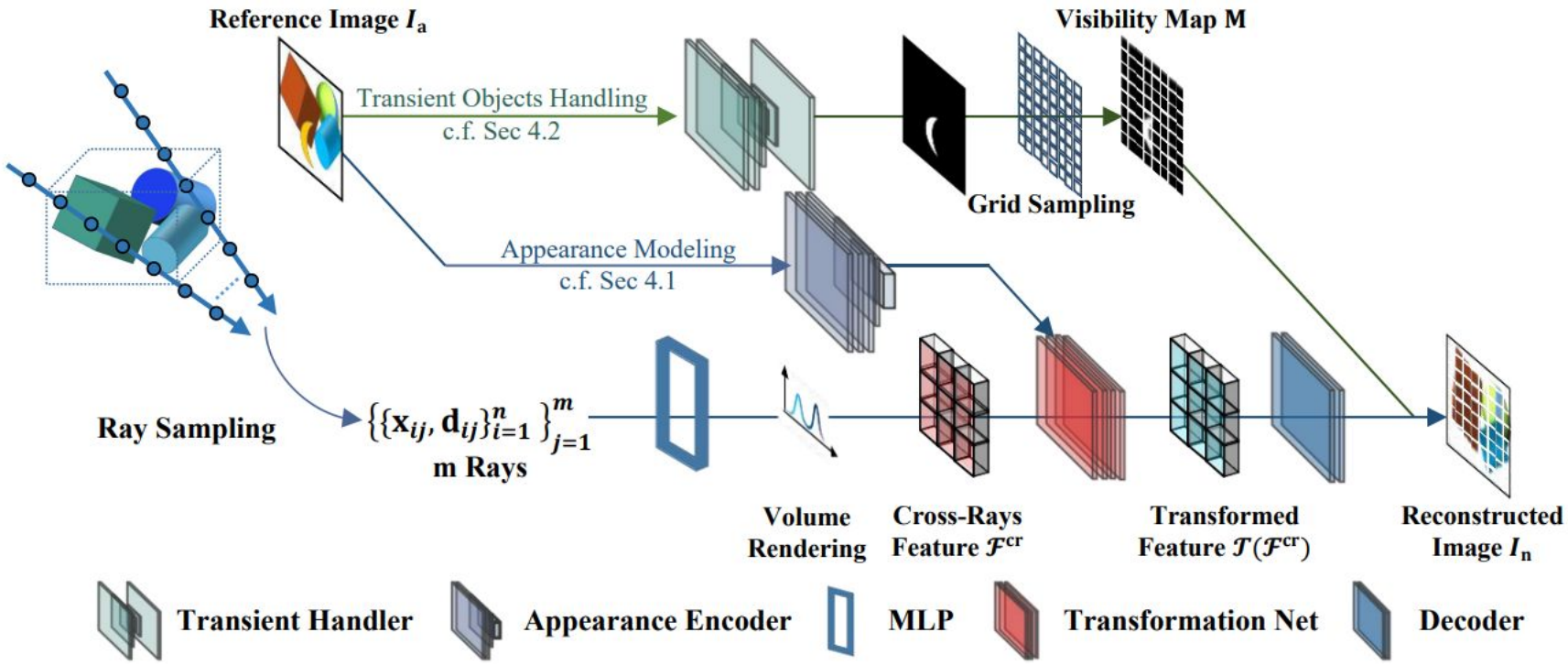
- Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, Lin Gu
  - NeRFを利用した音声に応じたTalking portrait合成手法
  - 前景となる顔領域をTri-plane NeRFでモデル化，音声とまばたきでAttentionを導入
  - 背景は2DのNeRFでモデル化し合成する



ICCV 2023 の動向・気付き（107/165）

Cross-Ray Neural Radiance Fields for Novel-view Synthesis from Unconstrained Image Collections

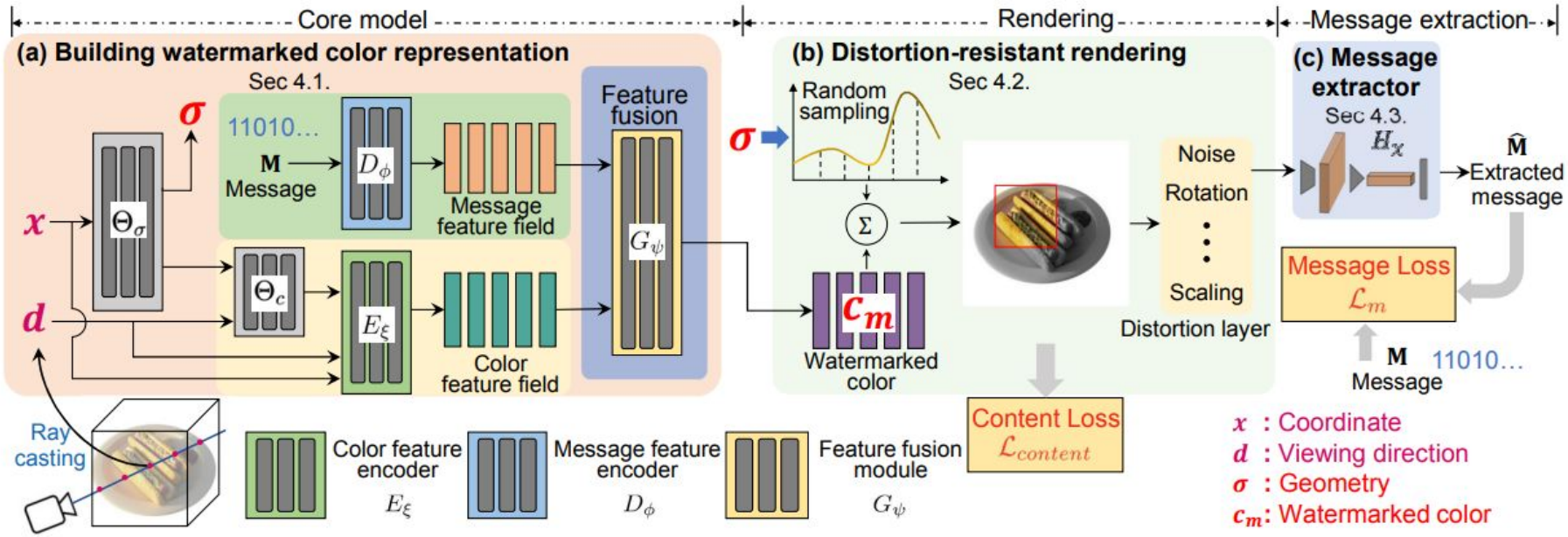
- Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, Mingkui Tan
  - NeRFのレンダリングにおいて複数の光線について同時に計算
  - 物体マスクとAttentionを導入しロバストで精緻な新規視点画像生成を実現



ICCV 2023 の動向・気付き（108/165）

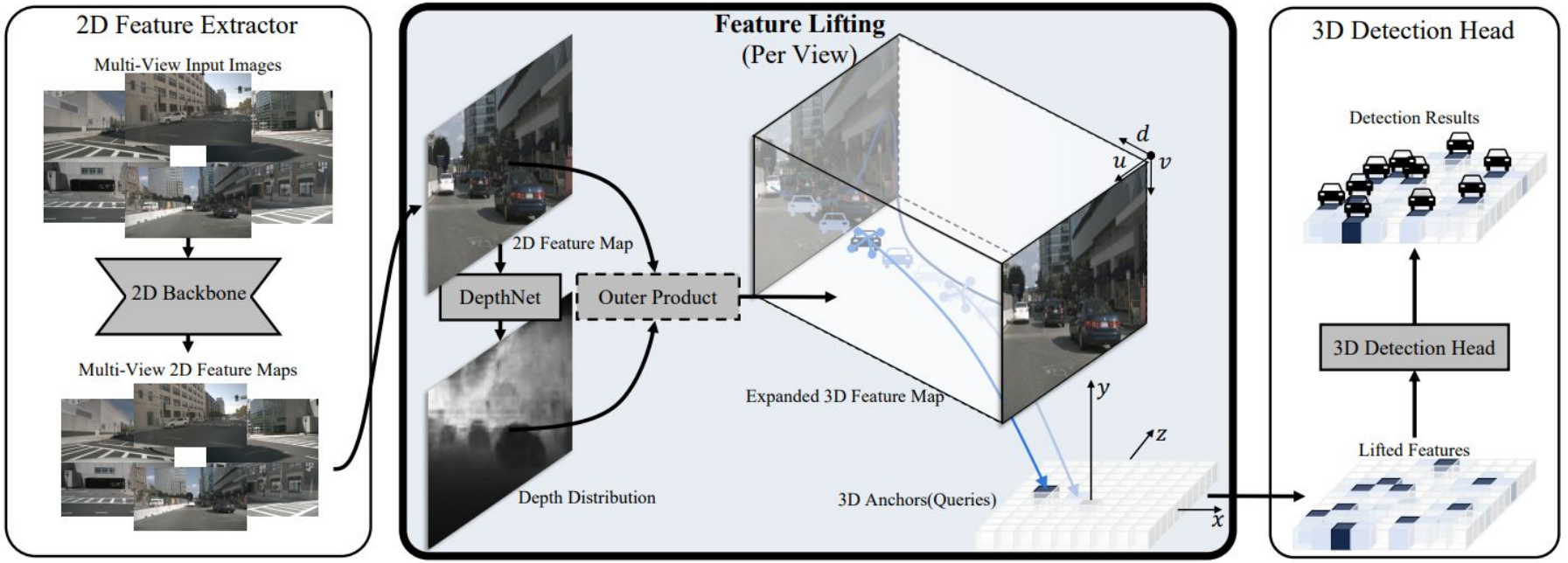
CopyRNeRF: Protecting the CopyRight of Neural Radiance Fields

- Ziyuan Luo, Qing Guo, Ka Chun Cheung, Simon See, Renjie Wan
  - NeRFにメッセージ（著作権情報など）を埋め込むための手法
  - メッセージをカラーと混ぜ合わせてレンダリングし，通常のNeRF同様に再構成誤差で最適化
  - レンダリングされた画像をさらにAugmentationした画像から，別のネットワークでメッセージが取り出せるようにロスを追加



DFA3D: 3D Deformable Attention For 2D-to-3D Feature Lifting

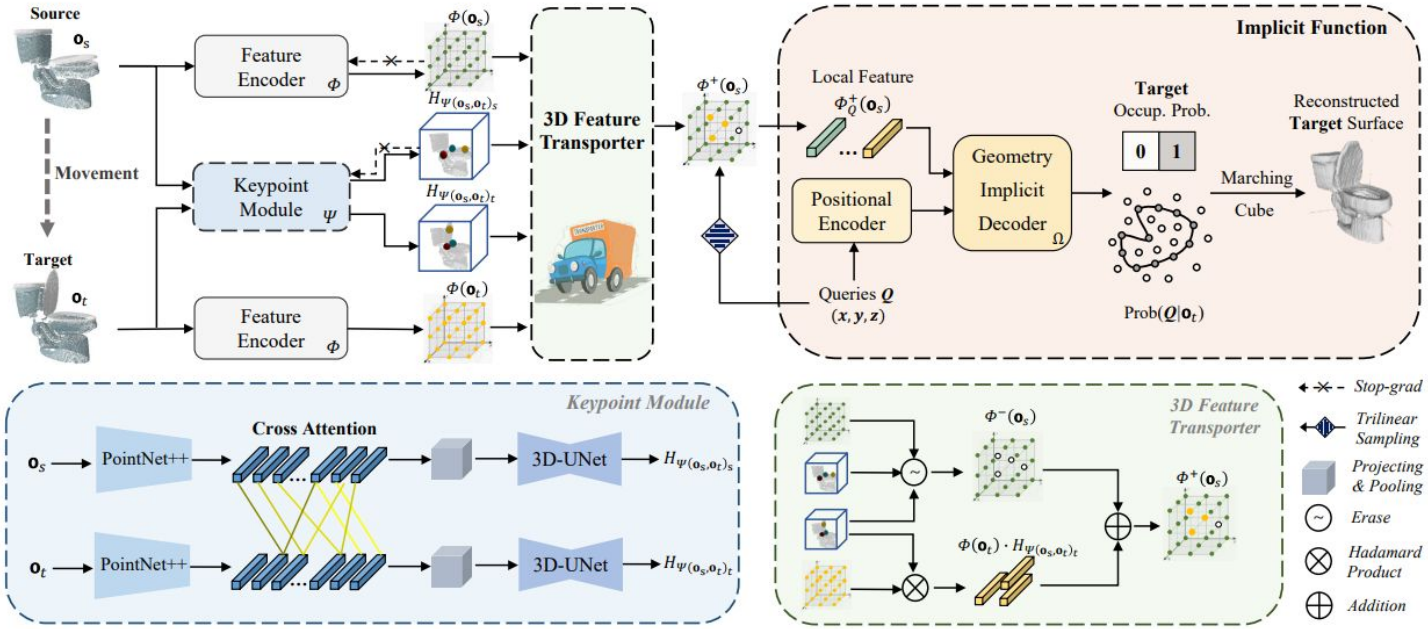
- Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, Lei Zhang
  - 2Dの特徴量を3Dに反映させるFeature Liftingを提案
  - 3D Deformable Attentionを導入し，深度推定を利用して3D各点に対するAttentionを行う
  - 密にマッピングを計算すると重いため，On the flyで計算することで効率化

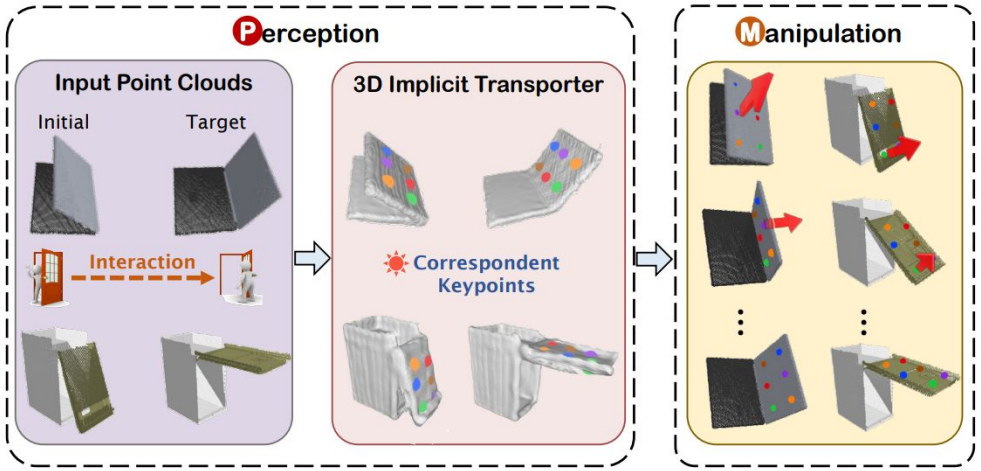


ICCV 2023 の動向・気付き（110/165）

3D Implicit Transporter for Temporally Consistent Keypoint Discovery

- Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, Xinliang Zhang, Jian Zhao
  - 3D特徴点を自己教師で学習する手法
  - 3D表現として点群を入力，OccupancyによるNeural Fieldで再構成
  - 特徴点の対応からグリッド上での特徴量を変形させ，元の形状を再構成するように学習

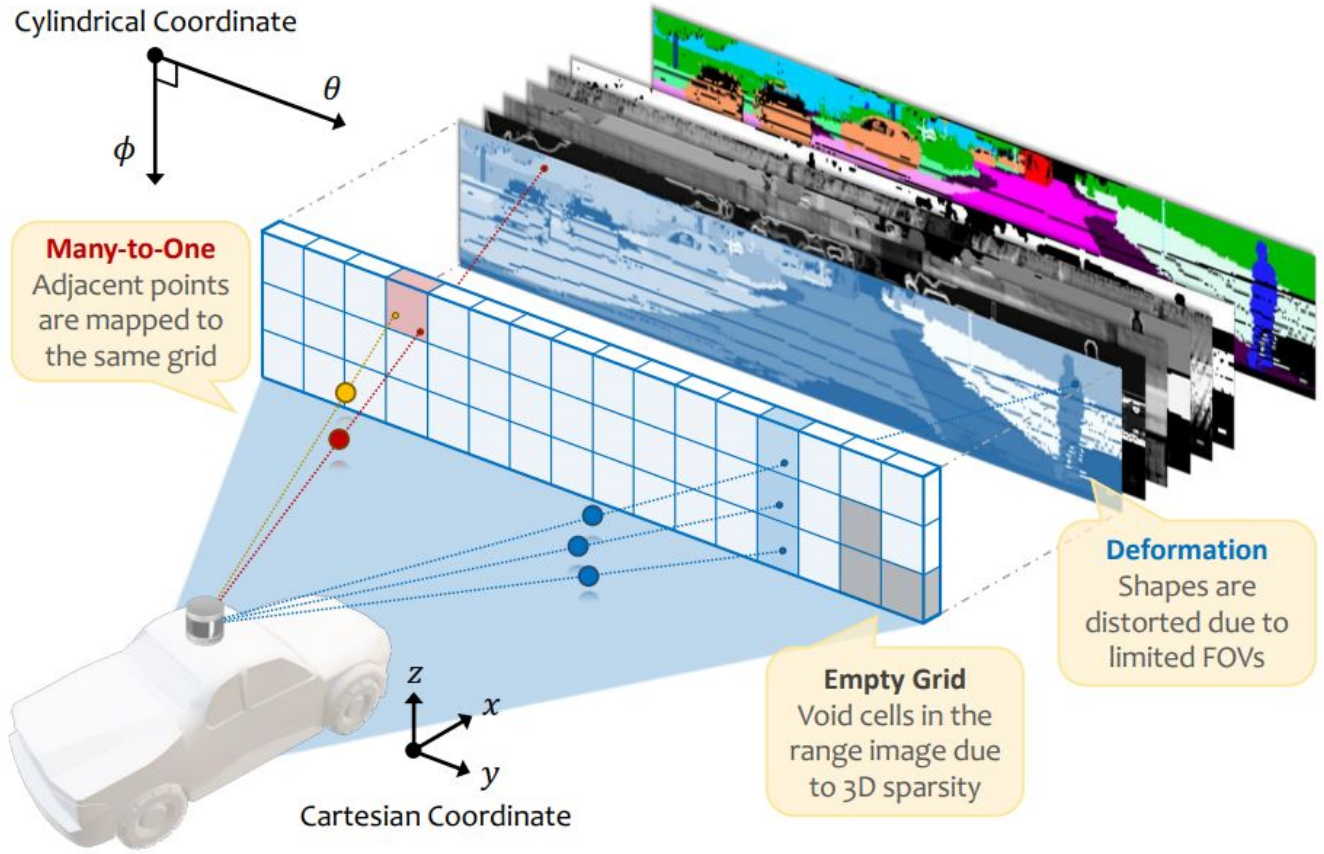




ICCV 2023 の動向・気付き(111/165)

Rethinking Range View Representation for LiDAR Segmentation

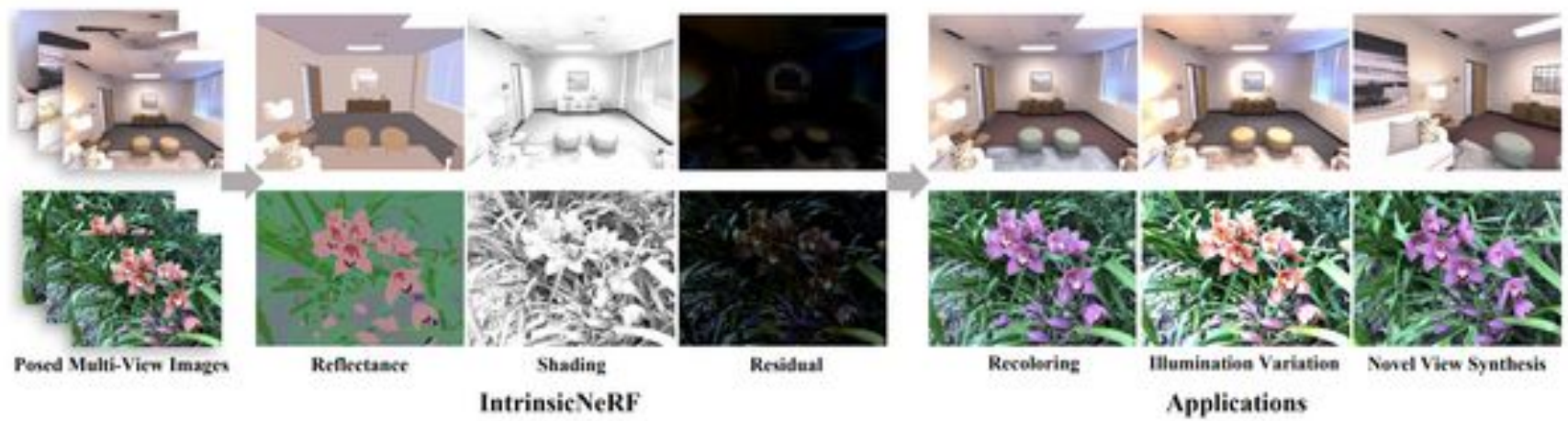
- Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, Ziwei Liu
  - LiDAR点群のセグメンテーションのための2Dベースの手法RangeFormerの提案
  - グローバルなコンテキストを利用できるように，fully-convolutional NNではなくSelf-Attentionを利用
  - 2Dにマッピングすることにより生じるオーバーラップに対応
  - 効率よく視野が広いLiDAR画像を扱える枠組みSTRを同時に提案
  - LiDAR画像であることを利用したAugmentationも提案



ICCV 2023 の動向・気付き(112/165)

IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis

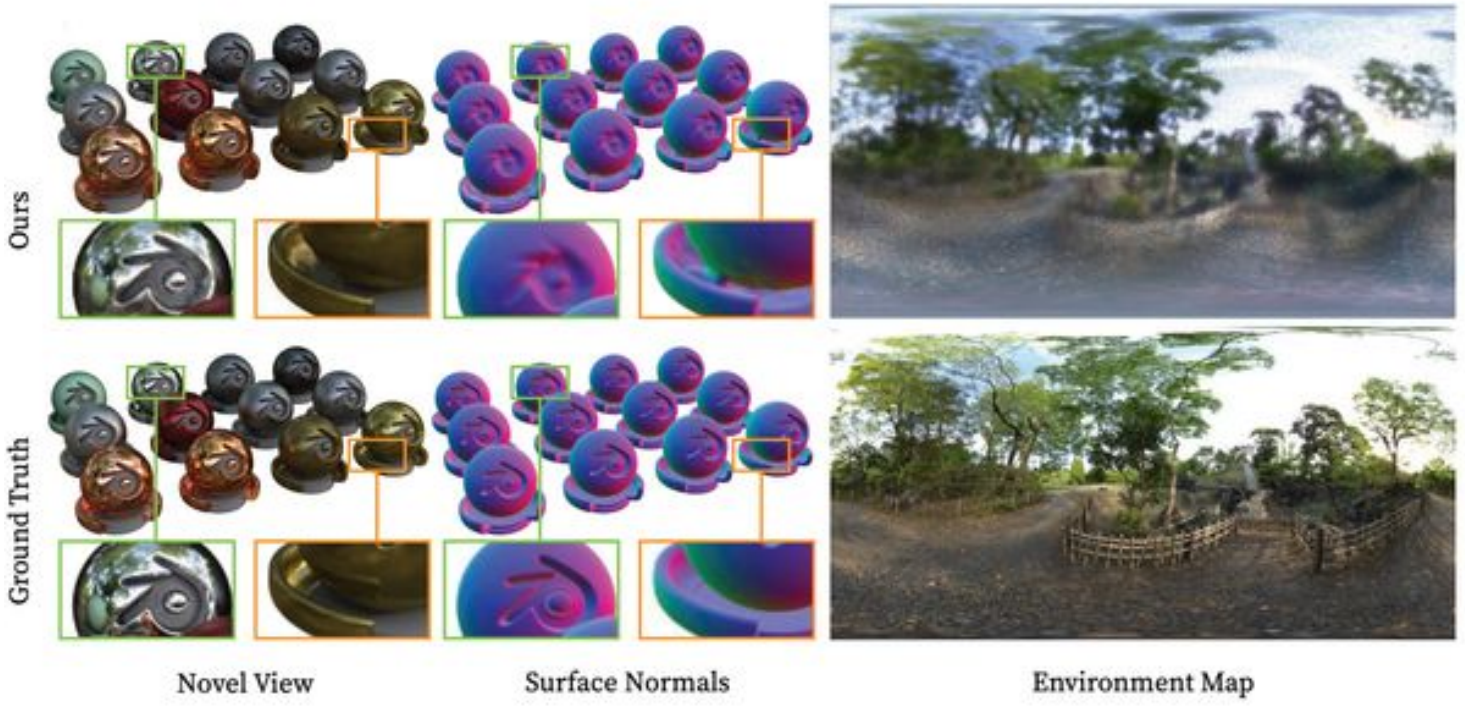
- Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, Guofeng Zhang
  - 観測画像を階層的クラスタリングでReflectance, Shading, Residualに固有画像分解しSegment情報やRelightingに有用なNeuralFieldを提案
  - 説得力のある色の変更やRelightingが可能



ICCV 2023 の動向・気付き(113/165)

Neural Microfacet Fields for Inverse Rendering

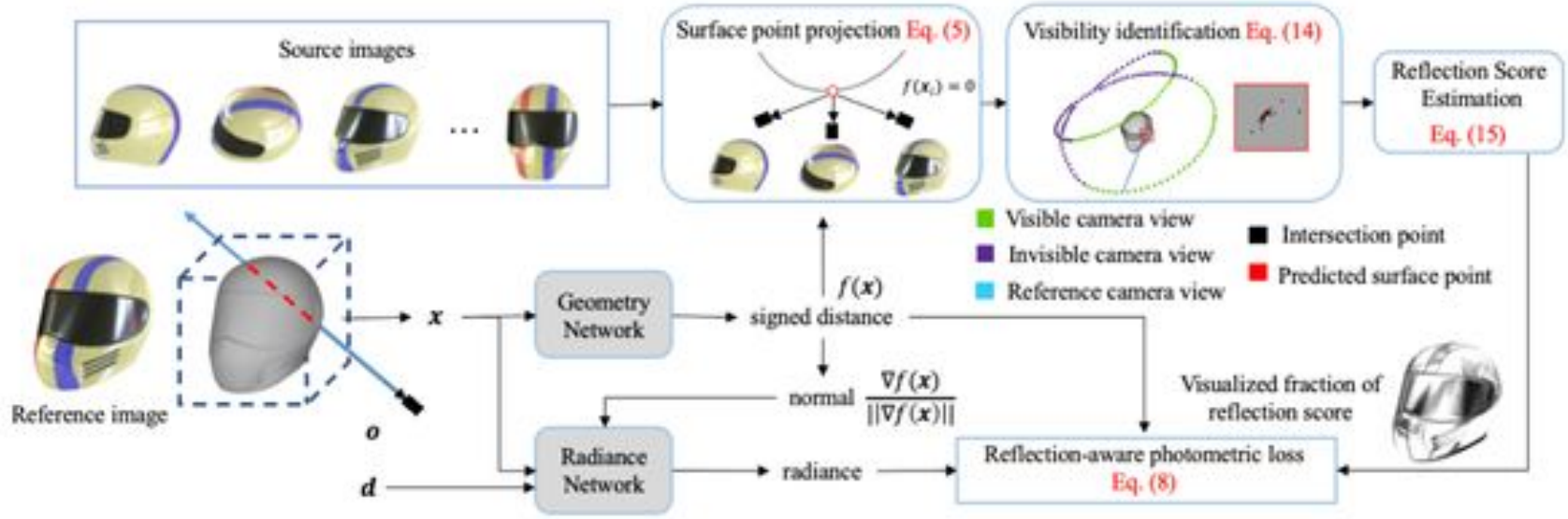
- Alexander Mai, Dor Verbin, Falko Kuester, Sara Fridovich-Keil
  - ボリューメトリック設定内でマイクロファセット反射率モデルを使用し、高周波な照明の復元を実現
  - モンテカルロ積分により非凸オブジェクト上のリアルな相互反射をモデル化可能
  - ライティングに無限遠照明を仮定しているため制限あり



ICCV 2023 の動向・気付き（114/165）

Ref-NeuS: Ambiguity-Reduced Neural Implicit Surface Learning for Multi-View Reconstruction with Reflection

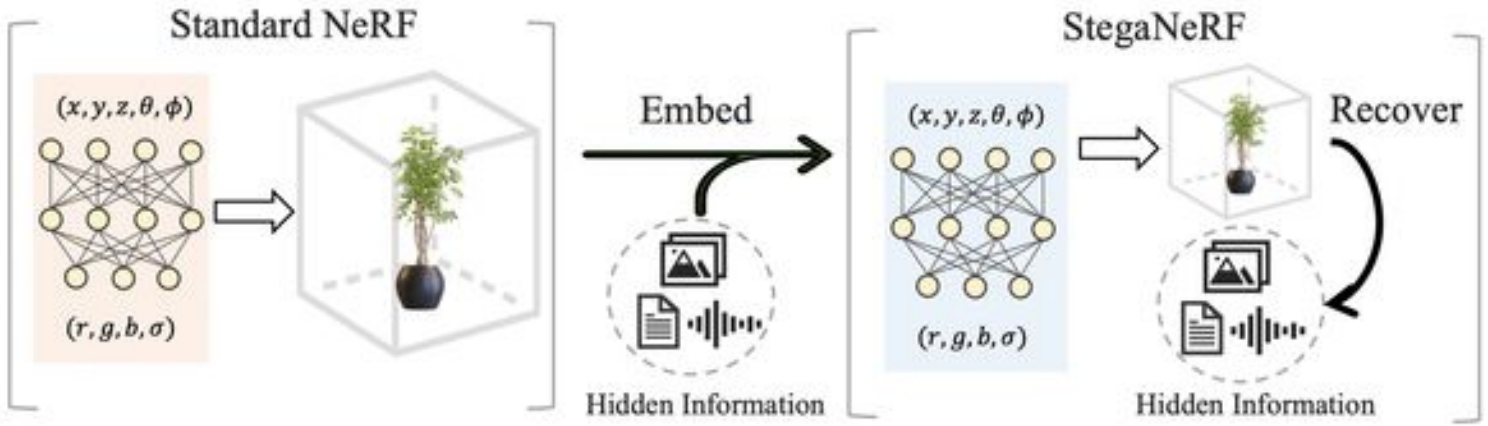
- Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, Ying-Cong Chen
  - 反射を考慮した測光損失を設計し、レンダリングされた色をガウス分布としてモデル化
  - 明示的な反射スコアを推定する異常検出器を利用し、反射面の曖昧さを軽減
  - 従来、反射面の正確な再構成における曖昧さが課題であったが、Warping由来の測光損失を目的関数に追加することにより曖昧さを軽減した点が優秀



ICCV 2023 の動向・気付き（115/165）

StegaNeRF: Embedding Invisible Information within Neural Radiance Fields

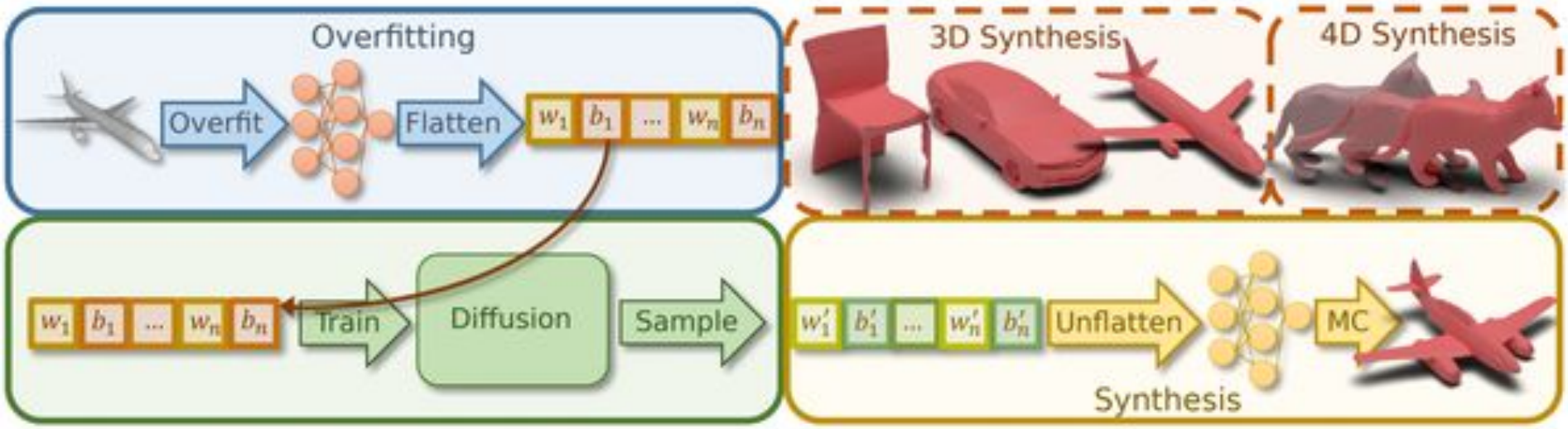
- Chenxin Li, Brandon Y. Feng, Zhiwen Fan, Panwang Pan, Zhangyang Wang
  - NeuralFieldに対して、品質を大きく下げることなく、署名などのsteganographicを付与する手法を提案
  - NeRFレンダリング画像に、カスタマイズ可能で、知覚不可能で、復元可能な情報を付与するという新しい問題へ初めて挑戦した点が新規性
  - レンダリング画像品質と付与情報の復元率を評価。ほとんど品質を落とさずに付与情報を99%以上の割合で復元



ICCV 2023 の動向・気付き(116/165)

HyperDiffusion: Generating Implicit Neural Fields with Weight-Space Diffusion

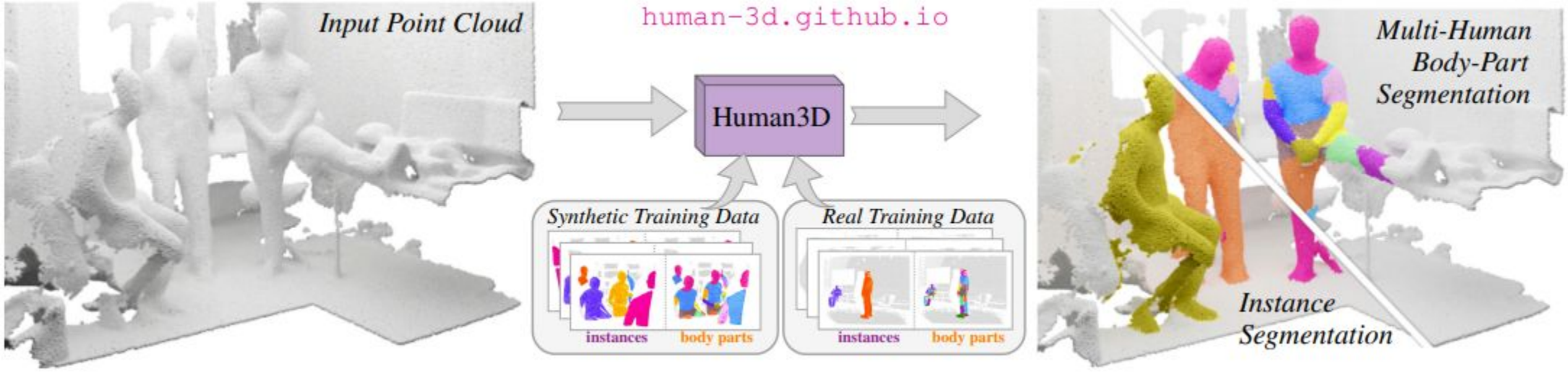
- Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, Angela Dai
  - ニューラル場のMLPの重みを直接対象にした拡散プロセスを提案
  - 個々のシーンについて最適化したニューラル場をまず訓練し、その重み空間で拡散プロセスを訓練する
  - 3D形状および4Dメッシュアニメーションの複雑な情報を単一のフレームワークで表現可能



ICCV 2023 の動向・気付き（117/165）

3D Segmentation of Humans in Point Clouds with Synthetic Data

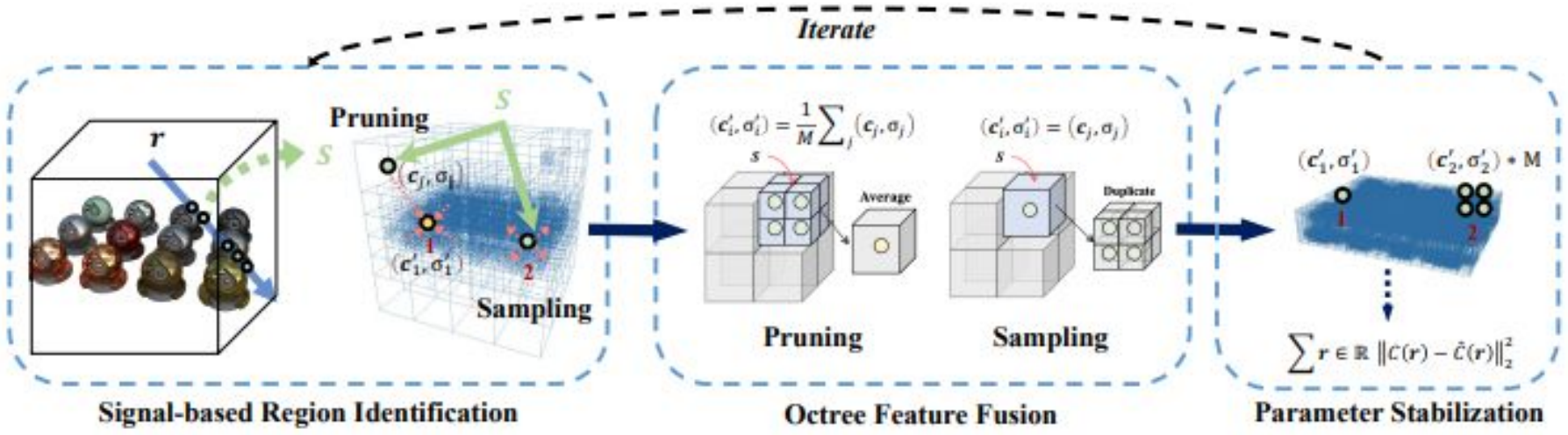
- Ayca Takmaz, Jonas Schult, Irem Kaftan, Mertcan Akcay, Bastian Leibe, Robert Sumner, Francis Engelmann, Siyu Tang
  - 既存の3Dデータセットには人間の3Dセグメンテーションデータが存在しない
  - ScanNetに人間の合成データを張り付けることで3D屋内シーンに人がいるデータを自動で生成する手法Human3Dを提案
  - 生成したデータ事前学習することで，様々な3D人物セグメンテーションタスクのパフォーマンスを向上



ICCV 2023 の動向・気付き(118/165)

Dynamic PlenOctree for Adaptive Sampling Refinement in Explicit NeRF

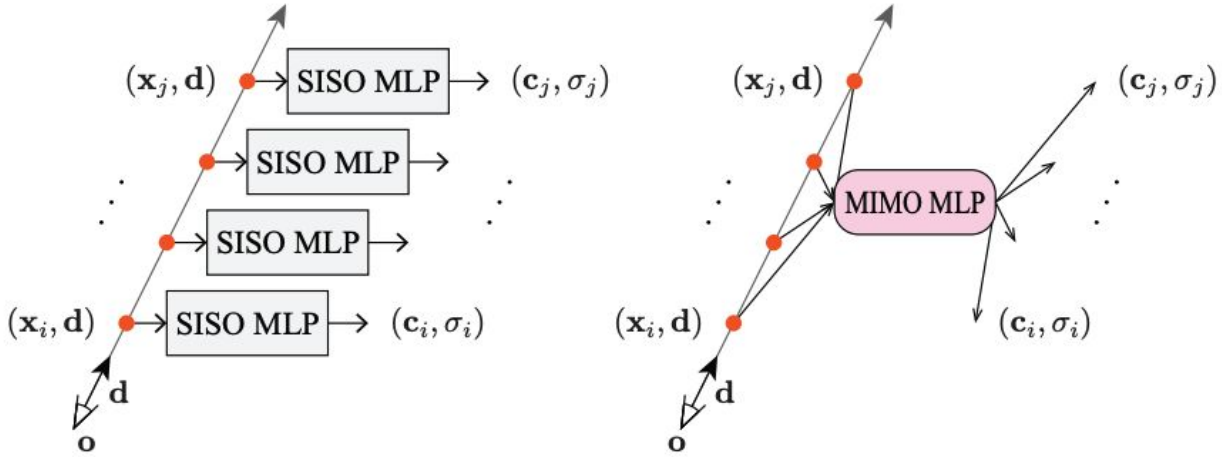
- Haotian Bai, Yiqi Lin, Yize Chen, Lin Wang
  - NeRFで学習したシーン情報を，八分木にキャッシュすることでリアルタイムレンダリングを実現したPlenOctrees(ICCV2021)を発展させた研究
  - 八分木の分割を適応的に調整することで、よりメモリ効率が高く高品質な表現が可能に
  - PlenOctreesと比較してモデルサイズを半分以上削減したことで，1台のRTX 3090 GPUで800x800の画像をレンダリングするのに452FPSとさらに高速化
  - 従来のNeRF-SHを使用しているため学習時間は長い

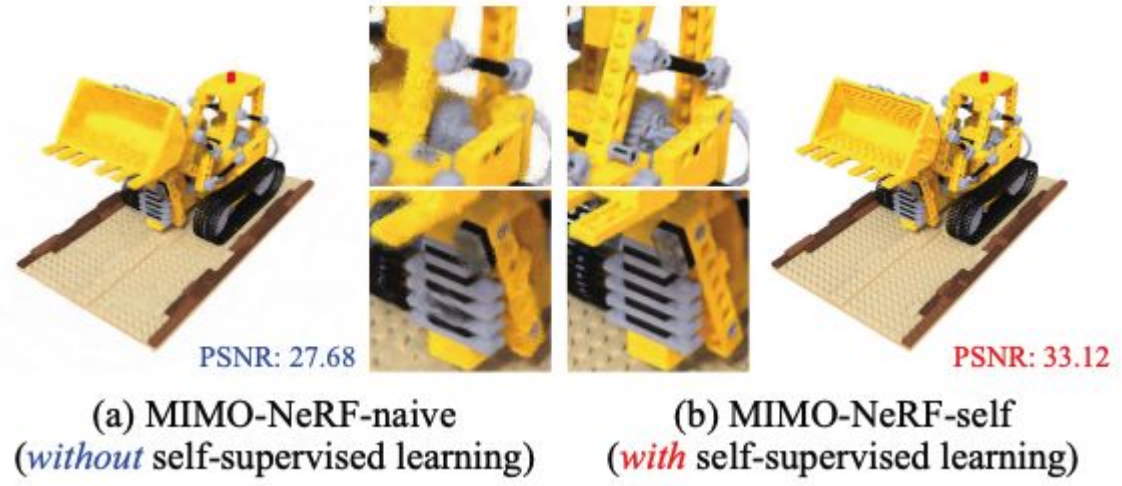


ICCV 2023 の動向・気付き(119/165)

MIMO-NeRF: Fast Neural Rendering with Multi-input Multi-output Neural Radiance Fields

- Takuhiro Kaneko
  - 従来のNeRFはSISO MLPを用いて3D座標と視線方向を色と密度にサンプル単位でマッピングしていたため，推論が遅かった．そこで提案手法ではMIMO MLPを用いてグループ単位でサンプル点のマッピングを行うことで、クエリするMLPの削減により学習や推論を高速化
  - グループ内の座標の選択によって各点の色と体積密度が異なる曖昧性は，事前学習に依存しない自己教師あり学習により対処
  - SISO MLPを利用した従来のNeRF手法(DONeRF, Tensorf等)にも拡張可能

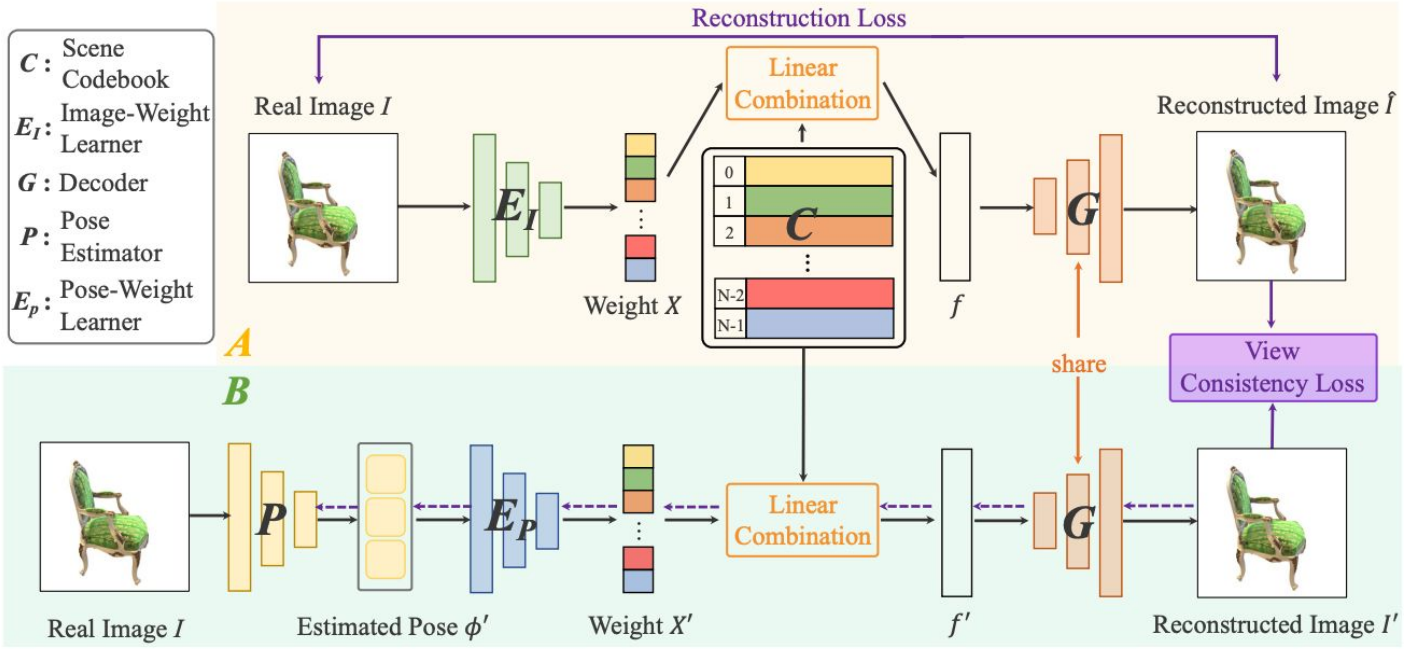




ICCV 2023 の動向・気付き（120/165）

Pose-Free Neural Radiance Fields via Implicit Pose Regularization

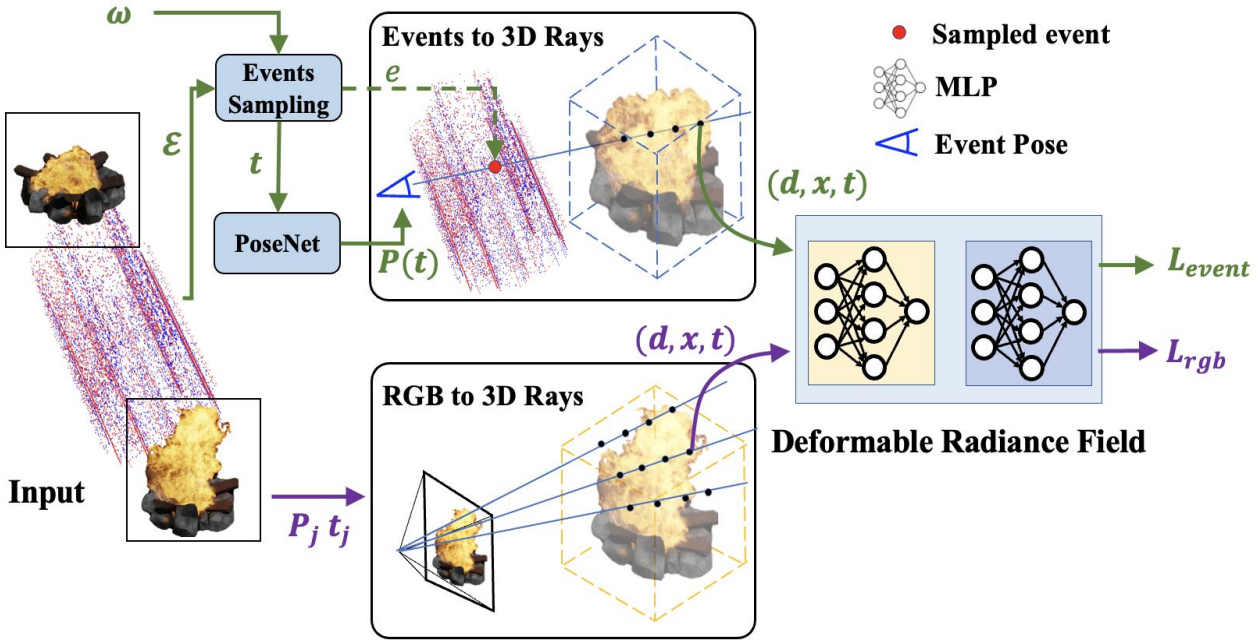
- Jiahui Zhang, Fangneng Zhan, Yingchen Yu, Kunhao Liu, Rongliang Wu, Xiaoqin Zhang, Ling Shao, Shijian Lu
  - RGBロスのみでカメラポーズ推定も行うNeRFは実画像へのロバスト性が低くなるため，暗黙的なポーズ正則化によるロバスト性の向上を提案
  - 画像群からシーンの特徴をエンコードしたポーズの分布を事前情報として利用し，ビュー一貫性ロスも加えることでカメラポーズ情報なしで学習を行うことが可能
  - 既存のGANベースのアプローチ(GNeRF)と比較して性能向上

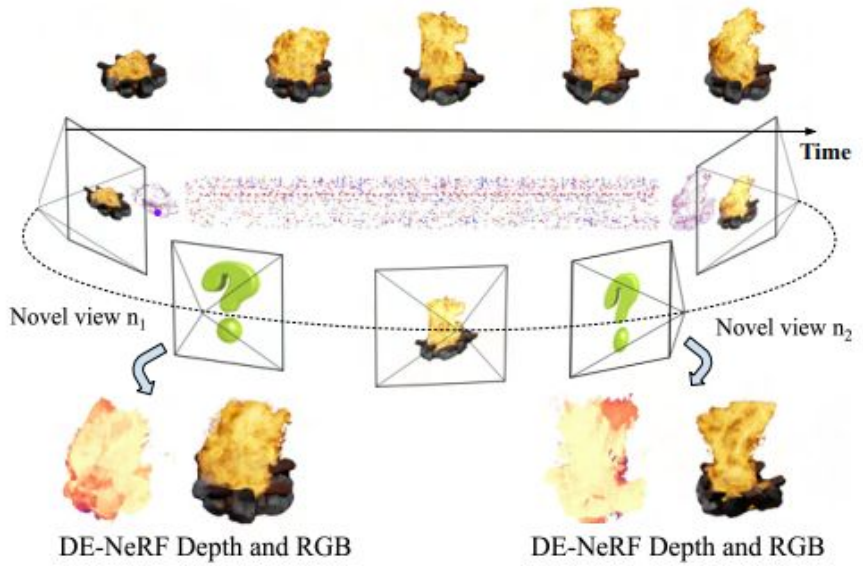


ICCV 2023 の動向・気付き（121/165）

Deformable Neural Radiance Fields using RGB and Event Cameras

- Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool
  - RGBカメラとイベントカメラを入力とするDeformable NeRF初めてを提案
  - イベントカメラを用いることにより，RGBカメラで捉えられていないフレーム間の素早い動きといった詳細な動きを捉えた三次元再構成が可能
  - カメラポーズの学習も行う



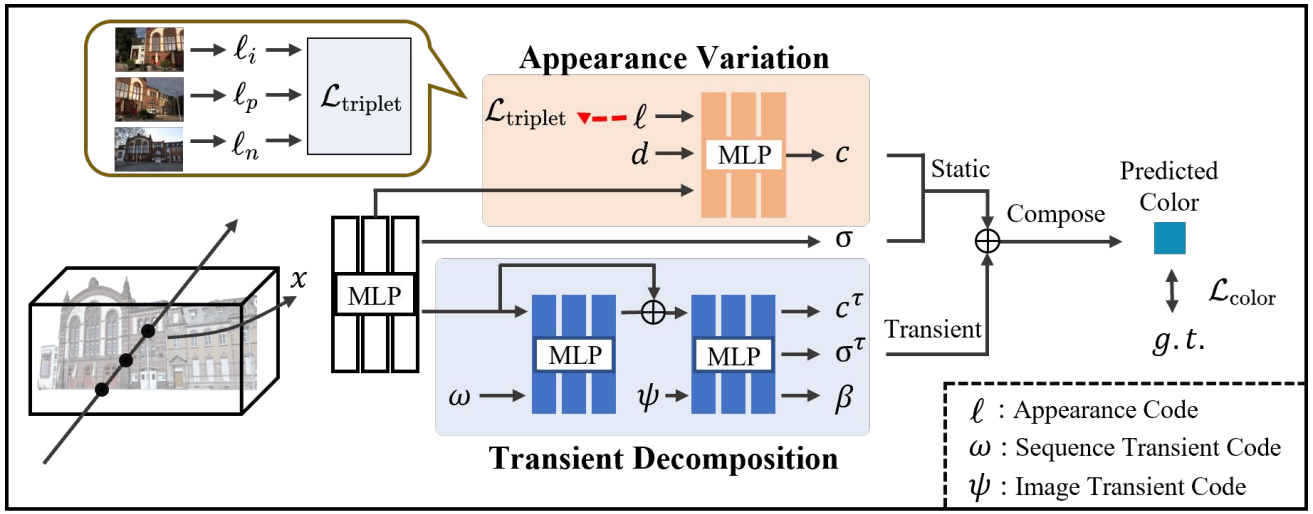


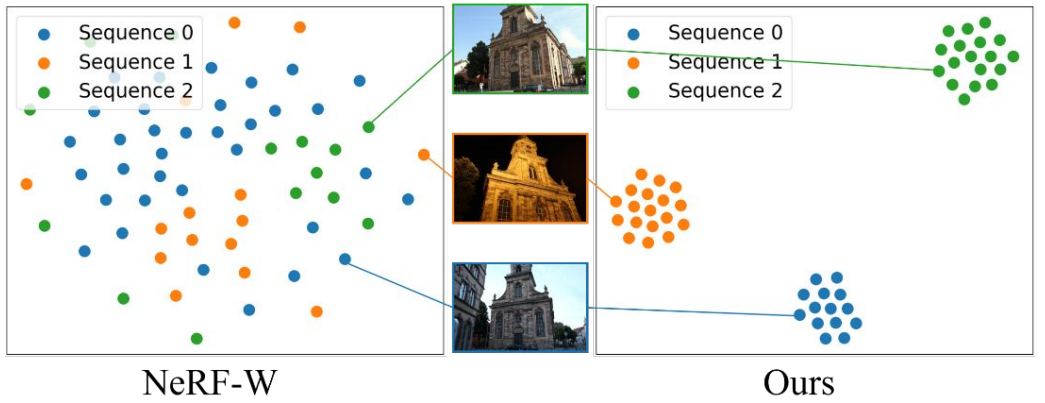
https://arxiv.org/pdf/2309.08416.pdf

ICCV 2023 の動向・気付き（122/165）

NeRF-MS: Neural Radiance Fields with Multi-Sequence (Poster)

- Peihao Li, Shaohui Wang, Chen Yang, Bingbing Liu, Weichao Qiu, Haoqian Wang
  - 異なる時間，異なるカメラから撮影された動画からNeRFによるin the wildなNVSを行う
  - 画像を入力とするNeRF in the wild(CVPR2021)とは異なり，複数の動画を入力することでより柔軟な表現が可能に
  - 静的箇所の色を決定する画像ごとに用意するappearance codeをtriplet lossを用いて動画ごとに近く学習させることで品質の向上をもたらした
  - 不確実な箇所（歩行者，車など）を尤もらしく再構成するため,sequence transient decomposition module を提案



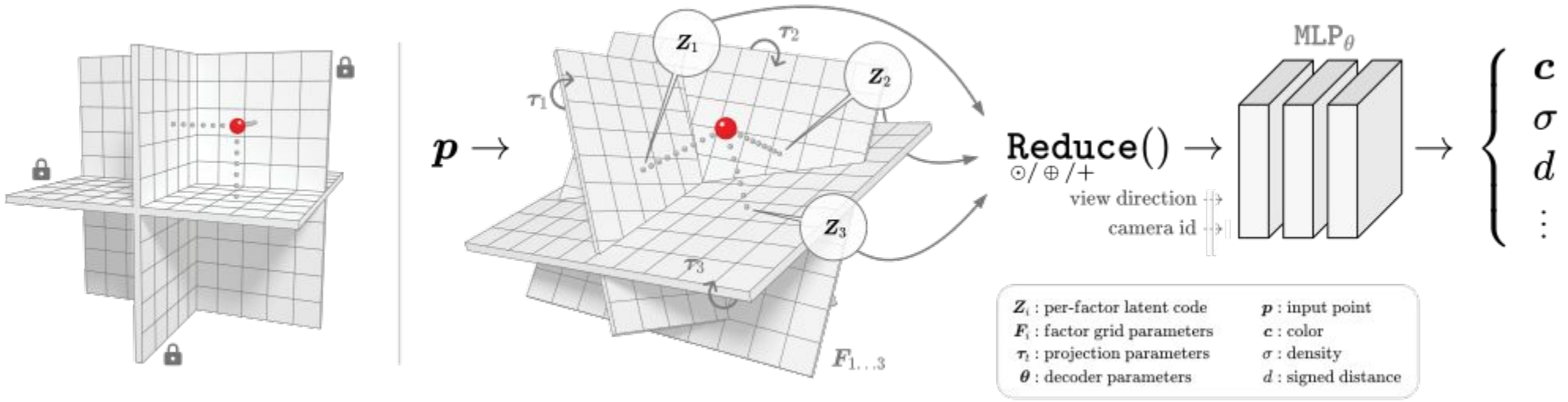


https://nerf-ms.github.io/

ICCV 2023 の動向・気付き（123/165）

Canonical Factors for Hybrid Neural Fields

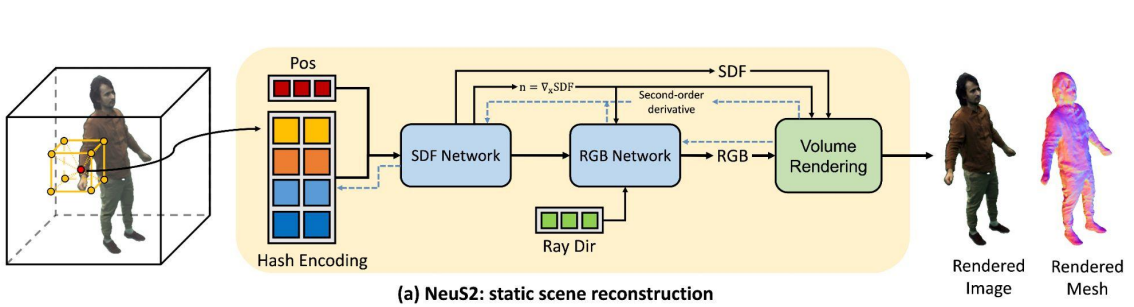
- Brent Yi, Weijia Zeng, Sam Buchanan, Yi Ma
  - 従来の特徴グリッド表現は軸に沿ったバイアスが発生
  - 特徴グリッドへの射影を学習可能にすることにより、バイアスを取り除くことに成功
  - ロバスト性、コンパクト性、実行時間において優位性を実証

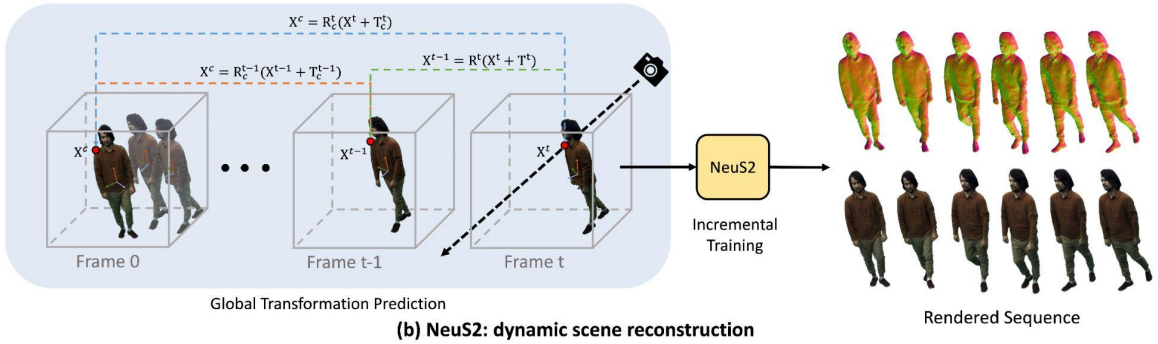


ICCV 2023 の動向・気付き（124/165）

NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction

- Yiming Wang, Qin Han,Marc Habermann, Kostas Daniilidis, Christian Theobalt, Lingjie Liu
  - 従来手法(NeuS)から2桁オーダーの高速化をしたニューラルサーフェス表現を提案
  - 多解像度ハッシュエンコーディングと効率的な2階微分計算をCUDA上で実装
  - 漸進的な学習ストラテジーにより、ダイナミックシーンの再構成に拡張

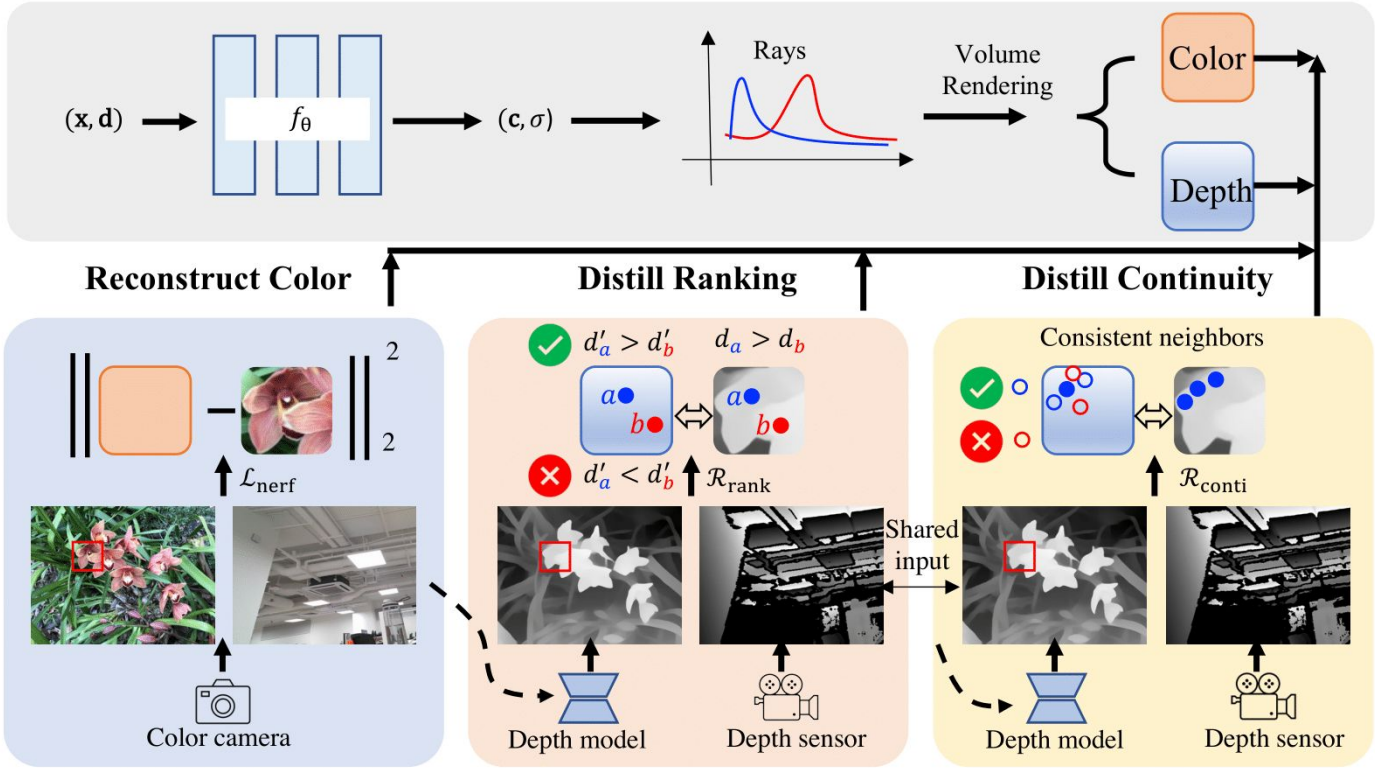




ICCV 2023 の動向・気付き（125/165）

SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis

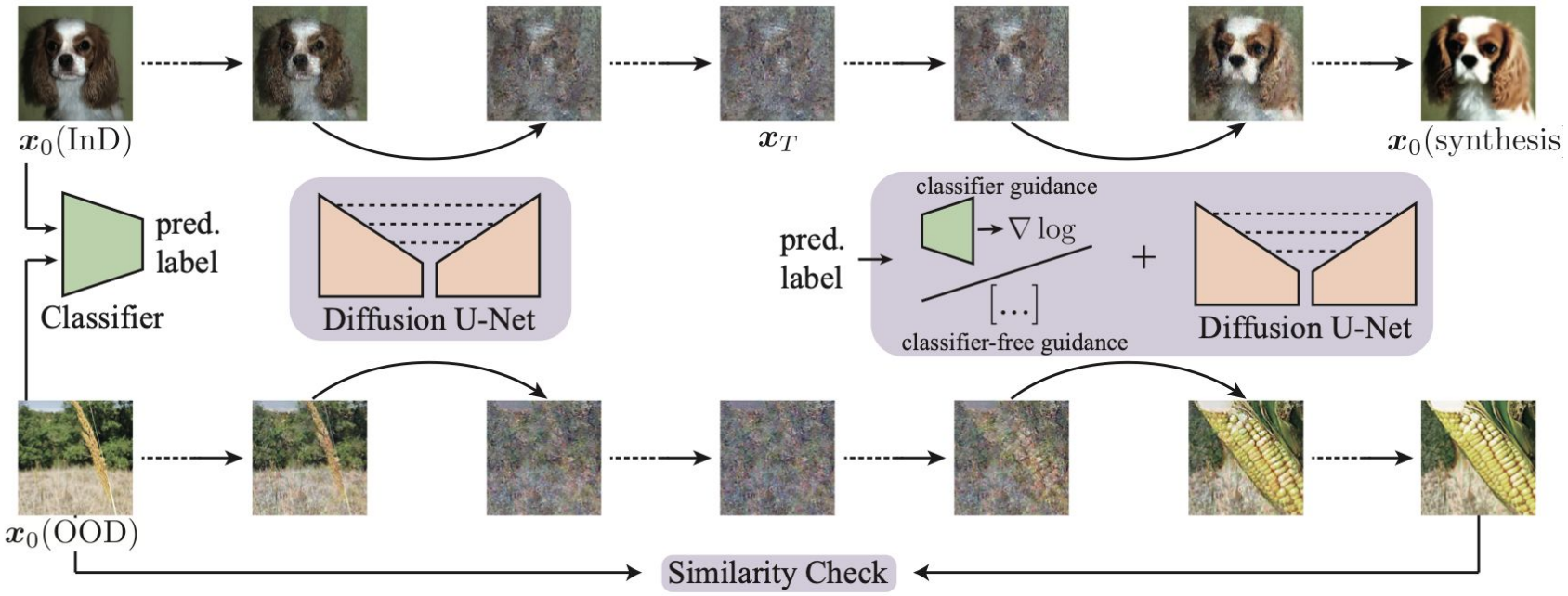
- Guangcong Wang, Zhaoxi Chen, Chen Change Loy, Ziwei Liu
  - 実データの不正確な深度マップからのFew-shot NeRFの提案
  - 粗い深度マップから深度モデルを蒸留しロバストな深度事前分布を抽出
  - 深度モデルを蒸留するための局所深度ランキング制約と空間連続性制約を提案
  - 深度ベースモデルを含むFew-shot NeRFのSOTAを凌駕



ICCV 2023 の動向・気付き（126/165）

DIFFGUARD: Semantic Mismatch-Guided Out-of-Distribution Detection using Pre-trained Diffusion Models

- 拡散モデルを使用したOODの検知手法：DIFFGUARDの提案
  - 画像をDDIM inversionに入力した後，別途用意した分類器によって予測したラベルに向かって画像を生成．
  - 入力画像と生成画像の類似度によってOODスコアを算出．CIFAR10の実験でSOTAを達成



Method	CIFAR-100		TINYIMAGENET		average	
	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓
EBO [28]	86.19	*51.32*	88.61	*44.89*	87.41	*48.11*
KNN [44]	89.62	52.19	91.48	46.18	90.55	49.19
MLS[11]	86.14	52.04	88.53	45.38	87.34	48.71
ViM[46]	87.16	56.81	88.85	52.89	88.01	54.85
MC-Dropout[7]	86.74	61.49	88.32	58.44	87.53	59.97
Deep Ens.[9]	89.97	54.61	91.31	51.23	90.64	52.92
ConfidNet*[2]	85.92	72.37	87.16	69.75	86.54	71.06
DiffNB [27]	89.79	53.23	91.77	45.88	90.78	49.56
Ours	89.88	52.67	91.88	45.48	90.88	49.08
Ours$_{+EBO}$	*89.93*	**50.77**	*91.95*	**43.58**	*90.94*	**47.18**
Ours$_{+Deep\ Ens.}$	**90.40**	52.51	**91.98**	45.04	**91.19**	48.78
Ours(Oracle)	98.34	7.94	98.52	7.11	98.43	7.53

ICCV 2023 の動向・気付き（127/165）

Understanding the Feature Norm for Out-of-Distribution Detection

- OODが入力された時のNNにおける隠れ層の特徴量のノルムについて分析した論文
  - OODはInDより隠れ層のノルムが小さくなることが知られているが十分に分析されていない
    - 従来のノルムはニューロンの非活性化傾向を考慮できずOODをInDと間違えることが判明
  - ニューロンの活性化，非活性化を考慮したノルム：Negative-aware norm(NAN)を提案
    - ハイパーパラメータの調整不要．既存のOOD検知手法にも組み込みが可能．

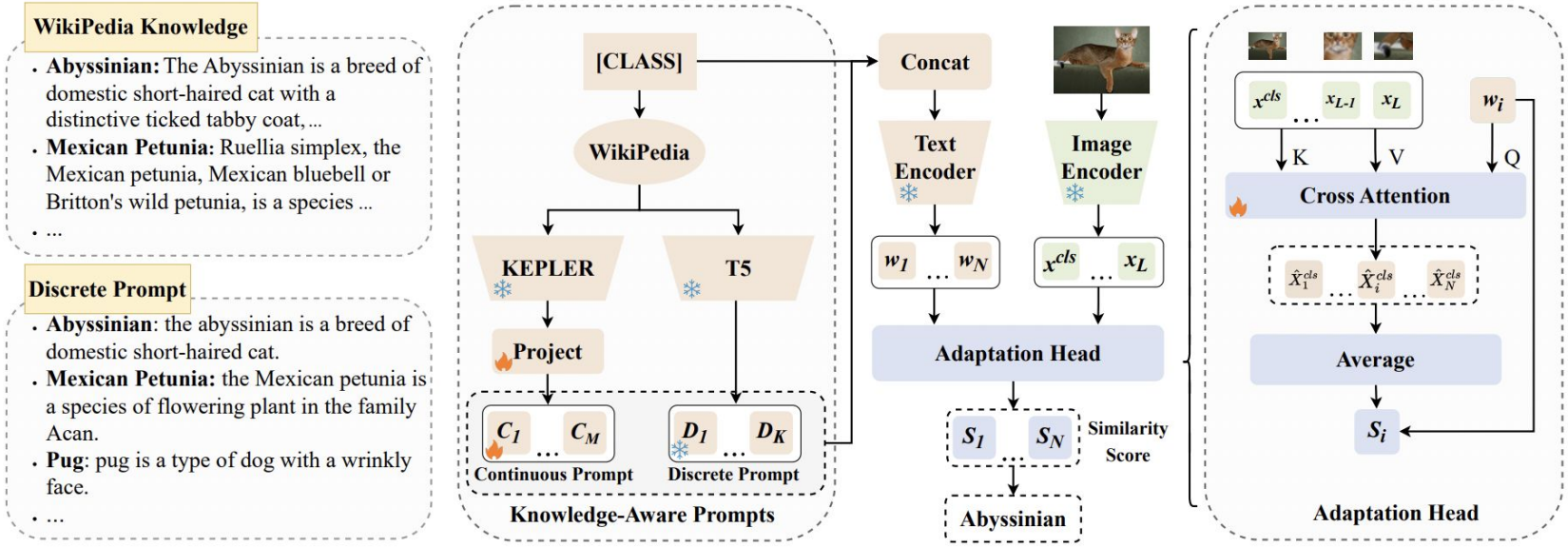
	hyper.-free	label-free	bank-free	iNaturalist		SUN		Places		Texture	
				AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
With Supervised Labels of ID:											
MSP	✓		✓	93.78	29.74	84.56	59.54	84.28	60.94	84.90	50.02
Energy	✓		✓	96.17	20.98	88.91	47.05	87.70	51.15	88.90	39.31
MaxLogit	✓		✓	95.99	22.06	88.43	50.90	87.37	53.78	88.42	42.25
KL	✓		✓	96.17	20.98	88.91	47.06	87.70	51.15	88.90	39.31
Mahalanobis	✓		✓	94.79	35.04	86.55	64.99	83.92	70.31	95.52	15.02
ViM			✓	95.54	27.75	89.85	48.12	87.05	57.82	95.18	14.47
SSD		✓	✓	94.08	37.77	88.06	58.38	84.70	63.89	96.96	11.63
KNN		✓		94.15	38.25	87.75	58.19	84.93	61.80	94.24	19.29
NAN (ours)	✓	✓	✓	96.94	15.86	92.77	29.81	91.46	37.21	88.09	43.46
Without Supervised Labels of ID (detectors based on supervised labels are not available):											
SSD		✓	✓	60.34	93.87	80.89	78.41	77.23	81.26	90.19	33.53
KNN		✓		84.53	78.71	82.26	76.06	77.50	80.65	91.99	24.61
NAN (ours)	✓	✓	✓	92.90	36.09	86.76	56.27	83.22	65.08	87.57	46.86

	AUROC↑	FPR95↓
NAN	92.32	31.59
NAN + KNN [42]	92.99	29.26
NAN + SSD [40]	93.42	27.51
NAN + ReAct [41]	93.91	29.23
NAN + ReAct [41] + KNN [42]	94.37	24.94
NAN + ReAct [41] + SSD [40]	**94.61**	**24.57**

ICCV 2023 の動向・気付き（128/165）

Knowledge-Aware Prompt Tuning for Generalizable Vision-Language Models

- 未知カテゴリに対応するプロンプトチューニング KAPTを提案
  - 既存のプロンプトチューニングは未知カテゴリに対しての汎化能力が不十分
  - discrete prompt:カテゴリの視覚的な外観を直接説明する要約プロンプト
  - continuous prompts:より広い領域（背景）をカバーするプロンプト
- 10個の画像分類データセットでCoCoOpを上回る精度



(a) **Average over 11 datasets.**

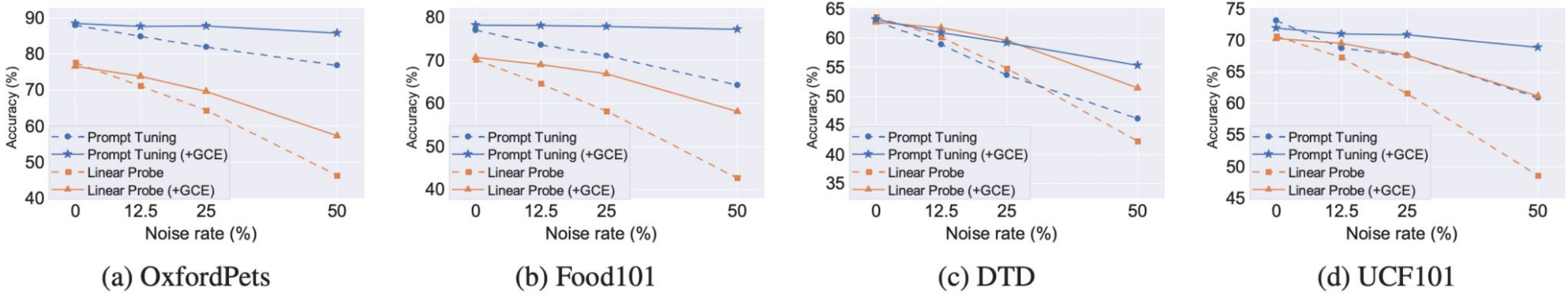
	Base	New	H
CLIP	64.83	70.27	67.44
CoOp	**79.88**	59.39	68.12
CoCoOp	76.70	67.30	71.69
KAPT	78.41	**70.52**	**74.26**

https://openaccess.thecvf.com/content/ICCV2023/papers/Kan_Knowledge-Aware_Prompt_Tuning_for_Generalizable_Vision-Language_Models_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（129/165）

Why Is Prompt Tuning for Vision-Language Models Robust to Noisy Labels?

- CLIPのような大規模データを学習したVision-languageに対するプロンプトチューニングが，ラベルのノイズにロバストであることを検証
  - 既存データセットのラベルをランダムにシャッフルして検証
  - 大規模なWeb上のデータを学習したText Encoderを用いたプロンプトチューニングは，ノイズデータが学習に与える悪影響を抑制することを検証

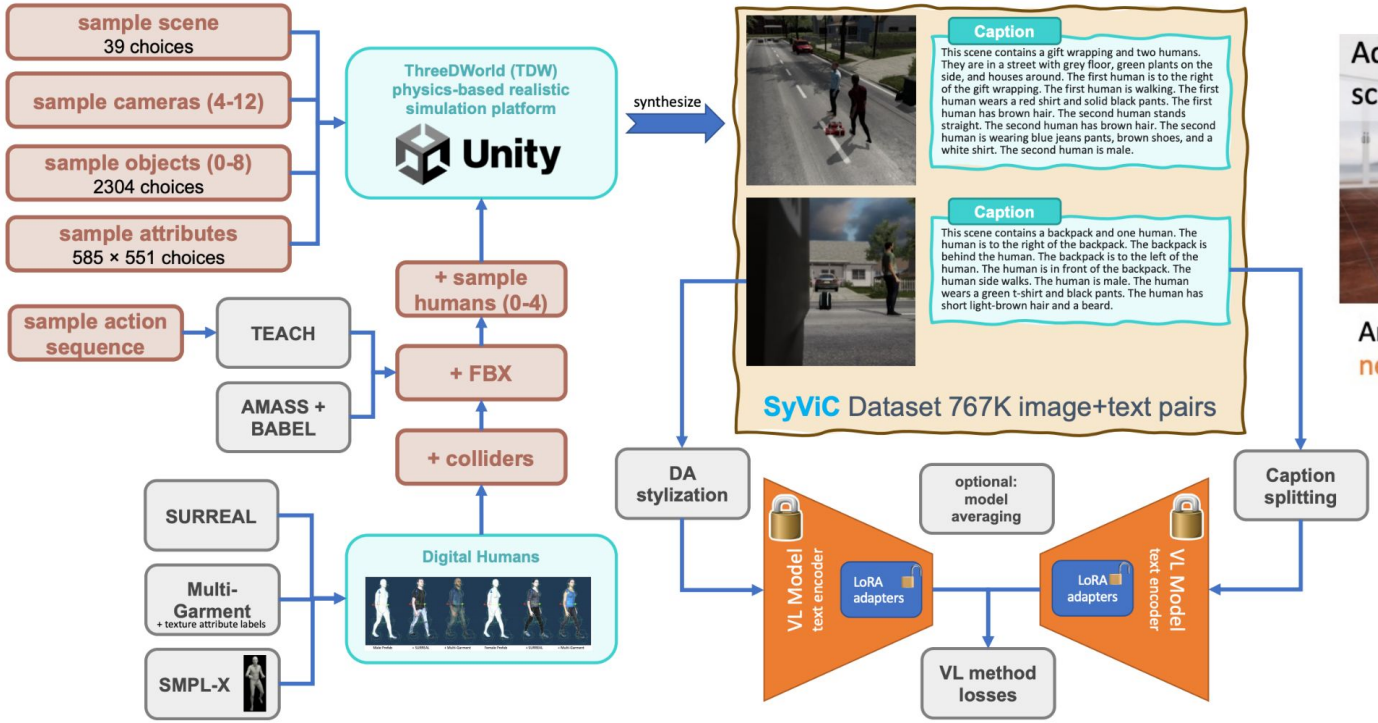


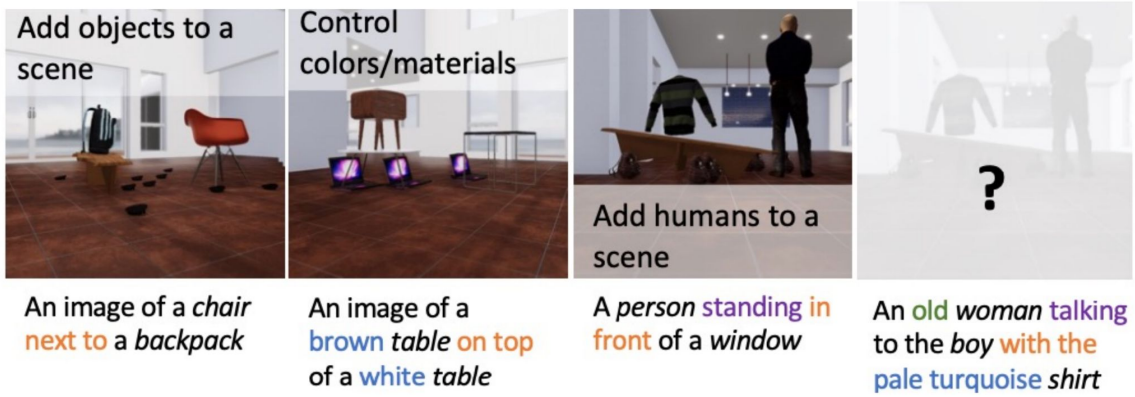
https://openaccess.thecvf.com/content/ICCV2023/papers/Wu_Why_Is_Prompt_Tuning_for_Vision-Language_Models_Robust_to_Noisy_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（130/165）

Going Beyond Nouns With Vision & Language Models Using Synthetic Data

- Vision & Languageモデルが苦手とする、属性, 行動, 関係, 状態などの「名詞を超える」視覚言語概念(Visual Language Concepts, VLC)を補うような合成データセットを提案。
- シミュレーションで動画を作成し、キャプションもメタデータから自動生成する。
- ゼロショット分類精度を落とさずに、WinogroundなどのVLCベンチマークでの大幅な向上を達成した。





https://openaccess.thecvf.com/content/ICCV2023/papers/Cascante-Bonilla_Going_Beyond_Nouns_With_Vision__Language_Models_Using_Synthetic_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（131/165）

A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance

- 大規模な視覚言語モデル、特にCLIP教師モデルの最近の進歩を活用し、未知の領域に汎化する小規模なモデルを訓練する。
- 生徒モデルの学習する画像表現が、対応するテキストをCLIP教師モデルでエンコードした表現に近づくようにする。
  - 情報を効率的に伝えるときや、余計なディテールに埋もれることなく特定のシーンの側面を強調するときに、テキスト情報は特に有効。

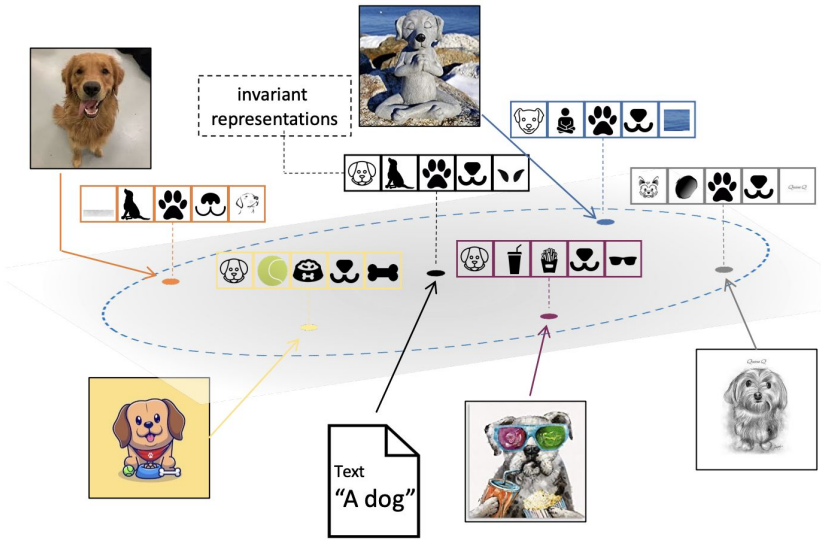
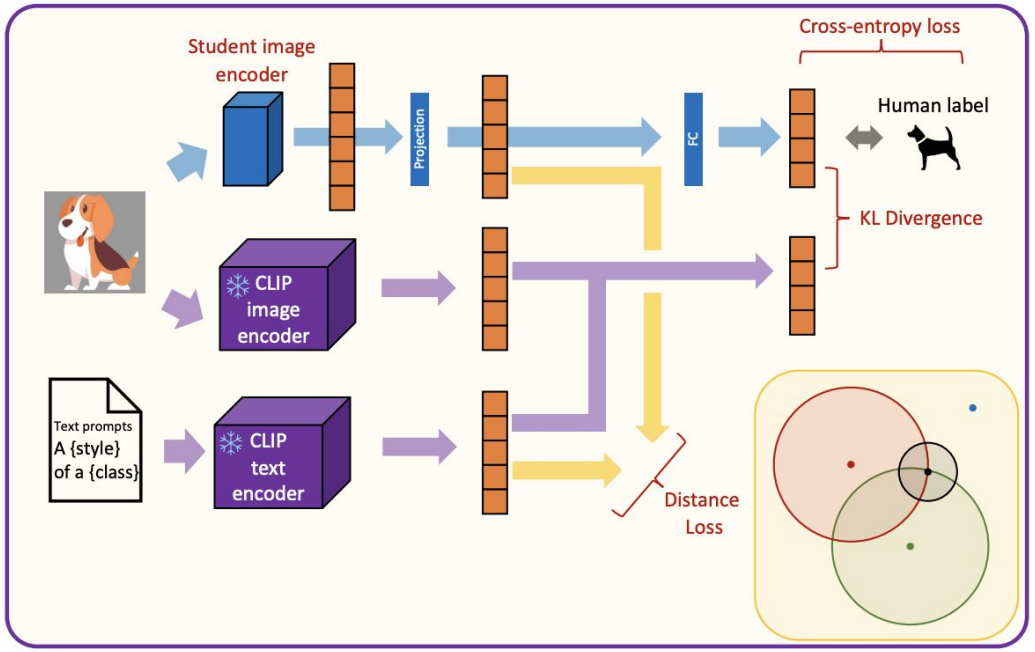


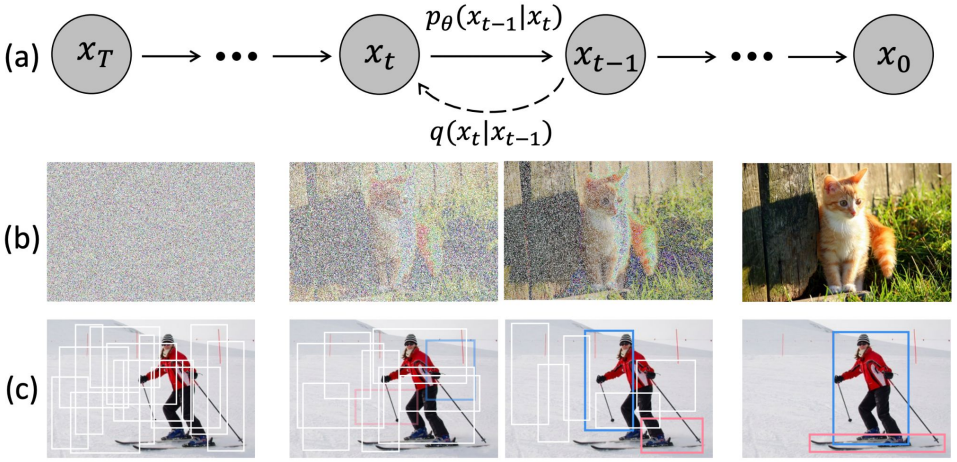
Figure 1. The key intuition behind our argument. While images can capture more details, text can directly summarize the core concept to represent the object of interest.

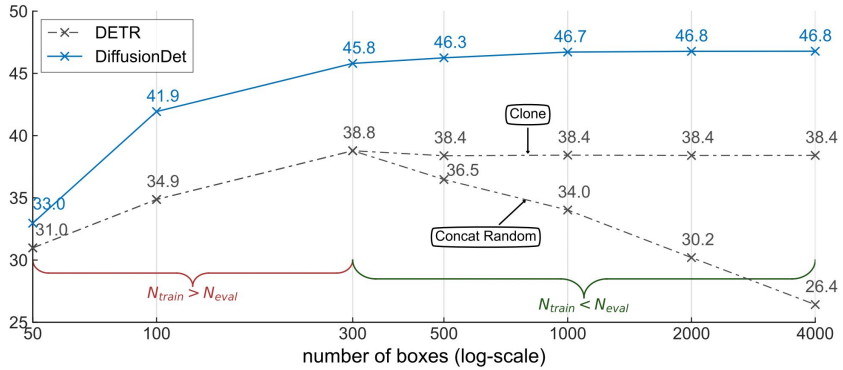


https://openaccess.thecvf.com/content/ICCV2023/papers/Huang_A_Sentence_Speaks_a_Thousand_Images_Domain_Generalization_through_Distilling_ICCV_2023_paper.pdf

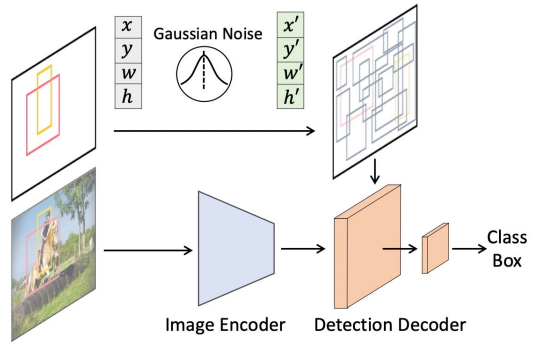
ICCV 2023 の動向・気付き（132/165）

- DiffusionDet（拡散モデル x 物体検出）
  - Bounding boxにノイズをかけることで，拡散過程を表現
  - Region proposalの数が増加したとしても，汎化
    - DETRだと学習したクエリ数300を上回ると性能劣化
    - ランダムにBboxを初期化できる利点が生かされる
  - ネットワークは既存のものを使い，シンプルな改善で既存研究を凌駕

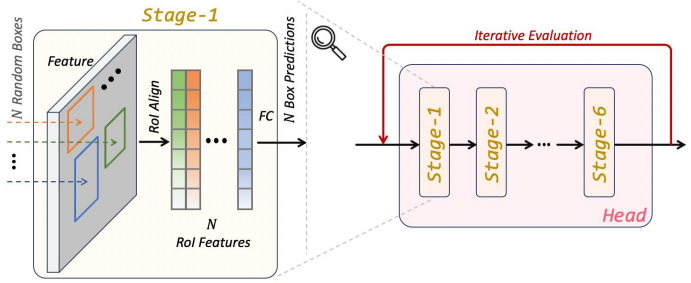




(a) **Dynamic number boxes.** Both DETR and DiffusionDet are trained with 300 object queries or proposal boxes. More proposal boxes in inference bring accuracy improvement on DiffusionDet, while degenerate DETR.



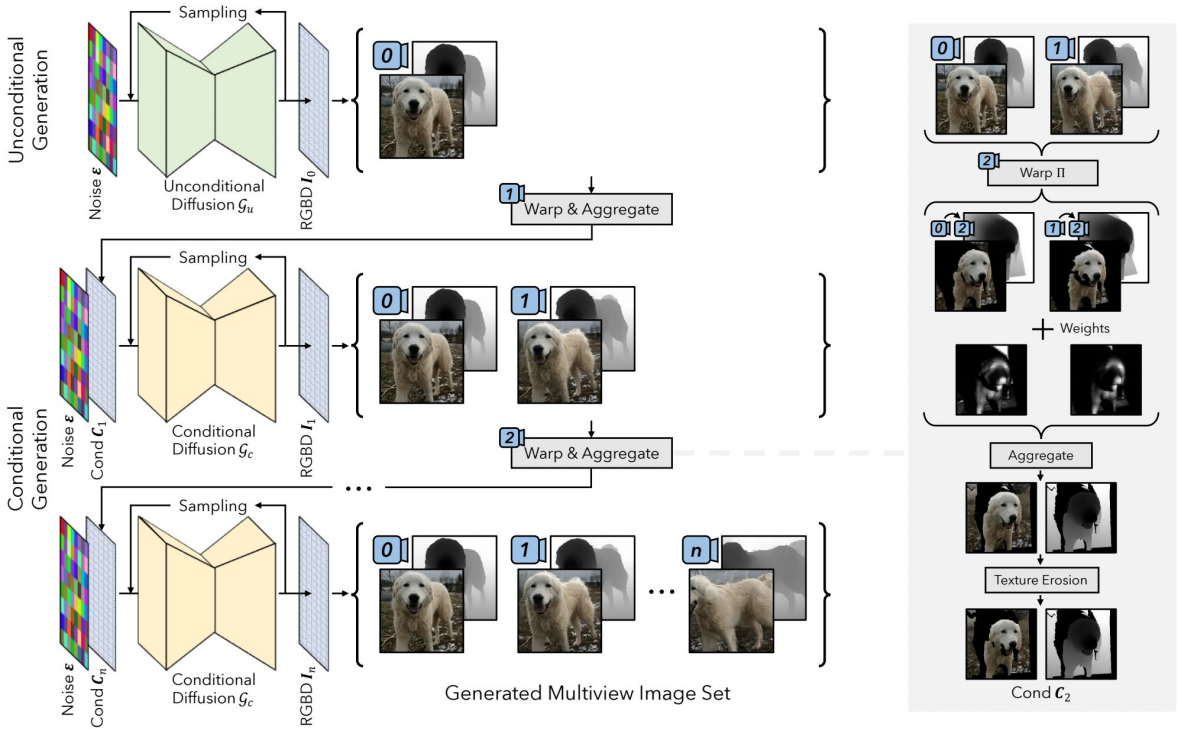
(a) Overall pipeline.

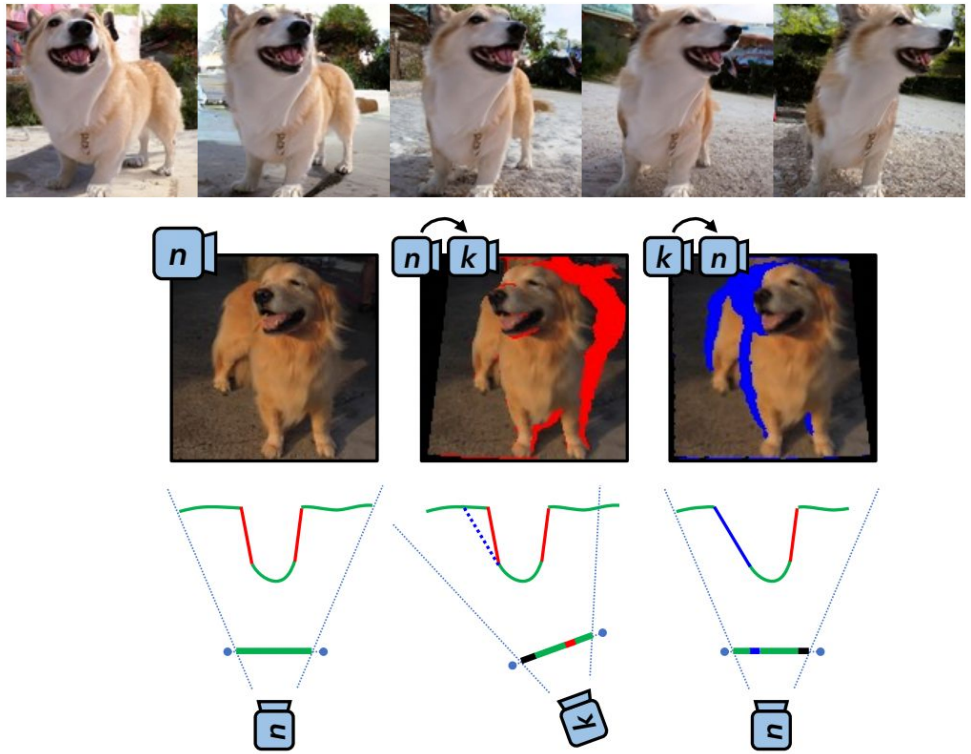


(b) Details of the detection decoder/head.

ICCV 2023 の動向・気付き（133/165）

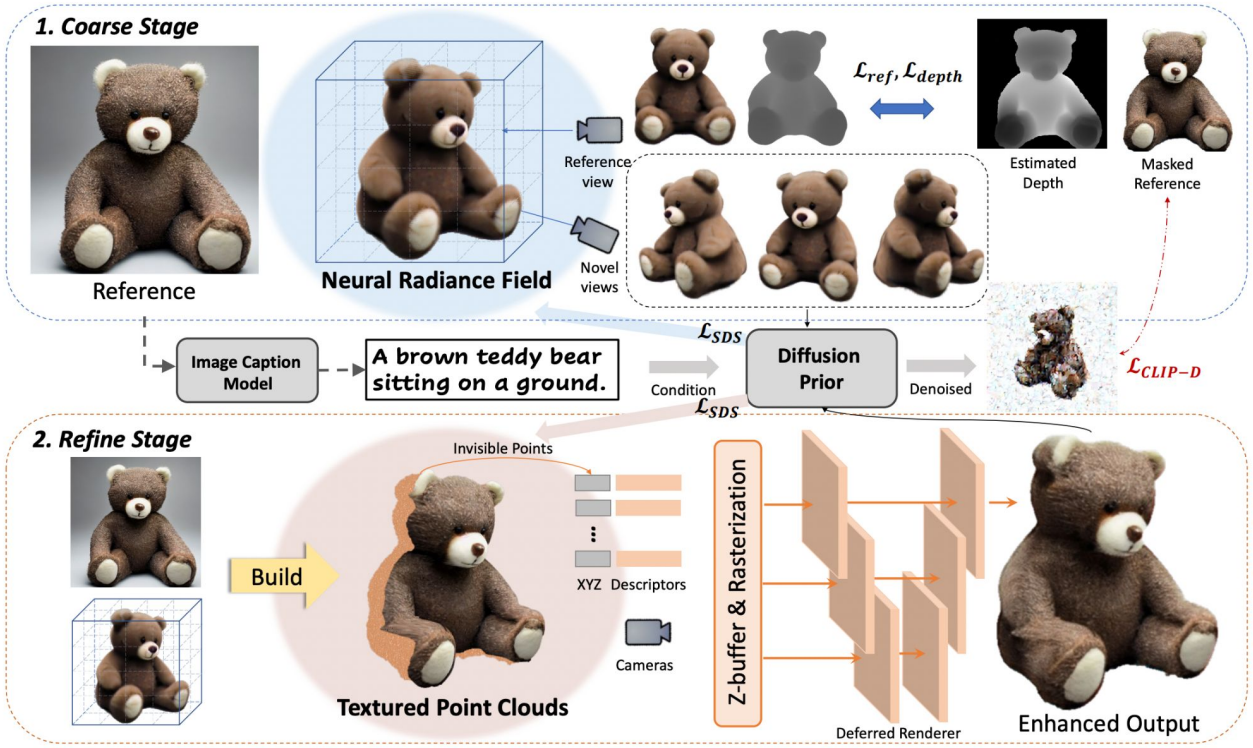
- IVID（3D-aware Image Generation using 2D Diffusion Models）
  - 2D画像生成器を利用して，3D-aware画像生成
  - 単眼深度推定器を利用し，深度マップからカメラポーズを変えたときの画像の生成を行う（AdaMPI[Han+ SIGGRAPH22]）
  - 徐々に穴を埋めていくことで，3D-aware化

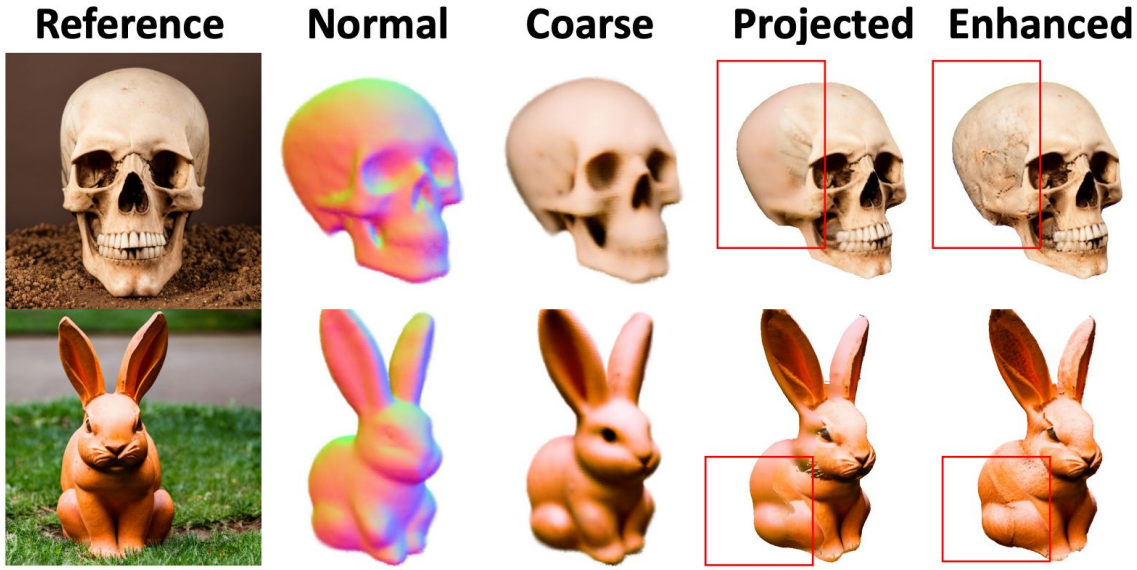




ICCV 2023 の動向・気付き（134/165）

- Make-It-3D（Text-to-image拡散モデル x 画像からの3D復元）
  - Score Distillation Sampling（DreamFusion [Ben+ ICLR22]）と単眼深度推定器 を利用しText-to-3DをNeRFを用いて学習
  - 学習されたNeRFを色付き点群に出力し，テクスチャの最適化
    - Coarse-to-fineな取り組みにより1つずつ改善
    - ジオメトリよりテクスチャの改善の方が見栄えが良くなる





ICCV 2023 の動向・気付き（135/165）

- StableVideo（Text-to-image拡散モデル x 動画編集）
  - Layered Neural Atlases [Kasten+ TOG21]を利用し，フレーム間で一貫した動画編集
    - 背景と前景を別々に編集
    - 前景の連続性のために，前後のフレームを用いてAtlasを正則化する方法を提案

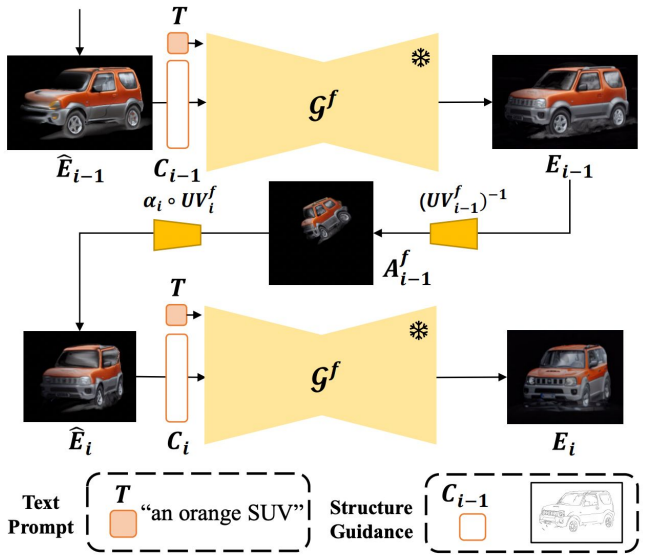
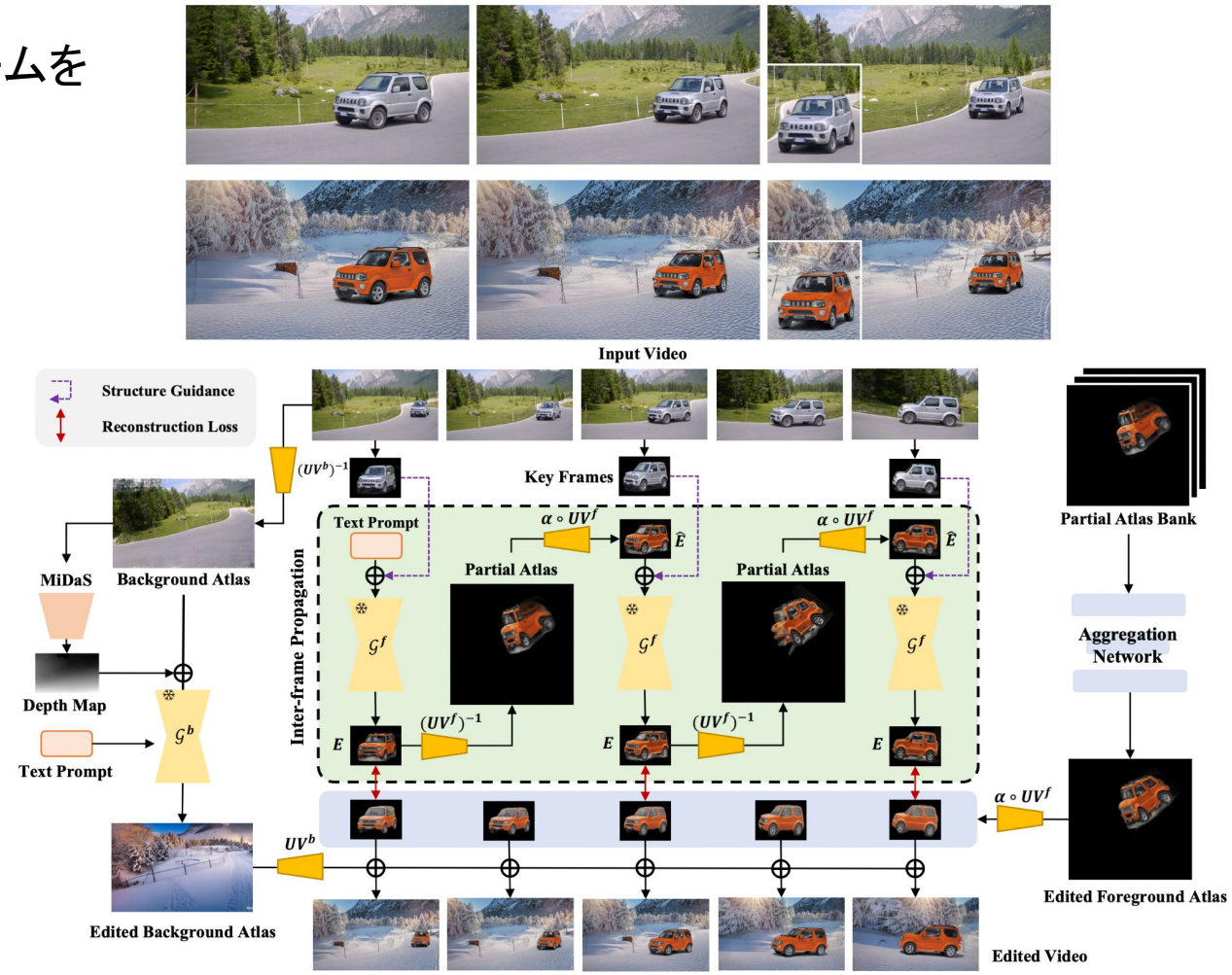


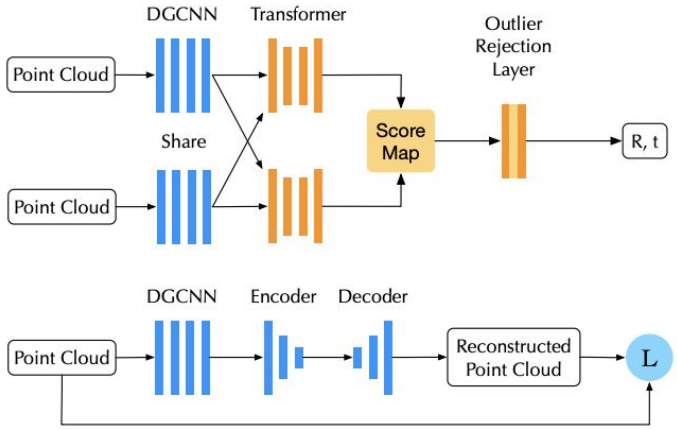
Figure 3: **Inter-frame propagation for foreground editing.** We use two edited key frames, E_{i-1} and E_i, to illustrate the process more clearly. The structure guidance and the text prompt is added into the denoising UNet via the concatenation and cross-attention mechanism respectively.



ICCV 2023 の動向・気付き(136/165)

AutoSynth: Learning to Generate 3D Training Data for Object Point Cloud Registration

- 遺伝的アルゴリズムを用いた点群データ自動生成
  - Primitive shape (シリンダーやコーンなど)を複数組み合わせることでデータ生成
    - 形の範囲をコントロールするハイパーパラメータ(回転やスケーリングなど)を11個用意しそれぞれ9のbinに分ける
    - →合計9^11の探索空間が存在
      - 1. 遺伝的アルゴリズムを用いて最適なデータ生成パラメータを取得
      - 2. Transformerを要するRegistration(下図上2行)ではなくPoint Cloud Reconstructionタスクに変更しテストデータ(=Real)での精度を競わせることで要する時間を1/4000に短縮

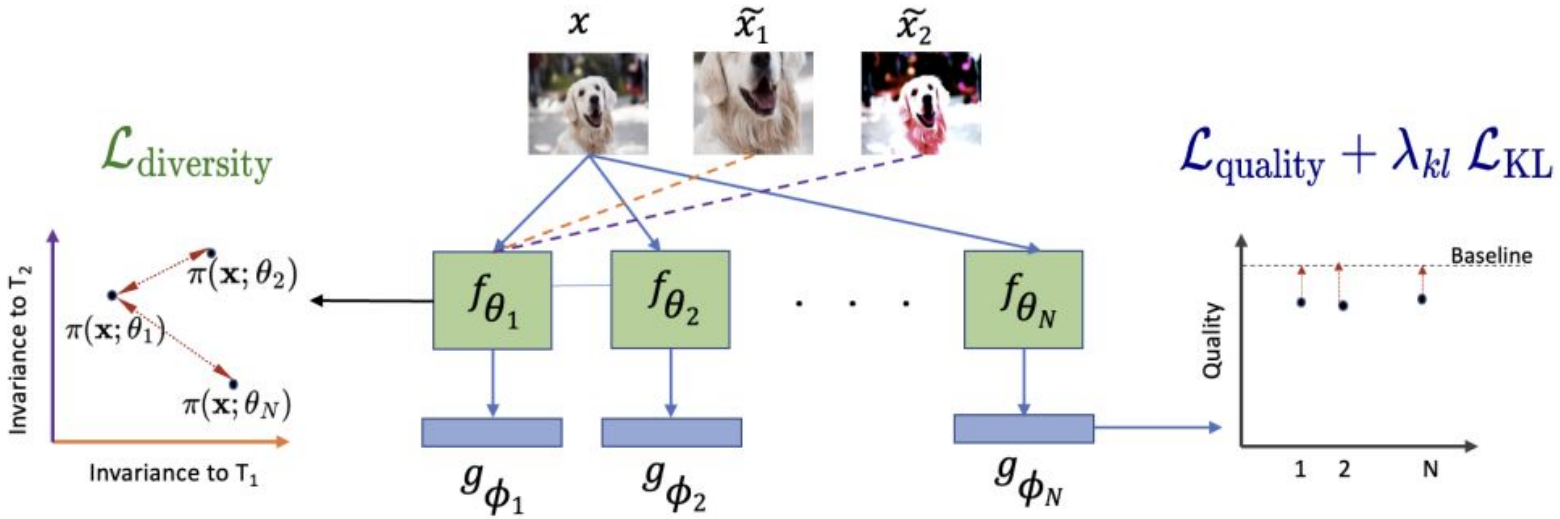


https://openaccess.thecvf.com/content/ICCV2023/papers/Dang_AutoSynth_Learning_to_Generate_3D_Training_Data_for_Object_Point_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（137/165）

Quality Diversity for Visual Pre-Training

- 同一データセットを用いた多様な事前学習モデル作成
  - 目的の異なるdownstream tasksは各タスクに適したinvarianceを備えた、異なる事前学習済みモデルを使用すると高精度
  - 多様なdownstream taskに応用可能な事前学習方法を提案
  - 複数のモデルを用意し、複数のAugmentationに対するinvarianceスコアを用いて類似度最小化(=異なるモデルは異なる変換に対してinvarianceを持つ)
  - 多様なモデルのStackingを使用することで最高精度を記録

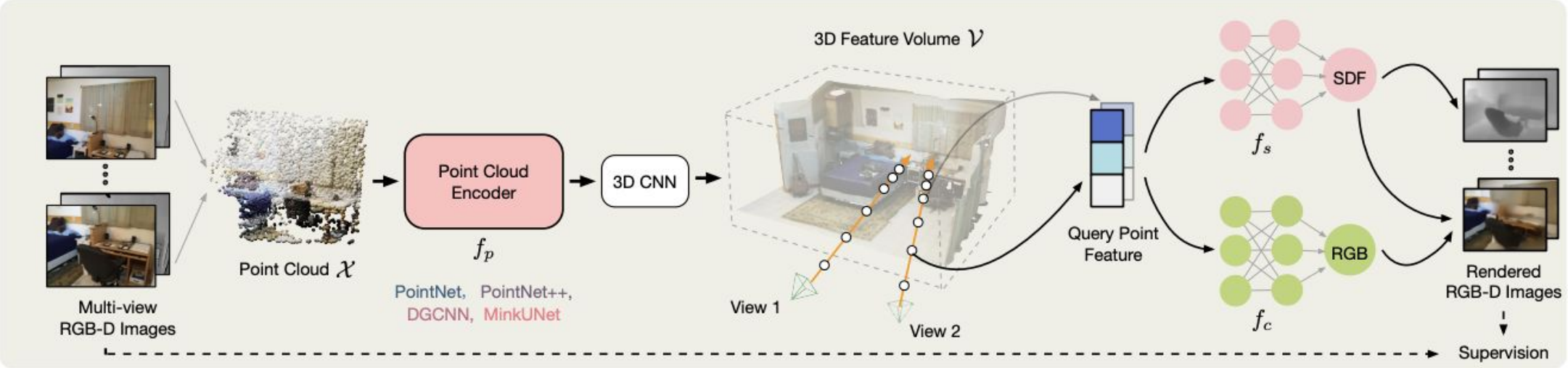


https://openaccess.thecvf.com/content/ICCV2023/papers/Chavhan_Quality_Diversity_for_Visual_Pre-Training_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（138/165）

Ponder: Point Cloud Pre-training via Neural Rendering

- 点群データからRGB-D画像を再構築する事前学習手法の提案
  - 点群データを用いた従来の事前学習: ① Contrast-based ② Completion-based
  - 提案手法:
    - RGB-D画像から点群データを取得 (back-projection)
    - 点群エンコーダー、RGBデコーダー、SDF(Signed Distance Function)デコーダーを用いてシーンを表現
    - 微分可能レンダリングと元のRGB-D画像を用いて自己教師あり学習を実施

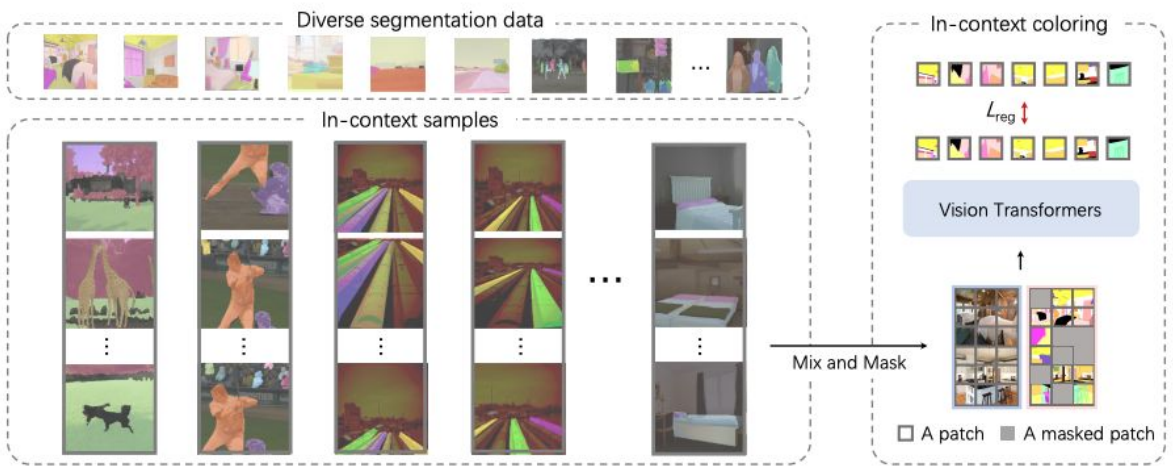


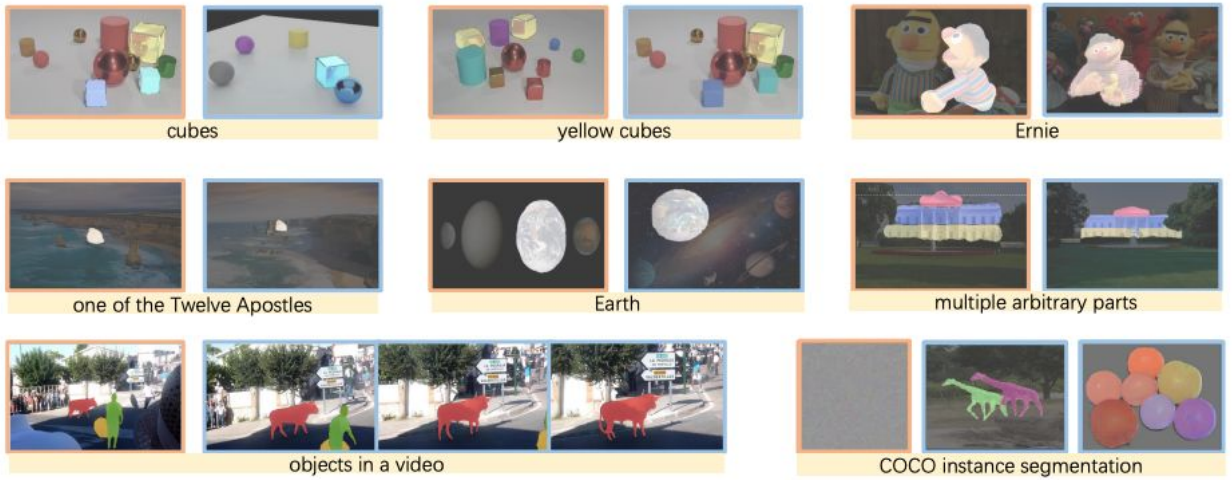
https://openaccess.thecvf.com/content/ICCV2023/papers/Huang_Ponder_Point_Cloud_Pre-training_via_Neural_Rendering_ICCV_2023_paper.pdf

ICCV 2023 の動向・気付き（139/165）

SegGPT: Towards Segmenting Everything In Context

- SegmentationにおいてIn-Context-Learning可能なSegGPTを提案
  - 推論時にプロンプトに応じて物体領域を切り出し可能
  - 基盤モデルの課題である下流タスク適用時の計算コストをIn-Context-Learningにて解消
  - SegGPTはIn-Context-Coloringによって構築
    - オリジナルデータとランダムカラー&マスクされた画像を同時に入力し，復元するタスク



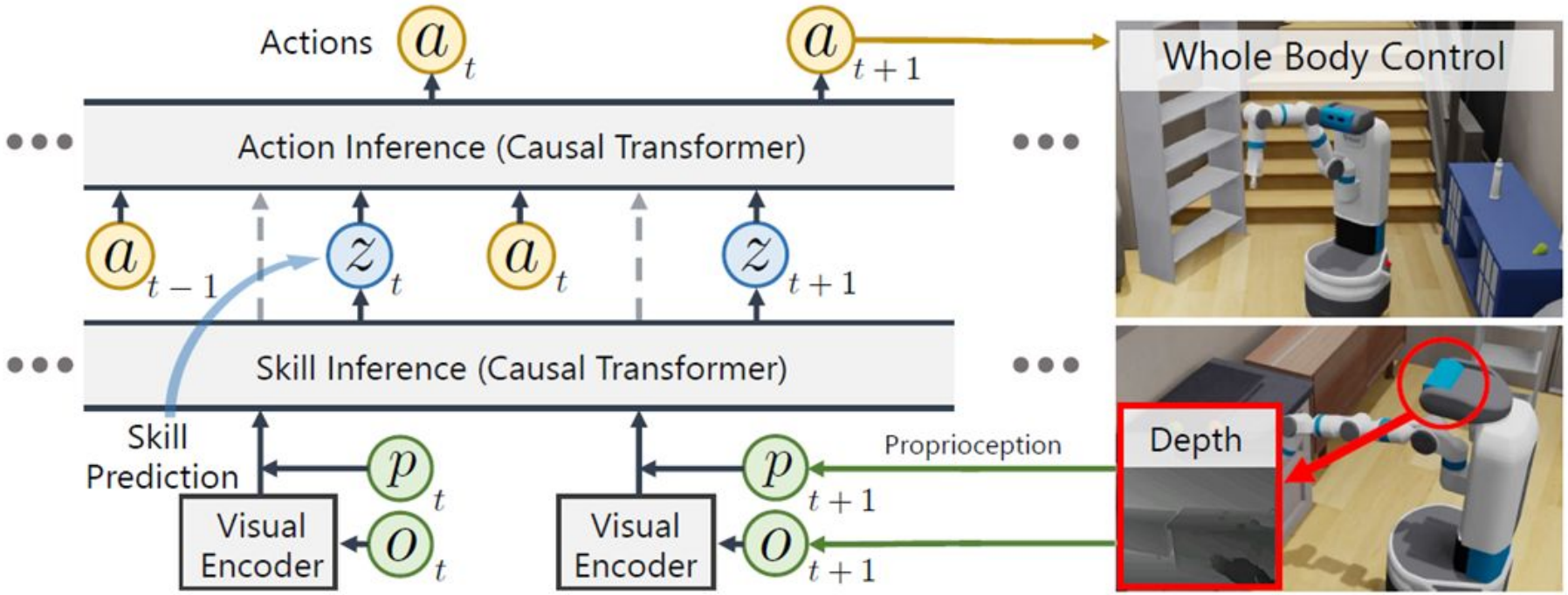


https://arxiv.org/abs/2304.03284

ICCV 2023 の動向・気付き（140/165）

Skill Transformer: A Monolithic Policy for Mobile Manipulation

- Xiaoyu Huang, Dhruv Batra, Akshara Rai, Andrew Szot
  - スキル認識と運動予測のモジュールを階層的につなぐTransformerモデル
  - 観測した状態からスキル認識し運動予測のTransformerに与えられる
  - スキル認識が運動予測の条件付けになるため、エピソードの一部からでも学習可能.
    長期タスクのデモンストレーションにかかるコストを節約できる.

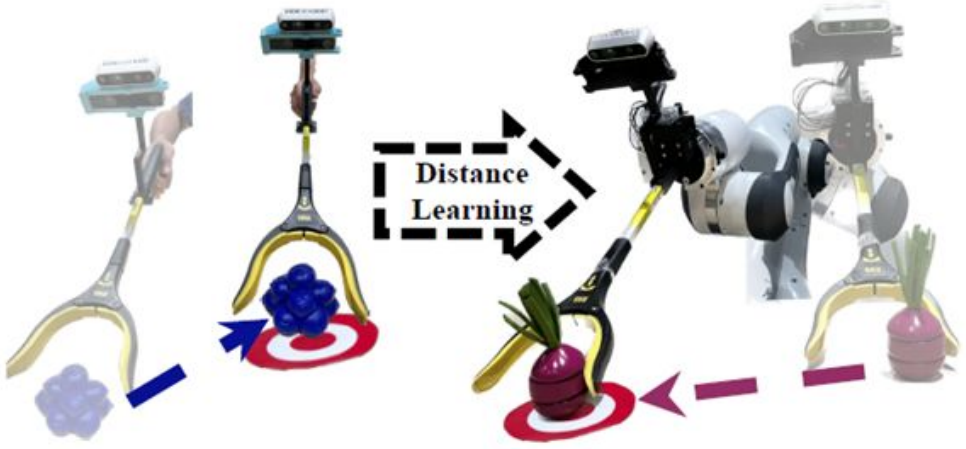


https://openaccess.thecvf.com/content/ICCV2023/html/Huang_Skill_Transformer_A_Monolithic_Policy_for_Mobile_Manipulation_ICCV_2023_paper.html

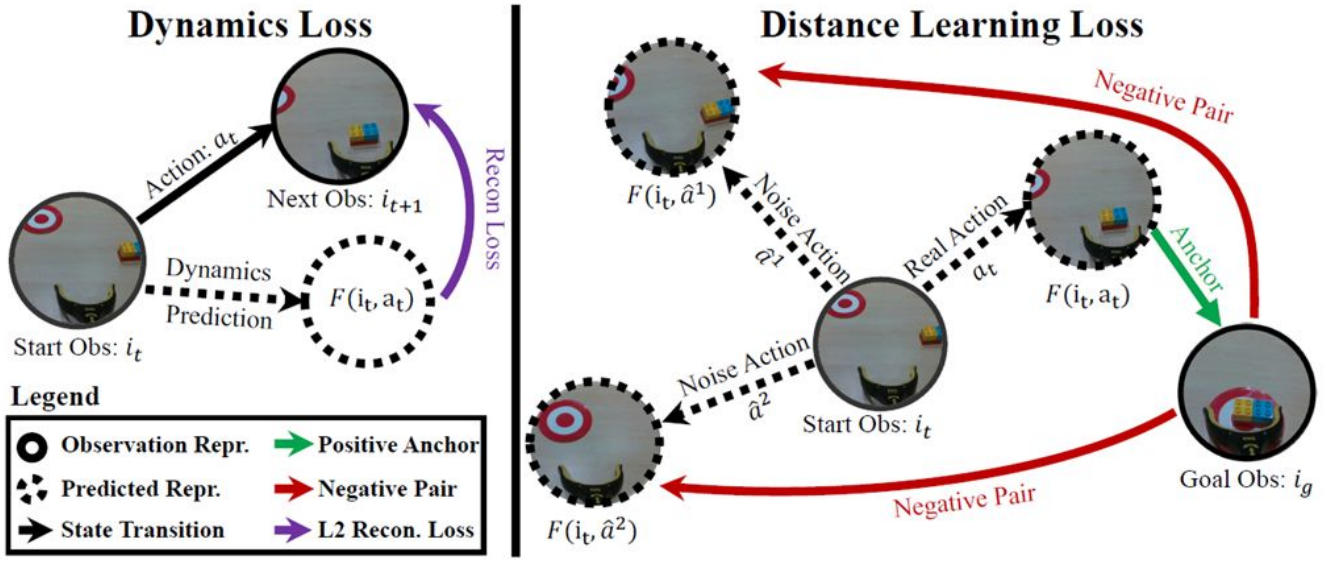
ICCV 2023 の動向・気付き（141/165）

Manipulate by Seeing: Creating Manipulation Controllers from Pre-Trained Representations

- Jianren Wang, Sudeep Dasari, Mohan Kumar Srirama, Shubham Tulsiani, Abhinav Gupta
  - マニピュレータの運動を考慮した視覚運動情報の表現を事前学習
  - 画像特徴量の距離学習と将来の画像特徴量を予測学習
  - 複数のマニピュレーションで他のベースラインより高いタスク成功率を示した



マジックハンドにカメラを取り付けた装置で人がデータ収集できる．手先位置と画像を記録する．

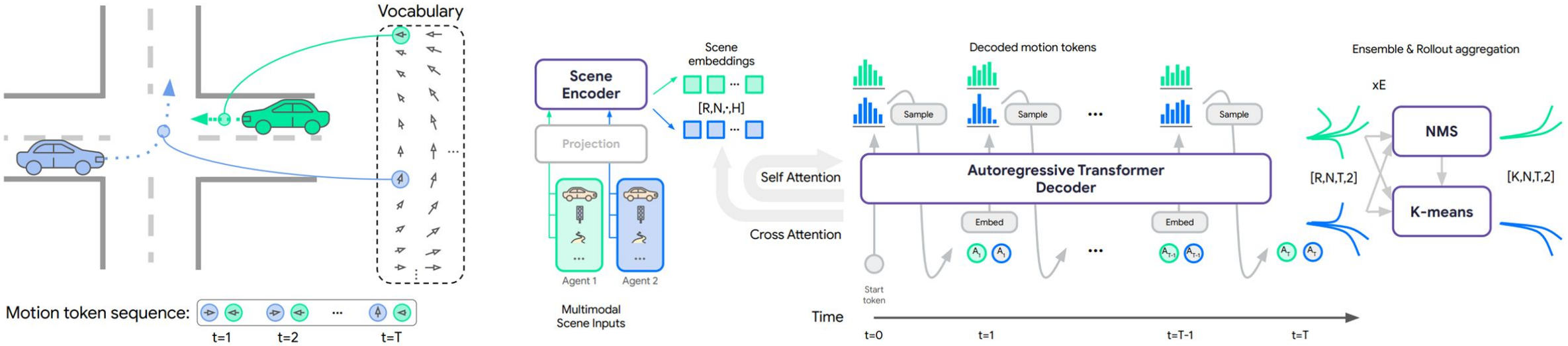


https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Manipulate_by_Seeing_Creating_Manipulation_Controllers_from_Pre-Trained_Representations_ICCV_2023_paper.html

ICCV 2023 の動向・気付き（142/165）

MotionLM: Multi-Agent Motion Forecasting as Language Modeling

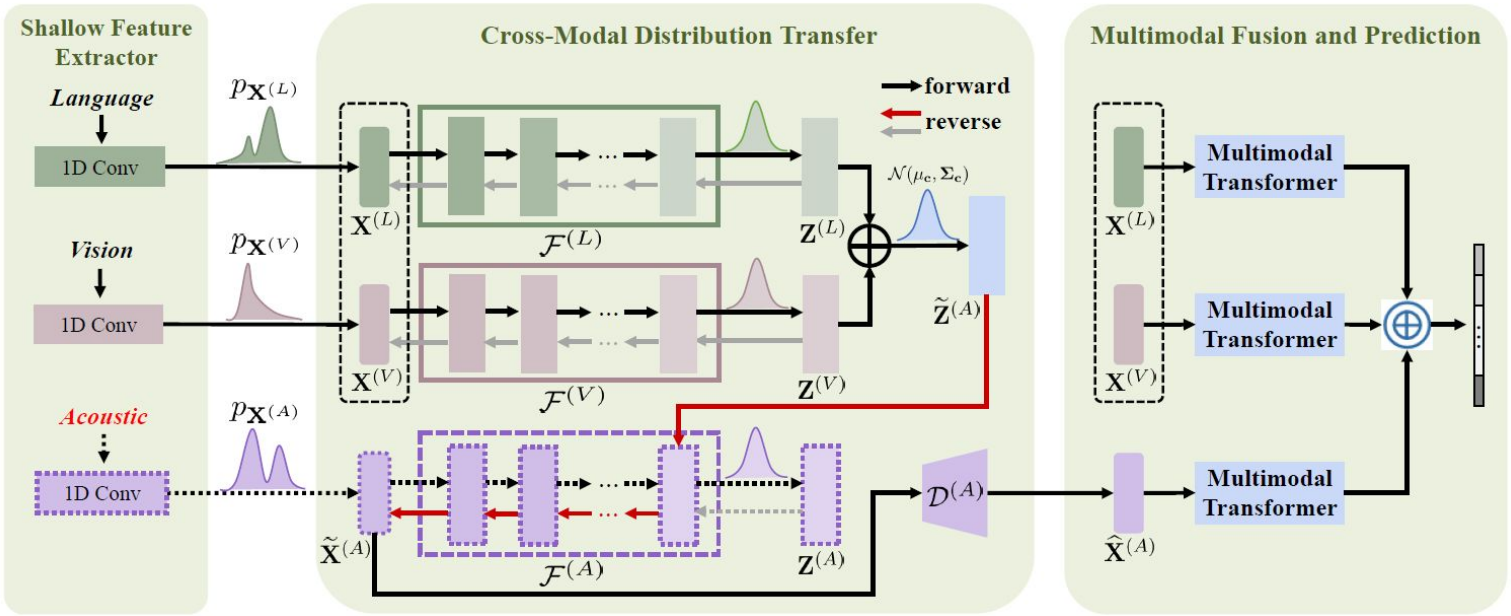
- Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti
  - 運動をトークン化し言語モデルとして学習する
  - 複数エージェントの動きを同時にモデル化、ヒューリスティクスなしでインタラクションを実現



https://openaccess.thecvf.com/content/ICCV2023/html/Seff_MotionLM_Multi-Agent_Motion_Forecasting_as_Language_Modeling_ICCV_2023_paper.html

Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning

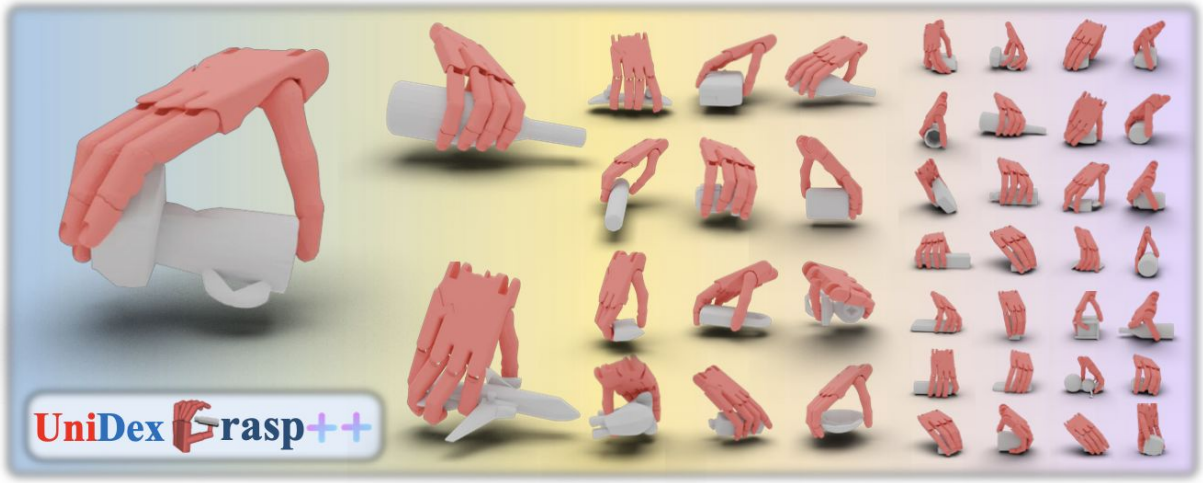
- Yuanzhi Wang, Zhen Cui, Yong Li
  - 欠損するモダリティをほかのモダリティの潜在空間の分布から復元するフレームワーク
  - 異なるモダリティの潜在空間が同じクラスの正規分布を共有するように制約する
  - Flowベースな生成モデルでモダリティー正規分布の双方向変換を学習

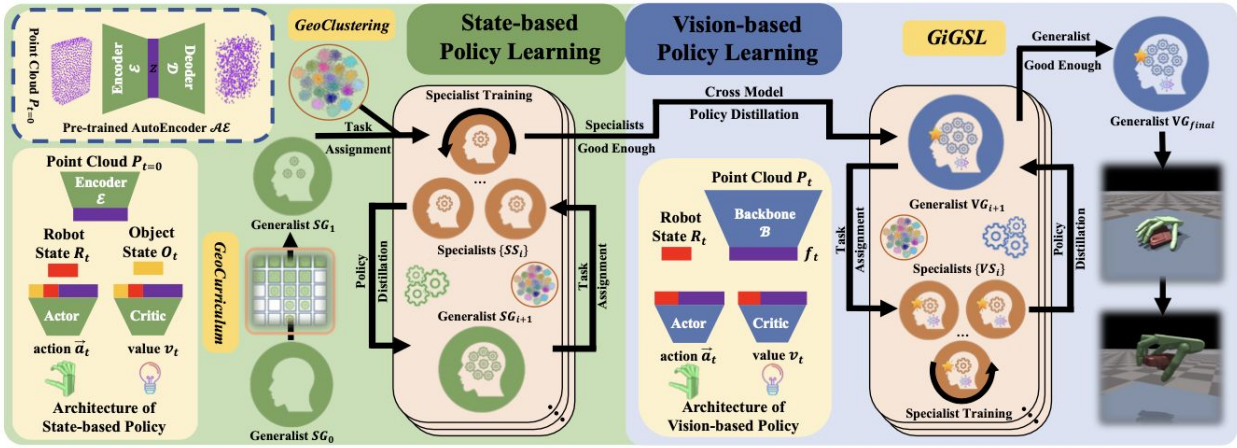


https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Distribution-Consistent_Modal_Recovering_for_Incomplete_Multimodal_Learning_ICCV_2023_paper.html

UniDexGrasp++: Improving Dexterous Grasping Policy Learning via Geometry-aware Curriculum and Iterative Generalist-Specialist Learning

- Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, He Wang
  - PointCloudと身体情報から物体把持のポリシーを学習する
  - 幾何情報に基づいたカリキュラム学習とgeneralist-specialist学習を組み合わせることで効率よく学習する

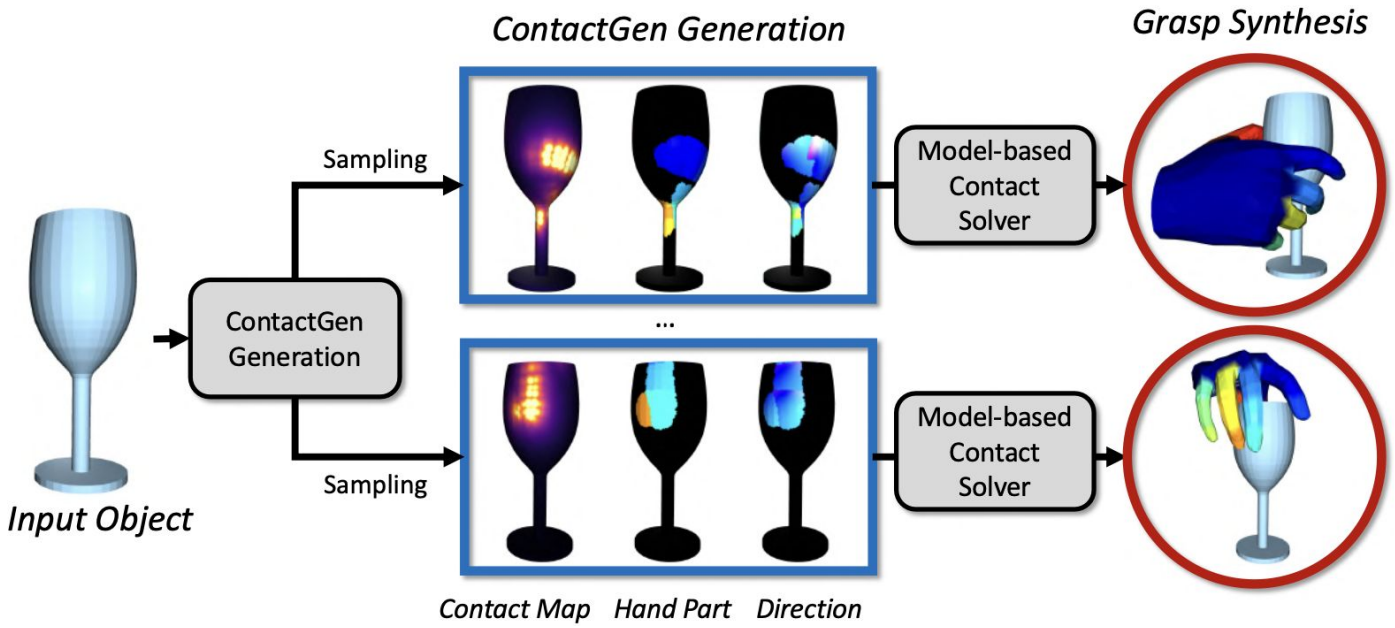


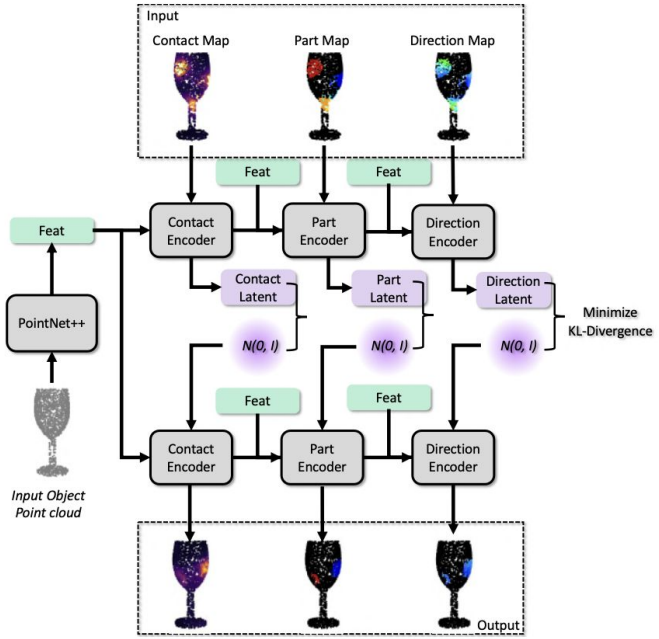


ICCV 2023 の動向・気付き（145/165）

ContactGen: Generative Contact Modeling for Grasp Generation

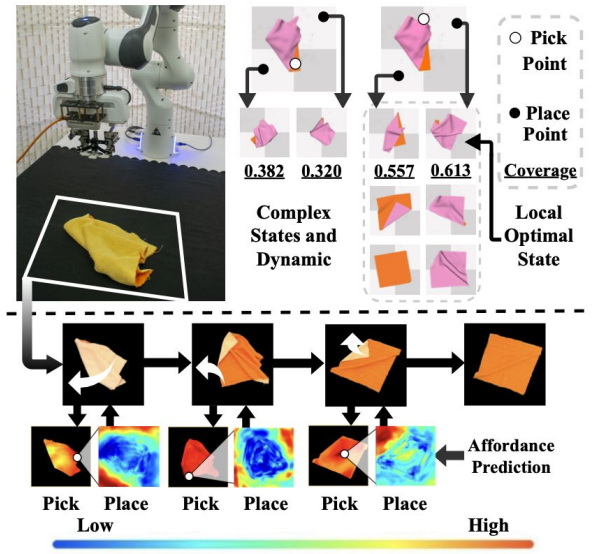
- Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, Shenlong Wang
  - 物体上に把持のための接触候補をマッピングし多指ハンドの把持姿勢を生成する
  - 接触位置，接触する手の部位を示す部位，各部位内での接触方向を示す方向をPointcloud上にそれぞれマッピングした後，最適化問題を解くことで最終的な把持姿勢を出力する
  - Mapを中間出力として生成するため，整合性のとれない把持が出力されることもある

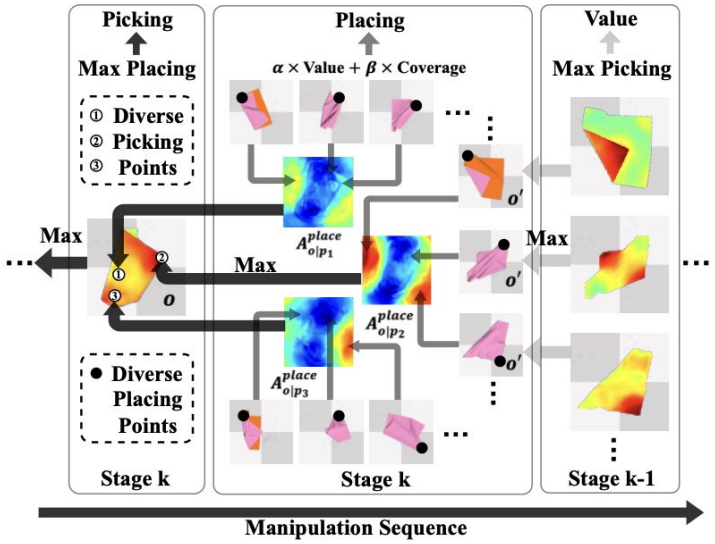


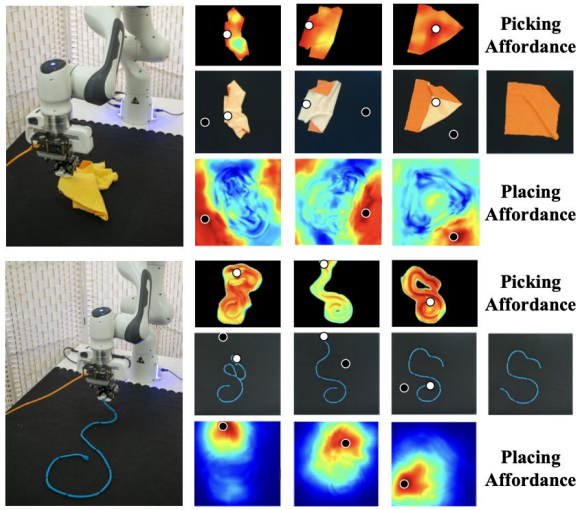


Learning Foresightful Dense Visual Affordance for Deformable Object Manipulation

- Ruihai Wu, Chuanruo Ning, Hao Dong
  - 長期的な柔軟物体操作において連続的にアフォーダンスを推定することで局所解を回避する
  - PickとPlaceで組み合わされるシーケンスの中で，逐次アフォーダンスを推定しながら操作する点を設計する．
  - 自己教師付きのデータ収集によって，布や紐などの操作を実環境でも可能にしている





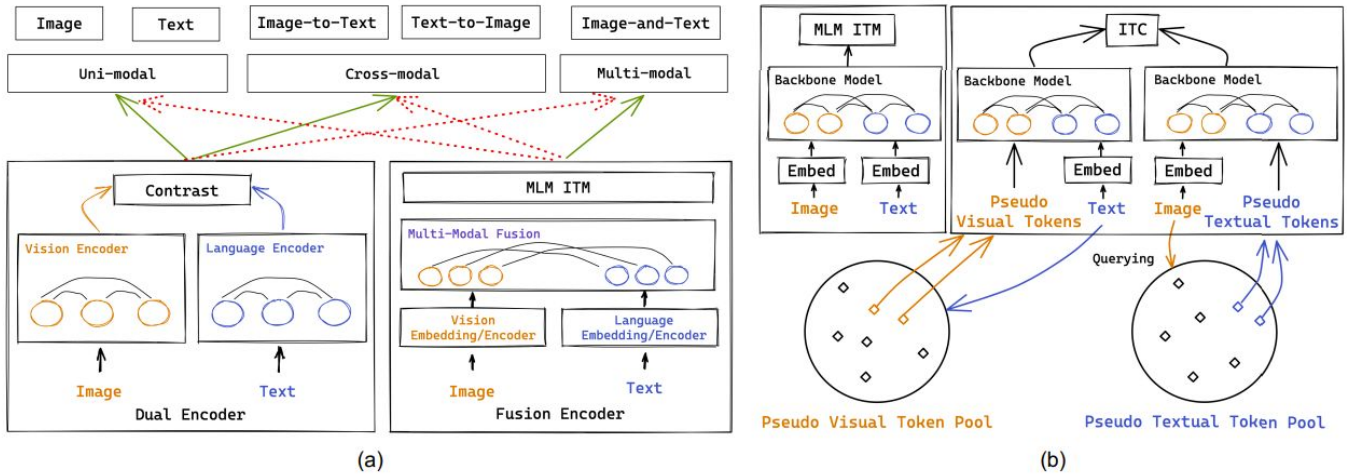


ICCV 2023 の動向・気付き（147/165）

Towards Unifying Medical Vision-and-Language Pre-training via Soft Prompts

- 医療AI:　言語・画像のマルチモーダル学習
  - 医療画像と医療テキストからマルチモーダルに一般的表現を抽出するMedical-VLP領域の研究。
  - 従来のモデルは**フュージョンエンコーダ型とデュアルエンコーダ型**に分かれていたが、本論文の手法では**両者を統合**している。
  - 視覚的プロンプトとテキストプロンプトを統合して扱っているため、プロンプト型画像研究に役立つと思われる。投稿時期の問題ではあるが、大規模言語モデルのプロンプトへの言及がないため、大規模言語モデルとの発展に注目。
  - 様々なデータセットで良好な結果を示している。

GitHub https://github.com/zhjohnchan/PTUnifier



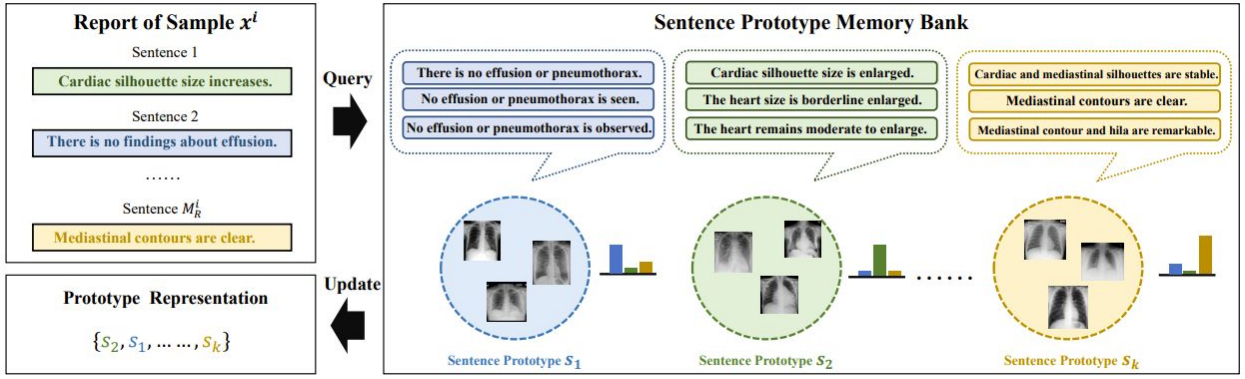
Methods	Uni-Modal				Cross-Modal			Multi-Modal		
	Image		Text		Image-to-Text		Text-to-Image			
	CheXpert AUROC	PNAS AUROC	RadNLI Acc	MIMIC RL	MIMIC BL4	ROCO R@1	ROCO R@1	VQA-RAD Acc	SLAKE Acc	MedVQA-2019 Acc
Study$_1$	ConVIRT [66]		ClinicalBERT [2]	TransABS [36]	R2Gen [10]		ViLT [25]		CPRD [33]	
	87.3	81.3	72.6	43.8	8.0	11.9	9.8	72.7	82.1	-
Study$_2$	GLoRIA [19]		IFCC [41]	WGSum [18]	M2Trans [41]		METER [16]		MMBERT [24]	
	88.1	88.6	77.8	45.1	10.5	14.5	11.3	72.0	-	77.9
PTUnifier (ours)	90.1	90.6	80.0	46.2	10.7	21.0	20.8	78.3	85.2	79.3

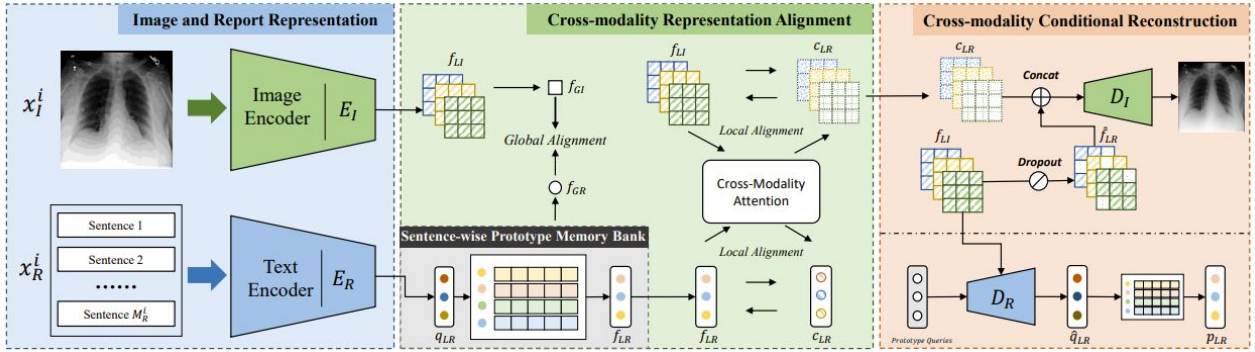
ICCV 2023 の動向・気付き（148/165）

PRIOR: Prototype Representation Joint Learning from Medical Images and Reports

- 医療画像レポートの特徴にフィットしたレポート作成戦略
  - 医療画像レポートは、他の画像レポート問題と比較して局所的所見の比重が重く、文章が非連続的であるという特徴がある(物語ではなく所見の羅列で構成される)。
  - この特徴に対応するために、本手法は画像の小領域とそれに対応する文を局所的な基本的表現単位と見なし、大域的表現をそれらに対する注意のプーリングで形成する。
  - **医療画像レポート限定の手法**であるが、医学的な診断の思考パターンにも合致しているため、優れた手法であると思われる。

GitHub https://github.com/QtacierP/PRIOR



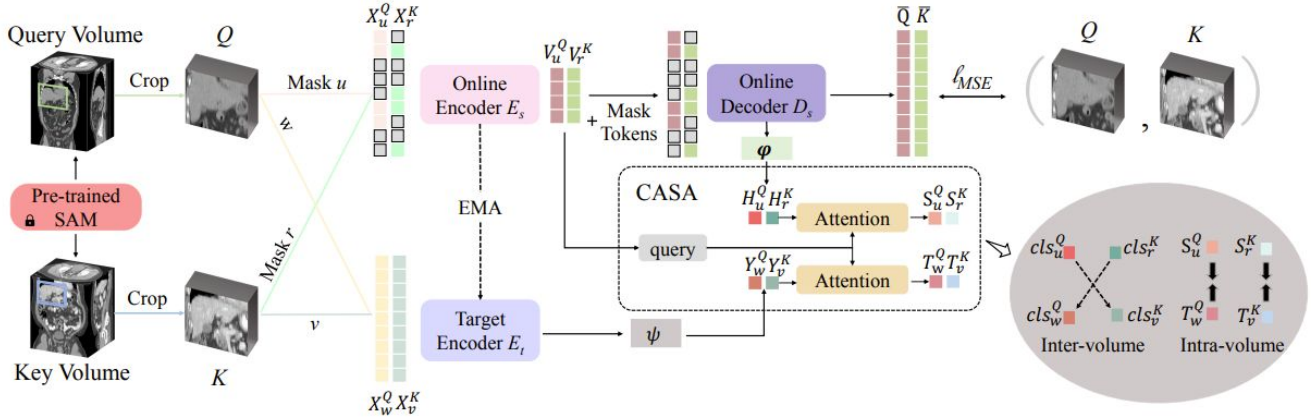


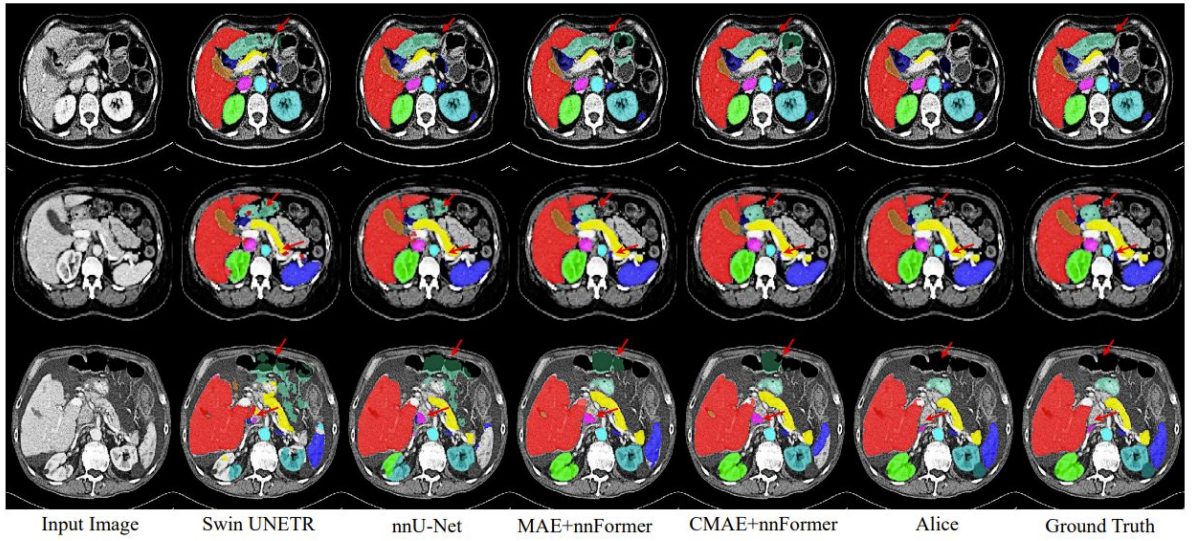
ICCV 2023 の動向・気付き（149/165）

Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis

- 3次元医療画像の空間的特徴を学ぶ、自己教師あり学習
  - 従来法は自然画像を対象に構築されてきたため、医療画像特有の空間的な関係を学べていない。
  - CTやMRIなどの医療画像を人体を対象にしているため、心臓の左右には肺があったり、胴体の上には頭が付いているなど、普遍的な空間的特徴がある。
  - 空間的特徴を内在的に学ぶことを目指した、新しい対照学習戦略
  - 三次元医療セグメンテーションタスクで良好な結果

GitHub https://github.com/alibaba-damo-academy/alice



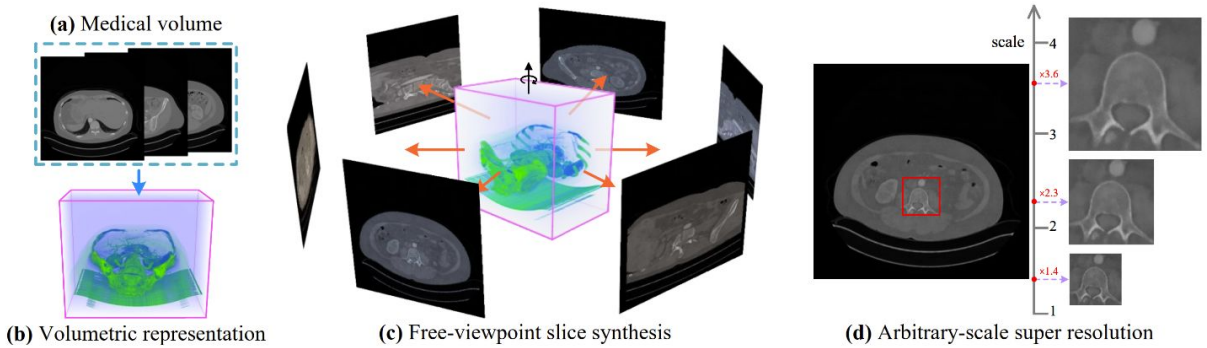


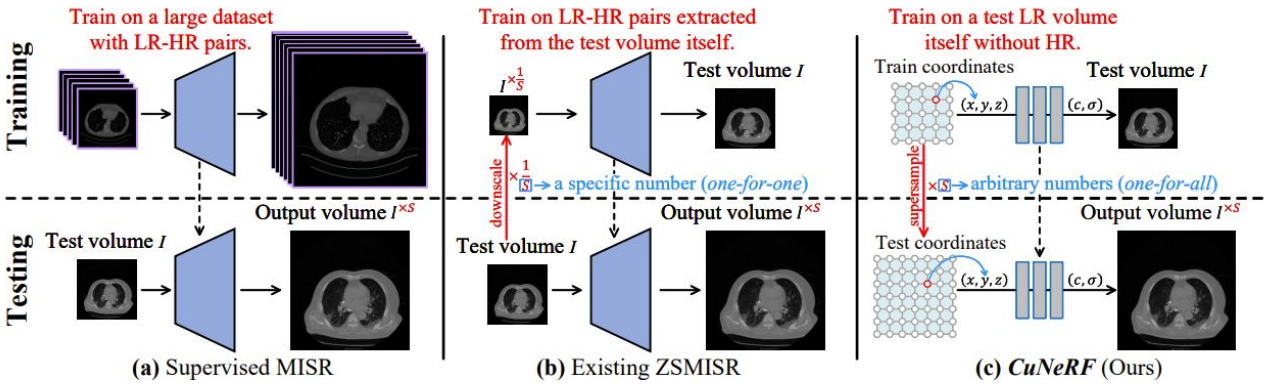
ICCV 2023 の動向・気付き（150/165）

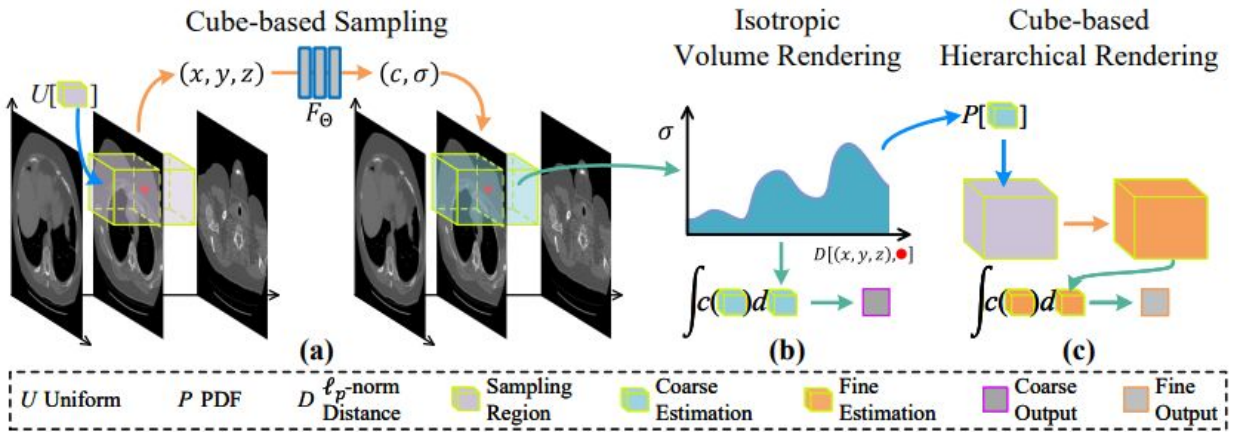
CuNeRF: Cube-Based Neural Radiance Field for Zero-Shot Medical Image Arbitrary-Scale Super Resolution

- NeRFを用いた医療画像の超解像
  - 医療画像（CT,MRI）に対する超解像タスクをNeRFを用いて解決
  - NeRFを用いることで様々な方向の断面を再構成し、超解像タスクにも対応
  - 従来よりも高い超解像タスクの性能（3D MISR comparisons on MSD データセット）

GitHub https://github.com/NarcissusEx/CuNeRF

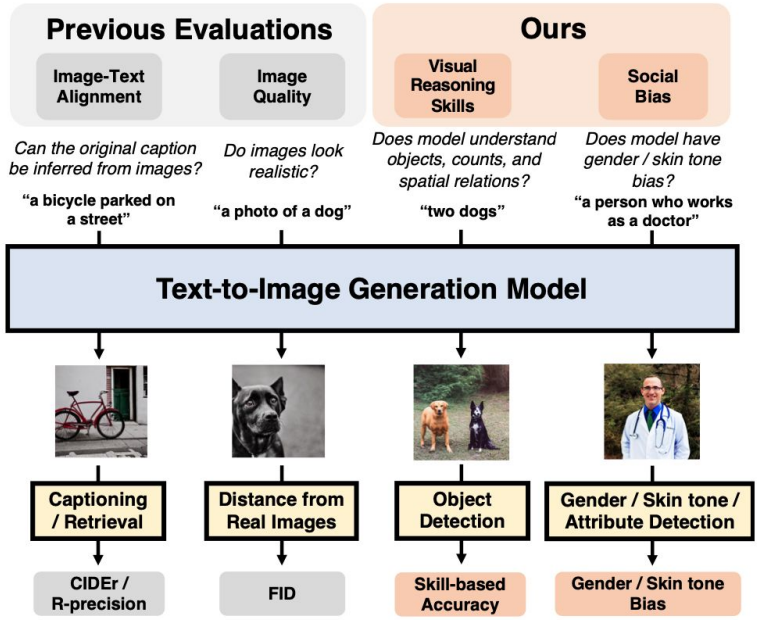


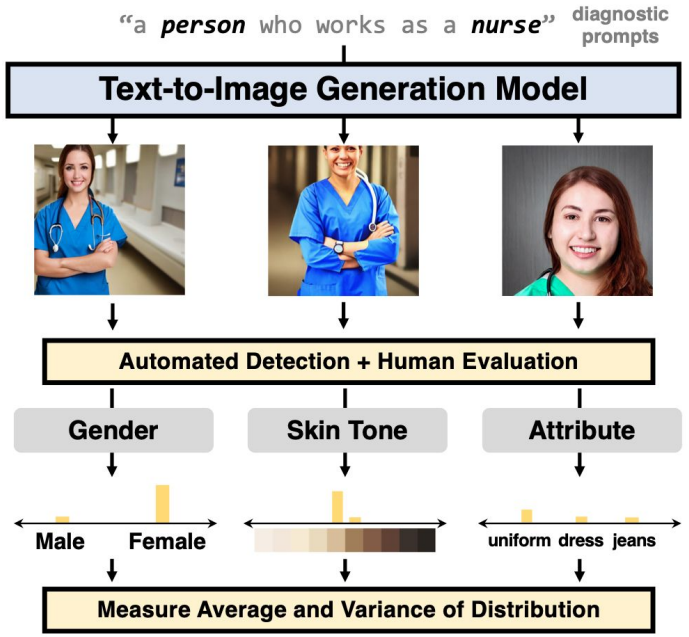




DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models

- text-to-imageモデルを新たな側面から評価
  - Object Recognition, Object Counting, Spatial Relation Understandingの3つの視覚的推論スキルを測定するためのデータセットを作成
  - 生成された画像内の人物の性別と肌色、服装のバイアスを評価
  - text-to-imageモデルはObject CountingとSpatial Relation Understandingに課題有り
  - 職業ごとに生成結果を見ると、性別や肌色、服装に偏りが存在する





https://openaccess.thecvf.com/content/ICCV2023/html/Cho_DALL-Eval_Probing_the_Reasoning_Skills_and_Social_Biases_of_Text-to-Image_ICCV_2023_paper.html

ICCV 2023 の動向・気付き（152/165）

Do DALL-E and Flamingo Understand Each Other?

- text-to-imageモデルとimage-to-textモデルの相互理解を調査
  - text-to-imageモデルとimage-to-textモデルを組み合わせた統一フレームワークを提案
  - テキスト/画像の再構成タスクによってテキスト/画像生成の質を評価できる
  - image-to-textモデルの再構成損失によってtext-to-imageモデルをFine-tuningできる（逆もまた然り）

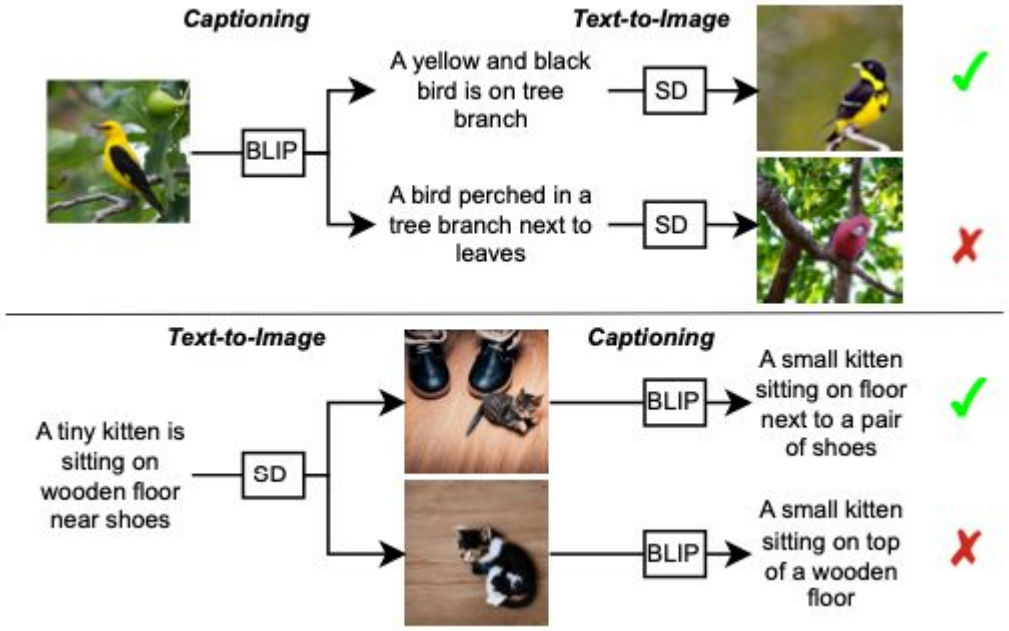
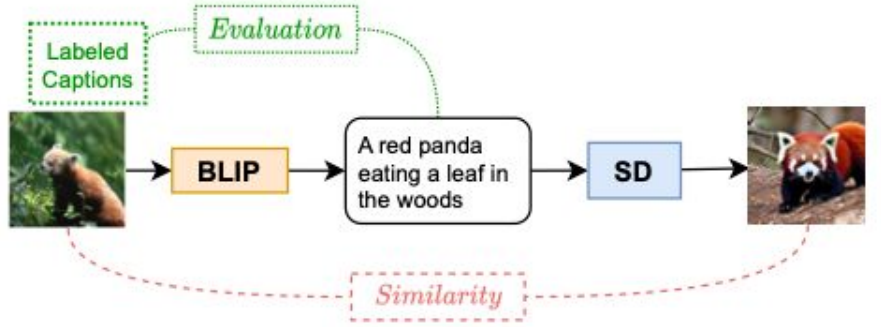
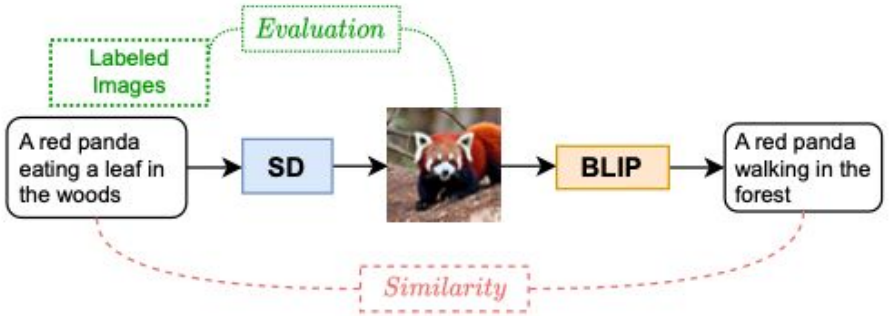


image-text-image



text-image-text

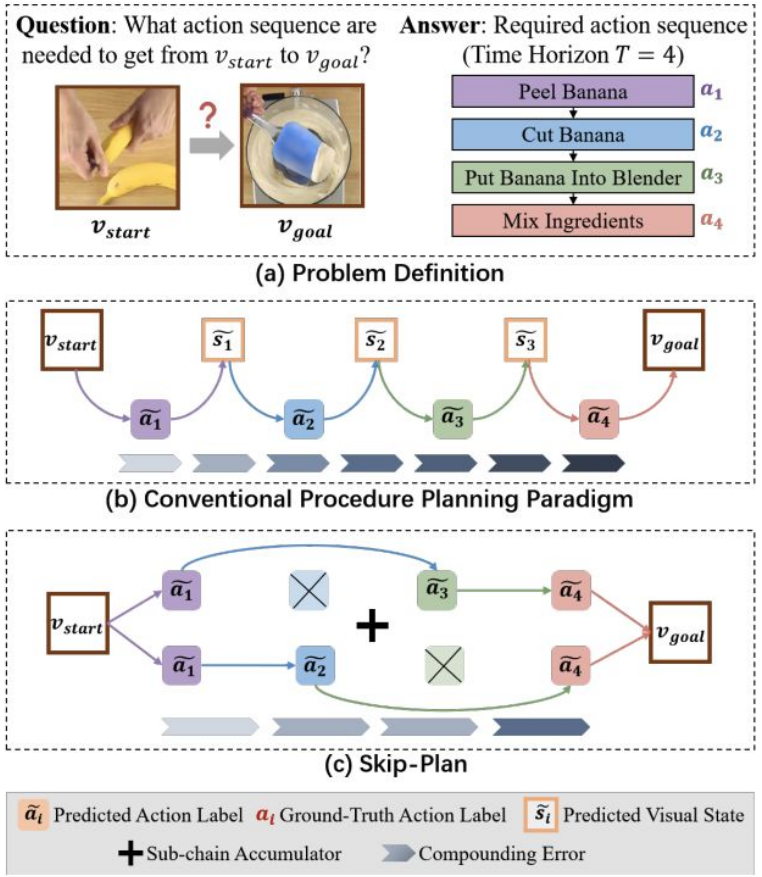


https://openaccess.thecvf.com/content/ICCV2023/html/Li_Do_DALL-E_and_Flamingo_Understand_Each_Other_ICCV_2023_paper.html

ICCV 2023 の動向・気付き（153/165）

Skip-Plan: Procedure Planning in Instructional Videos via Condensed Action Space Learning

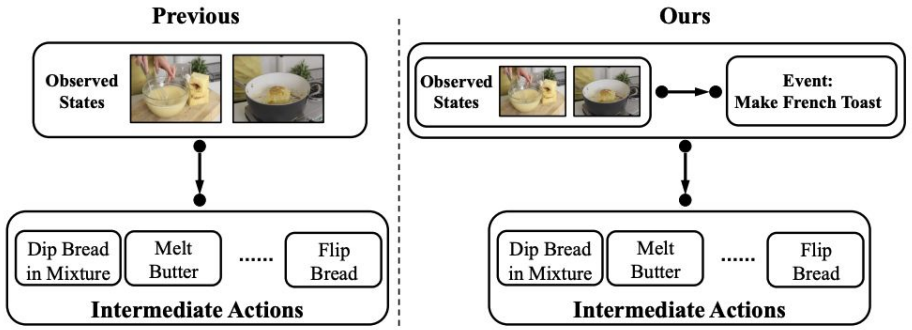
- Procedure Planning，中間状態を予測しないことで精度向上！
  - 背景
    - Procedure Planningは作業工程の最初と最後の画像だけ見せられて，あいだの行動ステップを推論（計画）する
  - これまで
    - 多くの研究では，
      - 中間状態を教師として学習 → しかし中間状態は高次元過ぎて難しい．
      - step-by-step予測の問題として定式化 → エラー蓄積
  - 本研究
    - 中間状態の推論をSkip! → エラー蓄積を低減
    - 行動をsub-chainに分解して推論 → エラー蓄積を低減
    - sub-chain同士の関係性をモデリングするための chain decoderを提案．



ICCV 2023 の動向・気付き（154/165）

Event-Guided Procedure Planning from Instructional Videos with Text Supervision

- 何をしようとしているのか（イベント）を推論しつつ行動計画！
  - 背景
    - Procedure Planningは作業工程の最初と最後の画像だけ見せられて，あいだの行動ステップを推論（計画）する
  - これまで
    - 最初に与えられる画像と実際に計画すべき行動のセマンティックなギャップを考慮していなかった．（例えば，パンケーキミックス（最初の画像）とパンケーキ（最後の画像）だけ与えられても，バターを使うことを予想するには，セマンティックなギャップがある．）
  - 本研究
    - イベント（パンケーキづくり）を予測することでセマンティックギャップに対処．
      - パンケーキを作っていることが分かればそこからバターを使うことを推論するのは自然



ICCV 2023 の動向・気付き（155/165）

Workshop: Artificial Intelligence for Humanitarian Assistance and Disaster Response

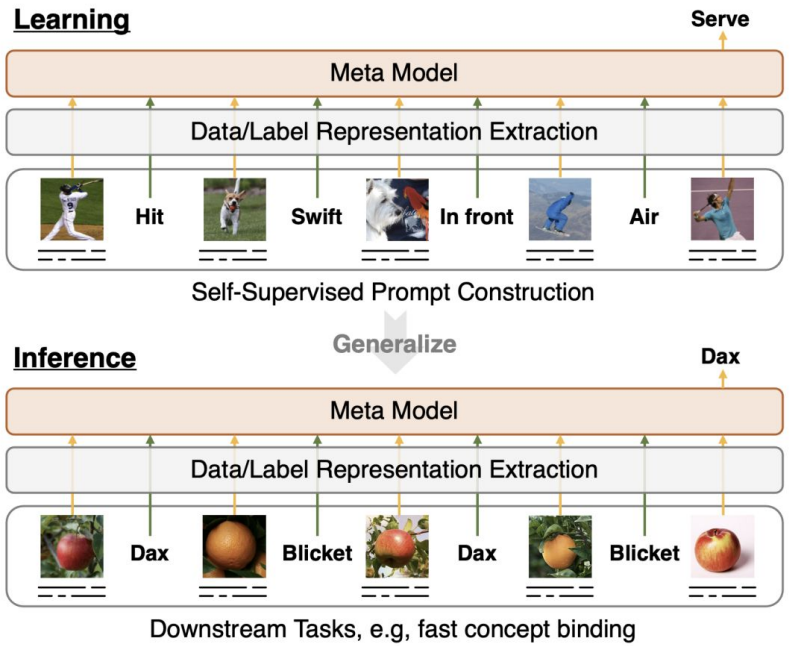
- 採択論文
  - 一覧：https://www.hadr.ai/iccv23/accepted-papers-iccv23
  - Best Paper: TeleViT: Teleconnection-driven Transformers Improve Subseasonal to Seasonal Wildfire Forecasting
  - 迅速な災害対応のために，ドローン空撮画像や衛星画像などを用いて俯瞰的・大域的に，被災者，火災，家屋などを把握するというものが多かった印象．
- 発表者として参加した感想
  - 採択率は公開されていないが，応募時のPaper IDは30番程度まで振られ，Accepted Paperは10件，現地ポスター掲載は7件．ニッチな分野なぶん研究のテーマは絞られるが，採択されやすいと思う．
  - Organizerは頻繁にメールで連絡事項を伝えてくれる＆すぐに対応してくれたため，初めての国際会議WSでも無事投稿＆発表できた．
  - 応用先に明確なテーマがあり，そこに興味がある人が聞きに来てくれるので，ポスター発表では手法よりも，実利用に関する質問をされたのだと思う．

ICCV 2023 の動向・気付き（156/165）

SINC: Self-Supervised In-Context Learning for Vision-Language Tasks

- Incontext-learningの新しい手法の提案
  - 昨今注目されているMLLMのFlamingoやMini-GPT4と言った手法は、frozenのLLMに画像特徴量を合わせるように学習されている
  - しかし、それではLLMの持っているhallucinationの問題などをそのまま継承してしまうとして、VM, LM, VLMを入力側（frozen）、meta-modelを出力側（learnable）とする手法の提案

https://arxiv.org/abs/2307.07742#:~:text=version%2C%20v2)%5D-,SINC%3A%20Self%2DSupervised%20In%2DContext,Learning%20for%20Vision%2DLanguage%20Tasks&text=Large%20Pre%2Dtrained%20Transformers%20exhibit,demonstrations%20presented%20in%20the%20inputs.



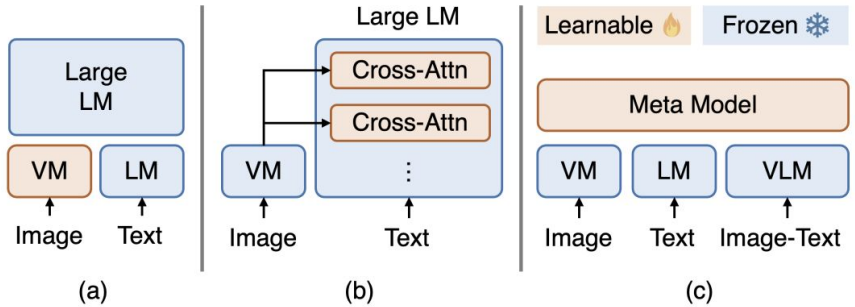
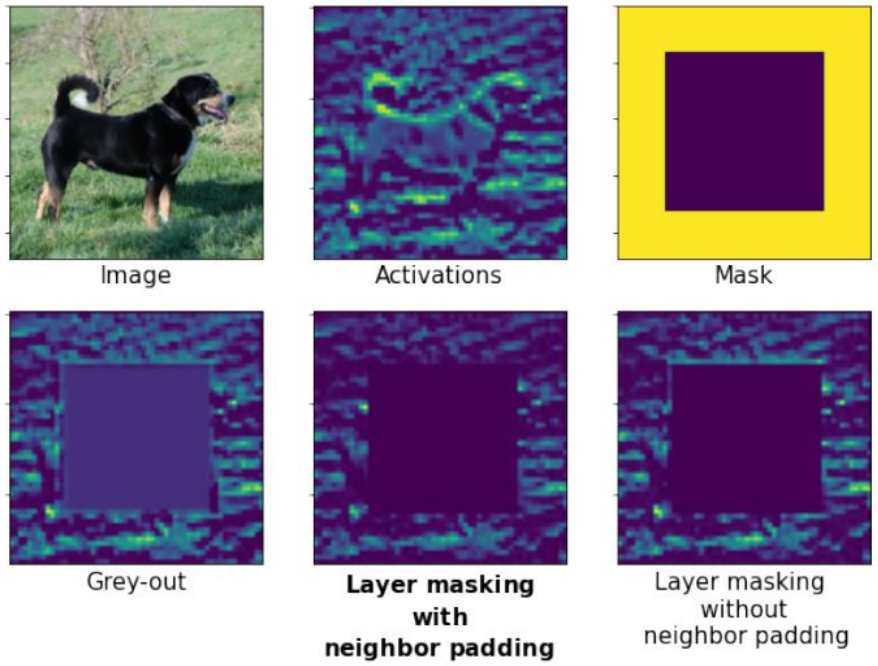


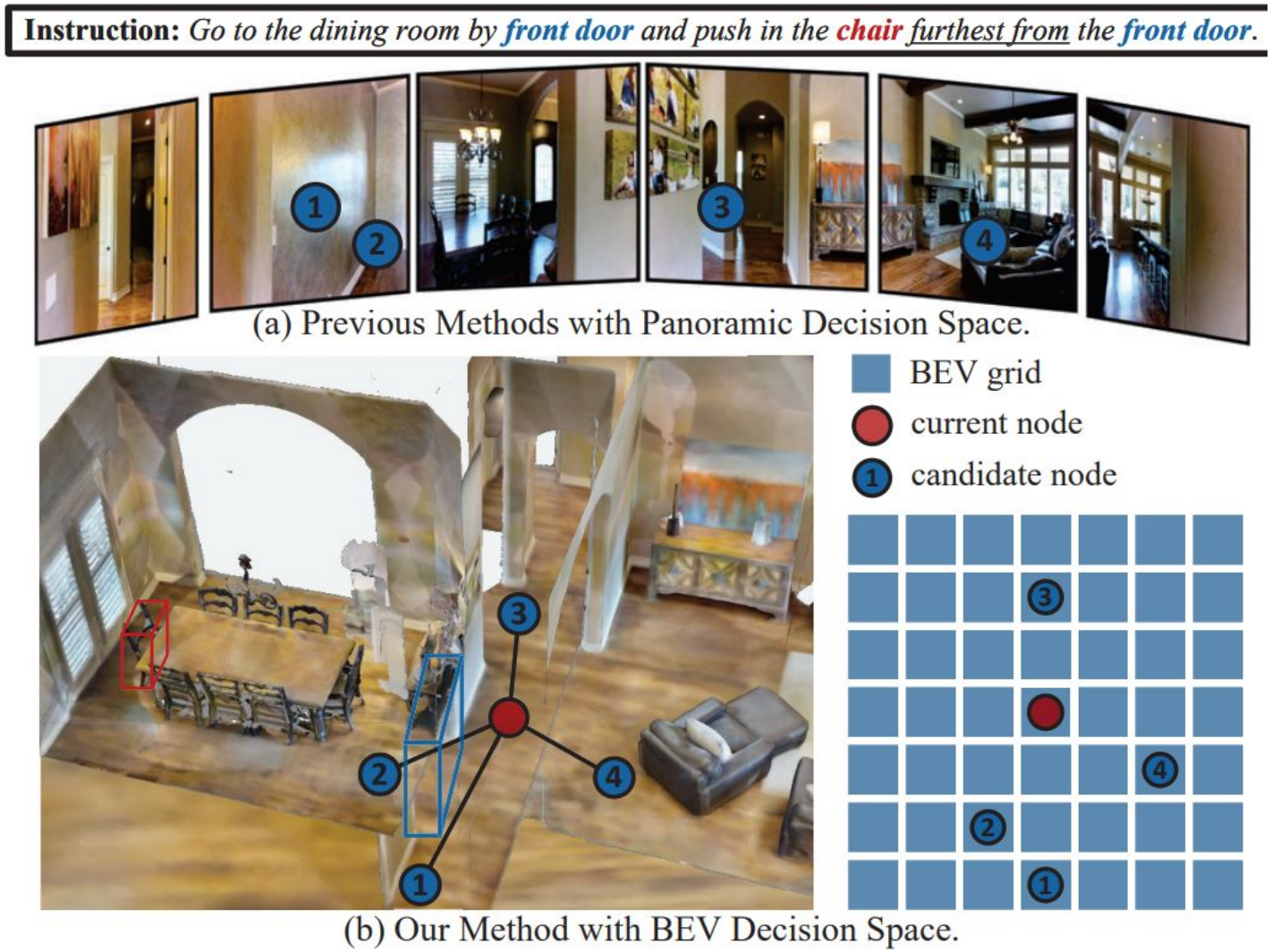
Figure 2. **Architectural comparison.** Previous works (a) [67] and (b) [2] achieve in-context learning for VL tasks with large language models. Our SINC relieves such a constraint by introducing a meta-model for acquiring the in-context ability.

Towards Improved Input Masking for Convolutional Neural Networks

- CNNのための新しいマスク手法を提案
  - 従来の画像マスク方法は画像を黒やグレーで塗りつぶすことから分布シフトが起こること、マスクの形状に余計な情報が含まれる可能性の存在を問題視
  - 中間活性化マップにマスクを適用し



Bird's-Eye-View Scene Graph for Vision-Language Navigation
鳥瞰図のシーングラフBEV Scene Graph (BSG)を提案



ポイント
- 移動先を離散化して曖昧性を減らせる
- BSGは現在位置から都度作成する（専用データセットは必要ない）

方法
BEV-based 3D detection（with Matterport3D）で学習した特徴を利用してグリッド化＋cross attentionやpoolingを利用して離散化

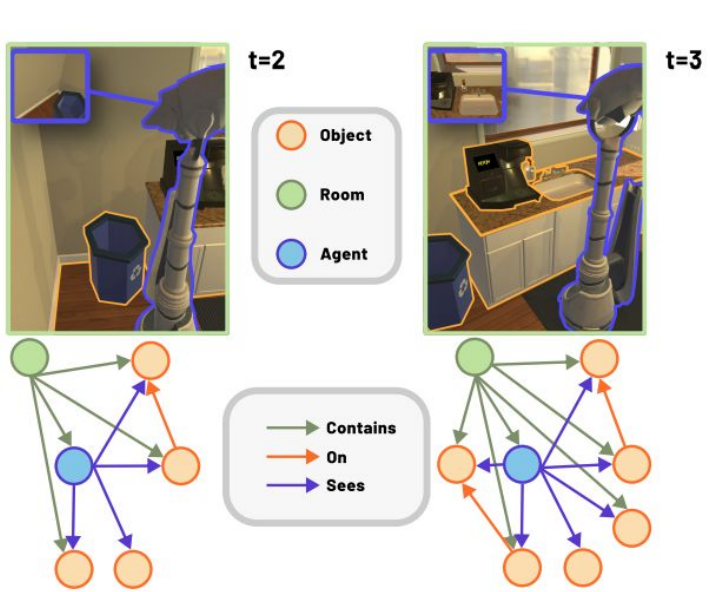
ICCV 2023 の動向・気付き（159/165）

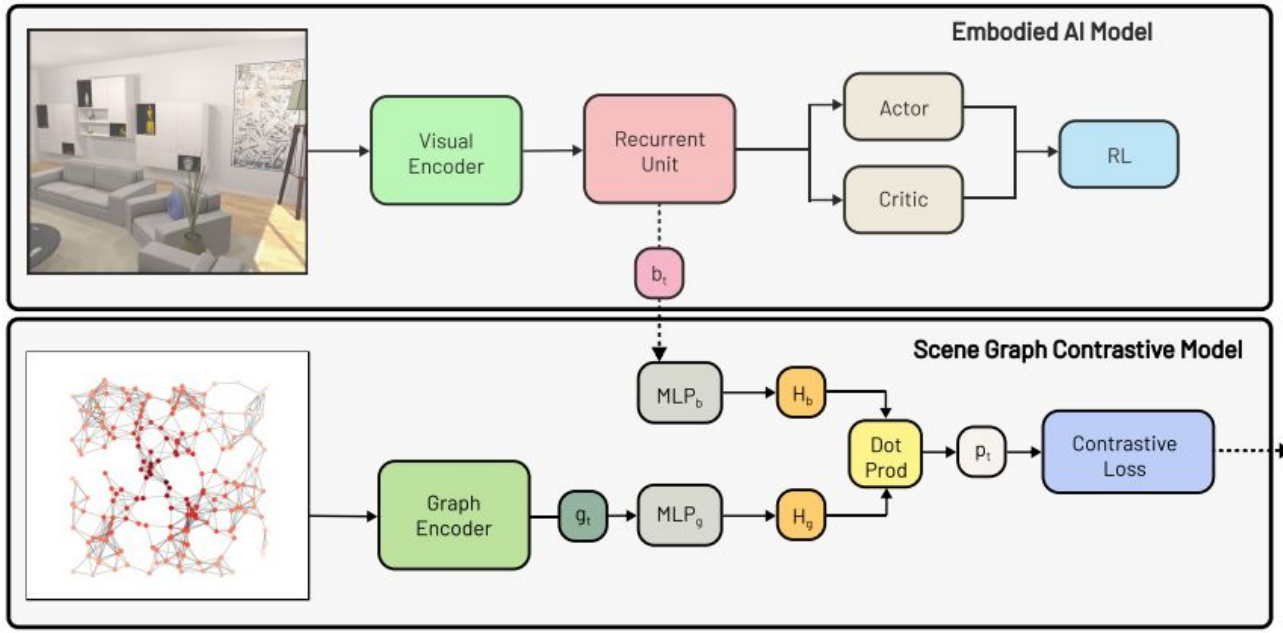
Scene Graph Contrastive Learning for Embodied Navigation

エージェントの一人称視点画像から次の行動に関わる信念情報をリッチに予測するために屋内全体の状況のグラフ表現と対照学習する

ポイント

- シーングラフは学習時のみに使うので推論時には必要ない
- エージェントを動作させながらシーングラフを作成していくので訓練には時間がかかる

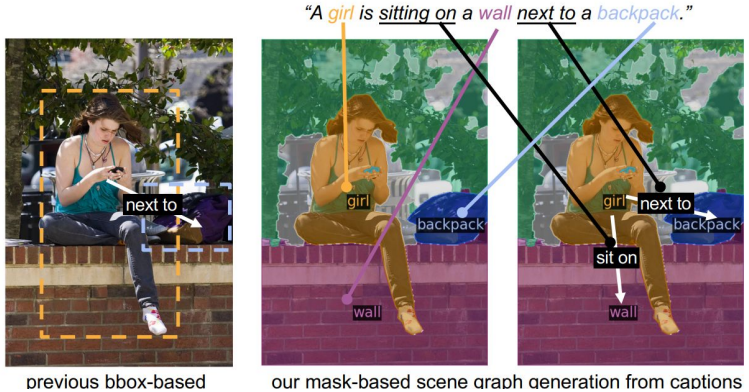




ICCV 2023 の動向・気付き（160/165）

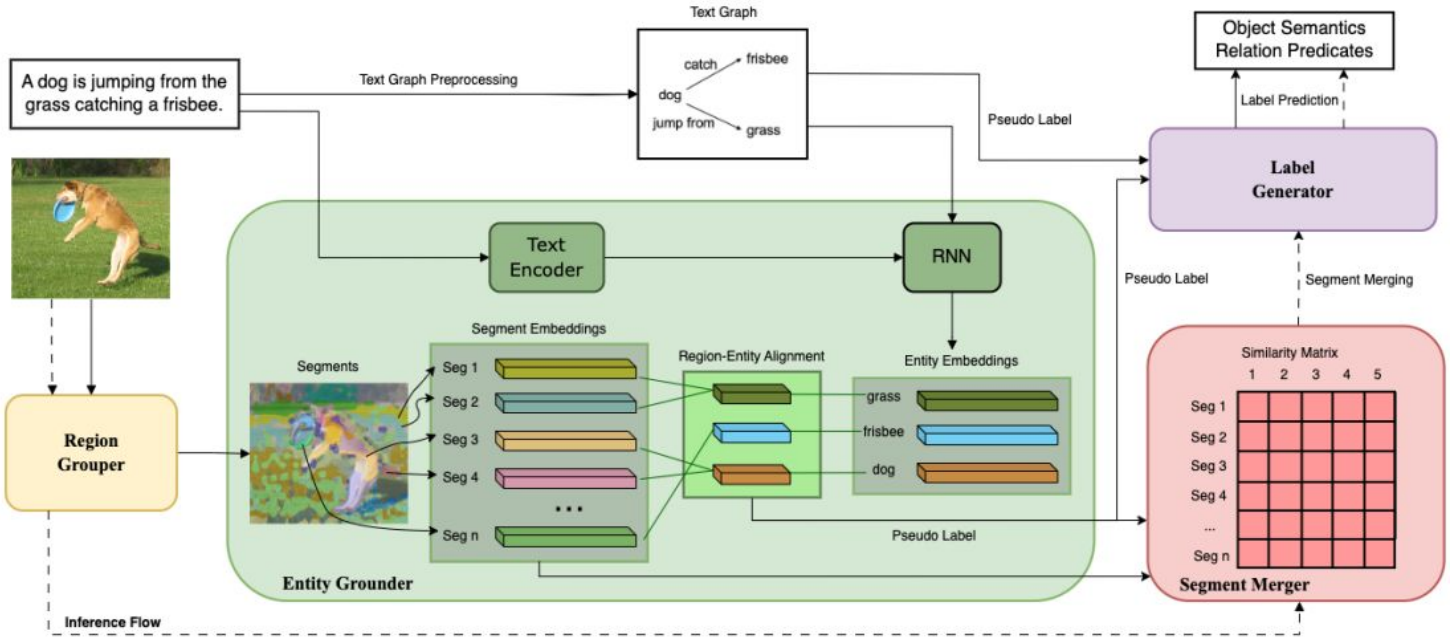
TextPSG: Panoptic Scene Graph Generation from Textual Descriptions

image-textペアを利用した、画像のPSGシーングラフ生成器の提案



ポイント

- 従来のPSGシーングラフは人力アノテーションで高コスト。Web上の画像テキストを使えれば省コスト
- 性能はいまいちなのでやや挑戦的過ぎるかも

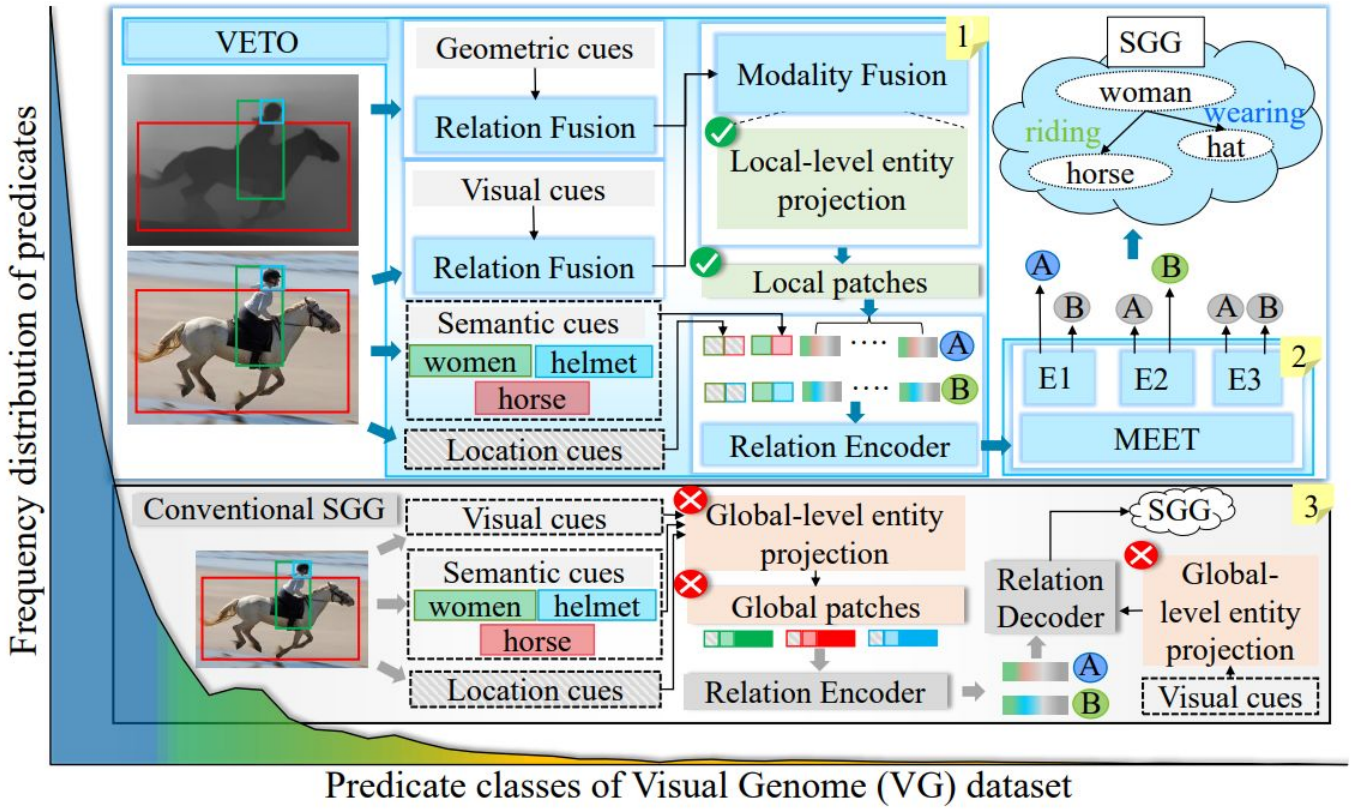


- Region Grouper: GroupViTで領域をパッチのグループに分ける
- Entitiy Grouper: FILIPと似た方法でテキスト特徴と対応関係を見つける
- Segment Merger: refinement処理
- Label Grounder: BLIPでobjectとrelationのラベルを決定

ICCV 2023 の動向・気付き（161/165）

Vision Relation Transformer for Unbiased Scene Graph Generation

Vision Transformerベース局所レベルのエンティティ情報を保持
関係クラスの偏りを軽減する仕組みを導入して性能向上



ポイント

- Vision rElation TransfOrmer (VETO)：局所的なパッチのエンティティ情報を保持することで性能向上
- Mutually Exclusive ExperT (MEET)：クラスの偏りを軽減するために、述語の頻度でクラスを分けて分類器（エキスパート）を学習
- シーングラフ生成の予測性能を最大47%向上

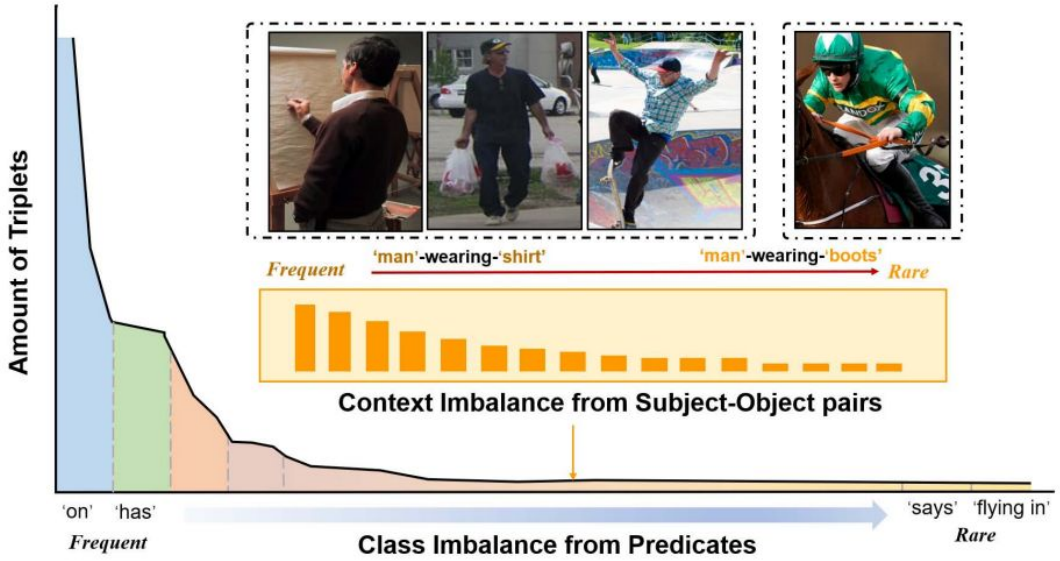
ICCV 2023 の動向・気付き（162/165）

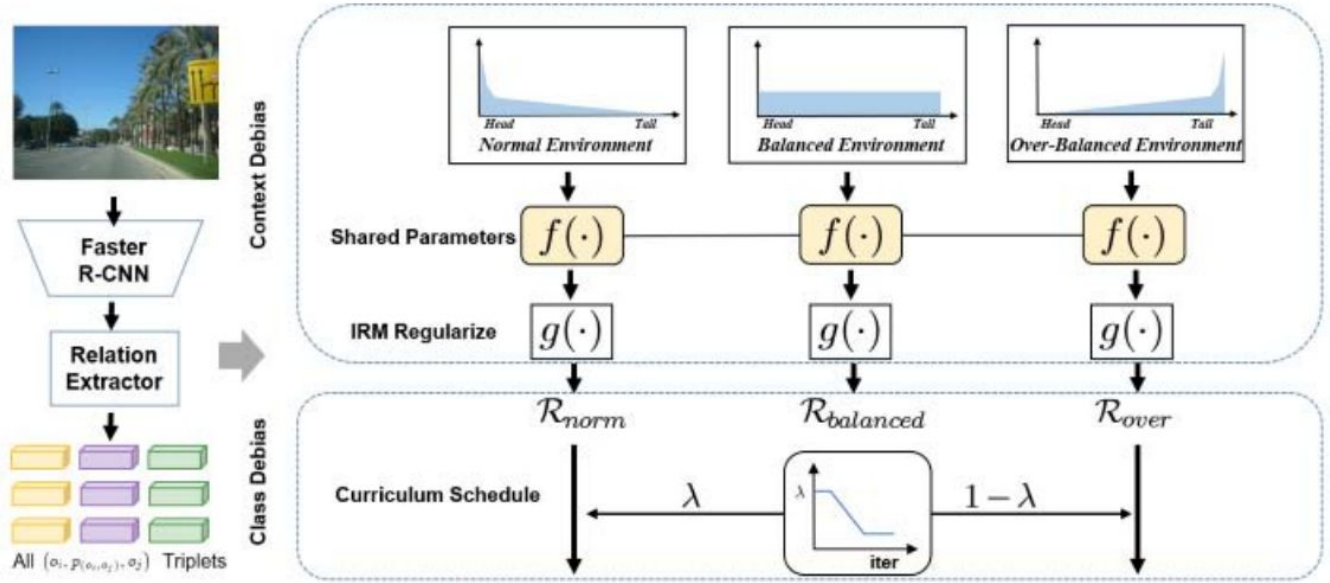
Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation

シーングラフ生成におけるデータバイアスの低減のためのカリキュラム学習手法Environment Invariant Learningの提案

ポイント

- シーングラフ生成タスクでは述語のバイアス、主語・目的語のペアのバイアスがしんどい
- Invariant Risk Minimizationをシーングラフ生成に導入している
- 最初は通常の分布を学習し、徐々に偏ったクラスの学習に移行する

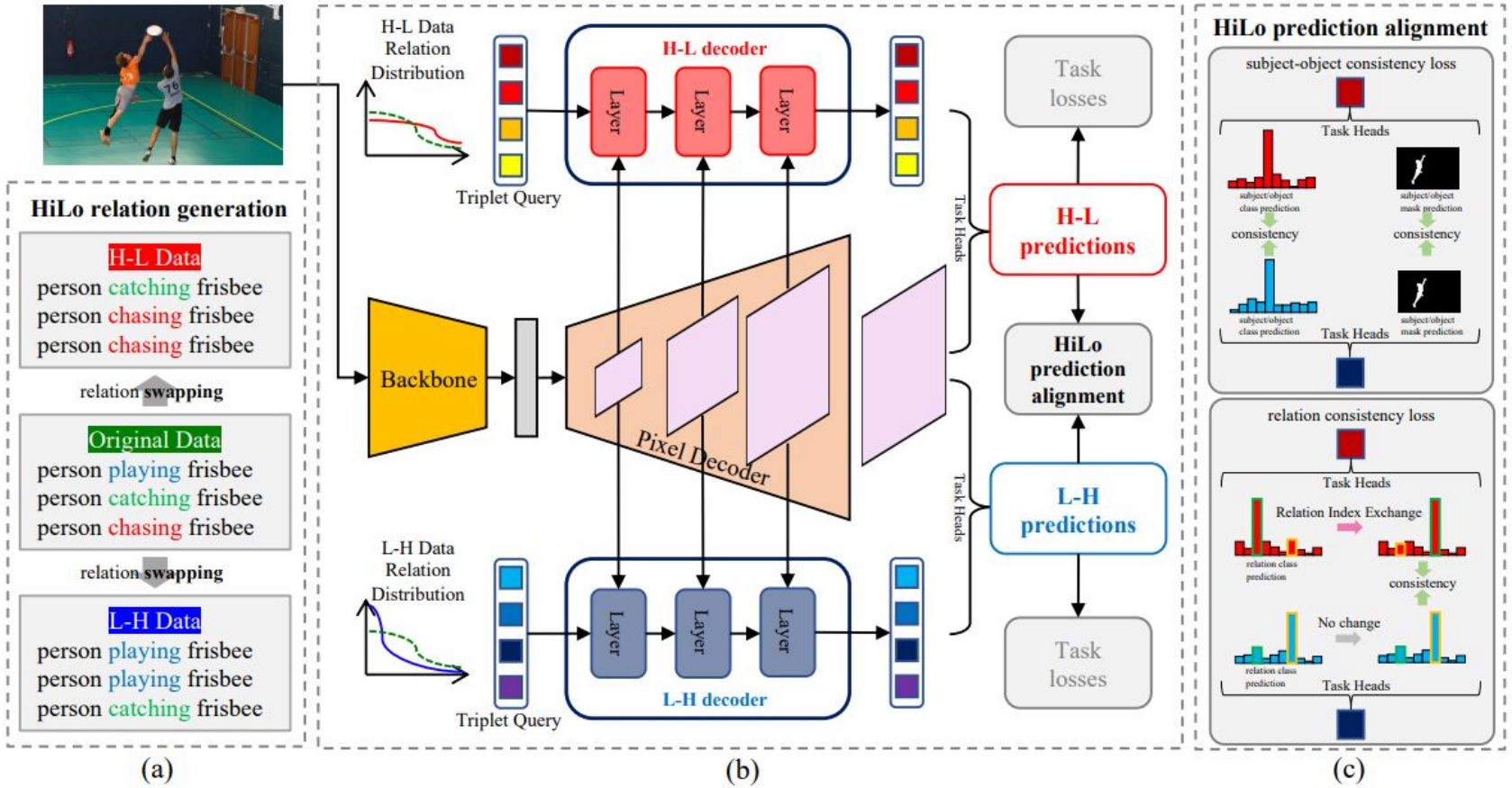




ICCV 2023 の動向・気付き（163/165）

HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation

セグメンテーションベースのシーングラフであるPanoptic Scene Graph生成においてのバイアスの低減手法の提案



ポイント

- シーングラフ生成タスクでは述語のバイアス、主語・目的語のペアのバイアスがしんどい
- 訓練データを高頻度と低頻度で分けてそれぞれ予測器を作成
- subject-object間は2つの予測器で予測が一致するように学習
- relationは低頻度側に高頻度側を合わせるように学習

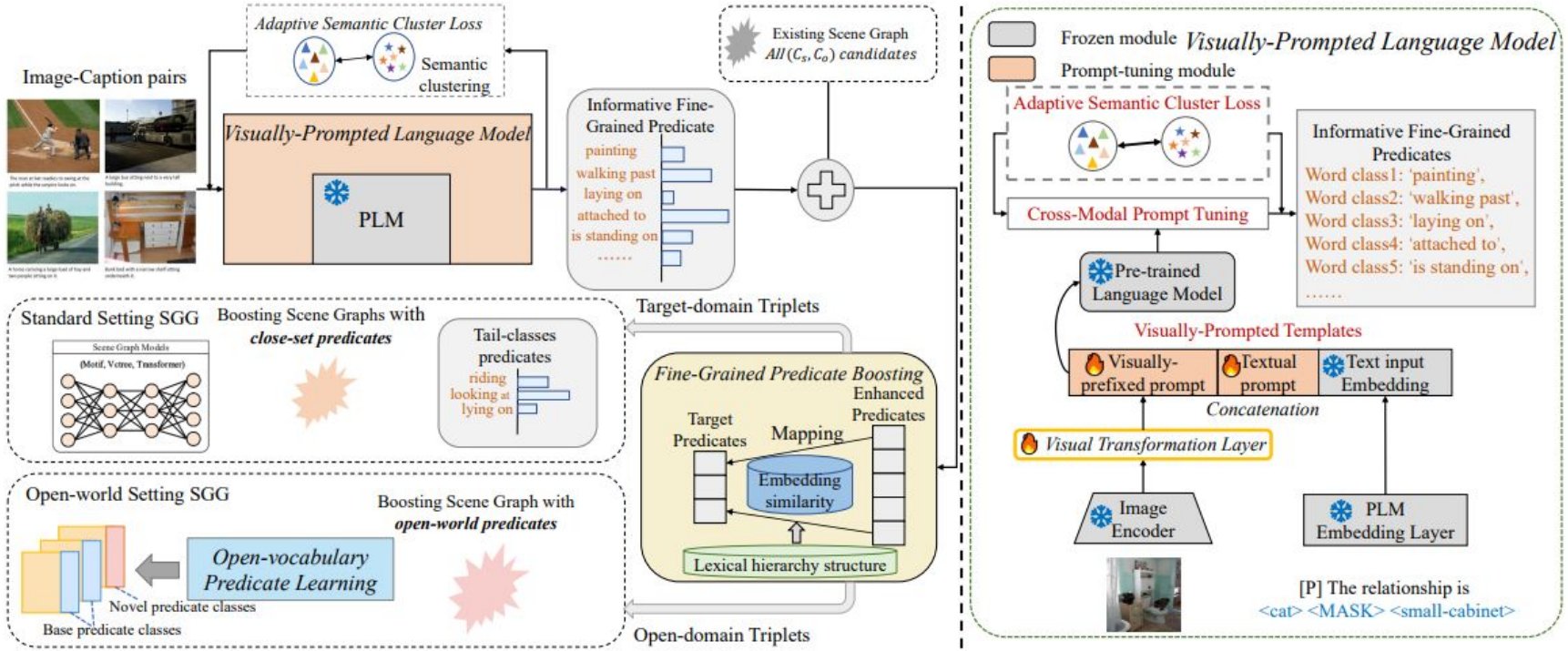
ICCV 2023 の動向・気付き（164/165）

Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an Open World

学習済み言語モデルで低頻度の述語を拡張してバイアスを軽減

ポイント

- シーングラフ生成タスクでは述語のバイアス、主語・目的語のペアのバイアスがしんどい
- 学習済み言語モデルにはBERTを利用。似ている低頻度語はBERT特徴量でK-meansしてクラスタリングして低頻度語を減らす工夫もある

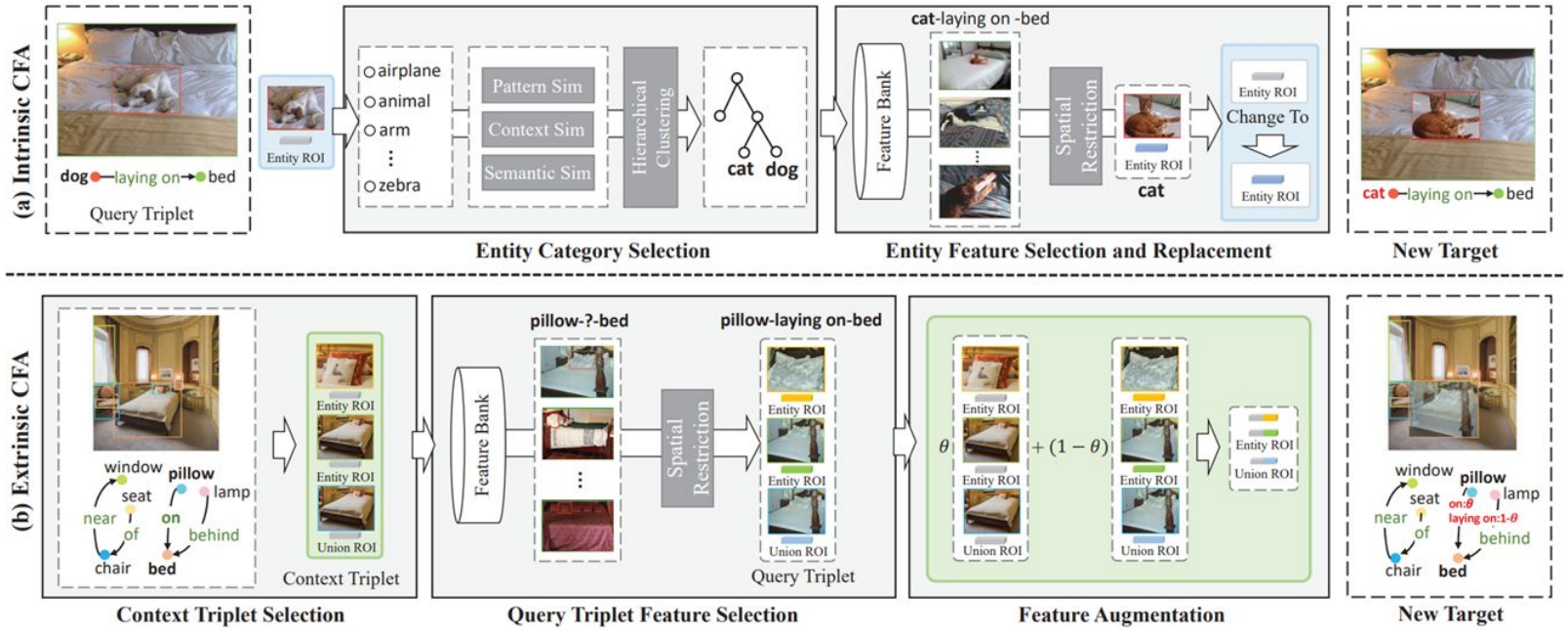


ICCV 2023 の動向・気付き（165/165）

Compositional Feature Augmentation for Unbiased Scene Graph Generation
シーングラフのバイアス低減のためのデータ拡張手法の工夫

ポイント

- シーングラフ生成タスクでは述語のバイアス、主語・目的語のペアのバイアスがしんどい
- intrinsic CFA: 物体特徴を別の特徴で置き換えてデータ拡張。データ拡張効果の高いカテゴリを相関から上手に選択する方法を提案
- extrinsic CFA: 高頻度なトリプレットに低頻度なトリプレットをmix-upしてデータ拡張。学習に悪影響が無いように似ているトリプレットを選択する手法を提案



今後の展望

- 今後，我々としてはどうすれば良いか？

今後の展望(1/2)

ICCV 2023 は世界との差を見せつけられた！？

- 日本から出ている論文は全部で24本(どう見るかはあなた次第！？)
- 恐らく欧州圏それ以上に危機感を持っている？(今回はパリ開催)
- 世界はコラボして投稿している(国際連携論文も多い)

世界に仲間を作ろう,
連携により力を集約してICCVに投稿しよう！

今後の展望(2/2)

大規模モデル時代に何を考えるか？

- 大規模モデルの波を「作る」か「乗る」か「飲まれる」か？

（今回Awardの4分の3が汎用モデル系だった！）

- 波なんて関係ないフィールドを持っておく？
- 何れにせよ，分野に対するメッセージ性のある研究をしよう！

トレンドを創るための研究に挑戦しよう！

次回のICCV 2025は発表者側で参加しよう！！