cvpaper.challenge

ECCV 2024 速報

鳥見晃平、山田亮佑、篠田理沙、Yanjun Sun、大谷豪、
田所龍、北田俊輔、柴田直生、守田竜梧、中原龍一、奥田萌莉、西村
和也、児玉憲武、松尾雄斗、綱島秀樹、大久保蓮、
Hao Guoqing、福原吉博、川村輝大、篠原崇之、阿部純、
井手康允、森稔、柳凜太郎、Yue Qiu、原健翔、片岡裕雄

ECCV 2024 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

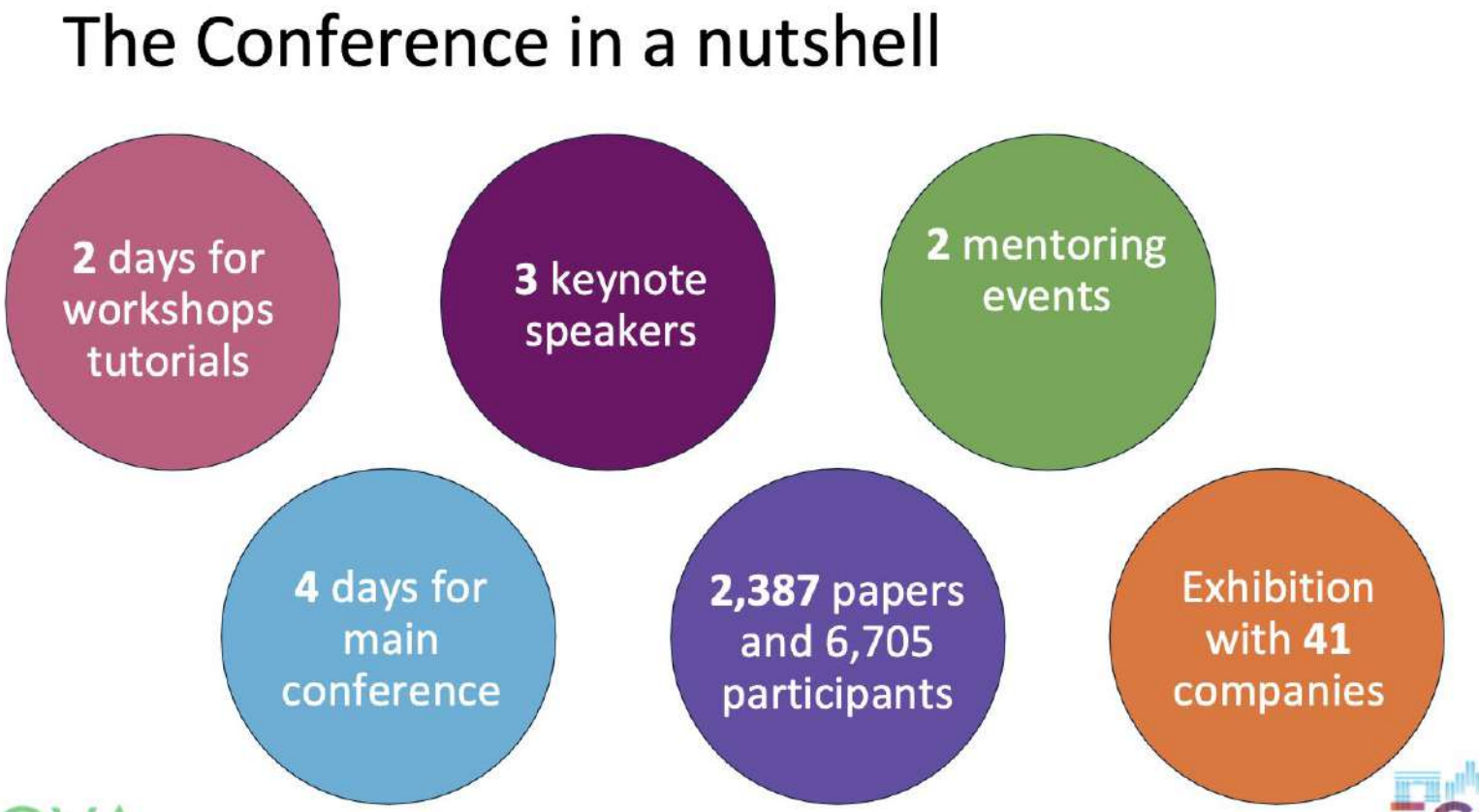
ECCV 2024 の動向・気付き(1/132)

Opening Slideより(1/19)

ECCV 2024 の動向・気付き(2/132)

Opening Slideより(2/19)

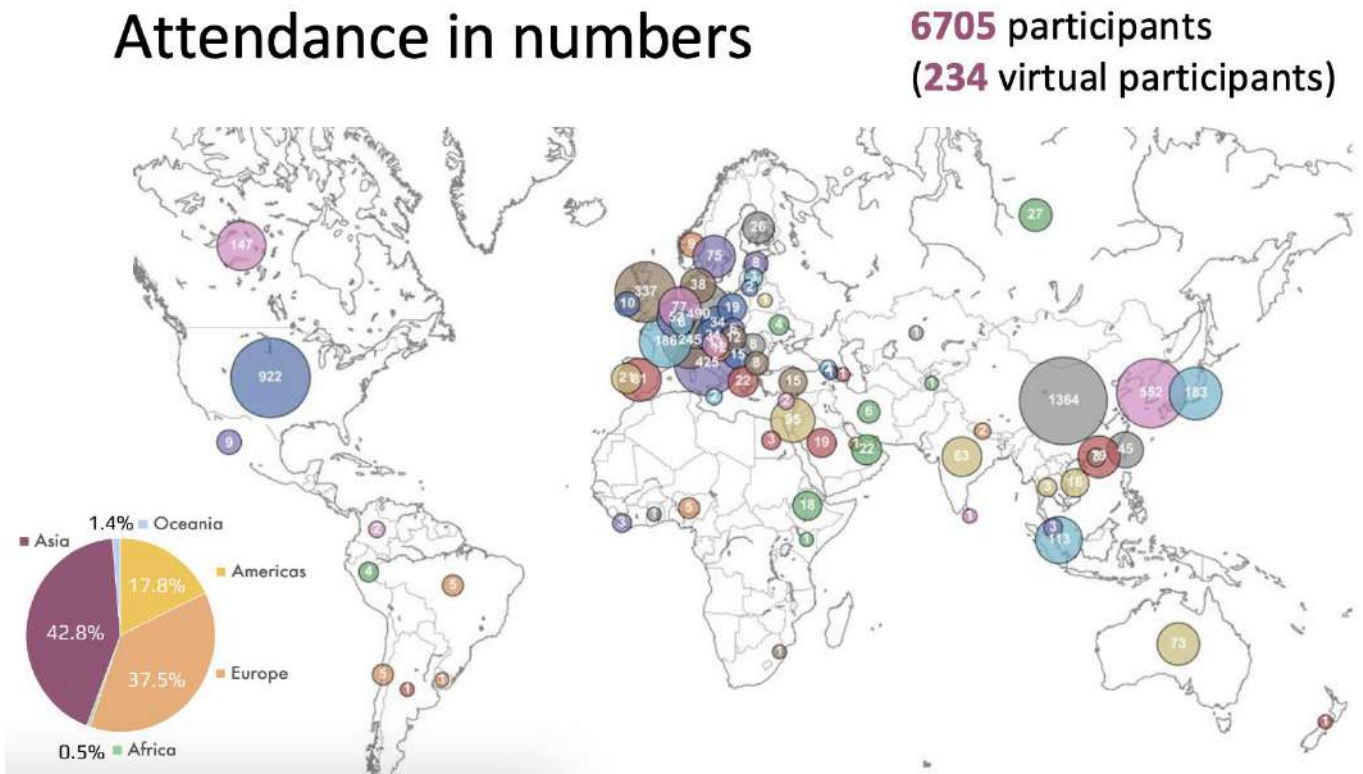
- ワークショップ2日間 / 本会議4日間の日程(本会議が通常より1日長め)
- 2,387件採択 / 6,705名参加はいずれも過去最多
  - だがいずれも夏開催のCVPRと比較すると若干見劣りしてしまう...



ECCV 2024 の動向・気付き（3/132）

Opening Slideより（3/19）

- ECCV（E = Europe）なだけありヨーロッパドメインの参加が多い（37.5%）
- それでもまだアジアドメインが割合は最多（42.8%）
- 南北アメリカは17.8%とCVPRよりは少なめ
  - アメリカ大陸からは発表なかったら行かない、CVPRは見にいこうという雰囲気？



ECCV 2024 の動向・気付き(4/132)

Opening Slideより(4/19)

- Workshop / Tutorialの提案と開催
  - Workshop: 131件中72件が採択(54.9%)
  - Tutorial: 18件中9件が採択(50.0%)

Beyond the main conference

Workshops
- Received proposals: **131**
- Accepted: **72**
- # papers in proceedings: **580**

Tutorials
- Received proposals: **18**
- Accepted: **9**

Workshop & Tutorial Chairs

Alessio del Bue | Cristian Canton | Jordi Pont-Tuset | Tatiana Tommasi

ワークショップ論文投稿側ではなく、運営側です。論文よりは高い確率で採択されます。

ECCV 2024 の動向・気付き(5/132)

Opening Slideより(5/19)

- 参加者の多様性確保のための予算も下りている
  - 旅費(29件)・参加費(58件)・バーチャル参加費(40件)が補助されている
  - 投稿数が前回ECCV 2022よりも4倍に跳ね上がった

Diversity grants

- **524** submissions (nearly 4x over ECCV 2022!)
- We granted:
  - **29** Travel Grants
  - **58** Registration Waivers
  - **40** Virtual Registration Waivers
- Awardees were well-represented across demographics and countries, and primarily presenting a paper for the first time.
- The applicants were reviewed by a panel of reviewers drawn from the community.

Diversity Chairs

David Fouhey

Rita Cucchiara

Diversity Chairがこの予算を運営、旅費や参加費のサポートとして申請すると通るかも？

ECCV 2024 の動向・気付き(6/132)

Opening Slideより(6/19)

- CV分野は特色のあるChairもある
  - スライドに記載なかったがLocal Arrangement Chairはミラノ？ボランティアの学生と見られる写真がある
  - Ethics Review Committeeが新しい → LLMによる査読や剽窃を細かくチェックしていた？
  - Program Chairは6名いる(単純に論文数が多い)
  - 最たる例はいわゆる Twitter(X) Chair でSNSを担当している(Social Media Chair)

Xでのポストが頻繁にあることが、CVPR / ICCV / ECCVの知名度を押し上げている？
※体感として

ECCV 2024 の動向・気付き（7/132）

Opening Slideより（7/19）

- Keynote speakerは3件4名
  - 生成AIによる動画生成、倫理考慮したAI、ドメインシフト

Keynote speakers

Lourdes Agapito & Vittorio Ferrari

"Synthesia: From computer vision research to real-world AI avatars"

Tuesday, 15:30-16:30

Sandra Wachter

"Fair, transparent, and accountable AI: What is legally required, what is ethically desired, and what is technically feasible?"

Wednesday, 15:30-16:30

Sanmi Koyejo

"Is distribution shift still an AI problem?"

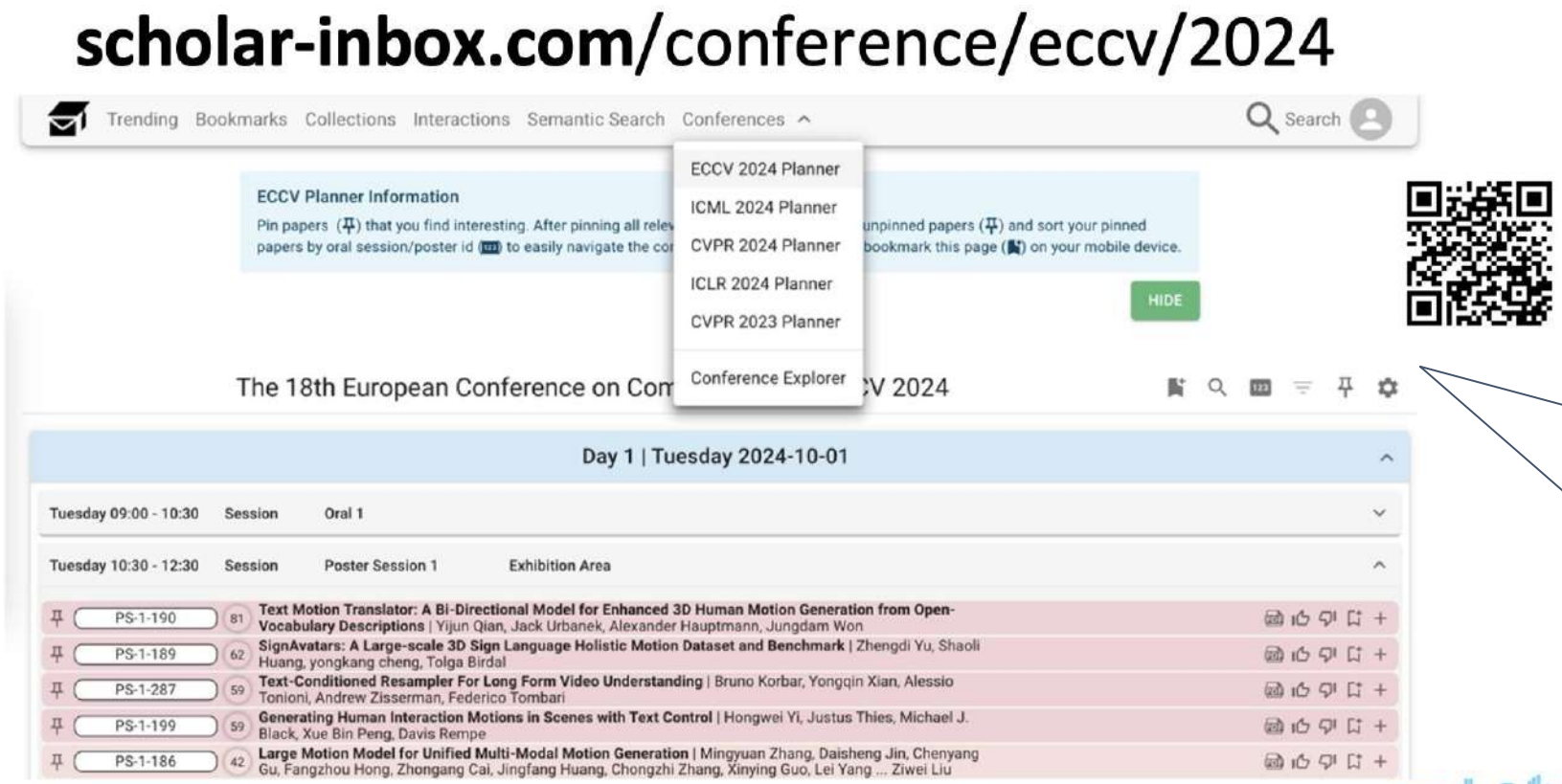
Thursday, 15:30-16:30

国際会議でのKeynoteは「今をときめく技術」と「AIの危険性」などバランスを考えて組まれることが多い

ECCV 2024 の動向・気付き（8/132）

Opening Slideより（8/19）

- Scholar Inbox（SI）による論文検索・アシスト
- SIは日々の論文チェックにとても使える！



みなさんScholar Inbox使っていますか？
cvpaper.challenge の研究メンバーでも増えていると伺います

ECCV 2024 の動向・気付き(9/132)

Opening Slideより(9/19)

- 8,585論文投稿・2,387論文採択 → 採択率27.9%
- 200論文がOral発表 → Oral採択率2.3%

Papers

8,585 valid submissions by **27,546** authors

Of the submissions:

- **2,387** accepted (**27.9%**)
- **200** accepted as orals (**2.3%**)
- **435** included a dataset as part of their contribution
- **173** desk-rejected due to policy violations
- **2,410** withdrawn by authors (at various stages of review process)
- **31** papers also reviewed by the ethics committee

ECVA

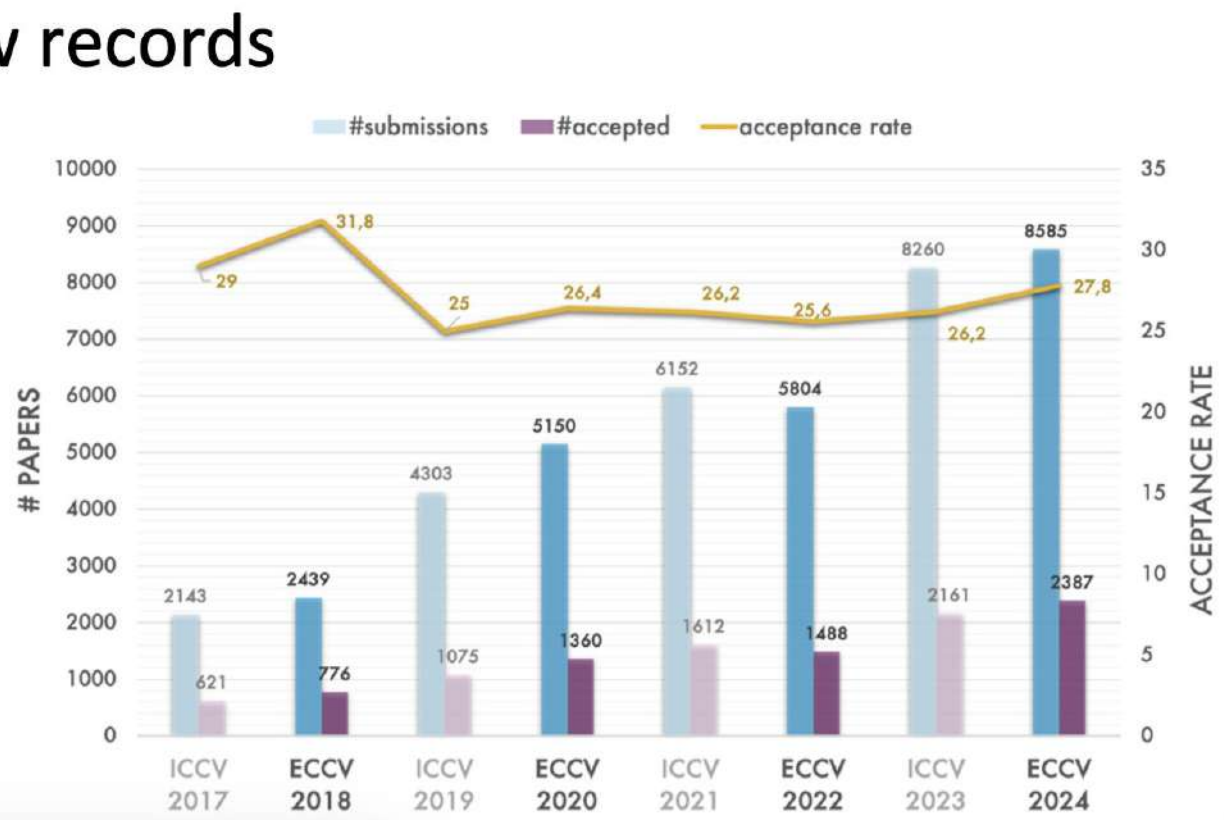
ECCV MILANO2024

cvpaper.challenge

ECCV 2024 の動向・気付き（10/132）

Opening Slideより（10/19）

- 投稿数で見ると最多！
  - 前回 ECCV 2022 から5,804 → 8,585論文投稿で大幅増加（47.9%増加）
  - 昨年同時期投稿のICCV 2023（8,260論文投稿）と比較しても増加中
  - 採択率は25.6% → 27.8%と微増で、当然採択数（2,387論文）もECCVでは最多



ECCVもどこまで伸びるのか？頭打ちはまだ見えていない

ECCV 2024 の動向・気付き（11/132）

Opening Slideより（11/19）

- 下図は国ごとの投稿数（参加者ではない）
- アジアドメインが54.5%と半分以上を占める（日本は273論文投稿）
- 北米アメリカは21.9%、ヨーロッパは20.4%

ECCVだけあり、ヨーロッパ圏内から意識的に投稿しようと考える？

ECCV 2024 の動向・気付き（12/132）

Opening Slideより（12/19）

- 査読プロセスは他の会議と同様だが、捌く本数が多くなっている
  - 92万件のメールが飛び交ったと書いている…
  - 各論文3件以上の査読を集めた上で、Area Chairが議論をしてAccept/Rejectを決定

Review Process

>920k emails through CMT

Double-blind review process managed with CMT

- Each paper received at least 3 reviews (> 26,000 reviews in total)
- Reviewers quota: 6 for senior researchers, 4 otherwise

Decisions made within triplets of ACs

- ACs' load: **18** papers on average
- **Lead ACs:** coordinating the triplet, offering guidance to less experienced ACs, handling emergency situations

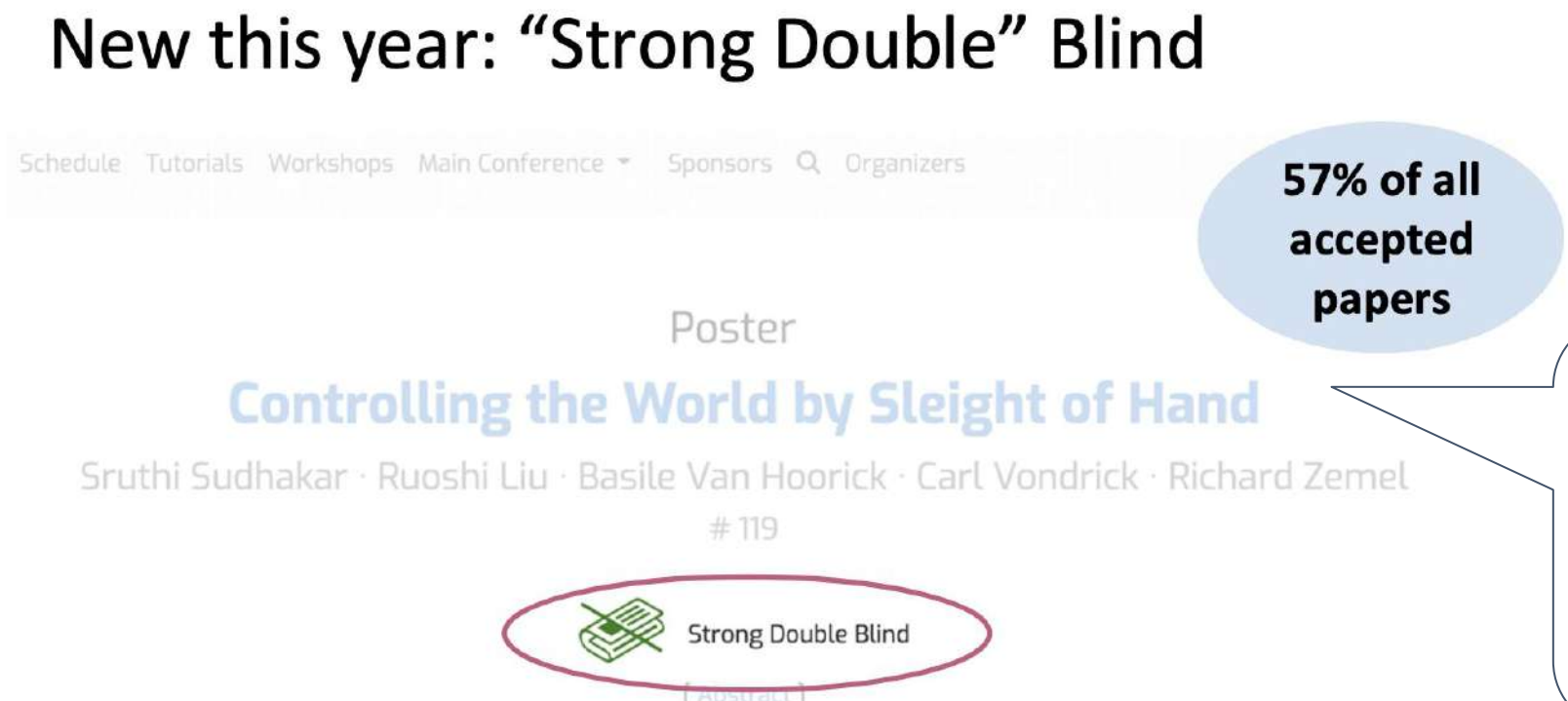
Paper assignments to ACs & reviewers

- Combining scores from CMT Subject Areas, TPMS, OpenReview (along with additional criteria, e.g., recommendations by ACs)
- For ACs also a **newly built affinity score** based on embedding similarity with recent papers on the ACs' Google Scholar profiles

ECCV 2024 の動向・気付き（13/132）

Opening Slideより（13/19）

- Strong Double Blindの導入
  - 論文投稿時に「arXivに投稿しない/していない論文は記載される」と書いてあった
  - 公に著者名と論文の対応関係が見えない状態で査読を通過したという証



やっぱり有名研究者の論文が通るよね…？という疑問を少しずつ解消しようとしている

ECCV 2024 の動向・気付き（14/132）

Opening Slideより（14/19）

- 大規模言語モデル（LLM）による査読は検出される
  - 論文著者側は使用OK、如何なる責任も著者側が背負う
  - 論文査読側は使用NG、公開前の論文をLLMの運営企業に出してしまうことになる
    - 文章そのままではなく、ワーディングを修正するくらいならLLMの使用OK

Large Language Models

ECCV 2024 Policy (same as CVPR *2024*)

- **Authors**:
  - permitted to use any tools, including LLMs, in preparing their papers,
  - fully responsible for any misrepresentation, factual inaccuracies, or plagiarism
- **Reviewers:**
  - strictly **prohibited from inputting submissions** into an LLM
  - **allowed to refine the wording** of their reviews with an LLM
  - ...but, **held accountable** for the accuracy of their content

ECCV 2024 の動向・気付き（15/132）

Opening Slideより（15/19）

- 実際に検出されたLLMによる査読
  - 64件の査読は著者により報告された、うち21件は詳細分析された
  - メタコメントも1件検出されている

Large Language Models

- **Numbers**:
  - **64 reviews** (out of >26,000) were reported by authors, ACs, and PCs
  - **21/64** reviews warrant further investigation
  - **1 AC** used LLM to write meta-reviews based on reviews (detected before meta-review release, replaced by emergency ACs)
  - At least **2** reviewers accused **authors** of using LLMs to generate paper
- **Issues detecting LLM-generated reviews:**
  - Automated tools (GPTZero etc.) just tell whether text was touched by LLM, not whether it was entirely generated, whether a paper was shown…
  - Hard to distinguish between inputting the submission to generate a review (very bad!) versus inputting a review draft for polishing
- **We will suggest submission bans to CVPR/ECCV/ICCV in certain cases**

ECCV 2024 の動向・気付き（16/132）

Opening Slideより（16/19）

- Desk rejectも発生している
  - 14件は二重投稿：ICML, NeurIPS, SIGGRAPH, EMNLPなどと連携
  - 6件は剽窃：自動検出システム（iThenticate）や人間チェック（おそらく査読者やエリアチェア等）

Desk rejections (the very troubling cases)

- **Dual submissions: 14 rejected**
  - Collaborated with concurrent venues: e.g., ICML, NeurIPS, SIGGRAPH, EMNLP, MICCAI, and more
    - **PSA**: Don't forget to withdraw if want to submit elsewhere!!
- **Plagiarism: 6 rejected** (after conditional acceptance)
  - Automated with iThenticate + manual verification
    - **PSA**: Don't plagiarize the related work!! Or anything else...

ECCV 2024 の動向・気付き（17/132）

Opening Slideより（17/19）

- 判定後に異議申し立てできる！
  - 全体で59件の申請
  - 5件がデータセット条件（公開必須？）が覆る
  - 1件が Reject から Accept に覆る（査読者/エリアチェアのポリシー違反）

Appeals

- Authors could fill a form with a valid reason to appeal a decision:
  - policy errors
  - clerical errors
  - significant misunderstandings by the reviewers or ACs
- **59** appeals received (CVPR 2024: 167)
  - PCs processed each appeal and consulted ACs where necessary
  - **5** conditionally accepted papers: dataset condition was removed
  - **1** paper decision changed from reject to accept due to clear policy error by AC & reviewers

SNSでポストされることもあるけど、疑問があったら正規のルートで申請しよう！

ECCV 2024 の動向・気付き(18/132)

Opening Slideより(18/19)

- 査読の数値
  - 7,293査読者(2,056緊急査読者)
  - 198のOutstanding Reviewers
  - 469のエリアチェア
  - 12のOutstanding Area Chairs

成果が多く出ている研究室はOutstanding Reviewerが何人もいる？逆も然りで、Chicken and Egg Problemですね

ECCV 2024 の動向・気付き(19/132)

Opening Slideより(19/19)

- 15件の賞候補論文

➢ **Robust Fitting on a Gate Quantum Computer** – Frances Yang · Michele Sasdelli · Tat-Jun Chin

➢ **Sapiens: Foundation for Human Vision Models** – Rawal Khirodkar · Timur Bagautdinov · Julieta Martinez · Zhaoen Su · Austin T James · Peter Selednik · Stuart Anderson · Shunsuke Saito

➢ **SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow** – Yihan Wang · Lahav Lipson · Jia Deng

➢ **LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning** – Bolin Lai · Xiaoliang Dai · Lawrence Chen · Guan Pang · James Rehg · Miao Liu

➢ **PointLLM: Empowering Large Language Models to Understand Point Clouds** – Runsen Xu · Xiaolong Wang · Tai Wang · Yilun Chen · Jiangmiao Pang · Dahua Lin

➢ **Integer-Valued Training and Spike-driven Inference Spiking Neural Network for High-performance and Energy-efficient Object Detection** – Xinhao Luo · Man Yao · Yuhong Chou · Bo Xu · Guoqi Li

➢ **Minimalist Vision with Freeform Pixels** – Jeremy Klotz · Shree Nayar

➢ **Latent Diffusion Prior Enhanced Deep Unfolding for Snapshot Spectral Compressive Imaging** – Zongliang Wu · Ruiying Lu · Ying Fu · Xin Yuan

➢ **PathMMU: A Massive Multimodal Expert-Level Benchmark for Understanding and Reasoning in Pathology** – Yuyuan Sun · Hao Wu · Chenglu Zhu · Sunyi Zheng · Qizi Chen · Kai Zhang · Yunlong Zhang · Dan Wan · Xiaoxiao Lan · Mengyue Zheng · Jingxiong Li · Xinheng Lyu · Tao Lin · Lin Yang

➢ **Expanding Scene Graph Boundaries: Fully Open-vocabulary Scene Graph Generation via Visual-Concept Alignment and Retention** – Zuyao Chen · Jinlin Wu · Zhen Lei · Zhaoxiang Zhang · Chang Wen Chen

➢ **Efficient Bias Mitigation Without Privileged Information** – Mateo Espinosa Zarlenga · Sankaranarayanan · Jerone Andrews · Zohreh Shams · Mateja Jamnik · Alice Xiang

➢ **Rasterized Edge Gradients: Handling Discontinuities Differentially** – Stanislav Pidhorskyi · Tomas Simon · Gabriel Schwartz · He Wen · Yaser Sheikh · Jason Saragih

➢ **On the Topology Awareness and Generalization Performance of Graph Neural Networks** – Junwei Su · Chuan Wu

➢ **Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models** – Vitali Petsiuk · Kate Saenko

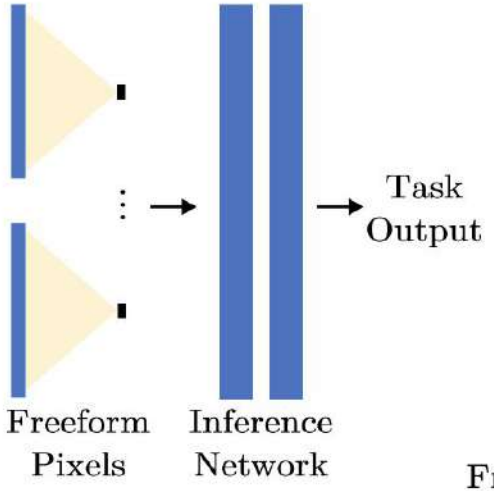
➢ **Controlling the World by Sleight of Hand** – Sruthi Sudhakar · Ruoshi Liu · Basile Van Hoorick · Carl Vondrick · Richard Zemel

ECCV 2024 の動向・気付き（20/132）

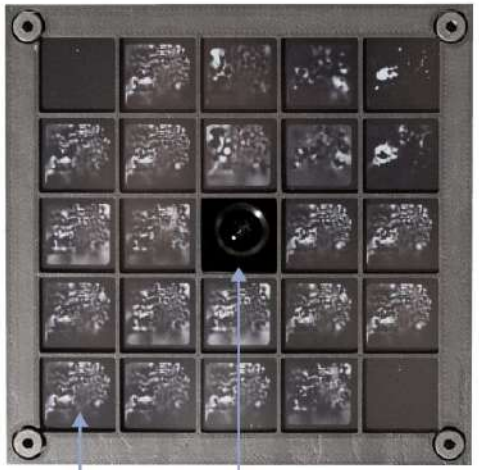
Best Paper Award: Minimalist Vision with Freeform Pixels

- タスクを解くのに必要な最小限の数の任意形状のsuperpixelを学習する。
- 任意形状のsuperpixelを撮像できるカメラを実装し、複数のタスクで評価。
- プロトタイプのカメラはソーラーパネルによる発電で自律動作が可能。
- プライバシーやサステナビリティの問題にアプローチしており、SDGs等に対する意識の高いヨーロッパらしいBest Paperの選定と言える。

(a) Workspace Monitoring



(b) Camera in a Network



(c) Minimalist Camera


発表中にカメラの実物を見せていた

ECCV 2024 の動向・気付き(21/132)

Best Paper Honorable Mention

- “Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models”
  - 著者らによるARC(ARithmetics in Concept space)によって、拡散モデルにおいて抑制する概念の再現性が向上した
    - Concept Inhibitionは、拡散モデルにおける重みの調整、修正を通じて特定の概念(倫理的・法的懸念のある概念等)の生成を防ぐアプローチ
    - 特定の概念の抑制が達成できているかは、敵対者に発見される脆弱性やバックドアを想定することで検証される
  - 推論の特性を利用した手法(ARC)で、抑制された概念が複数のプロンプトによって再構成されることを示した
    - 現在のConcept Inhibitionの限界、脆弱性を指摘している

Vitali Petsiuk, Kate Saenko. “Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models” , in ECCV 2024.
https://arxiv.org/abs/2404.13706

ECCV 2024 の動向・気付き（22/132）

Best Paper Honorable Mention

- “Rasterized Edge Gradients Handling Discontinuities Differentially”
  - ラスタライズベースのレンダリングは、不連続性とレンダリング近似のため正確な勾配の計算が困難
  - 提案手法“EdgeGrad”によって、ラスタライズされた画像を連続的に捉えた勾配計算の実現、計算の高速化、再現度の向上を達成

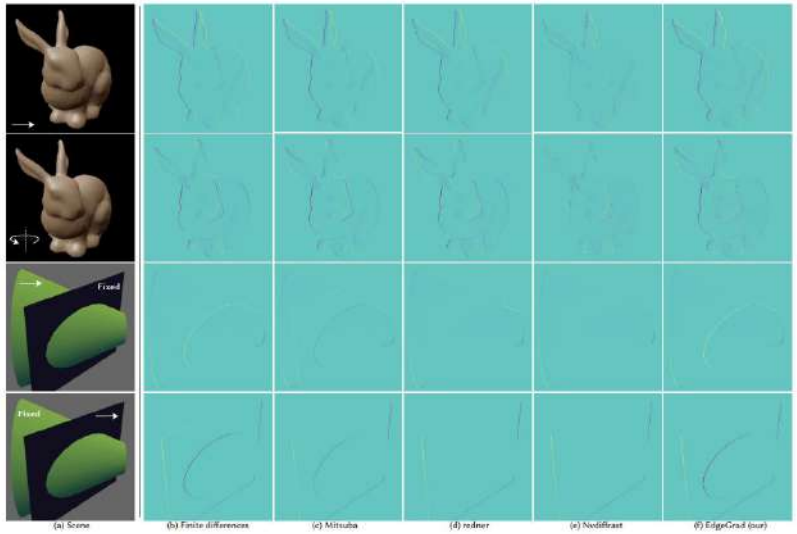


Fig. 5: Comparison of forward gradient on several test scenes with the numerical solution using finite differences. (a) Synthetic test scene, we show gradients with respect to a parameter. We use (b) finite differences as a reference, and compare with (c) Mitsuba 3 [15,36], (d) redner [21], (e) nvdiffrast [20], and (f) our edge gradient approach.

Scene			
redner [21]	1.20%	53.52%	37.94%
Mitsuba3 [36]	0.67%	75.83%	37.59%
Nvdiffrast [20]	45.91%	69.33%	39.42%
EdgeGrad(our)	6.01%	3.35%	8.35%

Table 1: Accuracy of backward gradients. Relative error, % (↓). This table shows relative errors in backward gradient computations for test scenes. Second and third scenes include geometry intersections, emphasizing our method’s advantage in managing these complexities.

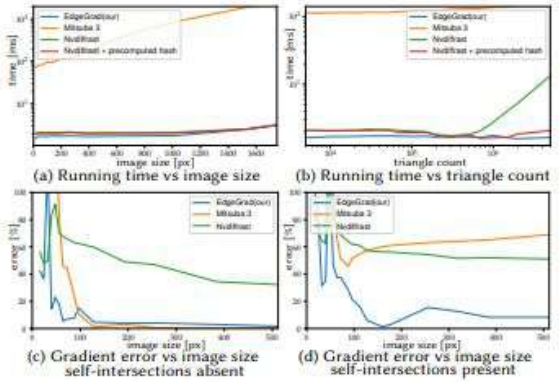


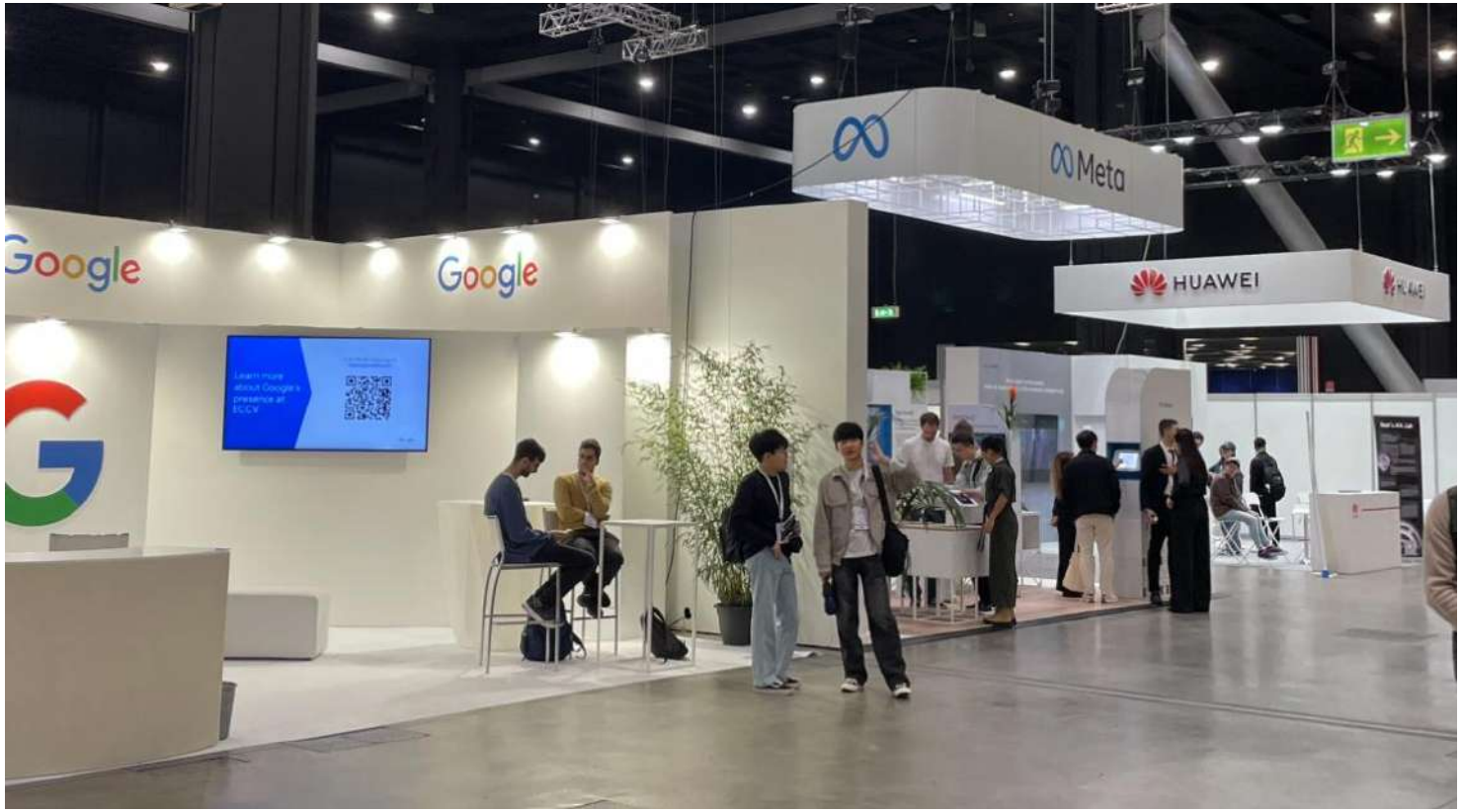
Fig. 6: Runtime performance and errors. a) Runtime [ms] by image size [px]; b) runtime [ms] by triangle count; c) gradient error [%] by image size [px] without self-intersections; d) gradient error [%] by image size [px] with self-intersections.

Pidhorskyi, S. et al. “Rasterized Edge Gradients: Handling Discontinuities Differentiably” in ECCV 2024 https://arxiv.org/abs/2405.02508

ECCV 2024 の動向・気付き(23/132)

Exhibits

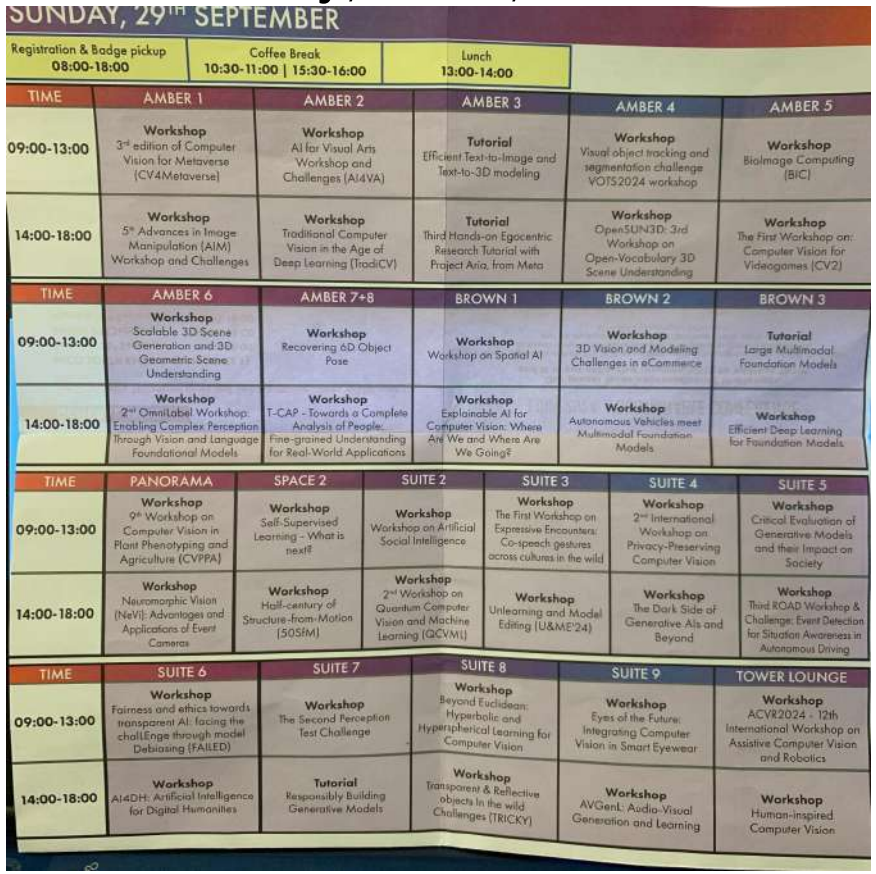
- 今年も企業によるdemoが賑わっていた



ECCV 2024 の動向・気付き（24/132）

Work Shopについて

- Multimodality,Audio,Biometricsなど、多岐にわたる知識が集約されていた



SUNDAY, 29TH SEPTEMBER

Registration & Badge pickup 08:00-18:00 | Coffee Break 10:30-11:00 | 15:30-16:00 | Lunch 13:00-14:00

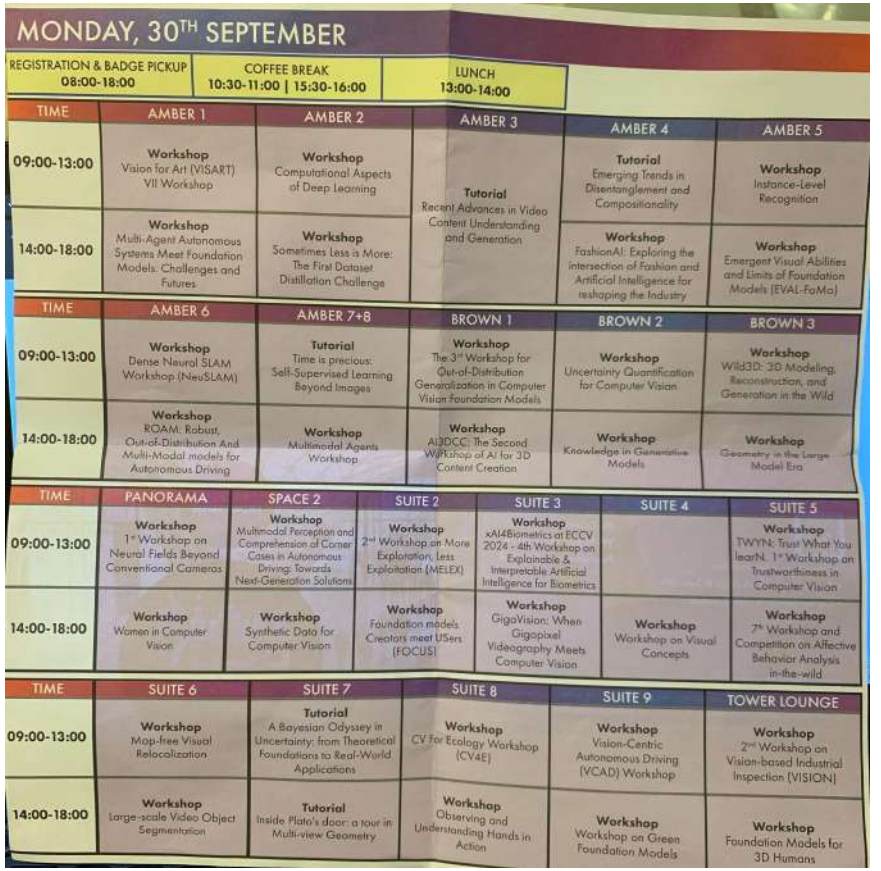
TIME	AMBER 1	AMBER 2	AMBER 3	AMBER 4	AMBER 5
09:00-13:00	Workshop 3rd edition of Computer Vision for Metaverse (CV4Metaverse)	Workshop AI for Visual Arts Workshop and Challenges (AI4VA)	Tutorial Efficient Text-to-Image and Text-to-3D modeling	Workshop Visual object tracking and segmentation challenge VOTS2024 workshop	Workshop BioImage Computing (BIC)
14:00-18:00	Workshop 5th Advances in Image Manipulation (AIM) Workshop and Challenges	Workshop Traditional Computer Vision in the Age of Deep Learning (TradiCV)	Tutorial Third Hands-on Egocentric Research Tutorial with Project Aria, from Meta	Workshop OpenSUN3D: 3rd Workshop on Open-Vocabulary 3D Scene Understanding	Workshop The First Workshop on: Computer Vision for Videogames (CV2)

TIME	AMBER 6	AMBER 7+8	BROWN 1	BROWN 2	BROWN 3
09:00-13:00	Workshop Scalable 3D Scene Generation and 3D Geometric Scene Understanding	Workshop Recovering 6D Object Pose	Workshop Workshop on Spatial AI	Workshop 3D Vision and Modeling Challenges in eCommerce	Tutorial Large Multimodal Foundation Models
14:00-18:00	Workshop 2nd OmniLabel Workshop: Enabling Complex Perception Through Vision and Language Foundational Models	Workshop T-CAP - Towards a Complete Analysis of People: Fine-grained Understanding for Real-World Applications	Workshop Explainable AI for Computer Vision: Where Are We and Where Are We Going?	Workshop Autonomous Vehicles meet Multimodal Foundation Models	Workshop Efficient Deep Learning for Foundation Models

TIME	PANORAMA	SPACE 2	SUITE 2	SUITE 3	SUITE 4	SUITE 5
09:00-13:00	Workshop 9th Workshop on Computer Vision in Plant Phenotyping and Agriculture (CVPPA)	Workshop Self-Supervised Learning - What is next?	Workshop Workshop on Artificial Social Intelligence	Workshop The First Workshop on Expressive Encounters: Co-speech gestures across cultures in the wild	Workshop 2nd International Workshop on Privacy-Preserving Computer Vision	Workshop Critical Evaluation of Generative Models and their Impact on Society
14:00-18:00	Workshop Neuromorphic Vision (NeVi): Advantages and Applications of Event Cameras	Workshop Half-century of Structure-from-Motion (50StM)	Workshop 2nd Workshop on Quantum Computer Vision and Machine Learning (QCVML)	Workshop Unlearning and Model Editing (U&ME'24)	Workshop The Dark Side of Generative AIs and Beyond	Workshop Third ROAD Workshop & Challenge: Event Detection for Situation Awareness in Autonomous Driving

TIME	SUITE 6	SUITE 7	SUITE 8	SUITE 9	TOWER LOUNGE
09:00-13:00	Workshop Fairness and ethics towards transparent AI: facing the chalLEnge through model Debiasing (FAILED)	Workshop The Second Perception Test Challenge	Workshop Beyond Euclidean: Hyperbolic and Hyperspherical Learning for Computer Vision	Workshop Eyes of the Future: Integrating Computer Vision in Smart Eyewear	Workshop ACVR2024 - 12th International Workshop on Assistive Computer Vision and Robotics
14:00-18:00	Workshop AI4DH: Artificial Intelligence for Digital Humanities	Tutorial Responsibly Building Generative Models	Workshop Transparent & Reflective objects in the wild Challenges (TRICKY)	Workshop AVGenL: Audio-Visual Generation and Learning	Workshop Human-inspired Computer Vision

WS：1日目



MONDAY, 30TH SEPTEMBER

REGISTRATION & BADGE PICKUP 08:00-18:00 | COFFEE BREAK 10:30-11:00 | 15:30-16:00 | LUNCH 13:00-14:00

TIME	AMBER 1	AMBER 2	AMBER 3	AMBER 4	AMBER 5
09:00-13:00	Workshop Vision for Art (VISART) VII Workshop	Workshop Computational Aspects of Deep Learning	Tutorial Recent Advances in Video Content Understanding and Generation	Tutorial Emerging Trends in Disentanglement and Compositionality	Workshop Instance-Level Recognition
14:00-18:00	Workshop Multi-Agent Autonomous Systems Meet Foundation Models: Challenges and Futures	Workshop Sometimes Less is More: The First Dataset Distillation Challenge		Workshop FashionAI: Exploring the intersection of Fashion and Artificial Intelligence for reshaping the Industry	Workshop Emergent Visual Abilities and Limits of Foundation Models (EVAL-FoMo)

TIME	AMBER 6	AMBER 7+8	BROWN 1	BROWN 2	BROWN 3
09:00-13:00	Workshop Dense Neural SLAM Workshop (NeuSLAM)	Tutorial Time is precious: Self-Supervised Learning Beyond Images	Workshop The 3rd Workshop for Out-of-Distribution Generalization in Computer Vision Foundation Models	Workshop Uncertainty Quantification for Computer Vision	Workshop Wild3D: 3D Modeling, Reconstruction, and Generation in the Wild
14:00-18:00	Workshop ROAM: Robust, Out-of-Distribution And Multi-Modal models for Autonomous Driving	Workshop Multimodal Agents Workshop	Workshop AI3DCC: The Second Workshop of AI for 3D Content Creation	Workshop Knowledge in Generative Models	Workshop Geometry in the Large Model Era

TIME	PANORAMA	SPACE 2	SUITE 2	SUITE 3	SUITE 4	SUITE 5
09:00-13:00	Workshop 1st Workshop on Neural Fields Beyond Conventional Cameras	Workshop Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving: Towards Next-Generation Solutions	Workshop 2nd Workshop on More Exploration, Less Exploitation (MELEX)	Workshop xAI4Biometrics at ECCV 2024 - 4th Workshop on Explainable & Interpretable Artificial Intelligence for Biometrics		Workshop TWYN: Trust What You learN. 1st Workshop on Trustworthiness in Computer Vision
14:00-18:00	Workshop Women in Computer Vision	Workshop Synthetic Data for Computer Vision	Workshop Foundation models Creators meet USers (FOCUS)	Workshop GigaVision: When Gigapixel Videography Meets Computer Vision	Workshop Workshop on Visual Concepts	Workshop 7th Workshop and Competition on Affective Behavior Analysis in-the-wild

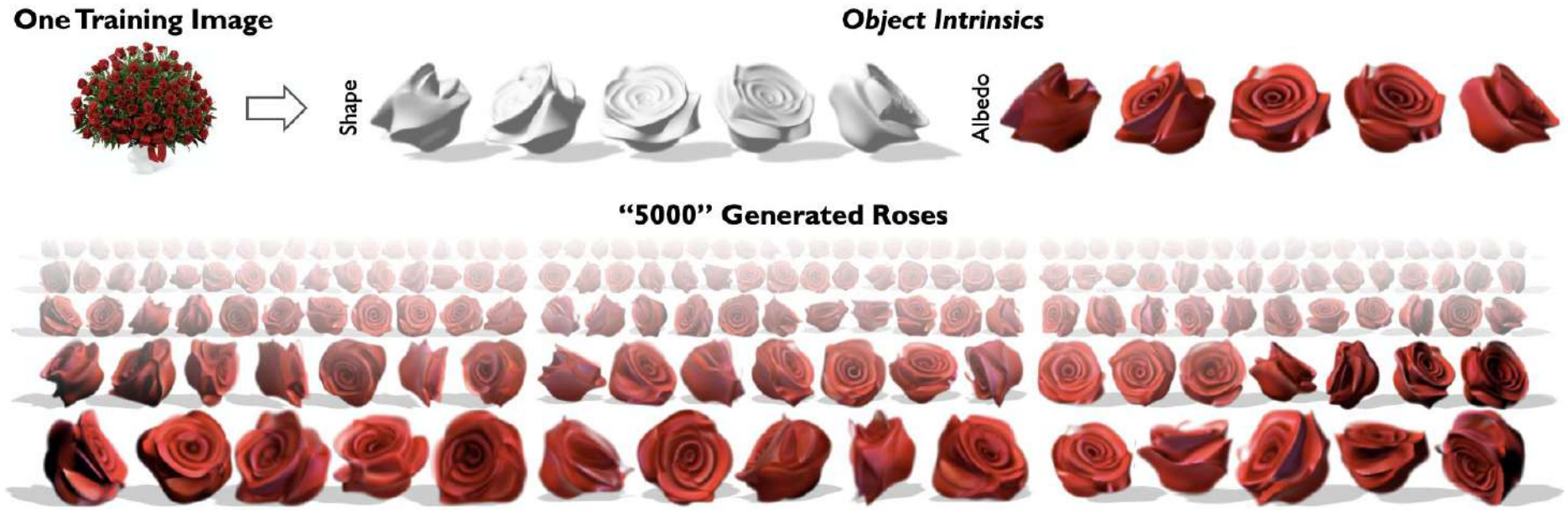
TIME	SUITE 6	SUITE 7	SUITE 8	SUITE 9	TOWER LOUNGE
09:00-13:00	Workshop Map-free Visual Relocalization	Tutorial A Bayesian Odyssey in Uncertainty: from Theoretical Foundations to Real-World Applications	Workshop CV for Ecology Workshop (CV4E)	Workshop Vision-Centric Autonomous Driving (VCAD) Workshop	Workshop 2nd Workshop on Vision-based Industrial Inspection (VISION)
14:00-18:00	Workshop Large-scale Video Object Segmentation	Tutorial Inside Plato's door: a tour in Multi-view Geometry	Workshop Observing and Understanding Hands in Action	Workshop Workshop on Green Foundation Models	Workshop Foundation Models for 3D Humans

WS：2日目

ECCV 2024 の動向・気付き(25/132)

Workshop: 1st Workshop on Scalable 3D Scene Generation and Geometric Scene Understanding

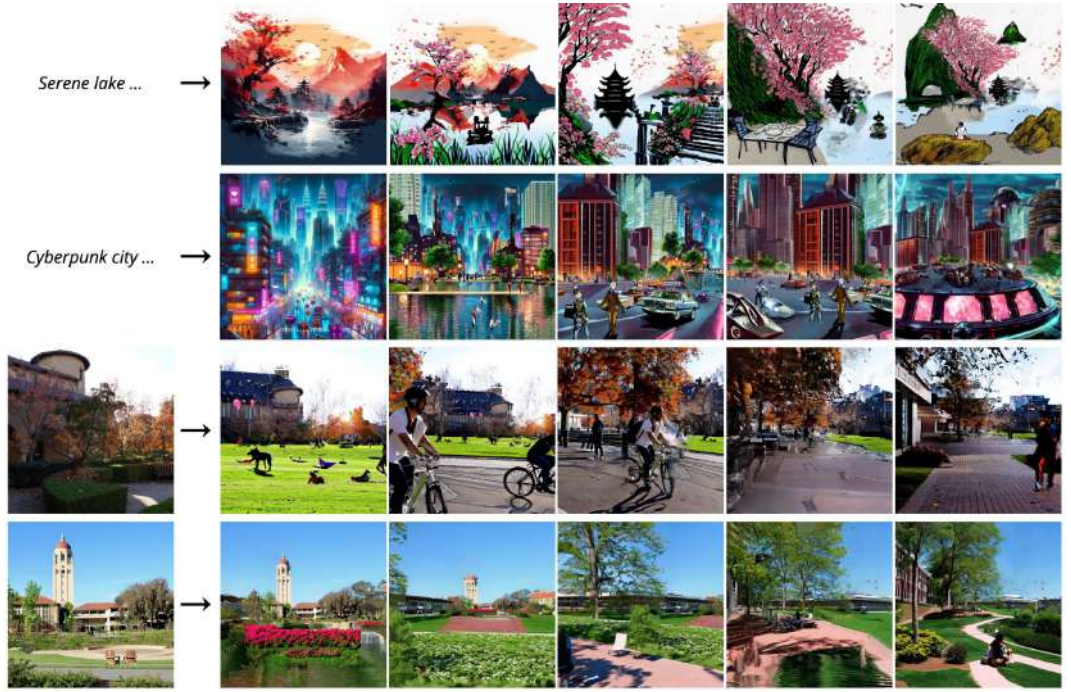
- Jiajun Wu先生の研究紹介 (1/3)
  - Seeing a Rose in Five Thousand Ways
    - バラのジオメトリ・テクスチャ・マテリアルなどの内部特性の知識を獲得，それらの知識を活用したレンダリング



ECCV 2024 の動向・気付き(26/132)

Workshop: 1st Workshop on Scalable 3D Scene Generation and Geometric Scene Understanding

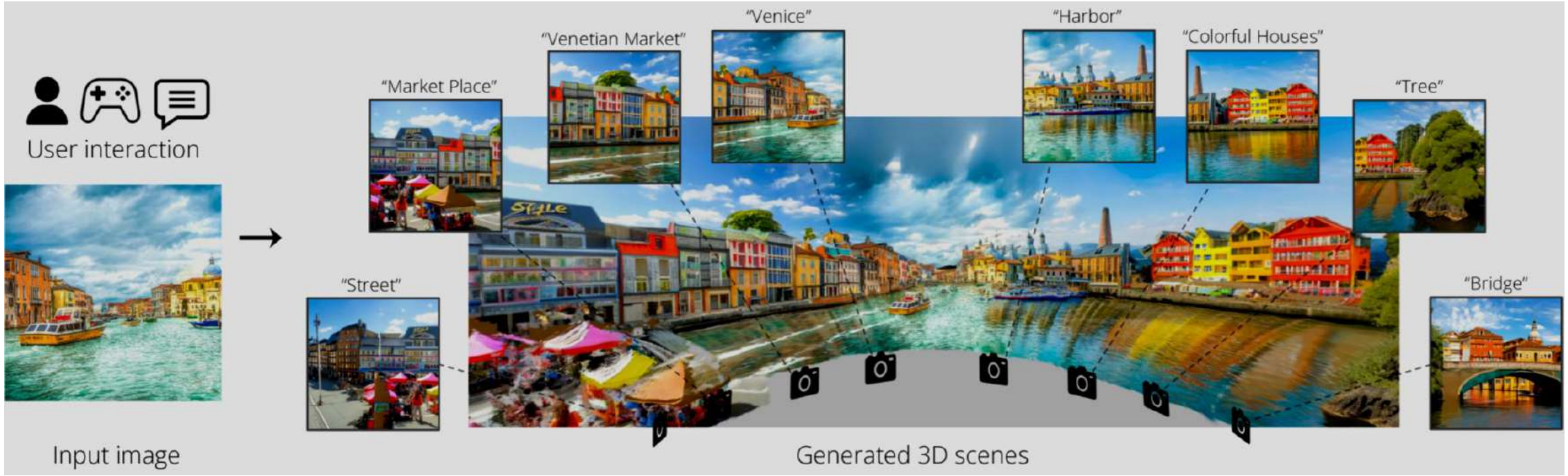
- Jiajun Wu先生の研究紹介 (2/3)
  - WonderJourney: Going from Anywhere to Everywhere … text or imageを入力として連続的で広大な3Dシーンを自動で生成



ECCV 2024 の動向・気付き(27/132)

Workshop: 1st Workshop on Scalable 3D Scene Generation and Geometric Scene Understanding

- Jiajun Wu先生の研究紹介 (3/3)
  - WonderWorld: A Framework for Interactive 3D Scene Generation … 画像からの3Dシーン生成について，単一画像のみを入力とする&高速&一貫性のあるシーン生成を提案



ECCV 2024 の動向・気付き(28/132)

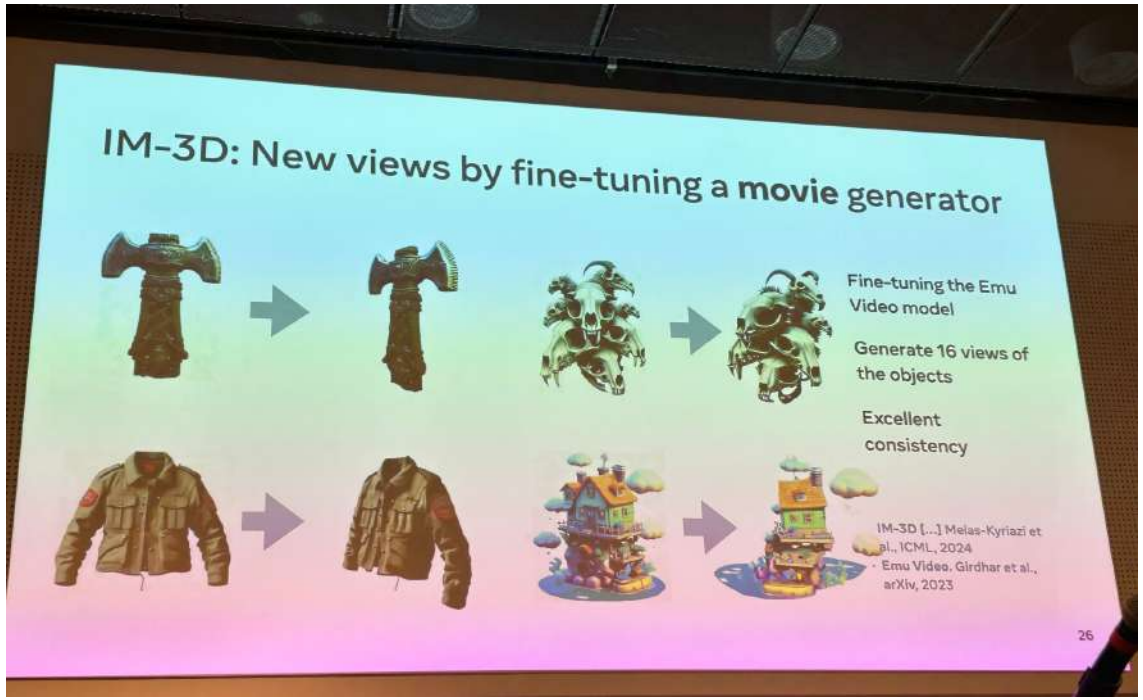
Workshop: 1st Workshop on Scalable 3D Scene Generation and Geometric Scene Understanding

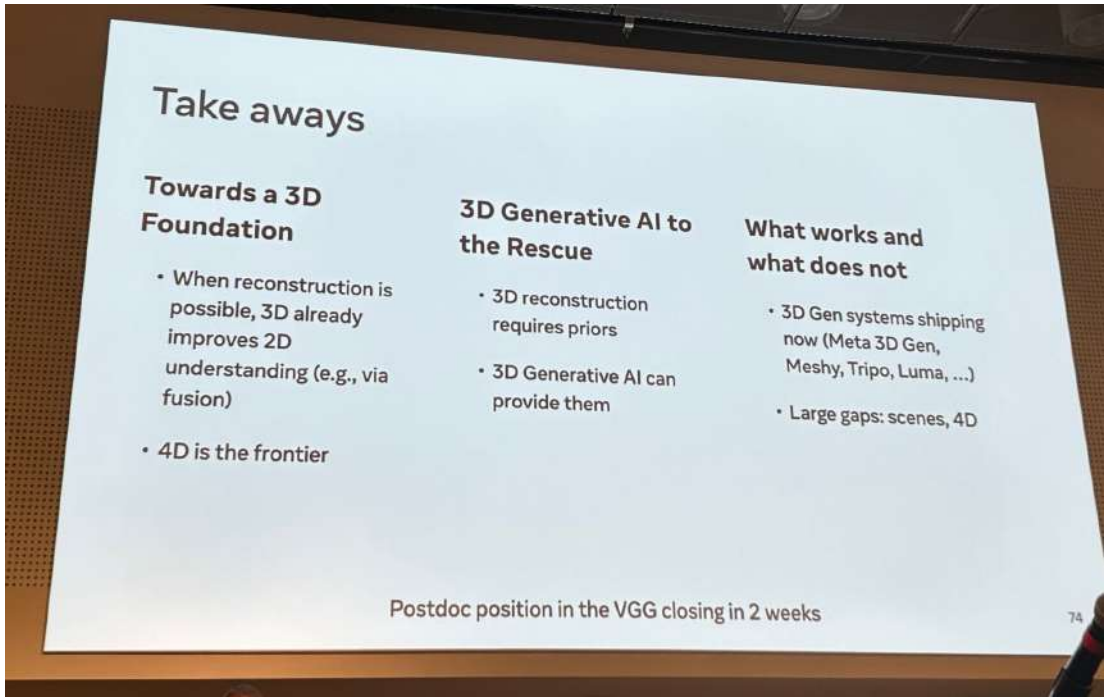
- Jiajun Wu先生の問いかけとキーワード
  - Learning vs. modeling
    - What are the minimal assumptions for 3D scene understanding?
    - How do these assumptions serve generation and perception?
  - Inverse rendering based on **casual, physical, and universal** object intrinsics
    - Effective use of in-the-wild data
    - Potential leverage of simulator/synthetic data
    - Interpretability and controllability
    - Generalization and compotisionality

ECCV 2024 の動向・気付き(29/132)

Workshop: Geometry in Large Model Era

- Title: Towards 3D Foundation with the Help of Generative AI (By Andrea Vedaldi)
  - Should 3D be one of the foundation of computer vision?
  - 3D生成モデルではmulti-view consistencyが非常に重要 -> Movieによる追加学習へ
  - 生成モデルのNew Frontierは3D Sceneと4D (空間+時系列)

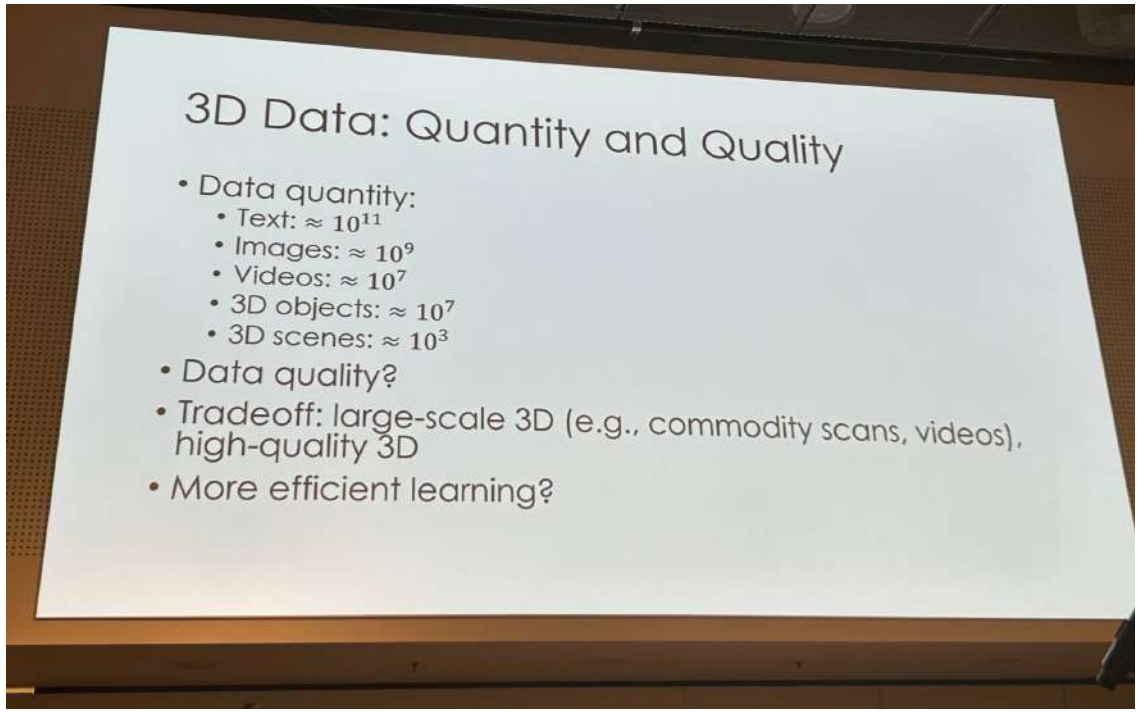


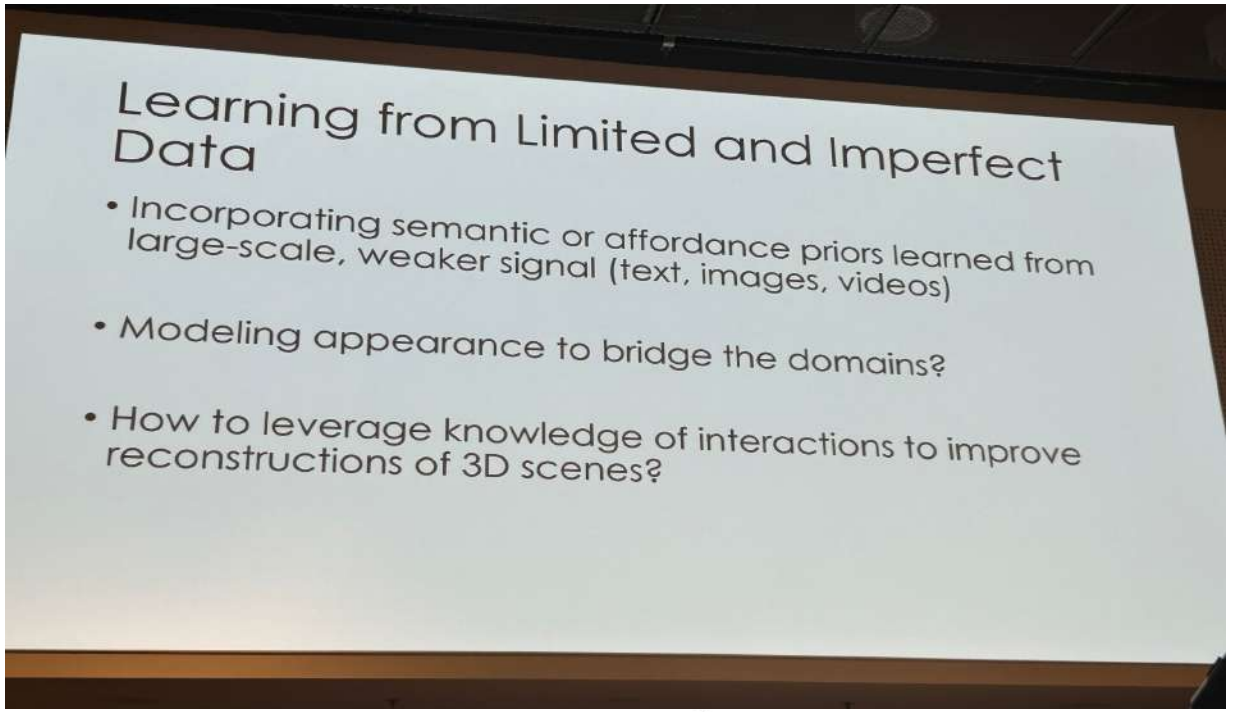


ECCV 2024 の動向・気付き(30/132)

Workshop: Geometry in Large Model Era

- Title: Generating 3D Geometry with Limited Data (By Angela Dai)
  - テキストや画像と比較して非常に収集が困難だ方こそ，限られたデータで効率的学習へ
  - Distilling information from larger models trains on weaker signal
  - ScanNet200を提案することで3D学習を新たなパラダイムへ
    - Angle Daiが3D scene understandingのトレンドを創出し続けている.

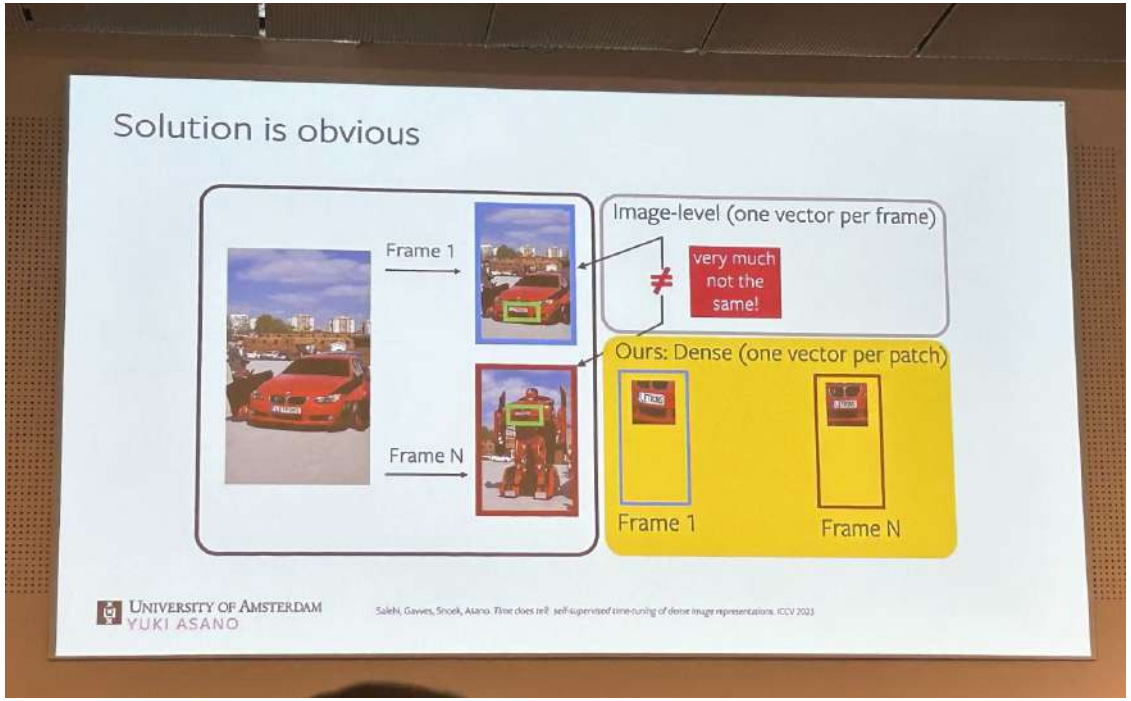


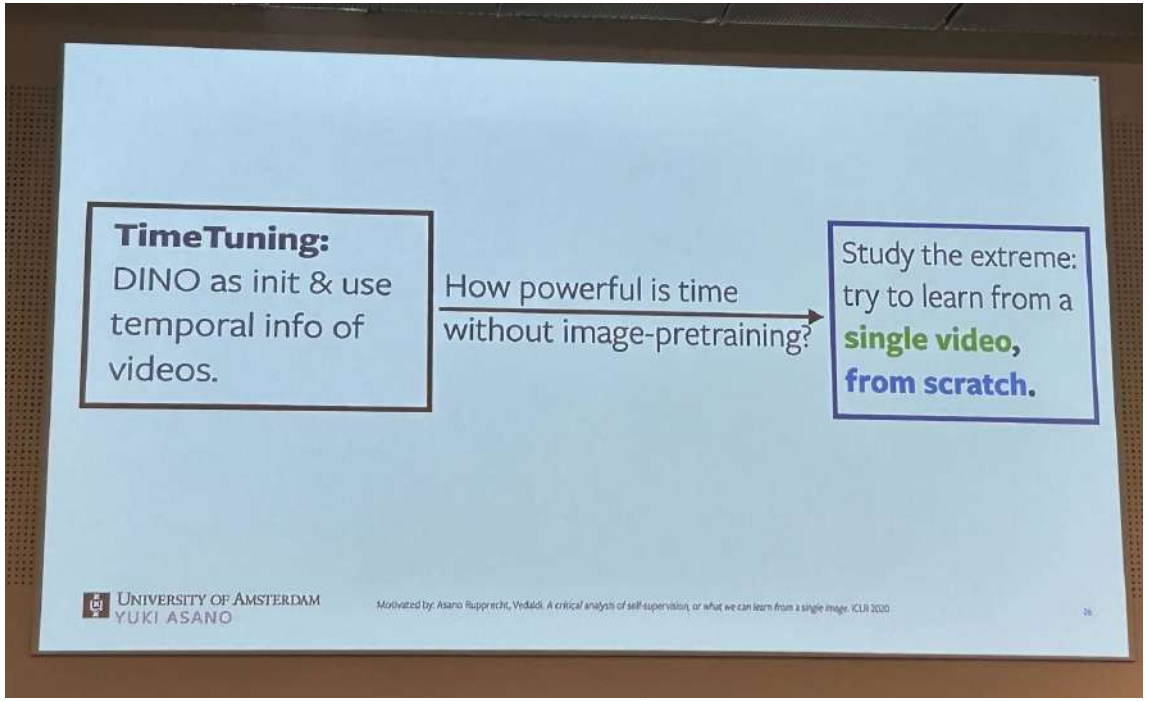


ECCV 2024 の動向・気付き(31/132)

Workshop: New Vision Foundation Models from Video(s)

- Title: 1-video pre-training, tracking image-pachs (by Yuki M. Asano)
  - トレンド: 自己教師あり学習のデータセットは画像から動画へ!?
  - 課題: 動画は時間とともに形状変化するため対照学習手法を改善する必要がある....
    - 研究1: 画像全体の対照学習ではなく, 画像パッチに注目した対照学習を考案
    - 研究2: Walking Tour(ベニス散策動画)を提案し, 完全に動画のみから自己教師あり学習

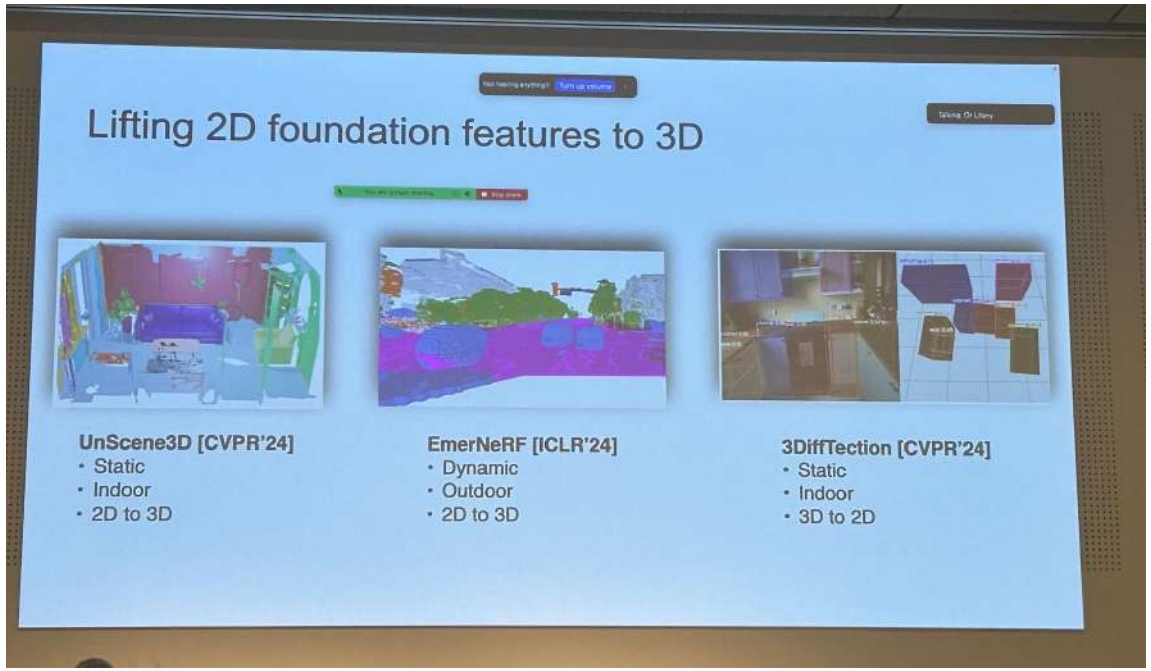


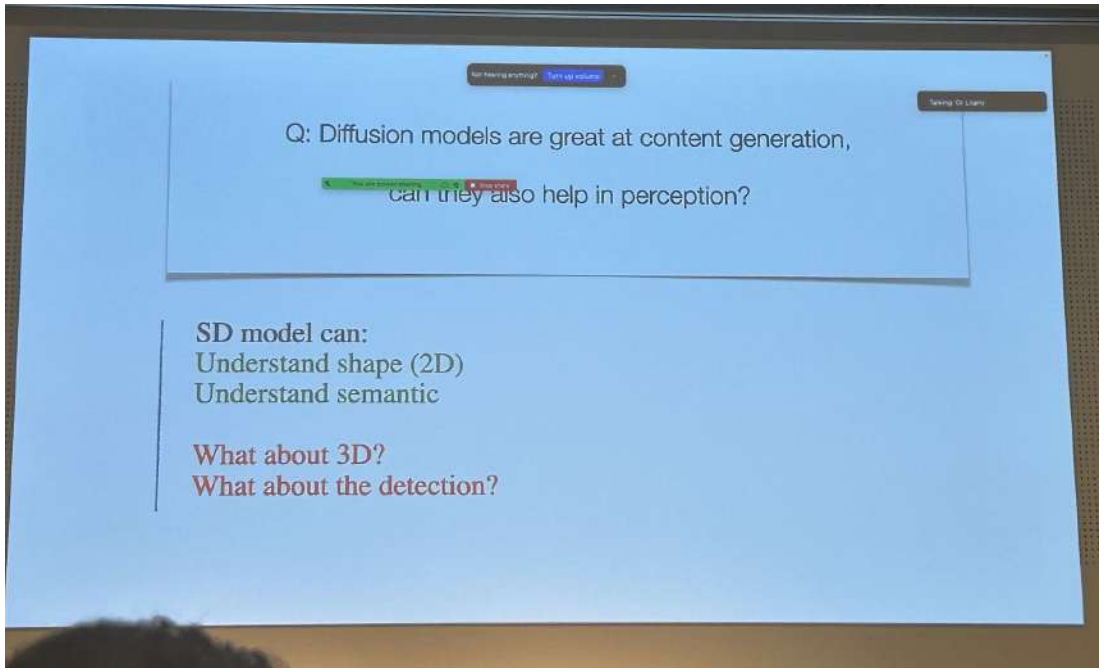


ECCV 2024 の動向・気付き(32/132)

Workshop in Open-Vocabulary 3D Scene Understanding

- Title: The Gentle Art of Feature “Lifting” (by Or Litany)
  - 3Dシーンを言語を介して理解するために，言語と3Dを繋げたいがデータ問題で難しい…
  - 2D特徴を活用することで3DシーンにてOpen-Word Segmentaion
    - どのような2D特徴を3Dに引き上げるべきか？
  - Stable Diffusionは知覚としては機能するか？3D幾何構造のPriorを獲得できているか？





ECCV 2024 の動向・気付き(33/132)

Workshop in Open-Vocabulary 3D Scene Understanding

- 植物フェノタイピングへのCVの応用に関するワークショップ
  - Full Paper 22件
    - セグメンテーション　:4件
    - 物体検出　:4件
    - データセットの構築系　:2件
    - XAI　:2件
    - 生成　:2件
    - 数カウント系　:1件
    - 3D　:1件
    - 蛍光　:1件
    - モニタリング　:1件
    - 自動アノテーション　:1件
    - マルチモーダル　:1件
    - 成長モデリング　:1件
    - メタ学習　:1件
  - Extended Abstracts 8件

植物フェノタイピングの未来図
(by ChatGPT)

植物系は、従来のタスク(セグメンテーション、検出)が目立つ。その一方で、XAIや生成に関する研究も進出し、その他は色々出てきている。今後もCVへの進出が期待される。

ECCV 2024 の動向・気付き(34/132)

Workshop: ACVR2024 – 12 th International Workshop on Assistive Computer Vision and Robotics.(1/2)

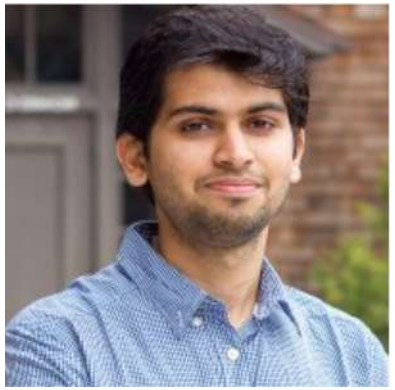
- コンピュータビジョンとロボティクスが現実の支援技術にどのような影響を与えるのか議論を行うワークショップ
- ロボティクスから支援技術、支援システムまで多岐にわたる分野から募集が来ている

https://iplab.dmi.unict.it/acvr2024/ より参照

ECCV 2024 の動向・気付き(35/132)

Workshop: ACVR2024 – 12 th International Workshop on Assistive Computer Vision and Robotics.(2/2)

- 招待講演では
  「Sensorimotor learning(感覚運動学習)に対するボトムアップアプローチ」
  「contextを考慮した視覚ナビゲーション」
  について扱われていた。
- ロボティクスx人間知覚のスケールが大きくなってきているように感じられた



Deepak Pathak, Carnegie Mellon University, US

Lamberto Ballan, University of Padova, IT

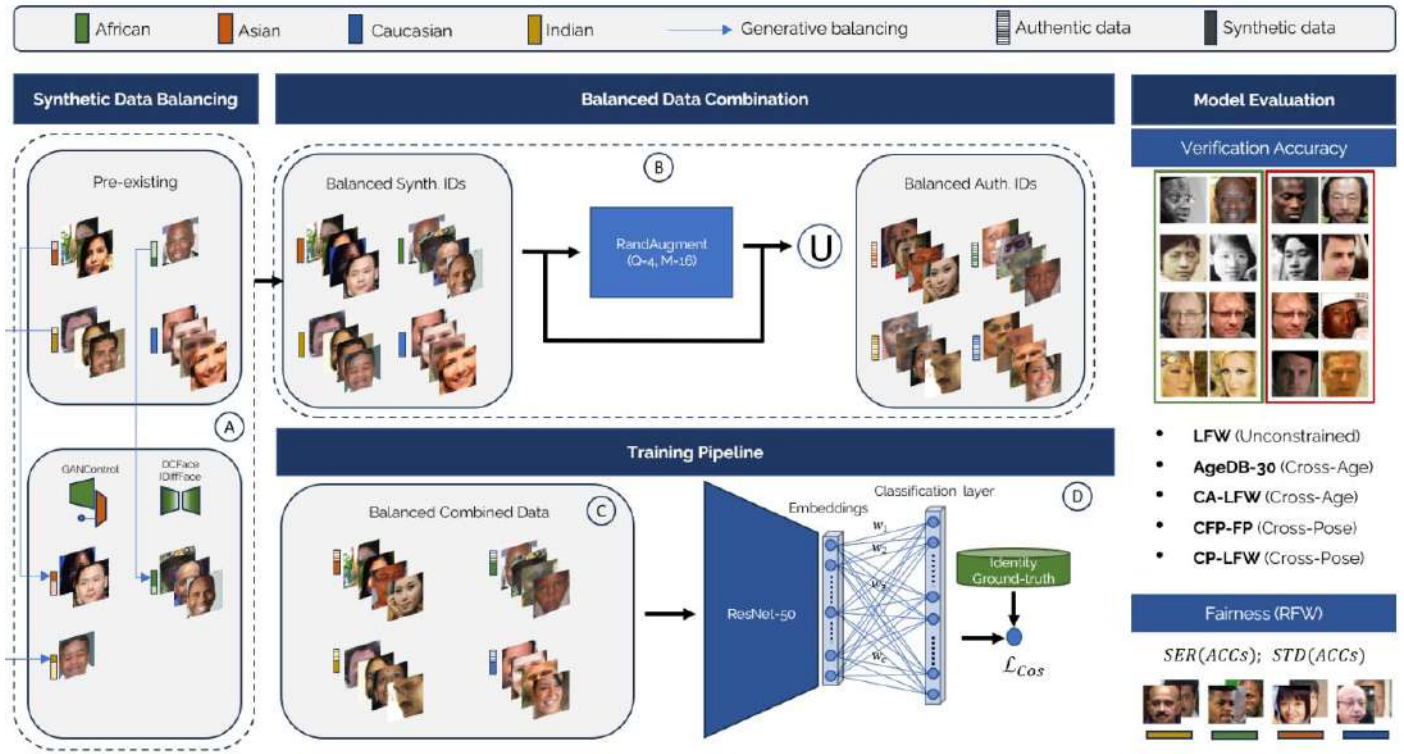
ECCV 2024 の動向・気付き(36/132)

Workshop: Synthetic Data for Computer Vision

- https://syntheticdata4cv.wordpress.com/acceptedpapers/
- 生成画像の学習効果について扱うWS.

- The Impact of Balancing Real and Synthetic Data on Accuracy and Fairness in Face Recognition
  https://arxiv.org/pdf/2409.02867

  - 結論：Face Recognitionタスクにおいて、学習データに実画像と生成画像が同じくらいの量含まれている方がGood
    - 画像引用：https://arxiv.org/pdf/2409.02867



ECCV 2024 の動向・気付き(37/132)

Workshop: 5th Advances in Image Manipulation (AIM2024) (1/3)

- 画像復元や操作などのlow-level visionタスクを扱うワークショップ
- 対象タスクのコンペも

Competition tasks

Efficient Video Super-Resolution
Depth Upsampling
Raw Burst Alignment
Sparse Neural Rendering - Track 1 - 3 views
Sparse Neural Rendering - Track 2 - 9 views
Video Saliency Prediction
Video Super-Resolution Quality Assessment
Compressed Video Quality Assessment
Pushing the Boundaries of Blind Photo Quality Assessment

https://cvlai.net/aim/2024/

ECCV 2024 の動向・気付き(38/132)

Workshop: 5th Advances in Image Manipulation (AIM2024) (2/3)

- 目的: 最新の技術動向や研究成果を共有して学術界と産業界を繋ぐ
- コンペでは結果の再現性を重視し、実世界のシナリオに焦点を当てている

Main Sponsors

Sony Interactive Entertainment

Many Thanks!!

ECCV 2024 の動向・気付き(39/132)

Workshop: 5th Advances in Image Manipulation (AIM2024) (3/3)

- 今後も同様のワークショップを開催予定

- 10th New Trends in Image Restoration and Enhancement (NTIRE)

  workshop and challenges, at CVPR 2025, https://www.cvlai.net/ntire/2024/

- 2nd Vision, Graphics and AI for Streaming (AIS)

  workshop and challenges, at CVPR 2025, https://ai4streaming-workshop.github.io/

- 6th Mobile AI (MAI)

  workshop, at CVPR 2025, https://ai-benchmark.com/workshops/mai/2024/

- 6th Advances in Image Manipulation (AIM)

  workshop and challenges, at ICCV 2025, https://www.cvlai.net/aim/2024/

ECCV 2024 の動向・気付き(40/132)

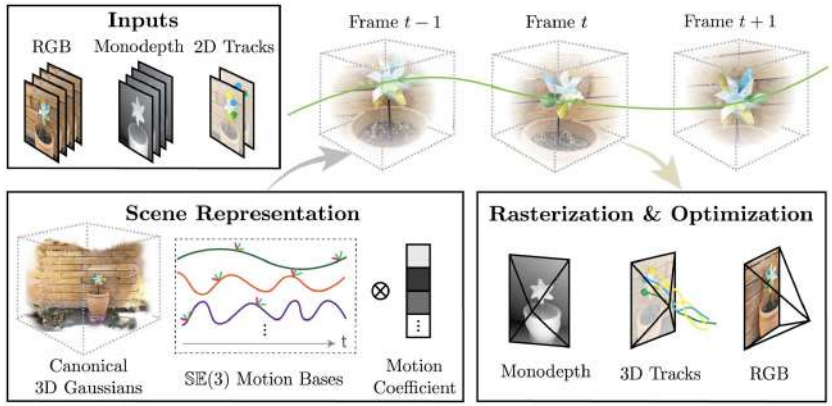
Tutorial: Recent Advances in Video Content Understanding and Generation

- Show-o: One Single Transformer to Unify Multimodal Understanding and Generation (Mike Z. Shou)
  - Comparison with previous studies on unifying image and text output using autoregressive model (AR):
    - Next-GPT, Seed-X: input LLM output to diffusion models (separate networks)
    - Large World Model, Chameleon: enable both image and text output in autoregressive model
    - Problem: AR is slow since it requires predicting token one by one -> using diffusion?
    - Diffusion model: continuous input/output vs. AR: discrete input/output
  - Show-o:
    - Design: discrete diffusion (inspired by masked generative transformers)
    - Show-o model: input output language and image, structure LLM (AR+diffusion)
    - Show-o characteristics:
      - Omni attention
      - Support both inpainting and outpainting, mixed modality generation
    - Training:
      - 1.3B model, training on various large-scale datasets
      - data scale, image resolution are important to obtain high accuracy
    - Results:
      - Recognition: comparative results with much larger models
      - Generation: better than much larger models
    - What is the Next?:
      - + video: structure is almost same with the one using images
      - Preliminary results: promising in video recognition and generation

ECCV 2024 の動向・気付き(41/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

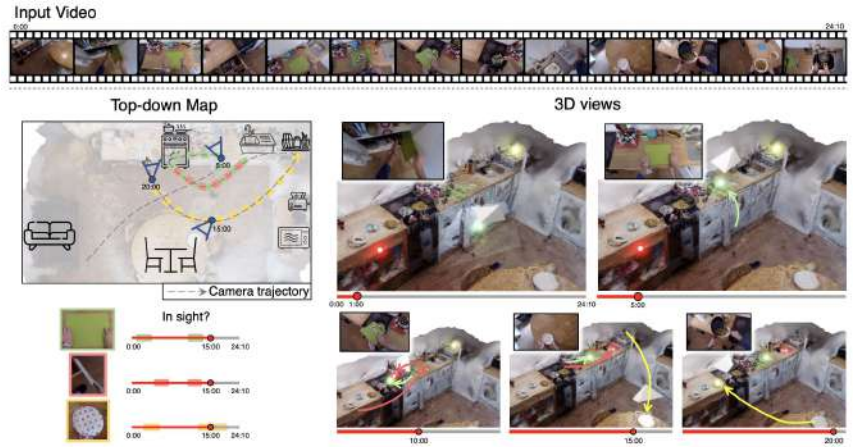
- Perceiving the World in 4D and thereafter(Angjoo Kanazawa)
  - Current 4D generation:
    - Recent video generation models learn depths, light transport, dynamics from watching videos effectively, but still not good at 3D reconstruction of body, body movement, etc.
    - Learning models with more data and enhanced models work
  - What we need:
    - 4D (space and time) physical understanding from interacting with world
  - Book recommendation: thinking fast and slow (Daniel kahneman)
- How to recover 4D from any video
  - Shape of Motion: 4D Reconstruction from a Single Video
    - Represent dynamic scene as a set of persistent 3D gaussians, represent motions across the video a compact set of shared SE3 motion bases; Utilize a set of data-driven priors, such as monocular depth maps and long-range 2D tracks.



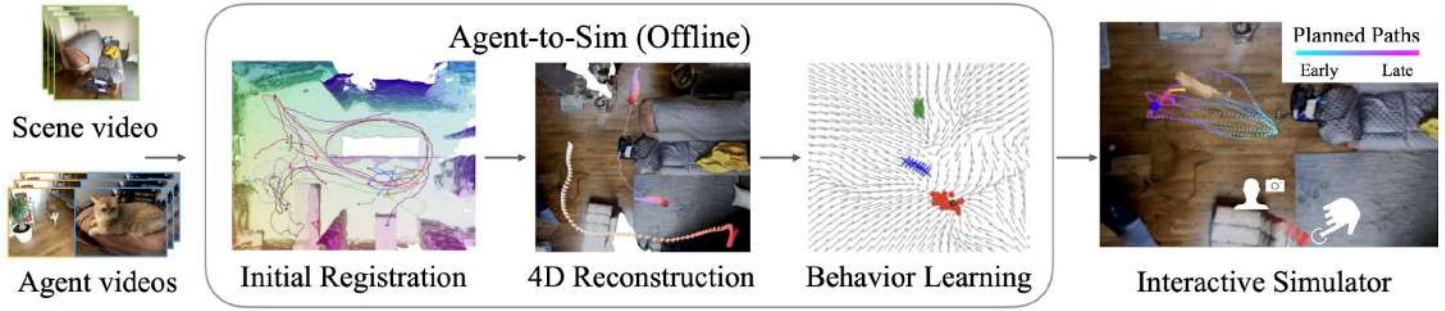
ECCV 2024 の動向・気付き(42/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

- What we can do using 4D(Angjoo Kanazawa)
  - Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind
    - Keep tracking objects while keeping in mind what objects are out of sight; lifts partial 2D observations to 3D world coordinates, matches them over time using visual appearance.



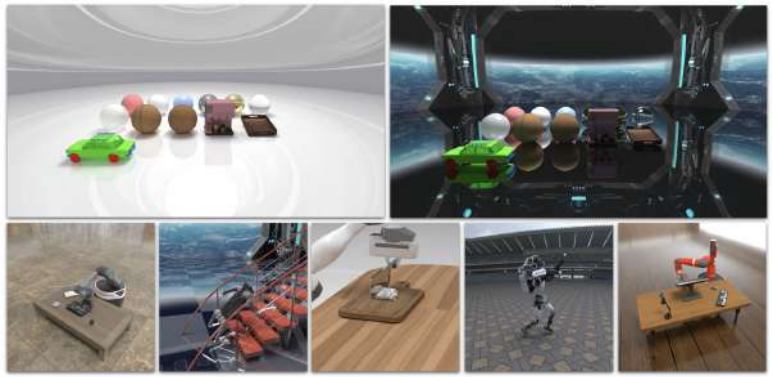
  - Agent-to-Sim: Learning Interactive Behavior Model from Casual Longitudinal Videos
    - A novel task and method showing the usefulness of 4D representation
    - persistent spacetime 4D reconstruction that contains the agent, the scene and the observer
    - Behavior generation (a diffusion model)

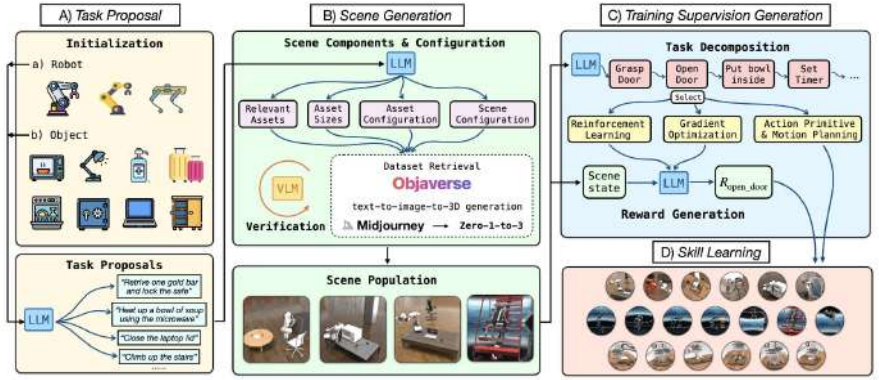


ECCV 2024 の動向・気付き(43/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

- Videos as World Models: Blending Visual and Physical Intelligence (Chuang Gan)
  - How about most advanced AI agents today compared with human babies?
    - Great advances in various real-world tasks
    - No embodied general intelligence (multimodal, multitask, multi-environment)
  - To achieve embodied general intelligence
    - Large video model as world models (physics, interactions are important)
  - Question1: How to generate large scale 3D interaction datasets?
    - Differentiable physics + renderer for learnable dataset generation
    - Gradient-based optimization is much efficient than RL
    - GeneSIS: general-purpose and differentiable physics simulator (vast range of robots) (left figure)
    - RoboGen: Towards Unleashing Infinite Data for Automated Robot Learning via Generative Simulation (right figure)
      - A generative agent that automatically proposes and learns diverse robotic skills at scale via generative simulation

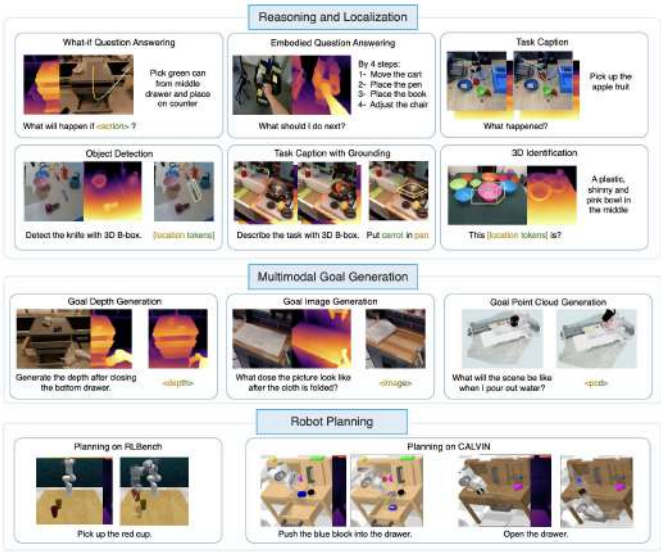




ECCV 2024 の動向・気付き(44/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

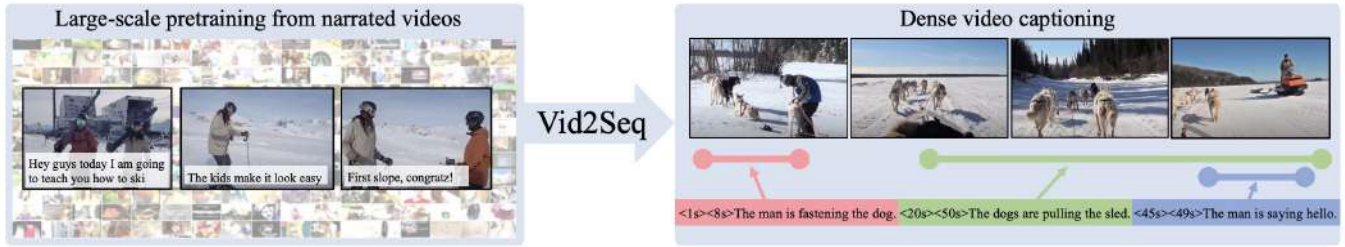
- Videos as World Models: Blending Visual and Physical Intelligence (Chuang Gan)
  - Question2: How to learn from 3D interaction data?
    - Inject 3D into LLM
      - Novel large-scale dataset by generation.
      - 3D features, combination of slam, nerf, etc.
    - How to really interact with real-world
      - Videos as world model for robot imagination
      - RoboDreamer:Learning Compositional World Models for Robot Imagination
        - learning a compositional world model by factorizing the video generation
      - 3D-VLA: A 3D Vision-Language-Action Generative World Model
        - train a series of embodied diffusion models and align them into the LLM for predicting the goal images and point clouds (right figure)
      - Compositional World Models for Embodied Multi-agent Cooperation
        - learn a compositional world model for embodied multi-agent cooperation
        - compositionally generate video to enable accurate simulation of multi-agent world dynamics



ECCV 2024 の動向・気付き(45/132)

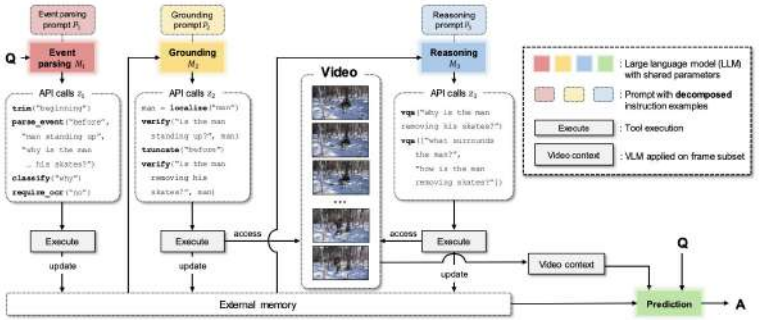
Tutorial: Recent Advances in Video Content Understanding and Generation

- Multimodal video representation (Cordelia Schmid)
  - Dense video captioning:
    - Current models can not reason over long videos
    - Vid2Seq:special time tokens; large-scale dataset pre-training; generative loss + denoising loss;



    - Dense Video Object Captioning from Disjoint Supervision
      - a new task for dense video object captioning – detecting, tracking and captioning trajectories of objects in a video.
      - a training strategy based on a mixture of disjoint tasks, which allows us to leverage diverse, large-scale datasets.

  - Video QA with reasoning:
    - Existing modular models often do not use image during visual program generation
    - MoReVQA: multistage, modular reasoning, simple baseline – JCEF outperforms ViperGPT,



ECCV 2024 の動向・気付き(46/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

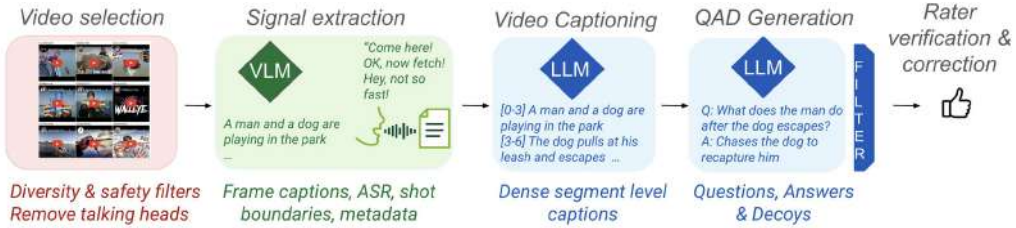
- Video Simulation and Holistic Understanding for Autonomous Driving: Systems and Backbones (Hao Zhao)
  - End-to-end autonomous driving
    - Issues: various corner cases
    - Solution: holistic video understanding
  - Holistic Video Understanding for Autonomous Driving
    - ADAPT: Action-aware Driving Caption Transformer: using language to explain various cases in AD
    - Hint-AD: holistically aligned interpretability for end-to-end autonomous driving : end-to-end AD model, alignment is the key
  - Video Simulation for AD
    - **Controllability and physics awareness are important**
    - Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability
    - SCA-WM: Structure-aware Collaboratively-aligned driving world model
    - SCP-Diff: Photo-Realistic Semantic Image Synthesis with Spatial-Categorical Joint Prior
    - MARS: an instance-aware, modular and realistic simulator for AD (compositional rendering)
    - Street Gaussians: Modeling Dynamic Urban Scenes with Gaussian Splatting (3DGS based)
    - Dynamic 3D Gaussian Fields for Urban Areas
  - What's next?
    - editable scene simulation via collaborative LLM-agents,
    - Drone-assisted road gaussian splatting with cross-view uncertainty
    - Pre-afford: universal affordance-based pre-grasping

ECCV 2024 の動向・気付き(47/132)

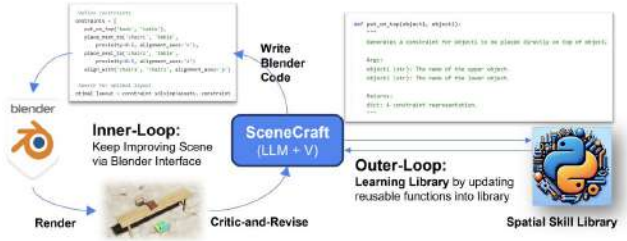
Tutorial: Recent Advances in Video Content Understanding and Generation

- Multimodal video representation (Cordelia Schmid)
  - Long video understanding benchmarks
    - VidChapters-7M: improve navigation in videos; extract chapter information from videos with transcriptions

    - Neptune: give models “human-level” capability to understand long videos; semi-automatic dataset generation;



  - 3D Scene Generation
    - Scene craft: an llm agent for generating blender code instead of animation

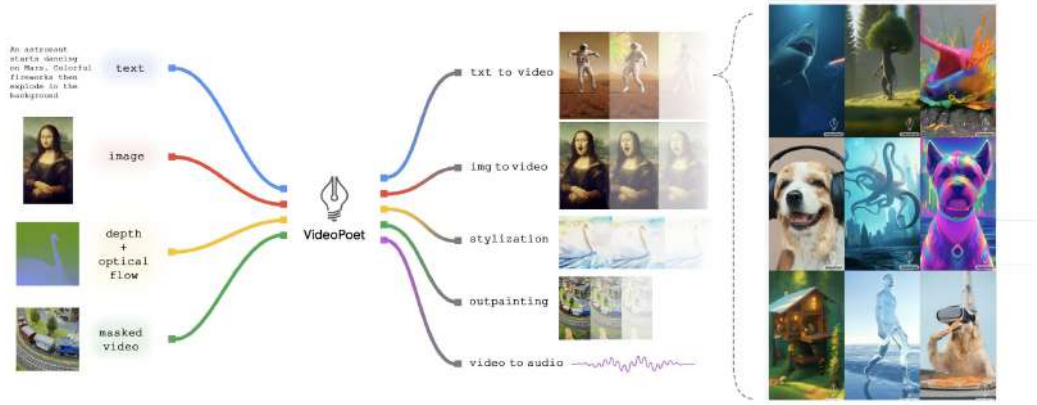


  - Conclusion:SOTA multimodal LLM can serve as strong reward signal for self-refinement

ECCV 2024 の動向・気付き(48/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

- LLMs Meet image and video generation (Yingqing He)
  - Image generation development: single domain -> open-domain -> allowing interactions
  - LLMs help the generation of images and videosin various aspects, such as as conditioner, planner, evaluator, captioner, agents
  - LLMs in image / video generation
    - How: encode different models all in discrete token space, generate the next token by using autoregressive generating multimodal tokens
    - VideoPoet: A Large Language Model for Zero-Shot Video Generation
      - a versatile video generator conditioning on multiple types of inputs and performs a variety of video generation tasks.

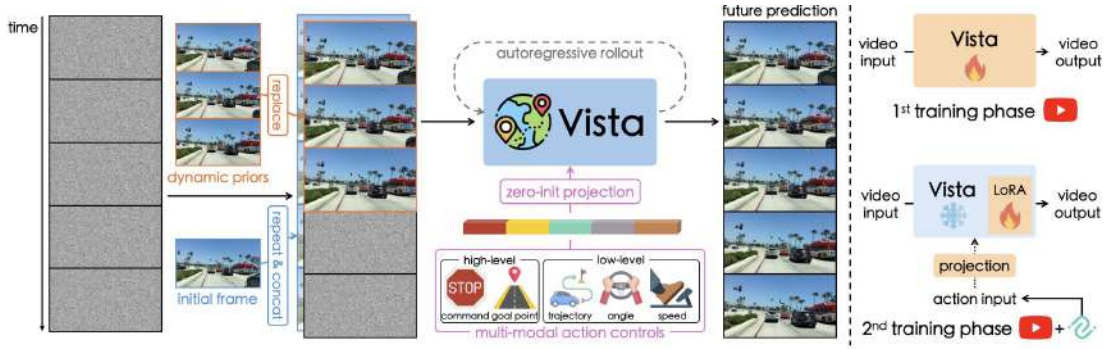


    - LLM as Planner: LayoutGPT
    - Videostudio: generating consistent-c ...Ms to generate lots of details for image generation forming the video;
    - LLM as captioner - sharegpt4v: improving image generation with better captions;

ECCV 2024 の動向・気付き(49/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

- Specializing Video Diffusion Models (Kashyap Chitta)
  - Video Latent Diffusion: where are we?
    - LDM recipe: step 1 – autoencoder with fixed-size latent; step 2 – latent denoiser;
    - Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models
      - Build on pre-trained image diffusion models, temporally video fine-tuning with temporal alignment layers
  - Building vista: can we specialize SVD for driving?
    - Vista: A Generalizable Driving World Model with High Fidelity and Versatile Controllability
      - Issues: dataset -> using online videos -> OpenDV-2k;
      - Adapting SVD for long rollouts: latent replacement, zero-init projections;

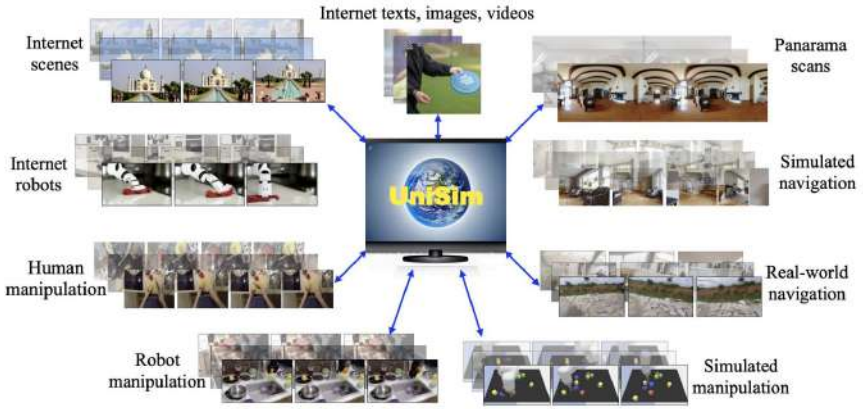


  - Practical Tips: what matters most during training?
    - Lots of resources and use them as smart as you can
    - EMA has a huge memory overhead but is essential
    - Offset noise improves temporal consistency
    - Domain-specific loss weights may be necessary
    - Iters/sec is the most important factor to scale

ECCV 2024 の動向・気付き(50/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

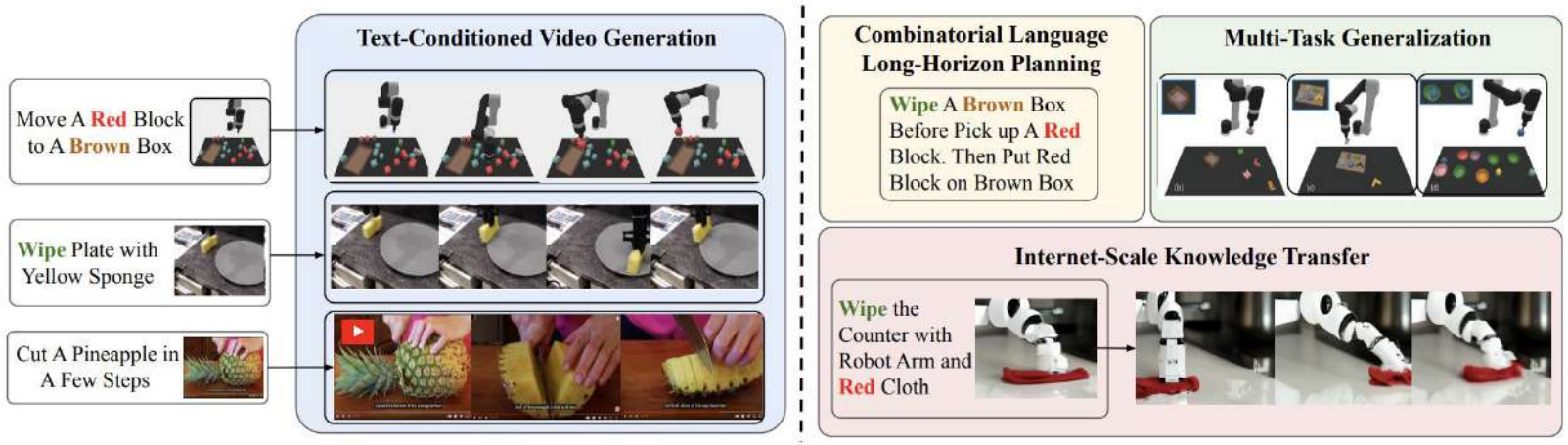
- Video Generation as Real-World Simulators (Sherry Yang)
  - Use internet-scale data to simulate the real world
  - World model from internet data, algorithms for decision making
    - Video as unified representation and task interface: Learn real-world physics, ego-centric movements, notions of objects/scenes
    - Adapting diffusion for world modeling: repeat the first frame for long-term consistency; condition on image & text for controllable generation; temporal super-resolution for flexible time horizon;
  - Video as the new language for real-world decision making
    - extend video generation to solve tasks in the real world
  - Learning interactive real-world simulators
    - explore the possibility of learning a universal simulator (UniSim) of real-world interaction through generative modeling
    - Planning with UniSim – using vision-language model as a reward model;
    - Why video generation for planning: video frames as intermediate goals, internet-scale data, temporal flexibility, search, plan



ECCV 2024 の動向・気付き(51/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

- Video Generation as Real-World Simulators (Sherry Yang)
  - Learning universal policies via text-guided video generation
    - Represent policies using text-conditioned video generation
    - Enable effective combinatorial generalization, multi-task learning, and real world transfer



  - Challenges and next steps
    - Challenge: need better world models
    - Steps: enhance world models by evaluation and feedback; good world models lead to successful robot execution? real-world feedbacks matter;

ECCV 2024 の動向・気付き(52/132)

Tutorial: Recent Advances in Video Content Understanding and Generation

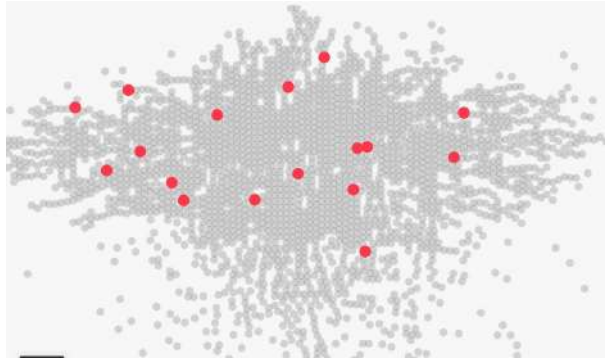
- Embodied Visual Perception in Unknown Environments (Ziwei Wang)
  - Indoor Embodied home assistant
    - HomeRobot: Open-Vocabulary mobile manipulation
    - Existing works: high-level task planners & low-level navigator
    - Challenges:
      - various objects and their arrangements
      - Offline reconstructions and online perception with poor quality
    - Solution:
      - Memory-based adapters for online 3D perception
      - Online scene map
      - Active exploration of unknown regions to find target objects
      - General memory-based adapters. Plug and play without any model and task-specific design
      - General online 3D perception framework – shift-based memory with 2D convolution and 3d-to-2d adapter
  - Reasoning for perception of unknown perception
    - Embodied instruction following in unknown environments
    - Embodied task planning with large language models
    - High-level planning and low-level control and navigation use LLMs
    - Use region attention and update scene map
    - SG-Nav: Online 3D Scene Graph Prompting for LLM-based Zero-shot Object Navigation
      - Zero-shot object goal navigation
      - Effectively using LLMs is critical

ECCV 2024 の動向・気付き(53/132)

医療AIの動向

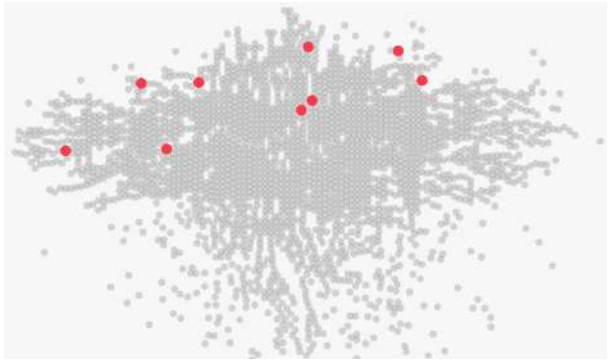
- 画像に対するレポート作成・病理画像解析が多い
- 内視鏡や心カテーテルなどの臨床的な手技の自動化を見据えた技術も

“medical” (医療)での検索結果 — 17件

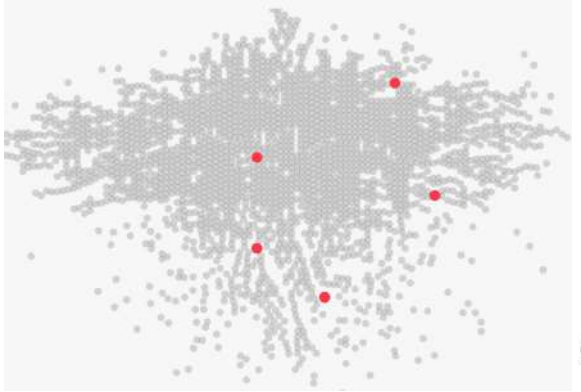


“endoscopy / endoscopic” (内視鏡) での検索結果 — 3件

“pathology” (病理)での検索結果 — 9件



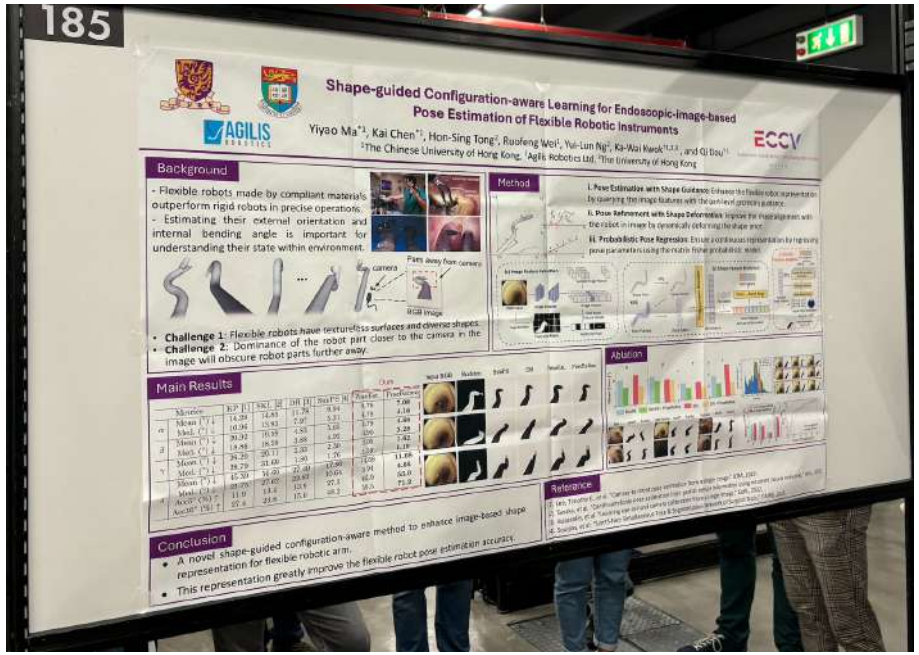
“X-ray” (X線画像) での検索結果 — 3件

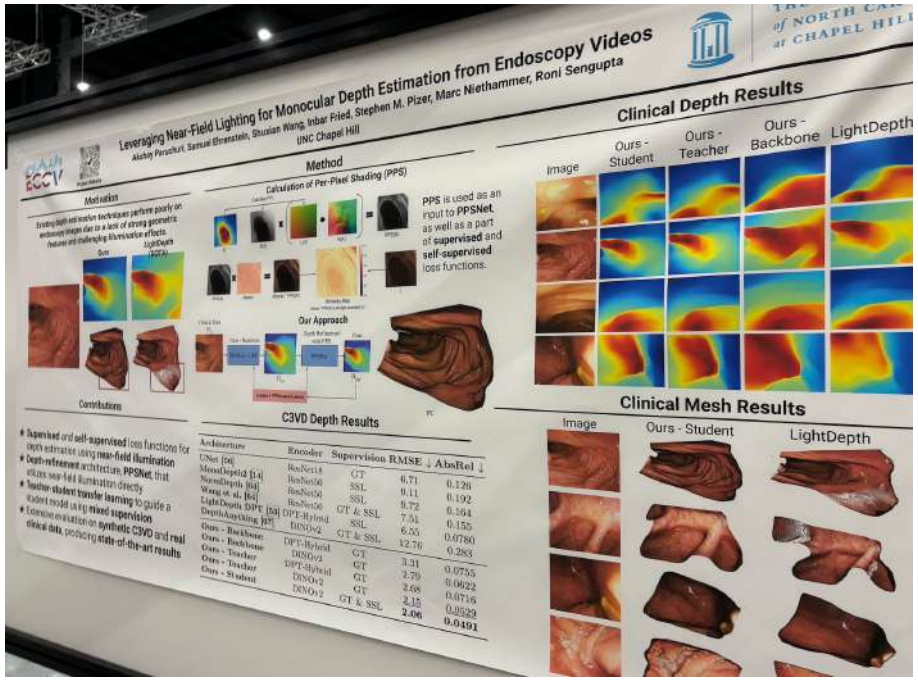


ECCV 2024 の動向・気付き（54/132）

医療AIの動向

- 内視鏡は、胃や大腸などの消化管に細長い管を挿しこんで内部を観察する手技
- カメラのポーズ推定や、消化管の深度推定などの研究が発表
- 疾患検出だけでなく、手技の自動化にも画像認識は欠かせない

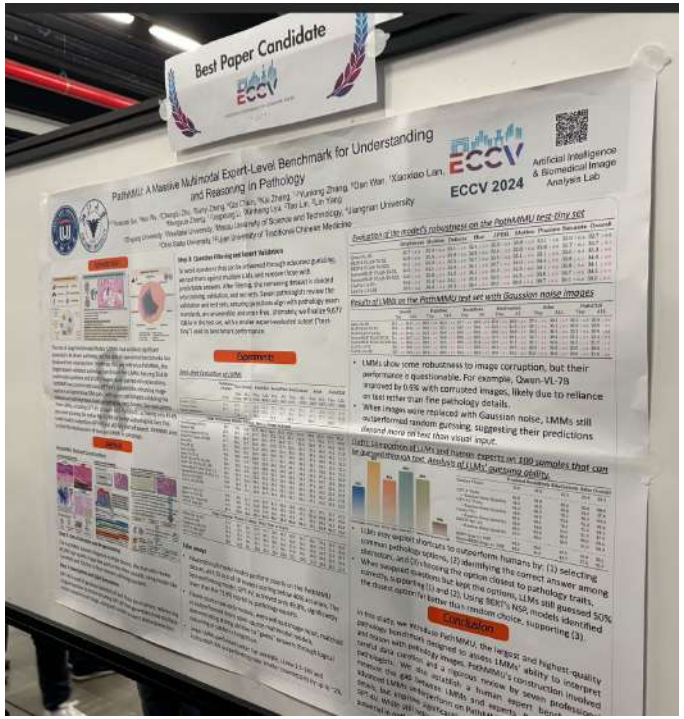




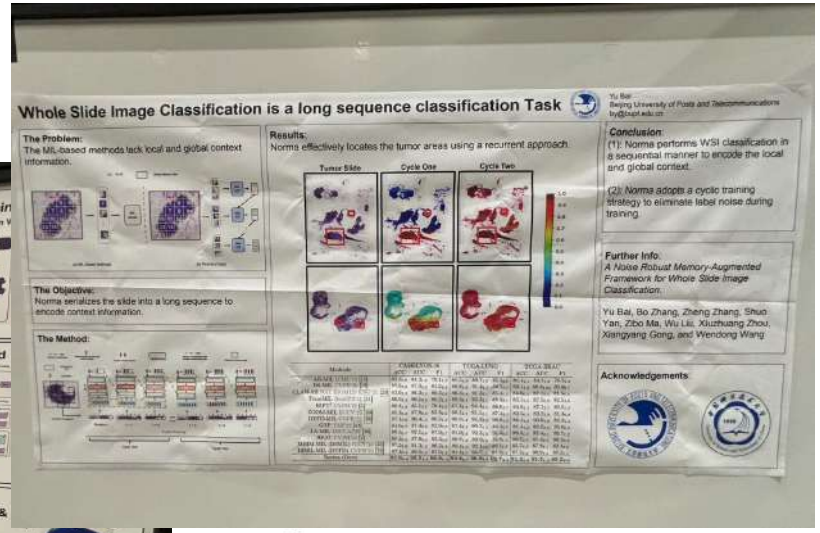
ECCV 2024 の動向・気付き（55/132）

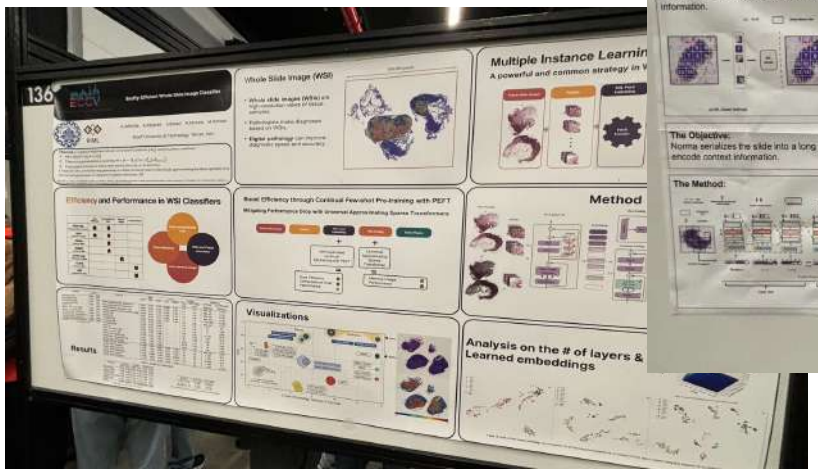
医療AIの動向

- 医療分野の中でも病理画像解析は最も盛んに発表されている印象
  - 病理画像とは臓器などの切片をプレパラートにして細胞などを観察するためのもの（がんなどの確定診断や進行度分類などに必要なことが多い）
- ”数万pixel×数万pixelの超高解像度画像をいかに解析するか”や”細胞情報をいかに解析に取り込むか”, “染色方法の違い（要はドメインの違い）にいかに対処するか”など様々な課題がある



← 病理画像のためのVision-Language大規模ベンチマークPathMMUはBest Paper Candidatesに



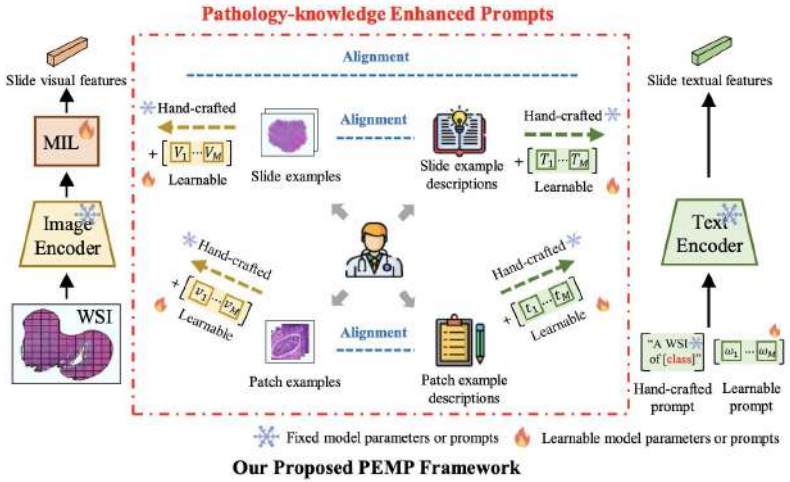


ECCV 2024 の動向・気付き（56/132）

病理医の専門知識により注目し，text に専門知識を反映する動向

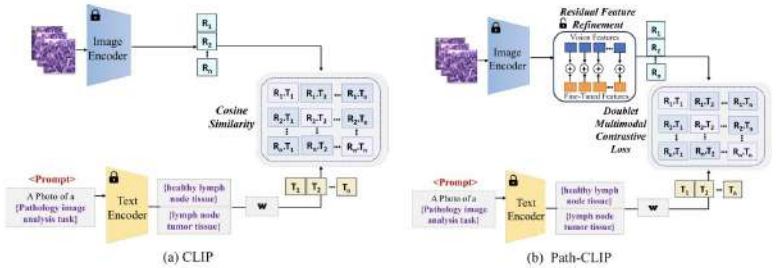
Pathology-knowledge Enhanced Multi-instance Prompt Learning for Few-shot Whole Slide Image Classification

病理学の視覚的，言語的知識をプロンプトに
組み込み，少数データでclassification



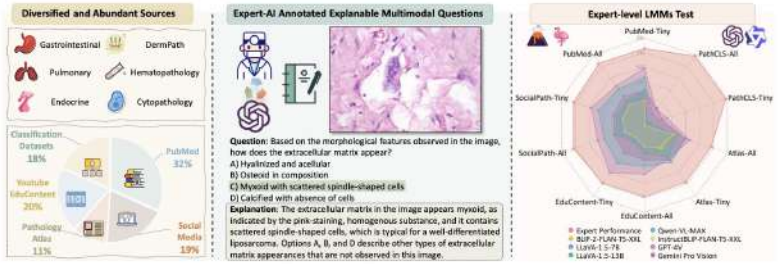
Bridging the Pathology Domain Gap: Efficiently Adapting CLIP for Pathology Image Analysis with Limited Labeled Data

CLIPの表現力を保ちつつ病理domain へ転移
Hidden representation perturbationを持つ
Residual feature refinement を提案し
病理画像認識への転移性能を効率的に！



PathMMU: A Massive Multimodal Expert-Level Benchmark for Understanding and Reasoning in Pathology

33,428の選択問題と24,067の画像のデータ
GPT-4V のcaption を病理医で精査し，
複雑な画像解析へ特化



ECCV 2024 の動向・気付き（57/132）

HARVAED の Mahmood Lab AI for Pathology に注目

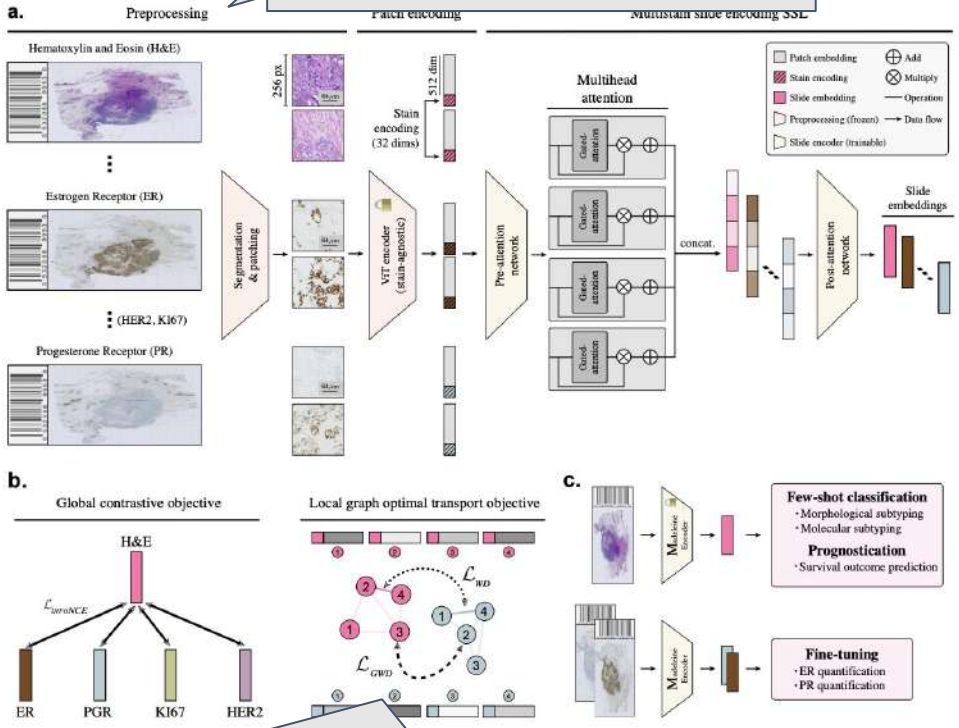
- Multistain を用いたrepresentation learning を発表
  - CVPR 2024 では transcriptomics を活用

FAISAL MAHMOOD

CONCHなどの病理画像向け
VLLM を次々に排出

[Multistain Pretraining for Slide Representation Learning in Pathology]

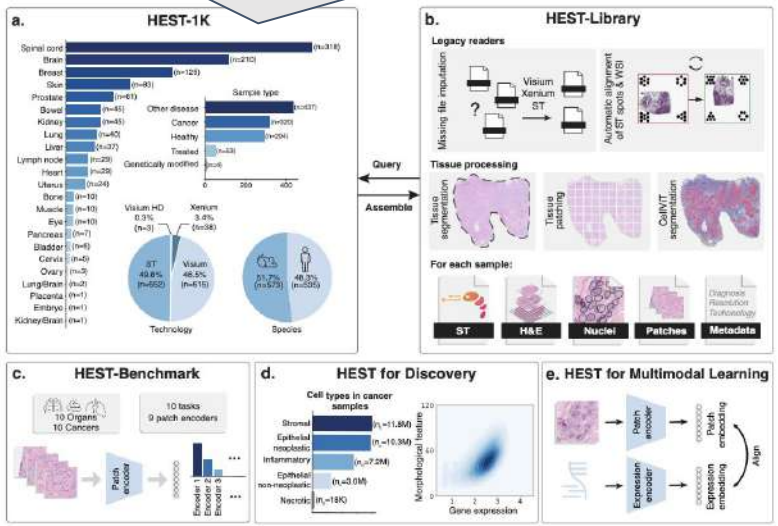
複数の染色方法で得られた画像の活用



Optimal transport によりmodal 間を
繋ぎ有効活用！

[HEST-1k: A Dataset for Spatial Transcriptomics and Histology Image Analysis]

ECCVではないがspatial transcriptomicsの
データも公開



ECCV 2024 の動向・気付き（58/132）

SLAM関係の論文

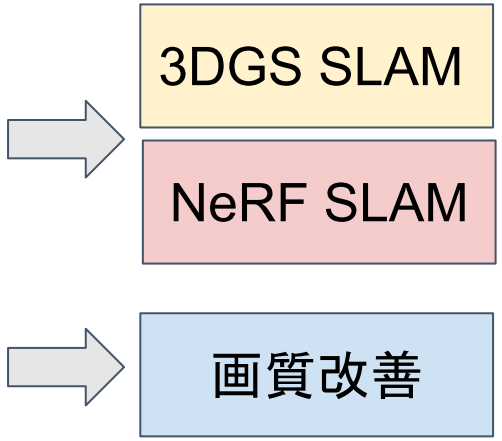
ID	title	name	領域	コメント	pdf
650	CG-SLAM: Efficient Dense RGB-D SLAM in a Consistent Uncertainty-aware 3D Gaussian Field	Jiarui Hu, Xianhao Chen, Boyin Feng, Guanglin Li, Liangjing Yang, Hujun Bao, Guofeng Zhang, Zhaopeng Cui*	3DGS SLAM	3DGSベース。トラッキングが15Hzと非常に早い	[pdf]
816	SGS-SLAM: Semantic Gaussian Splatting For Neural Dense SLAM	Mingrui Li, Shuhong Liu, Heng Zhou, Guohao Zhu, Na Cheng, Tianchen Deng, Hongyu Wang*	3DGS SLAM	3DGSを用いたセマンティックSLAM	[pdf]
952	RGBD GS-ICP SLAM	Seongbo Ha, Jiung Yeon, Hyeonwoo Yu*	3DGS SLAM	3DGSの利用 とても速い 107FPS	[pdf]
2049	Learn to Memorize and to Forget: A Continual Learning Perspective of Dynamic SLAM	Baicheng Li*, Zike Yan*, Dong Wu, Hanqing Jiang, Hongbin Zha*	NeRF SLAM	NeRF SLAM 動的シナリオ	[pdf]
2140	LRSLAM: Low-rank Representation of Signed Distance Fields in Dense Visual SLAM System	Hongbeen Park, Minjeong Park, Giljoo Nam, Jinkyu Kim*	NeRF SLAM	NeRF SLAM 低ランクテンソル分解 による計算高速化	[pdf]
703	I2-SLAM: Inverting Imaging Process for Robust Photorealistic Dense SLAM	Gwangtak Bae, Changwoon Choi, Hyeongjun Heo, Sang Min Kim, Young Min Kim*	SLAMによる画質改善	動画のBlurやホワイトバランスなどの画質調整を、 SLAMを逆に使って解決	[pdf]
2247	Self-Supervised Underwater Caustics Removal and Descattering via Deep Monocular SLAM	Jonathan Sauder*, Devis Tuia	SLAMによる画質改善	海中動画 SLAMを用いた水の模様などの除去	[pdf]
792	"Hyperion – A fast, versatile symbolic Gaussian Belief Propagation framework for Continuous-Time SLAM"	David Hug*, Ignacio Alzugaray, Margarita Chli	技術要素改善・他	連続的運動パラメータ化によるマルチセンサー統合、非同期・分散計算	[pdf]
51	Deep Patch Visual SLAM	Lahav Lipson*, Zachary Teed, Jia Deng	技術要素改善・他	高速で高精度なループクロージャー	[pdf]

ECCV 2024 の動向・気付き(59/132)

SLAMの種類

カメラ位置情報

Only Visual SLAM	計算	• 画像のみから計算 • 直接深度画像を作る単眼デプス推定 • NeRFや3DGSなどの2D⇒3D技術 • 画像のみをインプットにするため、画質修正にSLAMを用いる応用研究も増えている
Visual Inertial SLAM	IMU	• 画像+IMU（センサー）情報で計算
RGB-D SLAM	D	• 深度情報（RGB-D)から点群に変換するなどしてカメラ相互関係を計算



ECCV 2024 の動向・気付き（60/132）

SLAM技術： NeRFから3DGSへ

NeRF SLAM → **3DGS SLAM**

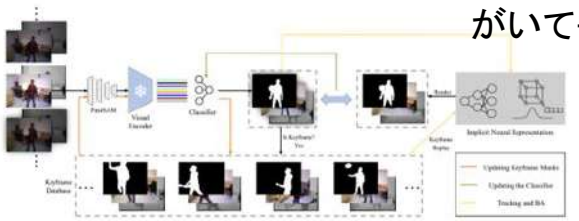
＊ 3D gaussian splatting は2023年8月に提案されたばかりの若い技術。今後の成長が期待される

○ コンパクトで連続的な表現
× 陰関数であるため明示的表現ではない
× 計算時間が長い・必要メモリ多い

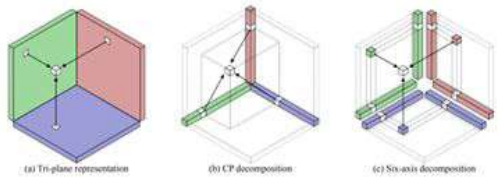
○ 計算速度が速い！
○ NeRFと違って明示的表現
△ 技術発展途上

Learn to Memorize and to Forget: A Continual Learning Perspective of Dynamic SLAM

動的環境（動いている人がいてもSLAM）

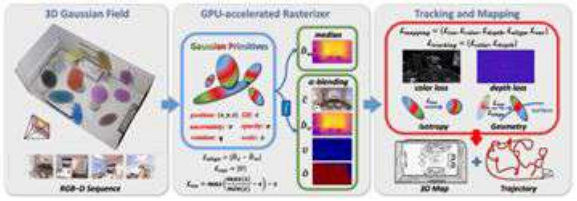


LRSLAM: Low-rank Representation of Signed Distance Fields in Dense Visual SLAM System



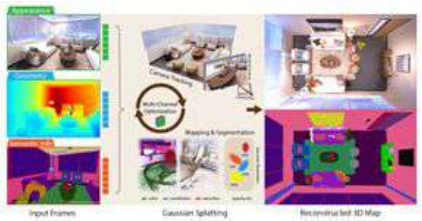
低ランク表現を用いた必要メモリ削減

CG-SLAM: Efficient Dense RGB-D SLAM in a Consistent Uncertainty-aware 3D Gaussian Field



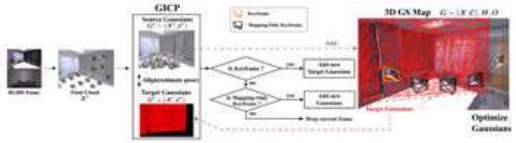
トラッキング速度が15Hz！はやい！

SGS-SLAM: Semantic Gaussian Splatting For Neural Dense SLAM



セマンティックSLAM

RGBD GS-ICP SLAM



最大で107FPSと非常にはやい

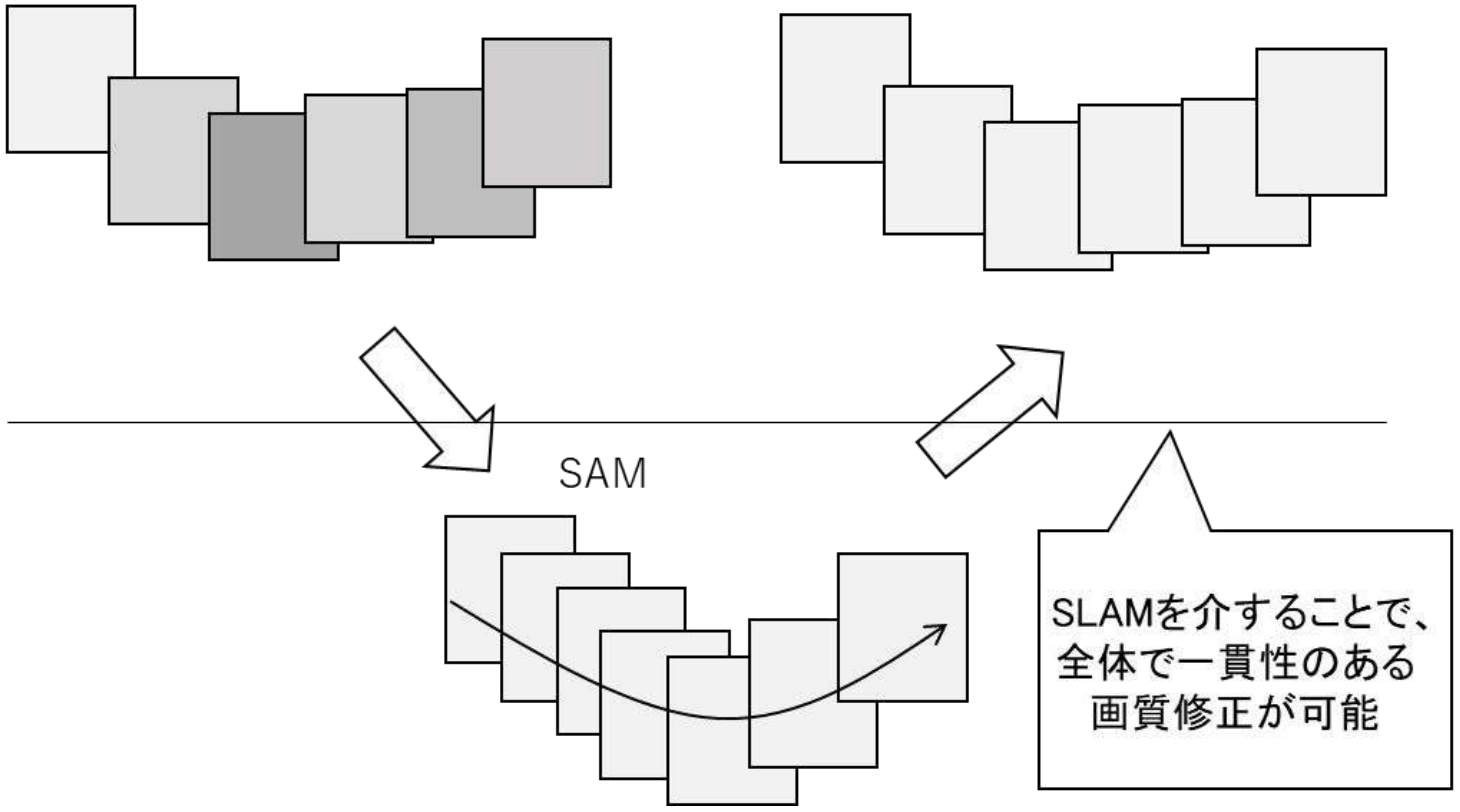
ECCV 2024 の動向・気付き（61/132）

SLAMを用いた画質改善（SLAMの他分野への応用）

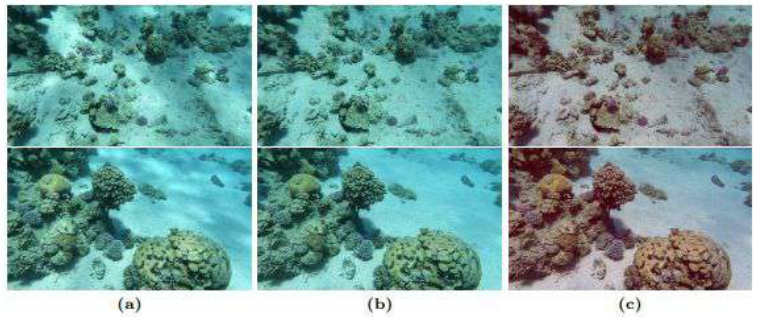
画質改善

元画像

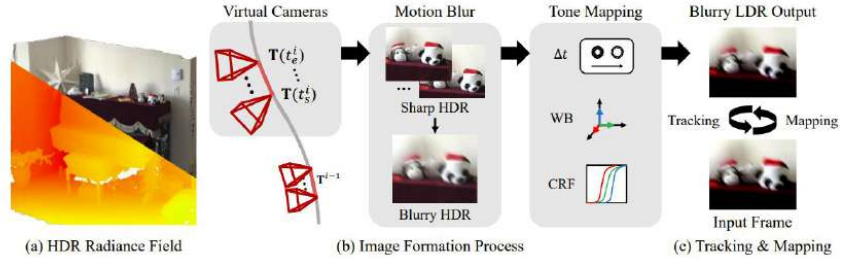
画像補正



Self-Supervised Underwater Caustics Removal and Descattering via Deep Monocular SLAM



I 2 -SLAM: Inverting Imaging Process for Robust Photorealistic Dense SLAM

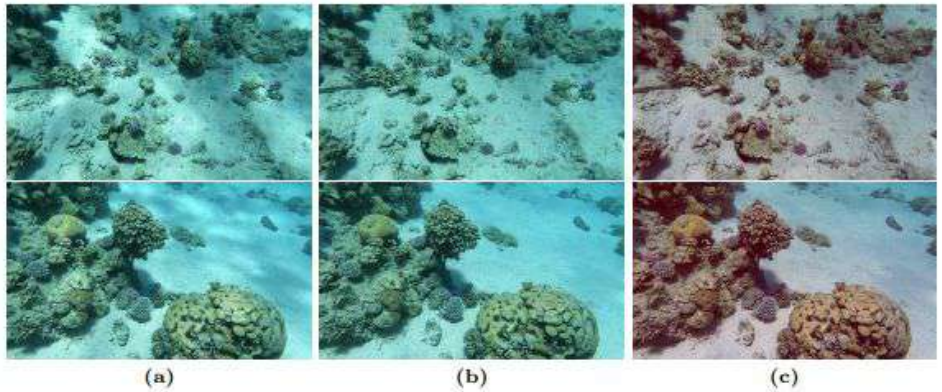


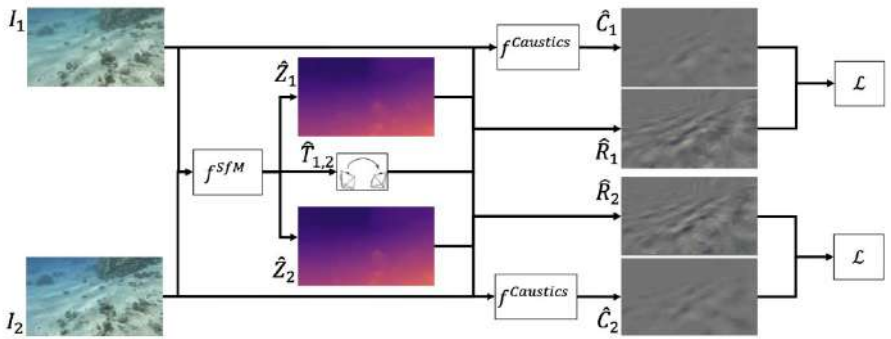
ECCV 2024 の動向・気付き（62/132）

SLAMを用いた画質改善（SLAMの他分野への応用）

Self-Supervised Underwater Caustics Removal and Descattering via Deep Monocular SLAM

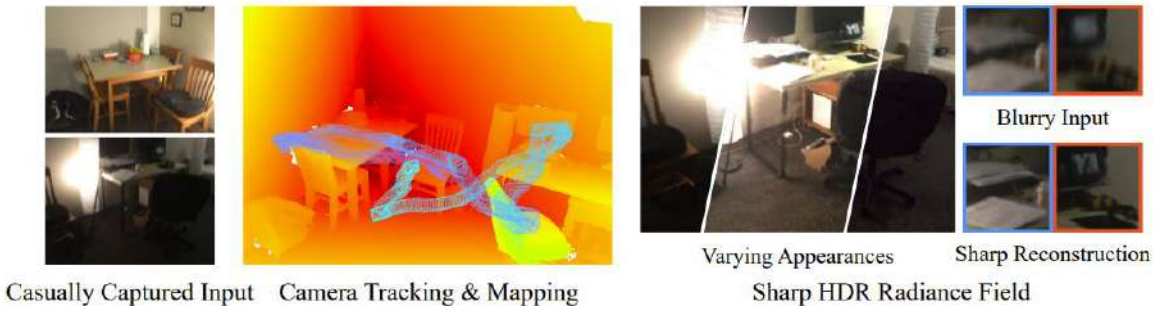
•水中画像特有の光の問題を、動画の SLAMをつくることで統一的に解決（a⇒c: 水の影が消えている）

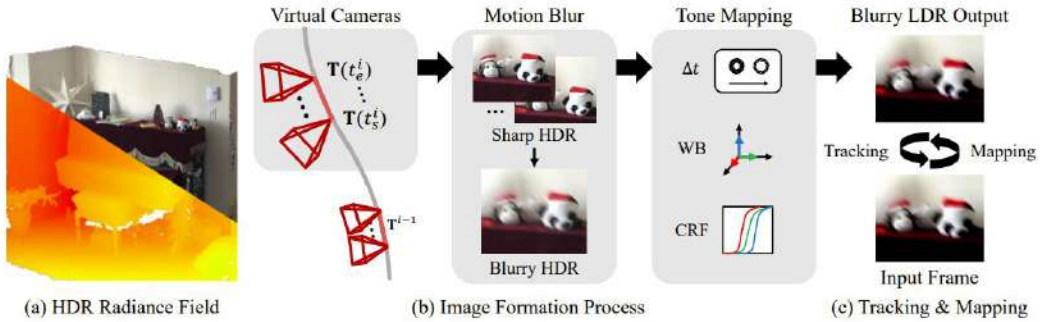




I 2 -SLAM: Inverting Imaging Process for Robust Photorealistic Dense SLAM

•動画のBlurやホワイトバランスなどの画質問題を、SLAMを逆に使って解決。





プロジェクトページ　動画がわかりやすい
https://changwoonchoi.github.io/i2slam/

SLAM関係の論文　気づきメモ

- ❑ タイトルにSLAMという単語があった論文9本を調査

- ❑ 3DGSを用いたSLAMが増加
  - ⇒ 計算の高速化がメリット　10ms以下の計算速度を誇る論文が多い
  - ⇒ NeRFと違って明示的である点もメリット
  - ⇒ 今後はNeRFベースのSLAMよりも3DGSベースのSLAMが増えるかもしれない
  - ⇒ 3DGS は2023年8月に出現したことを考えると驚異的なスピード

- ❑ SLAMを地図ではなく、画質補正に使う論文の増加
  - ⇒ 動画ごとにSLAMを作成し、バンドル調整するように各画像を補正する手法
  - ⇒ SLAMの応用として興味深かった

ECCV 2024 の動向・気付き(64/132)

超解像 (Super-Resolution) 関連の動向

- “Super-Resolution”に関連する投稿は30件
  - 内容
    - Real-World(任意スケール) 単一画像超解像: 6
    - 参照ベース単一画像超解像 : 2
    - ビデオ超解像 : 5
    - マルチビュー (3D) 超解像 : 3
    - 軽量化・高速化 : 7
    - テキスト超解像 : 1
    - センシング系超解像 : 3
    - データセット再考・ベンチマーク提案 : 3
  - ベースモデル(複数使用の場合両方カウント)
    - GAN : 3
    - Flow : 2
    - Diffusion : 11
    - Transformer : 8
    - CNN : 9

ECCV 2024 の動向・気付き（65/132）

超解像 (Super-Resolution) 関連の動向

- Bicubic縮小以外の劣化を考慮した研究がほとんどに
- Diffusionで高品質な超解像が可能になったが、特殊なタスクや軽量化の研究には未だにCNN等が使用されている
- 特定タスクに特化したモデル構造の改善が行われる傾向
- Diffusion系では一般的なモデル構造の改善の研究がしばらく続きそう

- 気づき・感想
  - 高解像度化を目的ではなく手段として捉え、実用化に向けた軽量化・高速化が注目されているのではないか
  - 論文採択のためにはPSNRの改善以上の付加価値が必要
  - 最近のreal-world超解像では、劣化の種類を特定する以外の手法が出てきている印象

ECCV 2024 の動向・気付き（66/132）

Sometimes Less is More: The First Dataset Distillation Challenge

- https://dd-challenge-main.vercel.app/#/
- 過去の採択論文までまとめられているGitHub：
  GitHub - Guang000/Awesome-Dataset-Distillation: A curated list of awesome papers on dataset distillation and related applications.
- ”Dataset distillation” がメイントピック
  - 超解像（Super Resolution）の文脈でもDataset distilationを
    している論文あり。（AAAI採択論文）
- 画像参考：What is Dataset Distillation Learning?
  https://arxiv.org/pdf/2406.04284
- 大量の実画像で学習するよりも、
  少量の生成画像（Distillated image）で
  学習するほうが高いAccuracyを達成。

Figure 1. **Real vs. distilled data.** Real images of airplane, car, and truck from CIFAR-10 (Krizhevsky et al., 2009) are shown on left and highly salient distilled images of the same classes are shown on the right. While distilled images *can* be used to train high-accuracy classifiers, *why* this is possible and *what* do they represent remains unclear.

- Applications
  - Continual Learning
  - Privacy
  - Medical
  - Federated Learning
  - Graph Neural Network
  - Neural Architecture Search
  - Fashion, Art, and Design
  - Recommender Systems
  - Blackbox Optimization
  - Trustworthy
  - Text
  - Tabular
  - Retrieval
  - Video
  - Domain Adaptation
  - Super Resolution
  - Time Series
  - Speech
  - Machine Unlearning
  - Reinforcement Learning
  - Long-Tail

ECCV 2024 の動向・気付き(67/132)

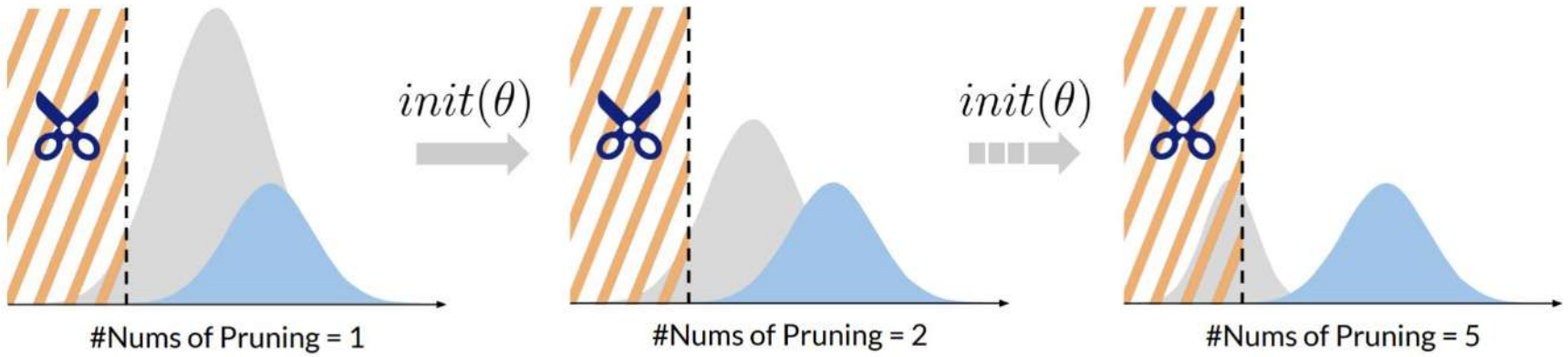
Watermark 関連の動向

- Image, Video, Dataset, Generative Model, NeRF など多岐にわたる
  - Diffusion Model にwatermarkを埋め込む手法が多い印象
    - 著作権侵害などの社会問題に対処するため？
  - Classification Model よりも Generative Model の方が圧倒的に多い.
- CVPR2024より多い論文数
  - 巨大モデルを作るためのコストの高さや生成モデルによる著作権問題などによってwatermarkの需要も増している？
- watermarkを埋め込む手法だけでなく，検知・抽出する手法も

ECCV 2024 の動向・気付き(68/132)

Finding needles in a haystack: A Black-Box Approach to Invisible Watermark Detection

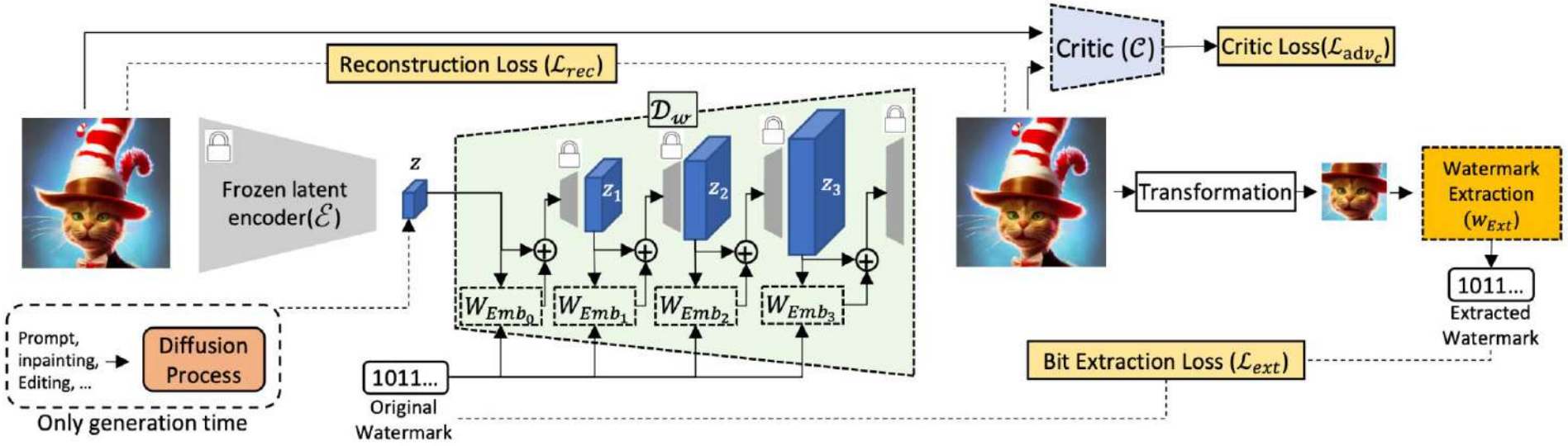
- “Finding a needles in a haystack”の名の通り，Datasetの中からwatermarkされている画像を検出する手法
- Datasetだけを用いるのでBlack-Boxで達成することができる
- cleanな画像とwatermarkされた画像のデータの分布の差を用いて，反復的にclean datasetを削ることによってwatermarked imageだけを抽出する(下図)



ECCV 2024 の動向・気付き（69/132）

LaWa: Using Latent Space for In-Generation Image Watermarking

- LDMのAutoencoderのデコーダをfine-tuneすることによって生成される画像にbinary messageを埋め込む手法
- Encoder-Decoderモデルと違い，アーキテクチャを変える事なくwatermarkすることが可能
- 様々なdistortionに頑健なwatermarkを実現

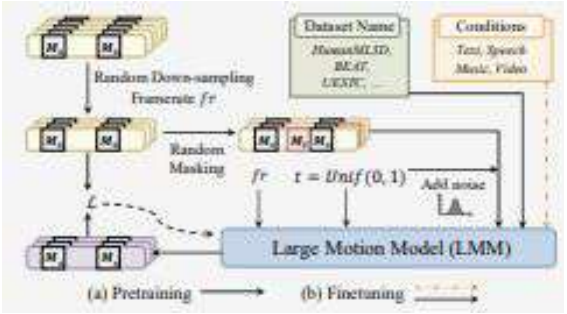


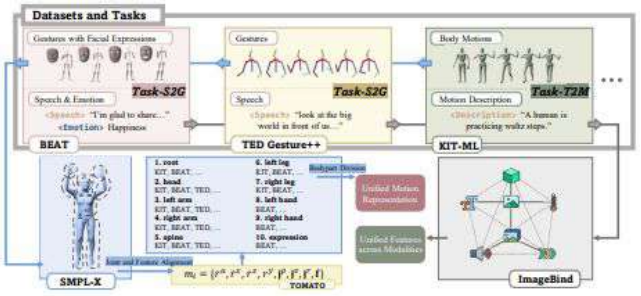
ECCV 2024 の動向・気付き（70/132）

採択論文のテーマからみた動向

- 3Dを対象とした研究が非常に盛んである。
  - 特に3D Object Detectionの研究が多い印象。
  - 拡散モデル等を用いたText-to-3D等の3D生成の研究も盛ん。
- 画像の枠を超えたマルチモーダルな研究も多い。
  - 言語と画像を対象とした研究が多い印象。
  - 一方で、言語・音楽・画像・動画と多様なモーダルを扱うような研究も見られた（下図）。

⇒より**現実に近く複雑性の高い問題** を対象とした研究が増加している。





単語	出現頻度（割合）
3d	372（14.5%）
diffusion	302（11.8%）
multi	198（7.7%）
video	196（7.6%）
generation	192（7.5%）
text	172（6.7%）
language	166（6.5%）

採択された論文タイトルに含まれる
単語の出現頻度とその論文数に対する割合
（上位の中から一部をピックアップ）

M.Zhang et al. “Large Motion Model for Unified Multi-Modal Motion Generation”, in ECCV 2024.

ECCV 2024 の動向・気付き（71/132）

3D Gaussian Splatting関連の動向

- **3D表現の質と効率性の向上に関する研究は依然として多い**
  - 高品質なレンダリングと効率的な最適化に関する研究が多数
  - アンチエイリアシングやGaussian特有の密度制御の改善など
- **動的シーンへの対応**
  - 静的なシーンだけでなく、動的オブジェクトや変形可能な対象への適用研究も多数見られる
- **実用性と応用範囲の拡大に関する研究**
  - モデル圧縮とレンダリング効率向上に関する研究
  - マルチタスク学習（セグメンテーション、デプス推定等）への展開など…
- **入力の多様化と制約の緩和**
  - 少数視点からの再構成や特殊な撮影条件への対応研究が増加
  - テキストプロンプトによる編集など、インタラクティブな操作の研究も登場

⇒より現実的で複雑なシーンや条件に対応可能な3D Gaussian Splattingの実現に向けた研究が増加している

他にも大規模シーンの再構成，生成モデルとの統合，メッシュ化への取り組みなども見られた

ECCV 2024 の動向・気付き(72/132)

3D Gaussian Splatting研究の分類例

- 3D表現(画質/効率)
  - View-Consistent 3D Editing with Gaussian Splatting
  - SAGS: Structure-Aware 3D Gaussian Splatting
  - Analytic-Splatting: Anti-Aliased 3D Gaussian Splatting via Analytic Integration
  - Pixel-GS: Density Control with Pixel-aware Gradient for 3D Gaussian Splatting
  - Revising Densification in Gaussian Splatting
  - On the Error Analysis of 3D Gaussian Splatting and an Optimal Projection Strategy
  - 3iGS: Factorised Tensorial Illumination for 3D Gaussian Splatting
  - GeoGaussian: Geometry-aware Gaussian Splatting for Scene Rendering
- 動的シーン
  - SWinGS: Sliding Windows for Dynamic 3D Gaussian Splatting
  - TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting
  - DynMF: Neural Motion Factorization for Real-time Dynamic View Synthesis with 3D Gaussian Splatting
  - Per-Gaussian Embedding-Based Deformation for Deformable 3D Gaussian Splatting
  - Topo4D: Topology-Preserving Gaussian Splatting for High-Fidelity 4D Head Capture
- 特殊な撮影条件への対応
  - BAD-Gaussians: Bundle Adjusted Deblur Gaussian Splatting
  - Gaussian in the wild: 3D Gaussian Splatting for Unconstrained Image Collections
  - Deblurring 3D Gaussian Splatting
  - BAGS: Blur Agnostic Gaussian Splatting through Multi-Scale Kernel Modeling
  - Gaussian Splatting on the Move: Blur and Rolling Shutter Compensation for Natural Camera Motion
- マルチタスク学習と応用拡大
  - VersatileGaussian: Real-time Neural Rendering for Versatile Tasks using Gaussian Splatting
  - SGS-SLAM: Semantic Gaussian Splatting For Neural Dense SLAM
  - ManiGaussian: Dynamic Gaussian Splatting for Multi-task Robotic Manipulation
- 少数視点からの再構成
  - FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting
  - MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-View Images
  - MVPGS: Excavating Multi-view Priors for Gaussian Splatting from Sparse Input Views
- 圧縮/レンダリング効率
  - CompGS: Smaller and Faster Gaussian Splatting with Vector Quantization
  - HAC: Hash-grid Assisted Context for 3D Gaussian Splatting Compression
  - GaussianImage: 1000 FPS Image Representation and Compression by 2D Gaussian Splatting

ECCV 2024 の動向・気付き（73/132）

クリエイティブ・グラフィックデザイン 関連の動向

- **デザイン生成**：クリエイティブな画像素材の作成支援や自動生成
  - ピュアな画像生成技術以外の、デザイン生成に関連する技術の動向を総括
  - 2019年頃から盛んになり始めた、まだ発展途上の研究領域（ref. SSII’24 技術マップ）
- **デザイン生成に関連する主要な設定**
  - Layout Generation: レイアウト生成
    - 画像・ロゴ・テキスト等の素材の配置を生成するタスク
    - スマホアプリの UI の配置もレイアウト生成の一部
  - Content-aware Layout Generation: コンテンツを考慮したレイアウト生成
    - ベースとなる画像を考慮してレイアウトを生成するタスク
    - ポスター、UI、広告、雑誌などの作成へ応用可能
  - Visual Text Rendering: 視覚的なテキスト生成
    - 生成画像に視覚的に審美性の高いテキストも同時にレンダリングする
  - Layer-aware Image Generation: レイヤーを考慮した画像生成
    - 前景、中間、背景のようなレイヤー構造を考慮した画像生成タスク

Image from https://sites.google.com/view/gdug-workshop/

ECCV 2024 の動向・気付き（74/132）

クリエイティブ・グラフィックデザイン 全体総括

- **分野としては現在進行形で発展中**
  - ECCV’24 ではクリエイティブ・グラフィックデザインに関する技術は約15本程度発表あり
    - 企業的には売上につながるような応用先のためあえて隠している可能性も否定できない
- **企業の発表が中心から大学へ裾野が広がる**
  - 初期は中国企業（Alibaba, Tencent, etc.）が分野をひっぱり、MSRA参入後大きく拡大
    - 日本では初期から CyberAgent が重要技術を提案し、今回も同様に複数研究を発表
  - ベンチマークデータセットや評価指標が固まり始めた
    - レイアウト生成等は比較的小さな計算資源で高速にPDCAを回せる
- **拡散モデルや大規模言語モデルの応用に注目**
  - ノイズから多様なレイアウトやデザインが生成可能であるために利用しやすい
  - 言語モデルが HTML や JSON 等のデータを出力できることをデザイン生成に応用
- **再投稿を経て ECCV’24 に採択された論文あり**
  - Layout-Corrector (CVPR’24 ➔ ECCV’24), LayoutDETR (CVPR’23? ➔ ICLR’23 ➔ ECCV’24), PosterLlama (CVPRW’24 ➔ ECCV’24), Dolfin (ICLR’23 ➔ ECCV’24), etc

ECCV 2024 の動向・気付き（75/132）

レイアウト生成 関連の動向

- **総括**
  - SoTA 技術の弱点の指摘・克服 (Layout-Corrector, Dolfin) や注目技術のレイアウト生成への応用 (LayoutFlow) が中心
- **個別事例**
  - Layout-Corrector: Alleviating Layout Sticking Phenomenon in Discrete Diffusion Model
    - 所属: 東北大学🇯🇵・LINEヤフー🇯🇵
    - 概要: 不調和なレイアウト要素を修正可能な離散拡散モデルに対する新たなモジュールの提案
  - Dolfin: Diffusion Layout Transformers without Autoencoder
    - 所属: 清華大学🇨🇳・カリフォルニア大学サンディエゴ校🇺🇸
    - 概要: 既存の AutoEncoder が不要な新たなモデルの提案
  - LayoutFlow: Flow Matching for Layout Generation
    - 所属: 東京大学🇯🇵・サイバーエージェント🇯🇵
    - 概要: レイアウト生成に対して Flow Matching を応用することで生成品質を維持しながら生成速度を大きく向上

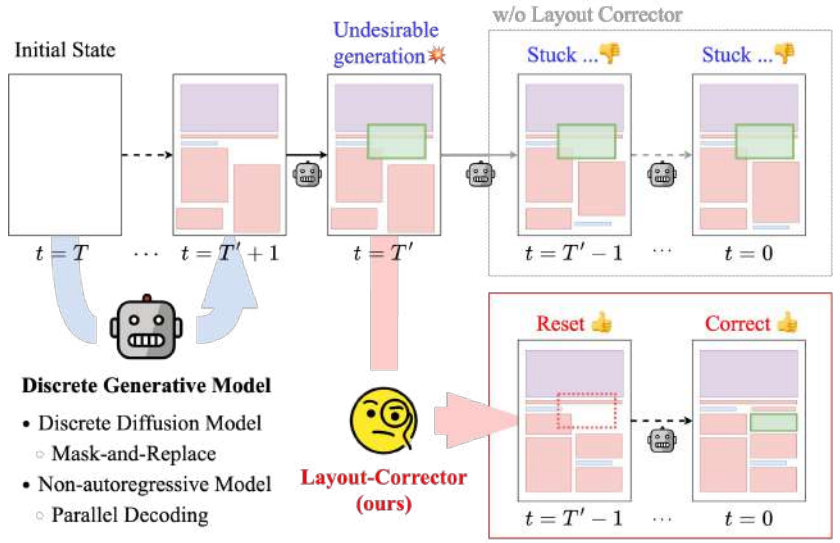


Image from Layout-Corrector

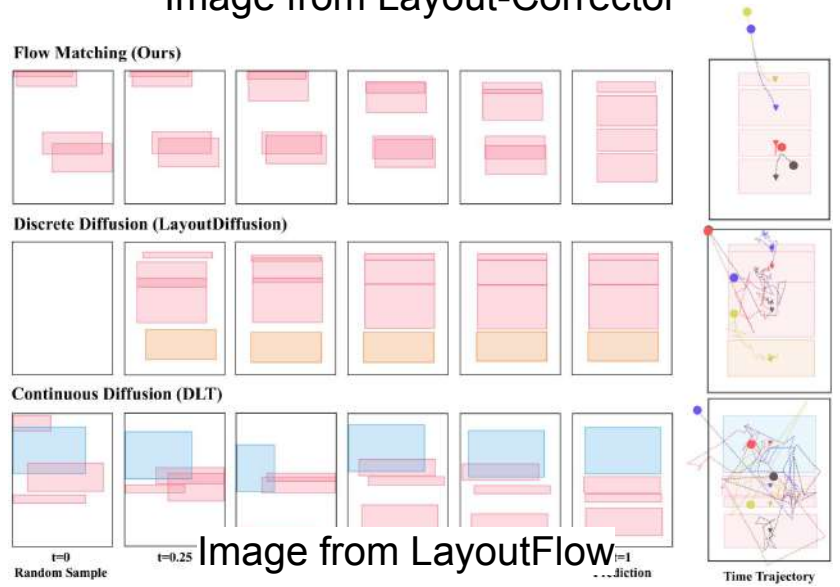


Image from LayoutFlow

ECCV 2024 の動向・気付き（76/132）

コンテンツを考慮したレイアウト生成 関連の動向

- **総括**
  - 大規模言語モデルや物体認識モデルを活用したデザイン生成に活路を見出す
- **個別事例**
  - PosterLlama: Bridging Design Ability of Langauge Model to Content-Aware Layout Generation
    - 所属: 蔚山科学技術大学校🇨🇳
    - 概要: レイアウト情報を HTML 表現として VLM 化した CodeLLaMA をもとに HTML 形式のレイアウトを生成
  - LayoutDETR: Detection Transformer Is a Good Multimodal Layout Designer
    - 所属: Salesforce Research🇺🇸
    - 概要: 背景画像を考慮したレイアウト生成を物体検出タスクのように定義しつつ様々なモデルを多数組み合わせたアーキテクチャの提案

Image from PosterLlama

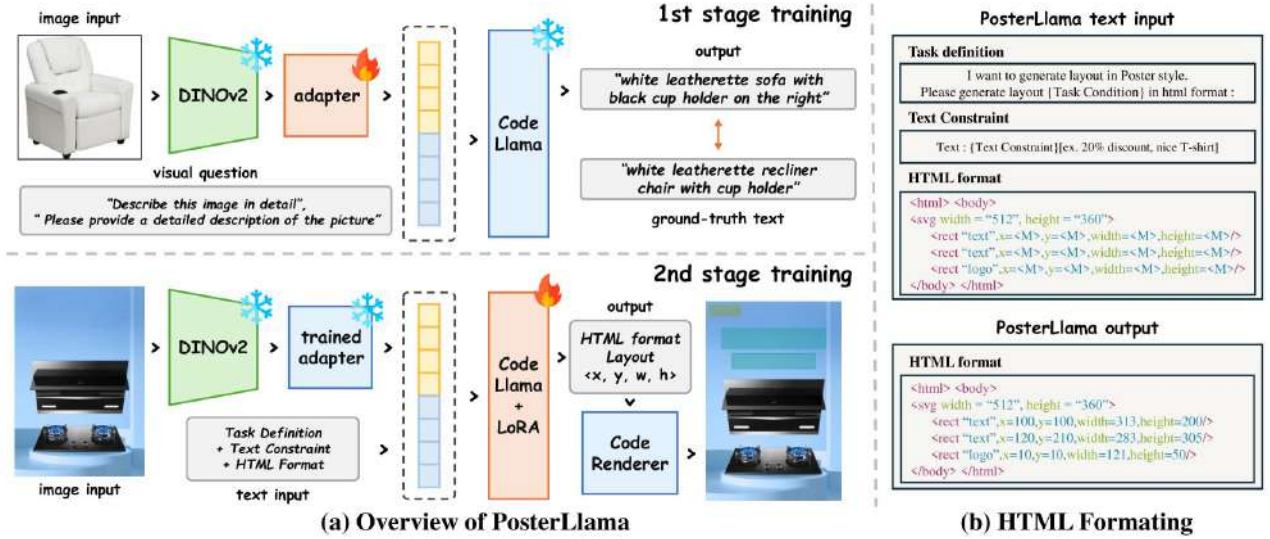
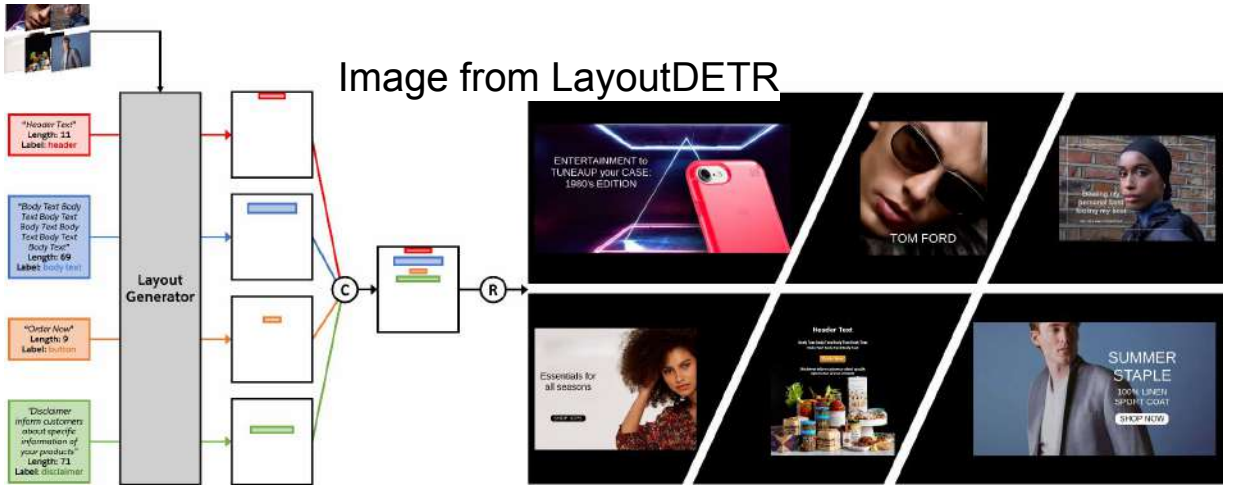


Image from LayoutDETR

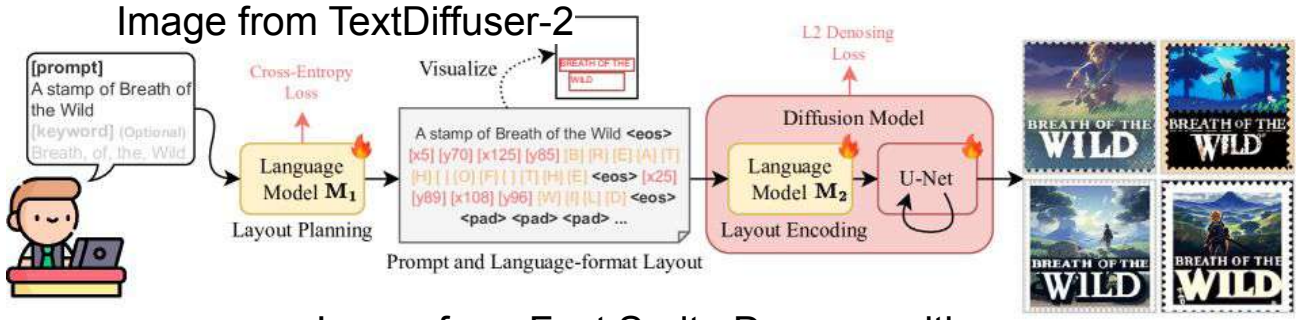


ECCV 2024 の動向・気付き（77/132）

デザイン生成 関連の動向 (1/3)

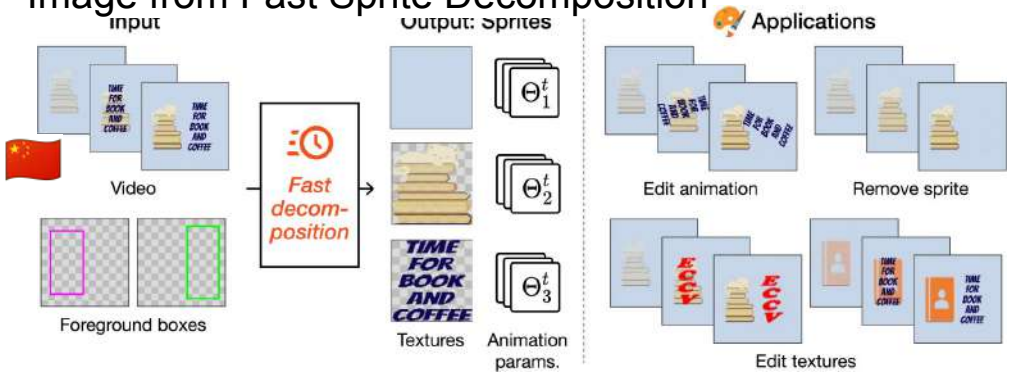
- **総括**
  - 画像中のテキスト生成やGIFアニメ分析
    Webデザインの自動化など多岐に渡る
- **個別事例**
  - TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering **(Oral)**
    - 所属: 香港科技大学🇨🇳・中山大学🇨🇳・Microsoft Research🇨🇳
    - 概要: 画像中に正しくテキストを描画できるように
      なおかつLLMでいい感じのレイアウトになるよう工夫

Image from TextDiffuser-2



  - Fast Sprite Decomposition from Animated Graphics
    - 所属: CyberAgent🇺🇸
    - 概要: アニメーションを構成要素へ分割する新たな視点の提案および新たなデータセットの構築

Image from Fast Sprite Decomposition



  - WebRPG: Automatic Web Rendering Parameters Generation for Visual Presentation
    - 所属: 浙江大学🇨🇳・Alibaba🇨🇳
    - 概要: HTMLをベースにWebページの
      自動生成に向けた新たな視点の提案

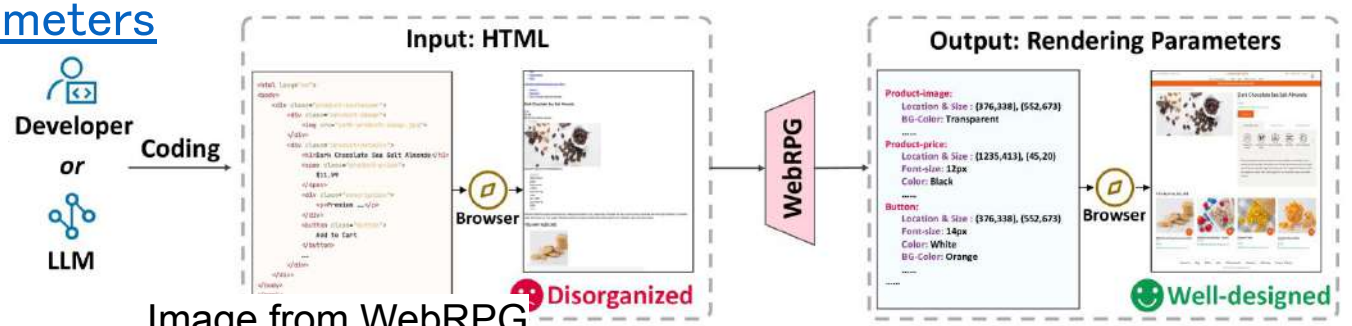


Image from WebRPG

ECCV 2024 の動向・気付き(78/132)

デザイン生成 関連の動向 (2/3)

- **総括**
  - 個別に制御がしやすいレイヤー構造を考慮した画像生成に今後期待
- **個別事例**
  - Layered Rendering Diffusion Model for Controllable Zero-Shot Image Synthesis
    - 所属: 北京航空航天大学🇨🇳・ブリストル大学🇬🇧・MBZUAI🇦🇪
    - 概要: 複数の視覚的要素をレイヤーで処理しそれぞれの要素を提案する視覚誘導手法で配置する画像生成手法の提案
  - LayerDiff: Exploring Text-guided Multi-layered Composable Image Synthesis via Layer-Collaborative Diffusion Model
    - 所属: 中山大学🇨🇳・Huawei🇨🇳
    - 概要: 背景と複数の前景からなる多層の画像生成の実現
  - Towards Reliable Advertising Image Generation Using Human Feedback
    - 所属: 華中科技大学🇨🇳・JD🇨🇳
    - 概要: 人間のフィードバックを反映した広告画像を生成可能なRFFTの提案

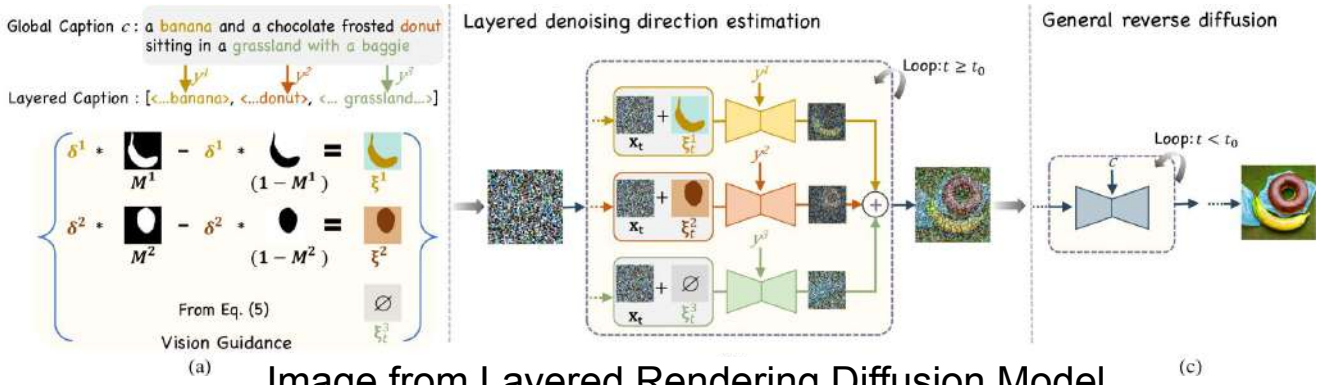


Image from Layered Rendering Diffusion Model

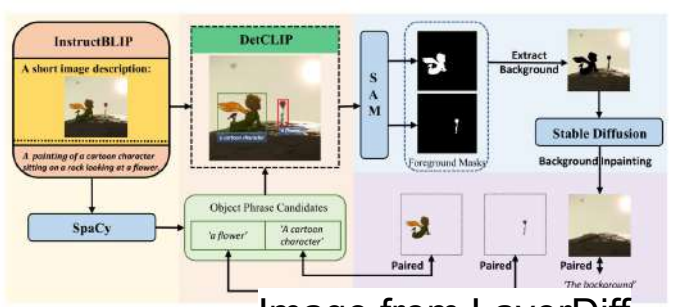


Image from LayerDiff

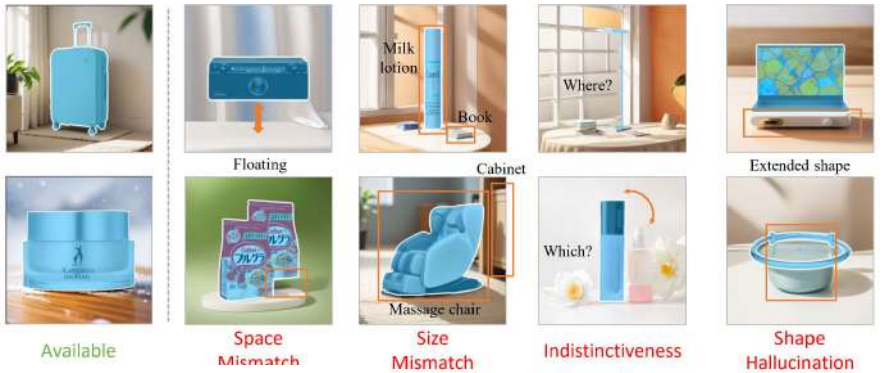
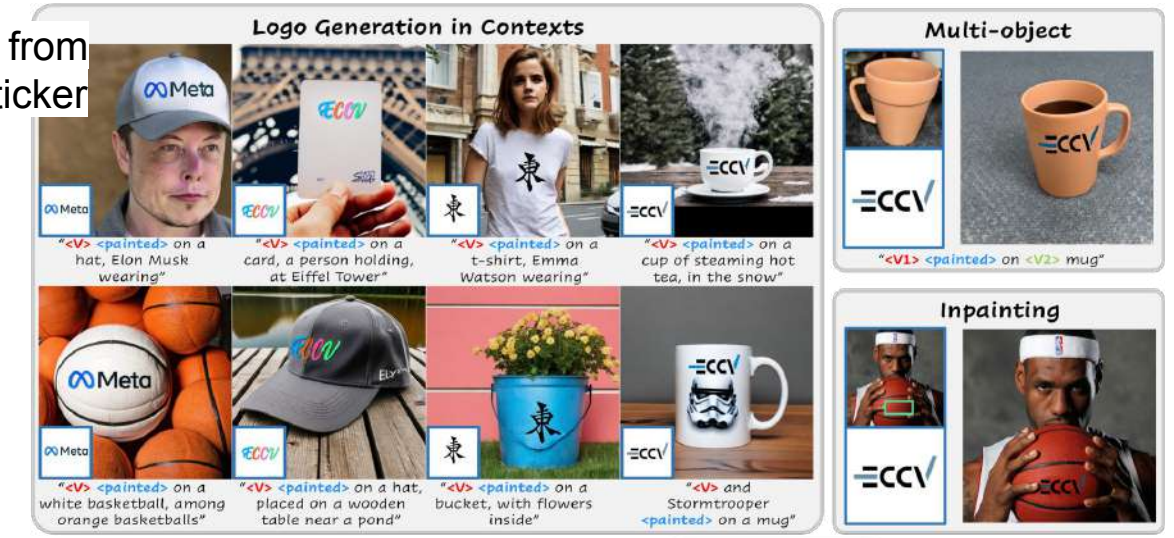


Image from RFFT

ECCV 2024 の動向・気付き（79/132）

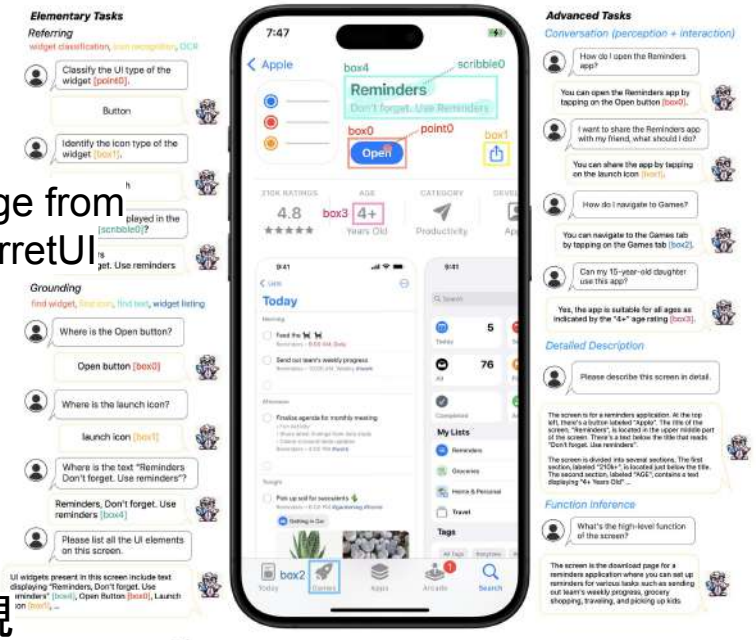
デザイン生成 関連の動向 (3/3)

Image from LogoSticker



- **総括**
  - Web/スマホUIに関するデータの整備やレイアウト生成手法の提案や生成画像に対するロゴの合成手法などが増えている
- **個別事例**
  - DreamStruct: Understanding Slides and User Interfaces via Synthetic Data Generation
    - 所属: CMU🇺🇸・テキサス大学オースティン🇺🇸・アールト大学🇫🇮
    - 概要: コード生成を利用したUI等の合成とラベル付与の方法の提案
  - Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs
    - 所属: Apple🇺🇸
    - 概要: 多様なUIデータの収集およびラベル付与の方法の提案
  - LogoSticker: Inserting Logos into Diffusion Models for Customized Generation
    - 所属: 香港中文大学🇨🇳・香港大学🇨🇳・Huawei🇨🇳・SmartMore🇨🇳
    - 概要: ロゴを拡散モデルの生成結果に忠実に挿入する方法の提案
      強化学習を応用することでさまざまな文脈で一貫性のある生成を実現

Image from FerretUI



ECCV 2024 の動向・気付き（80/132）

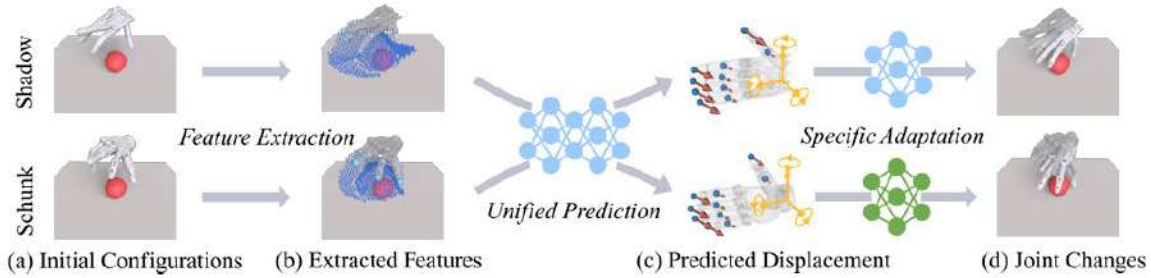
ロボットの身体によらず一般化できる手法がいくつかみられる

- ロボットのデータセットはスケールアップしにくい→ロボットに共通な表現（Trajectory，Grasp pose）で扱うモデルにする
  - Learning Cross-hand Policies of High-DOF Reaching and Grasping
    - グリッパーの形状を関節と指先の位置をKeypointとして単純化した上で，それぞれのキーポイントのDisplacementを予測する．
  - Track2Act: Predicting Point Tracks from Internet Videos enables Generalizable Robot Manipulation
    - ゴール画像から動作の軌跡の予測をおこなってロボットの動作に応用する．軌跡予測を多様な動画から学習し，軌跡予測に対して，動作コマンドを同時に学習させることロボット動作を可能にする．
  - Robo-ABC : Affordance Generalization Beyond Categories via Semantic Correspondence for Robot Manipulation
    - アフォーダンスをもとに把持を行う上で，事前のメモリーとして大量な画像にContact mappingをしたのち，データベースと最もマッチする物体の画像を探索して再度マッピングする．

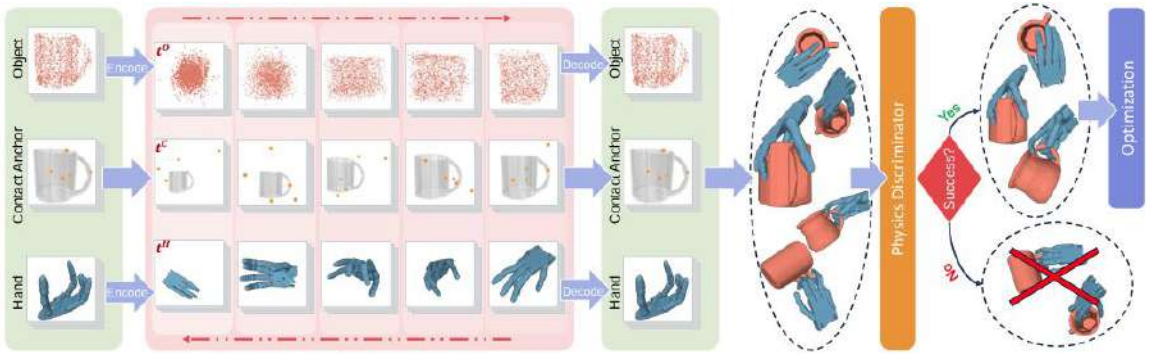
ECCV 2024 の動向・気付き（81/132）

ロボット（manipulation）応用に関する気づき

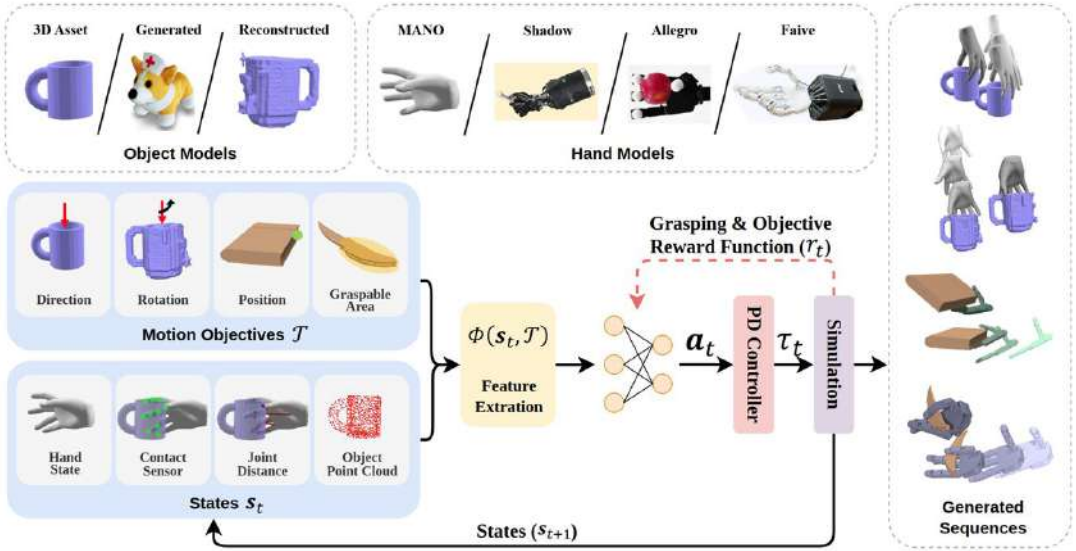
- Manipulationとの関連で発表が多いのはGrasp生成
  - それぞれに興味深い問題を解いていた
  - 扱いが難しい多指ハンドを複数使った研究が，いくつか見られた点が興味深い



grasp policyを異なるロボットハンドへ転移
[Q.She et al. "Learning Cross-hand Policies of High-DOF Reaching and Grasping"]



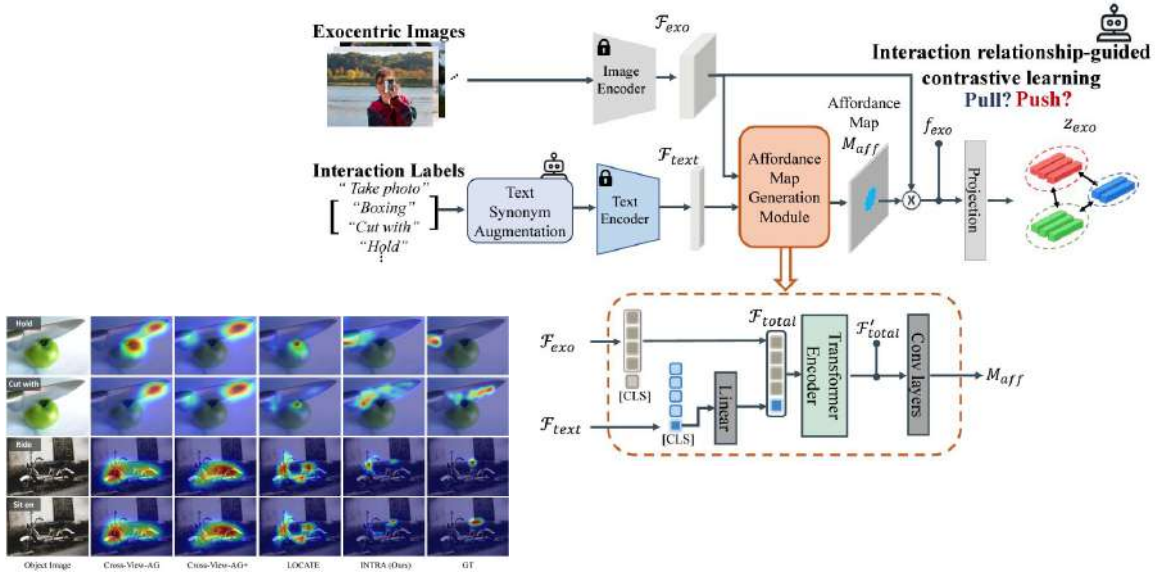
物体・ハンド・接触の情報を Diffusion modelで同時生成することで，多様かつ成功率が高い graspを生成
[J. Lu et al., UGG: Unified Generative Grasping]

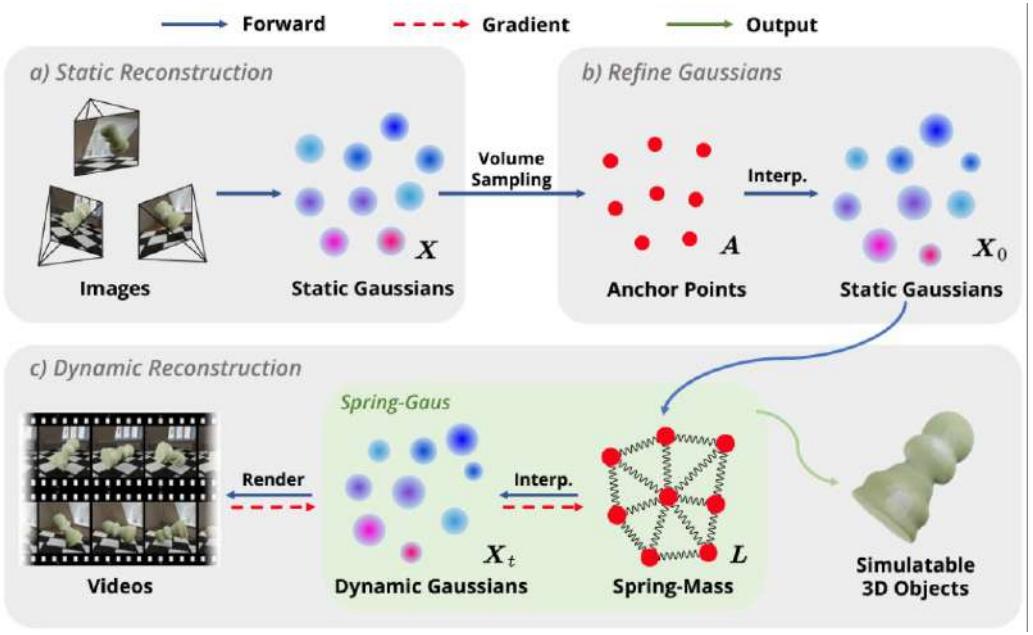


複数の多指ハンドを用いた大規模な Grasp生成
[H. Zhang et al.,
GraspXL: Generating Grasping Motions for Diverse Objects at Scale]

ECCV 2024 の動向・気付き（82/132）

ロボット（manipulation）応用に関する気づき





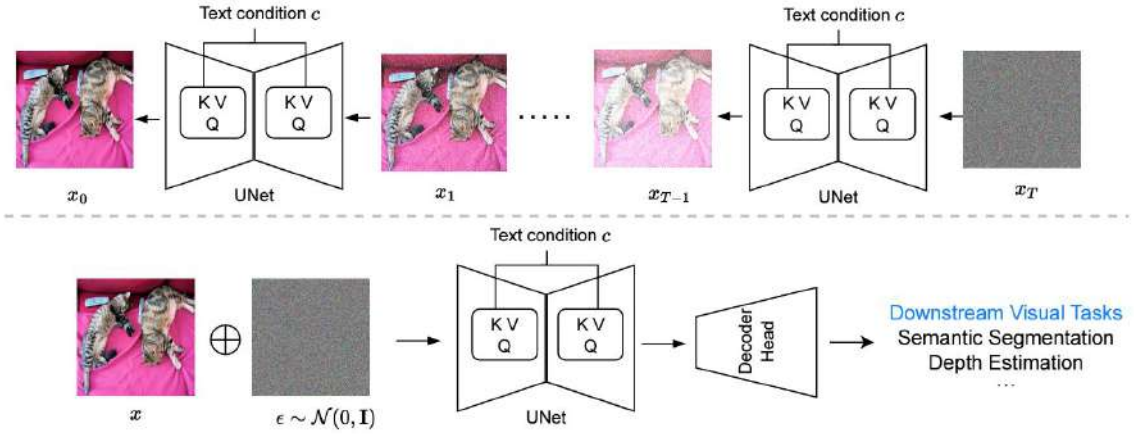
- 弱教師あり学習による affordance grounding
  ロボットによる物体操作でも affordance推定は重要．これまでは教師あり学習が多い印象．これからは弱教師あり学習が主流になるか．
  [J.H. Jang et al., INTRA: Interaction Relationship-aware Weakly Supervised Affordance Grounding]

- 視覚からの物理特性の推定
  CVの先端技術である3D Gaussianによる再構築に，物体操作で重要となる物理特性の推定を組み合わせている点が興味深い．
  [L. Zhong et al., Reconstruction and Simulation of Elastic Objects with Spring-Mass 3D Gaussians]

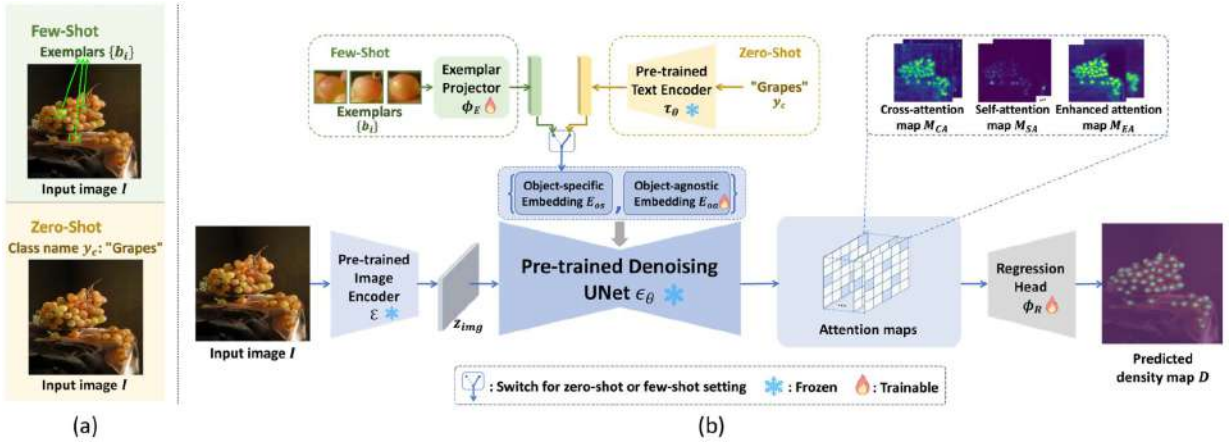
ECCV 2024 の動向・気付き（83/132）

ロボット（manipulation）応用に関する気づき

- ロボット分野でも重要な技術で発表件数が多いものには，以下があった
  - Diffusion model, state space model（特にMamba），マルチモーダルデータの学習
  - これらはロボット応用に比較的近いが，ロボットならではの情報（動作情報，力覚や触覚などのモダリティ）とビジョンを結びつけた研究がCV分野でもより広がることが期待される



- Diffusion modelのノイズ除去モデルは semantic segmentationなどの良いbackboneとして機能する．この場合の best practiceを提示．
  [M. Zhang et al., Three Things We Need to Know About Transferring Stable Diffusion to Visual Dense Prediction Tasks]



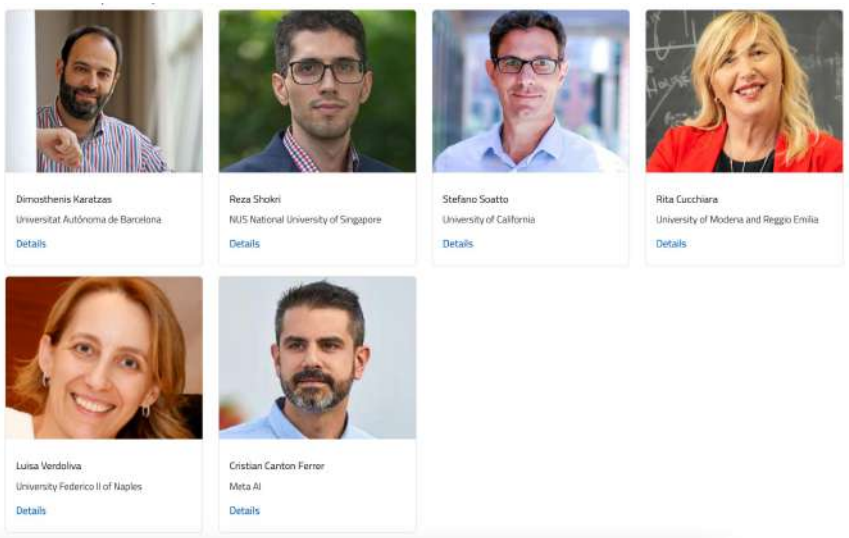
- 未知の物体を数える技術．日常生活では結構必要なタスク．こちらもノイズ除去モデルをbackboneとして利用．
  [X. Hui et al., Class-Agnostic Object Counting with Text-to-Image Diffusion Model]

ECCV 2024 の動向・気付き(84/132)

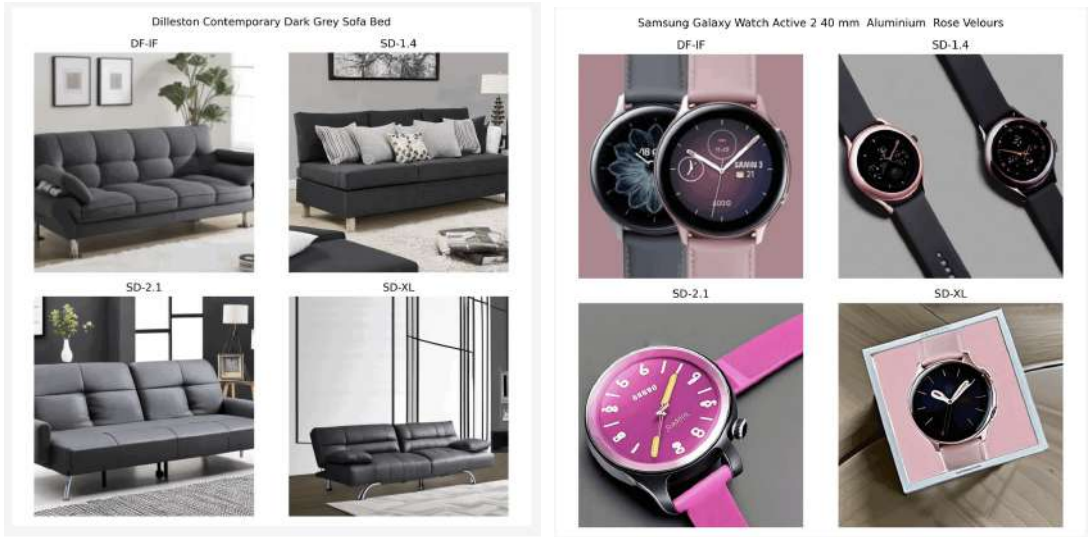
拡散モデルの発展に伴い難化するDeepFake/Synthetic image検出

- ECCV 2024が初開催のTrust What You learN (TWYN) Workshopではセキュアで安全なAI技術の開発を目指すヨーロッパの組織であるELSAのメンバーを中心として、UnlearningやDeepFake検出の技術が議論された。

- SoTAの拡散モデルから生成された画像を対象とした検出手法や、生成モデルの種類を超えた汎化を提案する手法が提案されていたが、DeepFakeの生成技術の発展速度が早くArms Raceの様相を呈している。



TWYN WorkshopのKeynote speakers



SoTAの拡散モデルを使用したベンチマーク(D3)中の画像

ECCV 2024 の動向・気付き(85/132)

ViTのfeature mapにおけるアーティファクト問題の原因と対策

- Vision Transformers Need Registersの論文で改善方法が既に提案されているが、モデルの再学習はコストが大きく、アーティファクトの原因にもまだ議論の余地が残されている。

- 今回のECCVでも複数の論文がこのトピックに取り組んでいる。
  - SINDER: Repairing the Singular Defects of DINOv2：重みの特異値ベクトルを原因とした説明
  - Denoising Vision Transformers：Posisional Encodingを原因とした説明

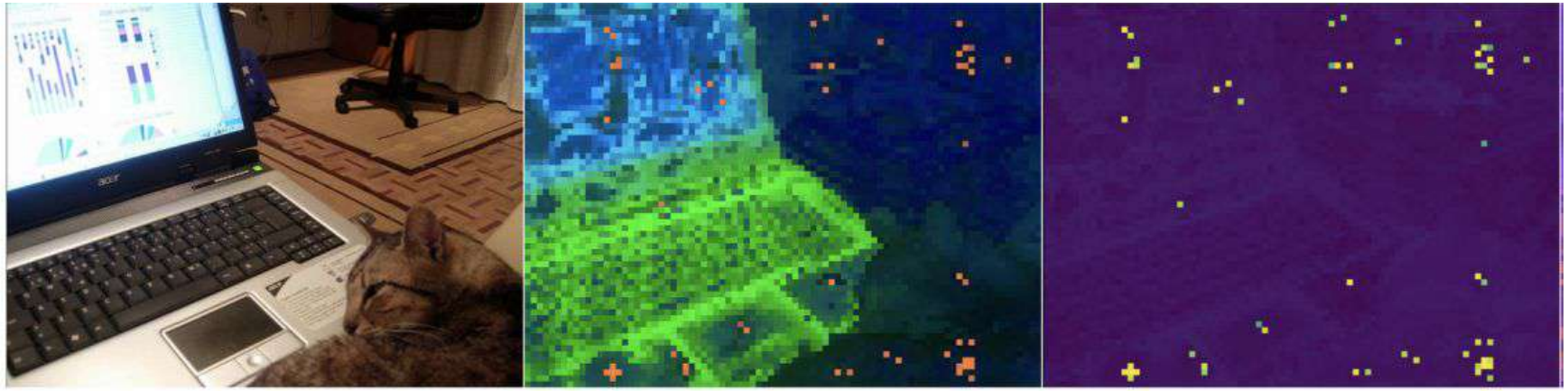


Image　　PCA of DINOv2　　Norm of DINOv2

ECCV 2024 の動向・気付き（86/132）

Model Merge関連の動向

- Model Mergeの概要
  - 複数のモデルを1つのモデルに統合する手法
  - 推論時間・メモリ量を抑制しつつ、アンサンブル効果を得ることが目的
  - Visionに限る手法ではないが、近年注目

- 主なアプローチ
  - 重みベクトル上での（重み付き）平均化（代表例：Model Soups [Wortsman+, 2022]）
    - 同じ基盤モデルから学習されるなど、同じ損失の盆地に属するという前提あり
  - 各モデルのレイヤーを取捨・選択して統合（代表例：Sakana AIの進化的モデルマージ）

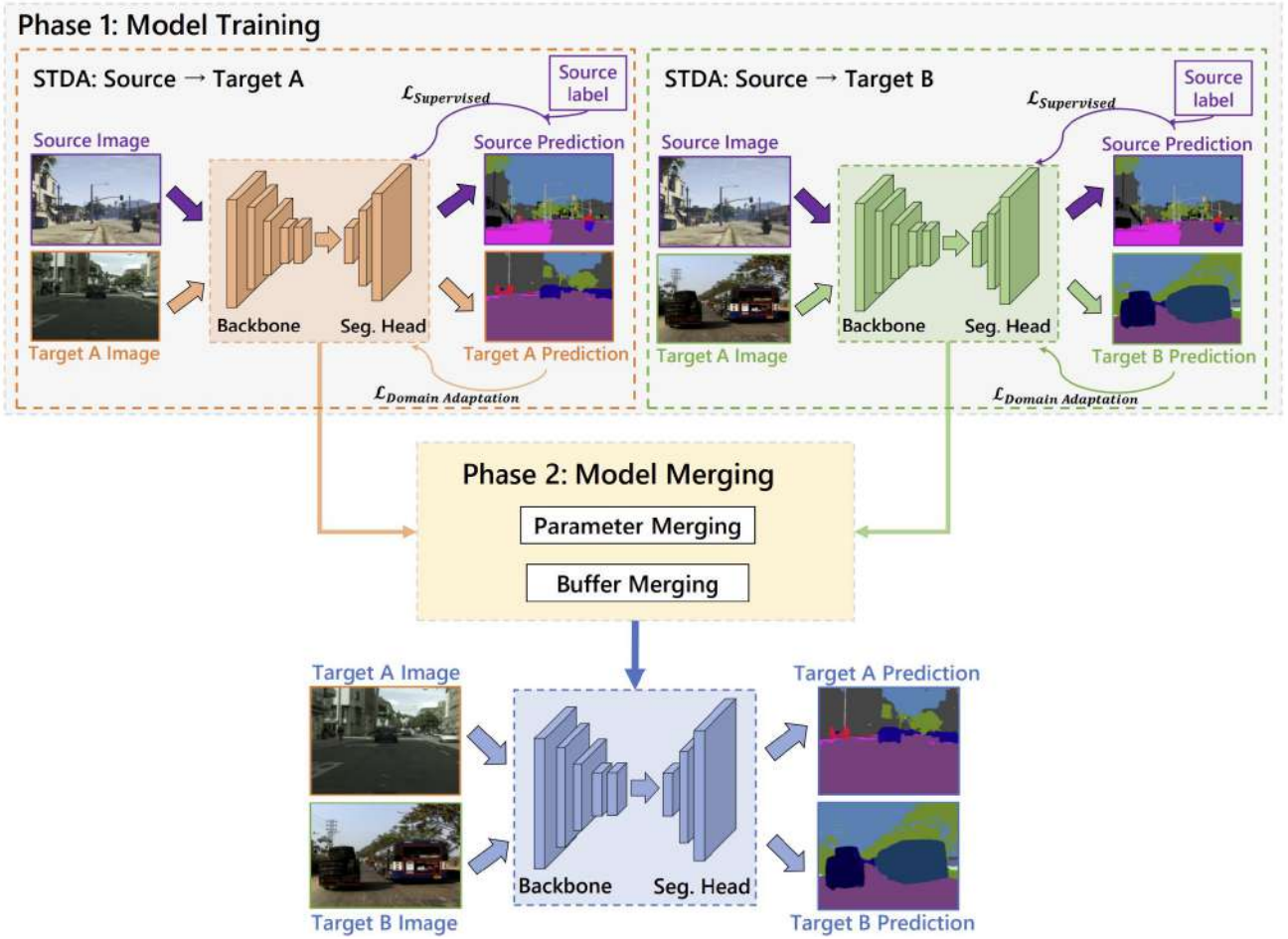
- タイトルにModel Mergeがある論文は以下の4件（全てベクトルの平均化）
  - Training-Free Model Merging for Multi-target Domain Adaptation
  - Diffusion Soup: Model Merging for Text-to-Image Diffusion Models
  - Model Breadcrumbs: Scaling Multi-Task Model Merging with Sparse Masks
  - MagMax: Leveraging Model Merging for Seamless Continual Learning

ECCV 2024 の動向・気付き（87/132）

Training-Free Model Merging for Multi-target Domain Adaptation

- 基盤モデルからターゲットデータでチューニングするドメイン適用において、マルチターゲットを目的とする場合、全ターゲットデータにアクセス可能という前提は非現実的

- 各ターゲットでチューニングしたモデルマージ及びバッファマージを適用することで、各ターゲットデータへのアクセス無しでマルチターゲットドメイン適用を実現
  - バッファマージは、具体的にはBNにおける平均と分散を、モデル間で重み付きの平均化する処理

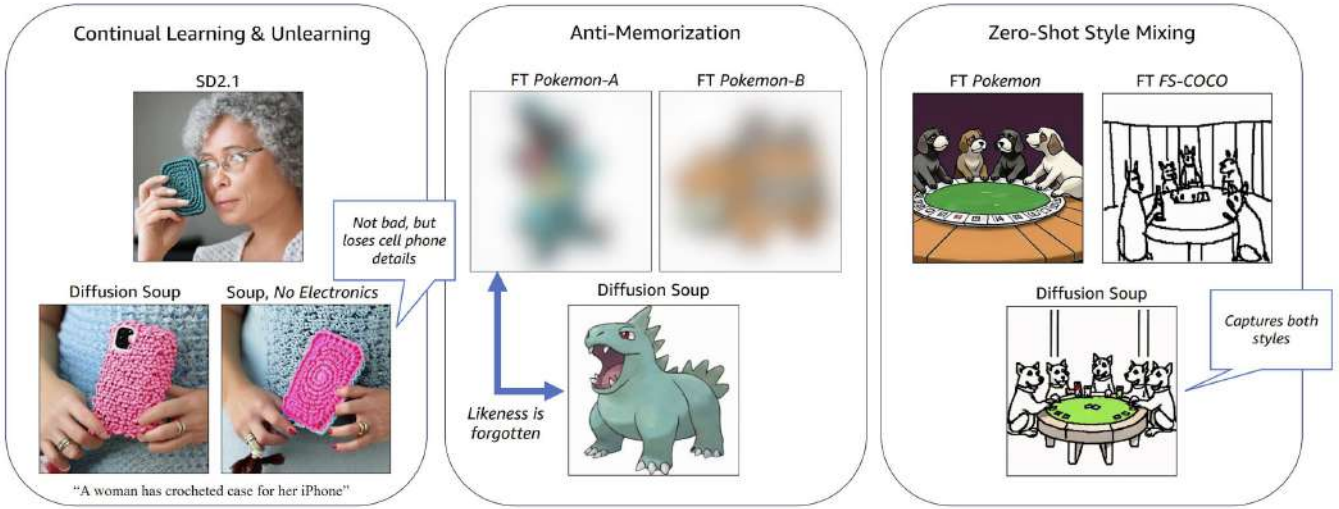


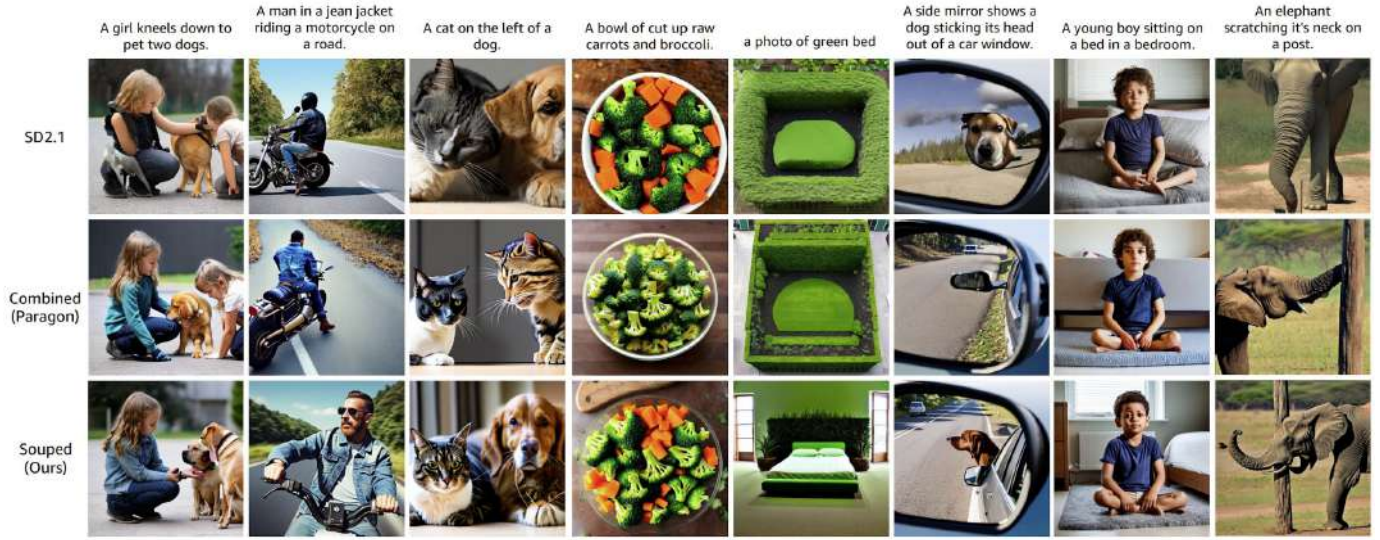
ECCV 2024 の動向・気付き(88/132)

Diffusion Soup: Model Merging for Text-to-Image Diffusion Models

- 拡散モデルによる画像生成において、複数のドメイン別にモデルを学習し、目的に応じてモデルマージでのモデル生成を提案

- 全ドメインデータで学習した単一の拡散モデルより高品質の画像を生成可能



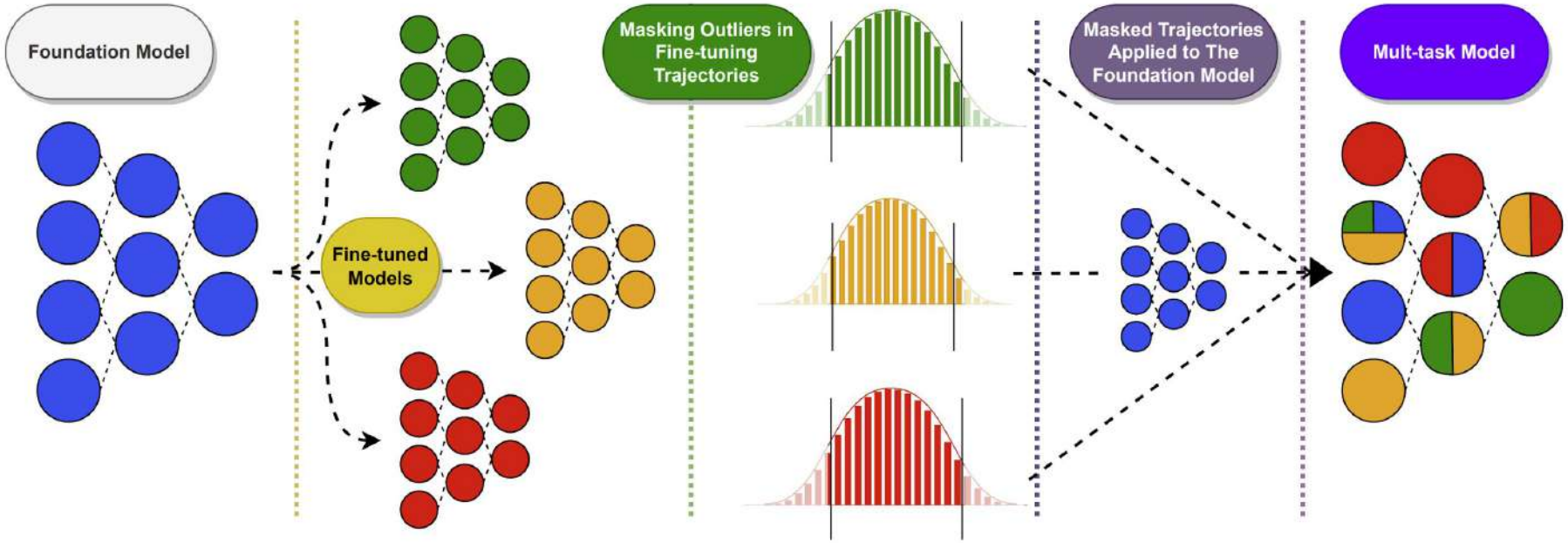


ECCV 2024 の動向・気付き(89/132)

Model Breadcrumbs: Scaling Multi-Task Model Merging with Sparse Masks

- タスクベクトル [Ilharco+, 2023]
  - 基盤モデルとfine-tunedモデル間の差分ベクトル
  - タスクベクトルの加減算でタスクの加減が可能

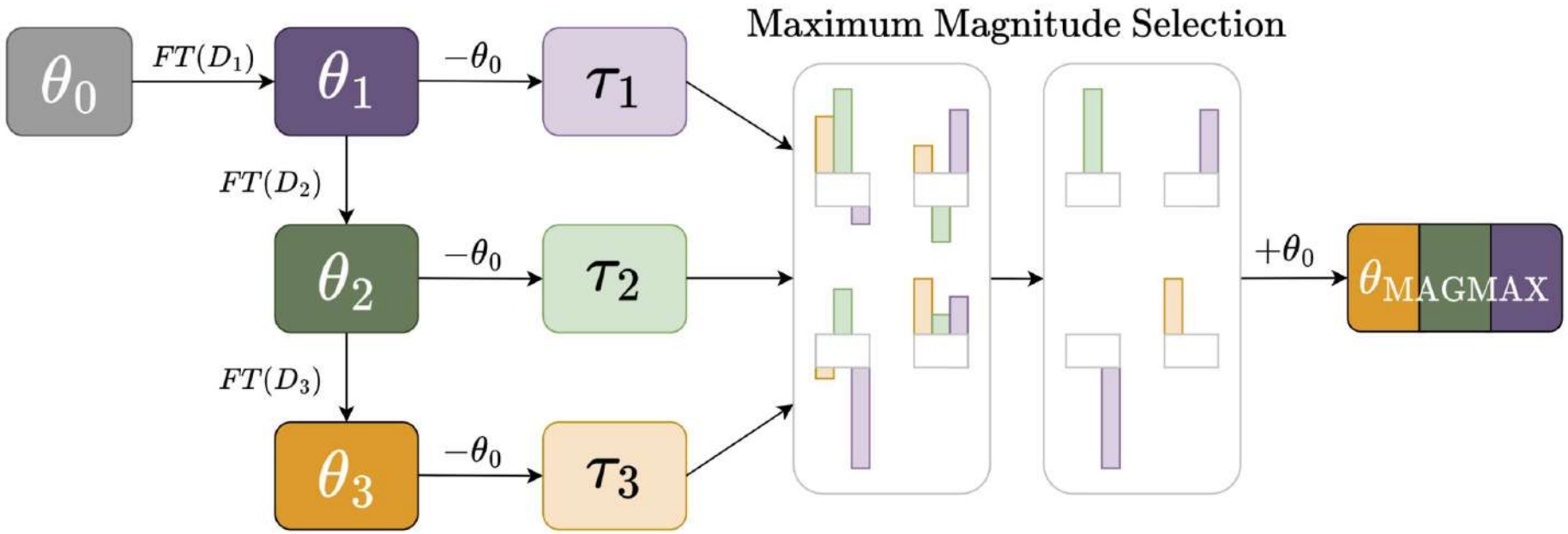
- タスクベクトルにはノイズが蓄積されている為、レイヤー別にベクトル値の大小のノイズ・外れ値を削除することで、より良いタスクベクトルを作成



ECCV 2024 の動向・気付き（90/132）

MagMax: Leveraging Model Merging for Seamless Continual Learning

- 継続学習の課題の1つが破滅的忘却。従来は正則化などで忘却を抑制
- タスク毎にモデル及びタスクベクトル作成。モデルマージによりタスクを追加
- マージの際、各値をそのままマージするのではなく、最大マグニチュードの値を選択してマージすることで、タスク追加時の精度を継続学習より向上

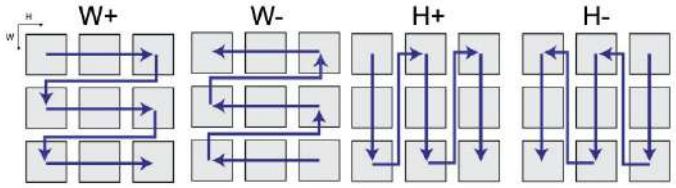


ECCV 2024 の動向・気付き(91/132)

状態空間モデル (SSM) や線形注視機構はCV分野でも注目されている

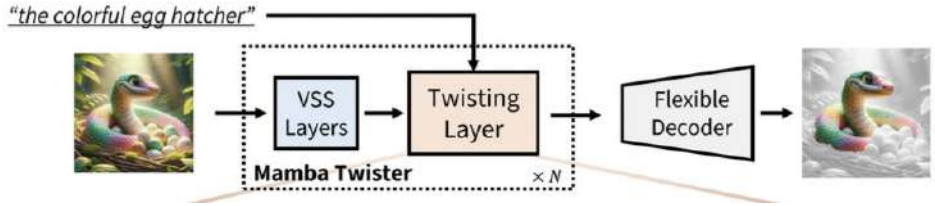
- セグメンテーションや映像認識など画像分類以外のタスクへの応用研究が多数採択

Mamba-ND: 多次元データ向けのSSM, 映像データでの評価
https://github.com/jacklishufan/Mamba-ND

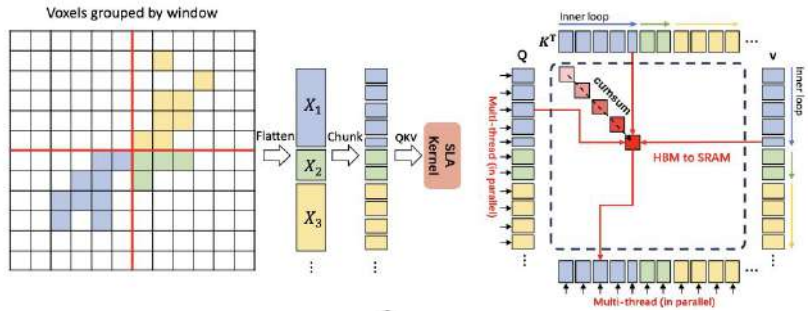


MTMamba: マルチタスク向けのSSM, セグメンテーションと深度推定
https://github.com/EnVision-Research/MTMamba

ReMamber: 参照セグメンテーション向けのSSM
https://github.com/yyh-rain-song/ReMamber



ScatterFormer: 三次元物体認識向けの線形注視機構
https://github.com/skyhehe123/ScatterFormer

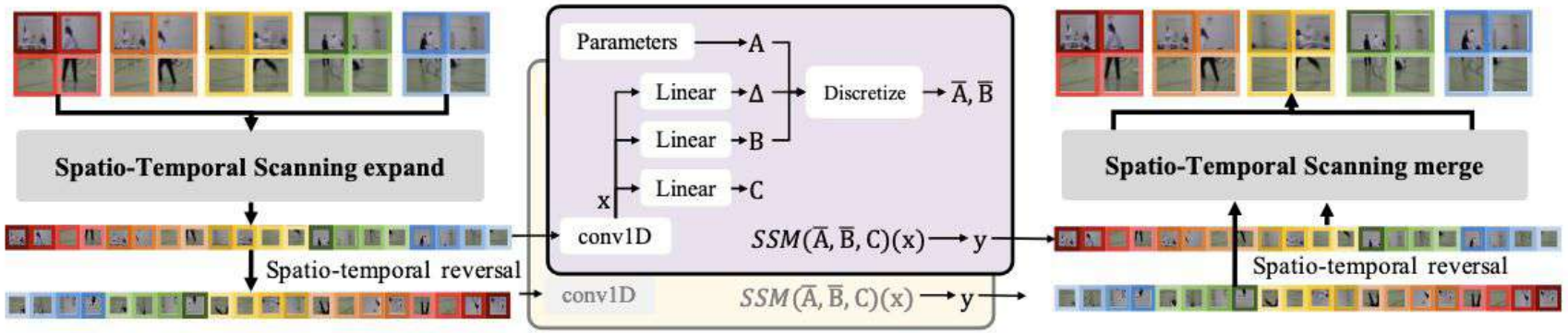


ECCV 2024 の動向・気付き（92/132）

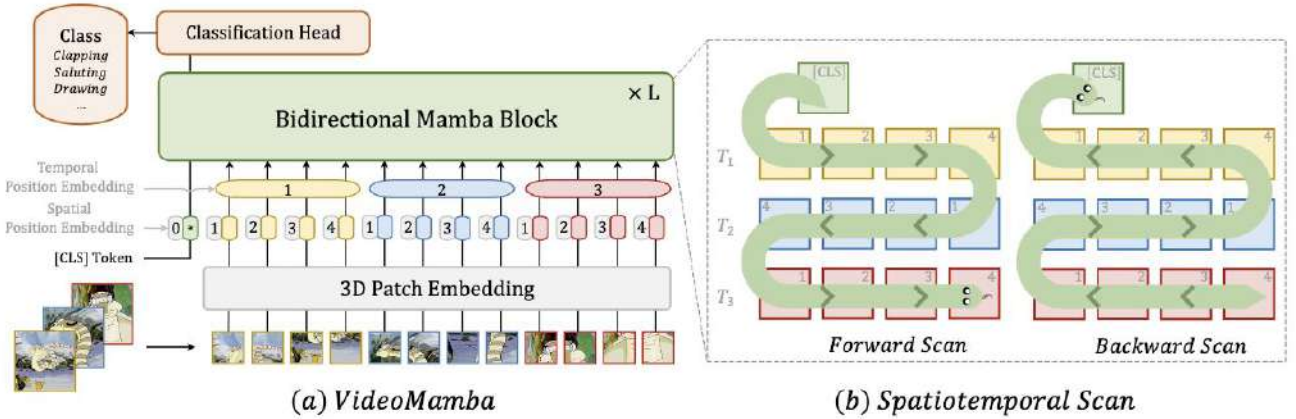
状態空間モデル（SSM）や線形注視機構はCV分野でも注目されている

- 映像処理のための“VideoMamba”という名前のモデルは2件ある
  - どちらも時系列データ向けのspatial-temporal scanningを提案

VideoMamba
(J. Park+)



VideoMamba
(K. Li+)



J. Park, et al., VideoMamba: Spatio-Temporal Selective State Space Model, ECCV, 2024.
K. Li, et al., VideoMamba: State Space Model for Efficient Video Understanding, ECCV, 2024.

ECCV 2024 の動向・気付き（93/132）

Mamba関連の動向

- Mambaの概要
  - 状態空間モデル（State Space Model: SSM）を基にした手法
    - 入力サイズに対する計算容量が線形なので長い入力が可能（Transformerは2次）
    - 以前から注目されていたが、Transformerを性能で上回れなかった
  - 各種改良したMamba [Gu&Dao,2023]が様々なタスクでTransformerを超え一躍注目
    - attentionの代わりにselection mechanismと呼ばれる機能を組み込み
    - Transformerとは得意・不得意があり、最新LLMでは両者を混在して使用（例：Jamba1.5）

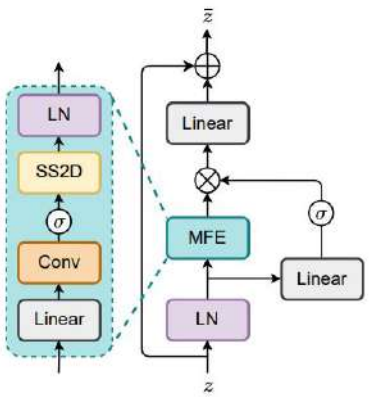
- MambaをベースとしたVision向けの手法や応用は続々提案されている状況
  - Vision Mamba [Zhu et al., 2024]： MambaをVision向けに利用
  - MambaOut [Yu & Wang, 2024]： Vision向けに不要な部分を削除して簡素化・高精度化
  - VSSD [Shi et al., 2024]： 改良されたMamba2[Gu&Dao,2024]をVision向けに利用

- タイトルにMamba（State Space）がある論文は8件
  - 1次元向けMambaにおける多次元データの入力（スキャン）の工夫
  - 各種応用に対する工夫（Transformerベースの手法との比較）

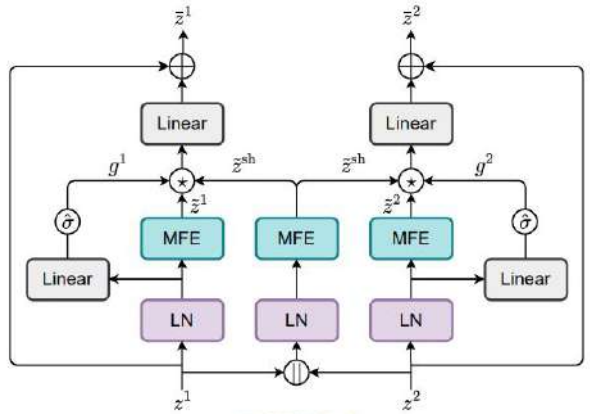
ECCV 2024 の動向・気付き(94/132)

MTMamba: Enhancing Multi-Task Dense Scene Understanding by Mamba-Based Decoders

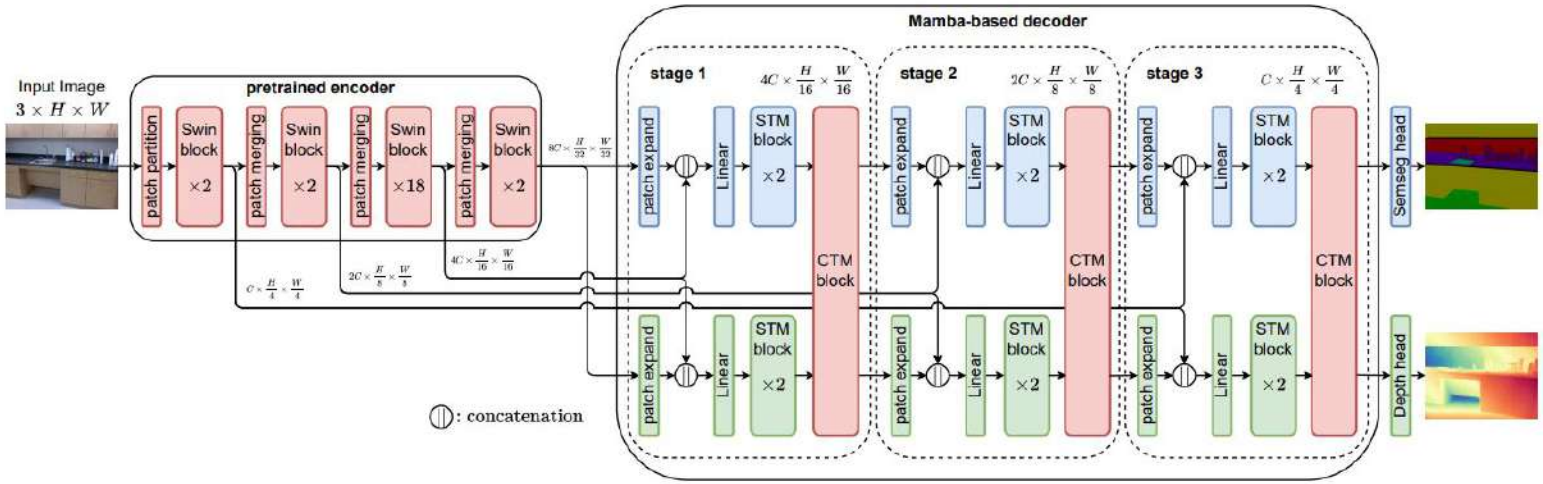
- マルチタスクのシーン理解に対し、Decoder部分に以下の2種のMambaを提案
  - 長距離依存性への対処：self-task Mamba(STM)
  - タスク間の情報共有：cross-task Mamba(CTM)
- Segmentation及びDepth推定で精度向上

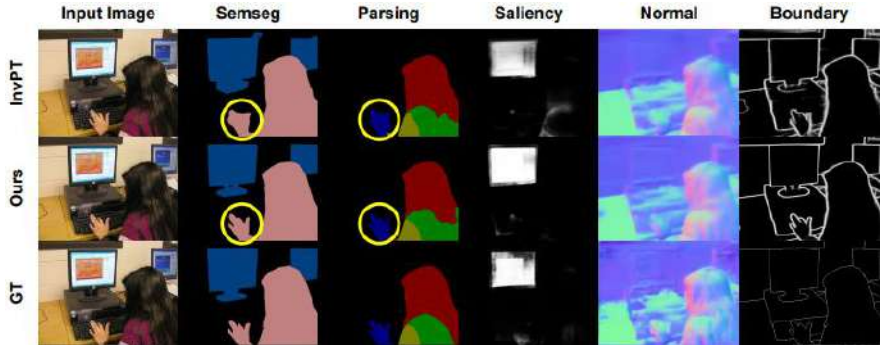


(a) STM block



(b) CTM block

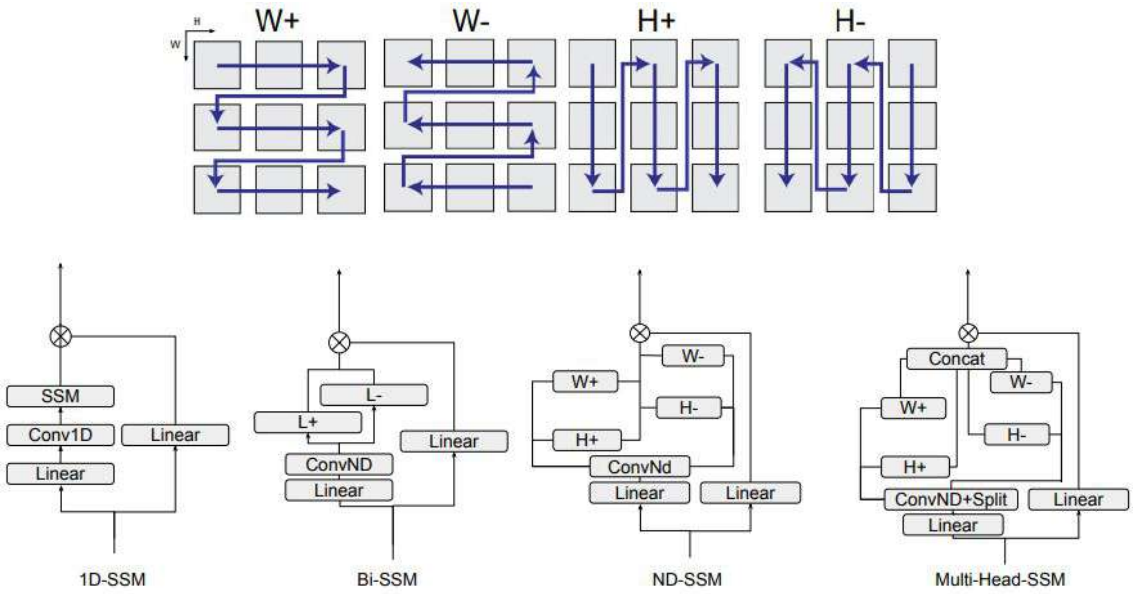


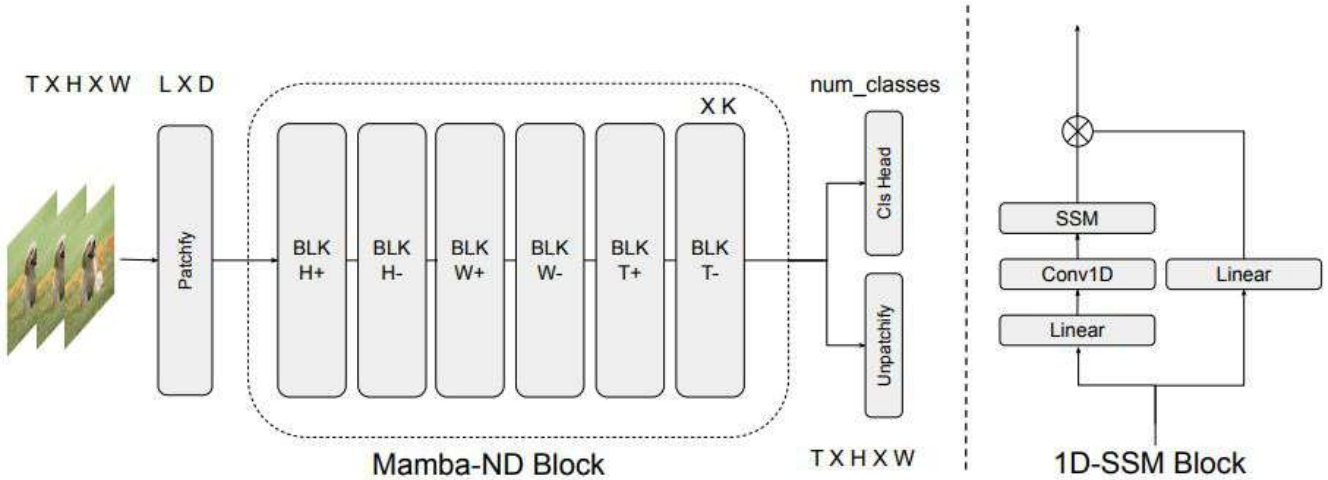


ECCV 2024 の動向・気付き(95/132)

Mamba-ND: Selective State Space Modeling for Multi-Dimensional Data

- 1次元処理をベースとするMambaを任意の多次元データ向けに拡張
- 1次元処理のブロックの処理内容は変更せず、多次元データのスキャン方向(順序)ごとの処理ブロックを並べることで対応
- 各種タスクでCNN・ViTより性能改善

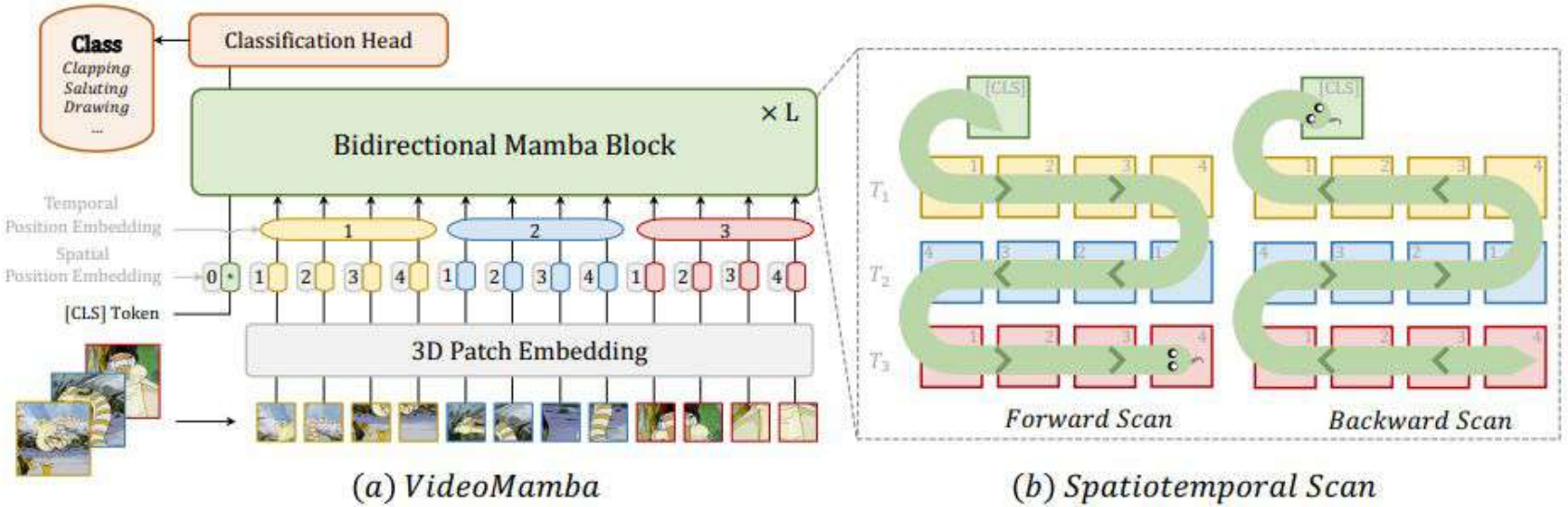




ECCV 2024 の動向・気付き（96/132）

VideoMamba: State Space Model for Efficient Video Understanding

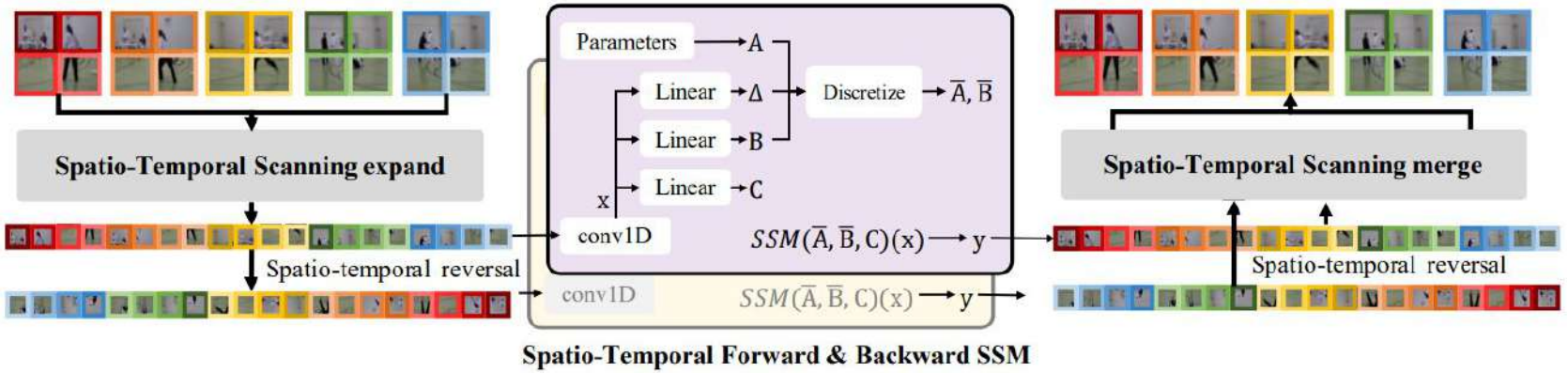
- Mambaを動画像（3Dデータ）向けに改良
  - データ（画像空間×時間）のスキャンは、画像空間×時間順のbidirectionalが一番効果的
    - 観測ブロック位置を固定した時間方向へのスキャン×位置を順番に変更、はあまり効果がない
  - Mambaのサイズを上げると過学習しやすい
    - 一段小さいサイズのMambaを用いた自己蒸留で対処
  - 計算容量が入力サイズに線形なので入力の解像度を上げやすく、上げるほど性能向上
    - 多くのタスクでVideoMambaがVideo向けCNNやTransformerを上回る



ECCV 2024 の動向・気付き（97/132）

VideoMamba: Spatio-Temporal Selective State Space Model

- Mambaを動画像（3Dデータ）向けに改良（手法名は前ページと同じ）
  - こちらの文献でもデータ（画像空間×時間）のスキャンは、画像空間×時間順のbidirectionalが一番結果が良好
  - 各種データでの検証でCNNやTransformerベースの手法に対しトップ若しくはそれに近い性能を提供
    - 必ずしもトップではないが、モデルサイズはこちらの方が小さい
    - 既にMamba-NDを引用・比較済み。こちらの方が結果が良い



Spatio-Temporal Forward & Backward SSM

Table 7: Comparison with previous work on HMDB51. ‡ denotes results from [29] and † is reproduced number for fair comparison. Magnitudes are Mega (10^6) for Param. The subscript denotes the trained epoch of the model. "N/A" indicates the numbers are not available for us.

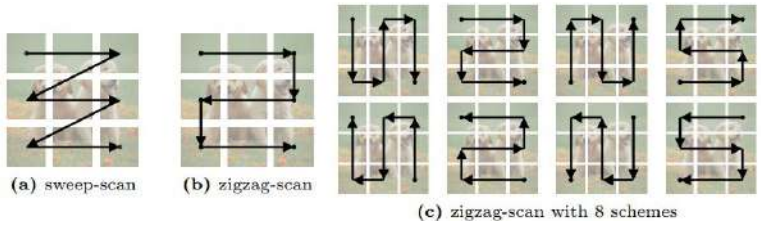
Method	Backbone	Pretrain	Frames	Param	Top-1
I3D [6]	Inception	IN-1K	30	25.0	49.8
SpeedNet [3]	S3D-G	K400	64	9.0	48.8
VTHCL [52]	SlowOnly-R50	K400	32	32.0	67.9
MemDPC [19]	R-2D3D	K400	40	32.0	54.5
CVRL [36]	SlowOnly-R50	K400	32	32.0	49.2
VideoSwin-T† [33]	Swin-T	IN-1K	32	27.9	54.4
VideoSwin-T† [33]	Swin-T	K400	32	27.9	69.9
VideoSwin-S‡ [33]	Swin-T	IN-1K	32	54.0	58.1
VideoMAE$_{\text{4800e}}$ [40]	ViT-B	-	16	87.0	62.6
VideoMAE$_{\text{4800e}}$ [40]	ViT-B	K400	16	87.0	73.3
S4ND-ConvNeXt-3D‡ [35]	ConvNeXt	IN-1K	30	29.0	55.2
Mamba-ND‡ [29]	Mamba	IN-1K	32	36.0	59.0
VideoMamba	Mamba	IN-1K	16	26.3	58.9
VideoMamba	Mamba	IN-1K	32	26.8	59.3
VideoMamba	Mamba	K400	16	26.3	68.6
VideoMamba	Mamba	K400	32	26.8	75.7

ECCV 2024 の動向・気付き（98/132）

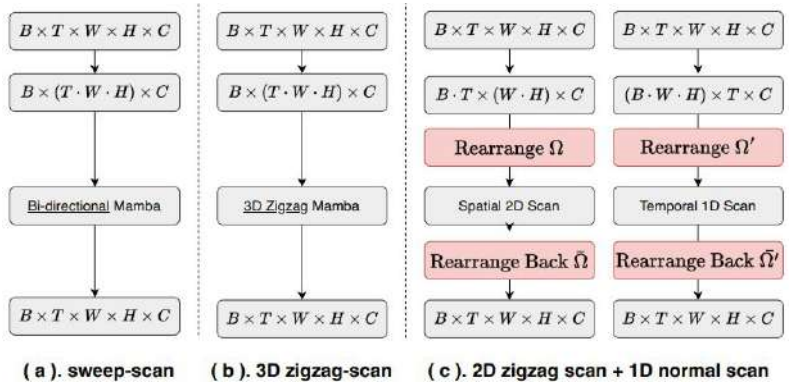
ZigMa: A DiT-style Zigzag Mamba Diffusion Model

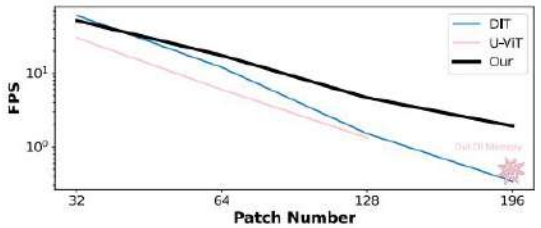
- Mambaを拡散モデルに適用
  - 多次元データのスキャンに対し、位置関係の帰納バイアスを取り込む為、一般的なsweepではなく、位置が連続したzigzagを複数スキーム（順番・経路）で適用
  - Transformerベースより、速度とメモリ使用量を改善

2D（Image）

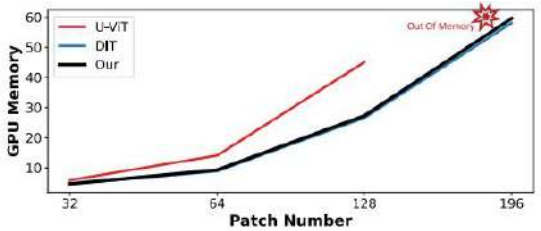


3D（Video）





(a) FPS *v.s.* Patch Number.

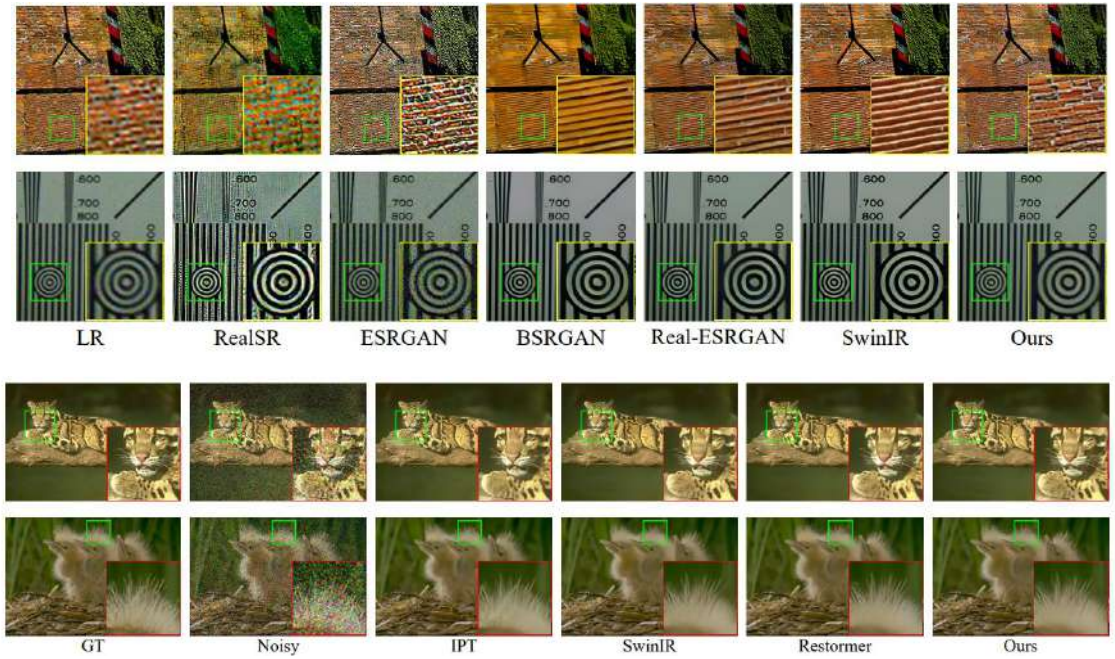


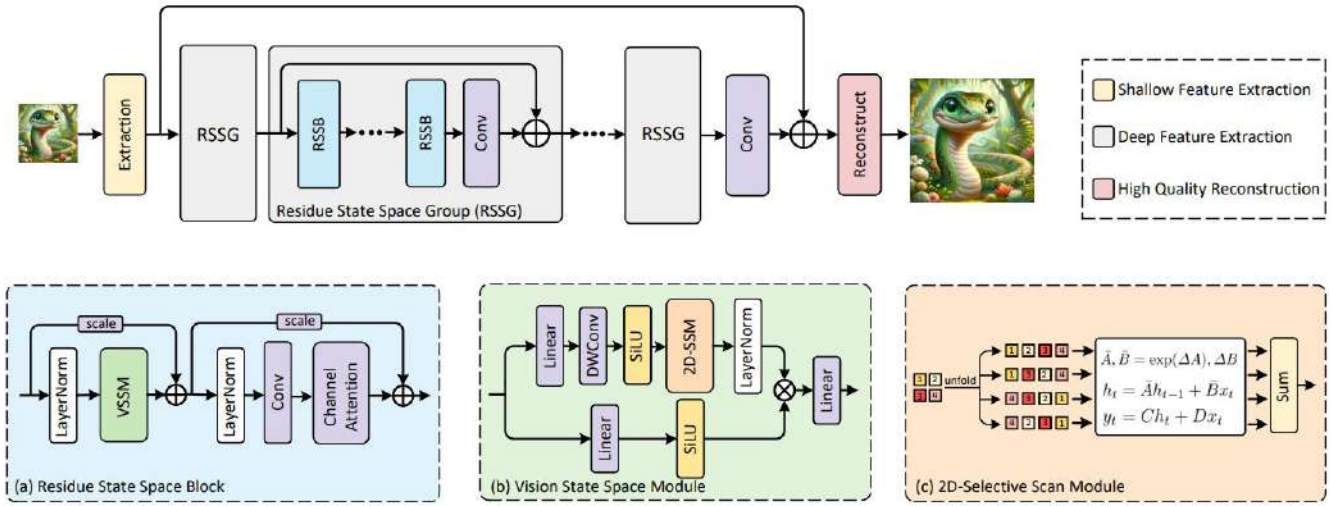
(b) GPU Memory *v.s.* Patch Number.

ECCV 2024 の動向・気付き(99/132)

MambaIR: A Simple Baseline for Image Restoration with State-Space Model

- Mambaを画像復元(超解像・ノイズ除去)に適用
  - 復元性能は受容野の大きさに依存するが、CNNでは小さく、Transformerでは大きく出来るが計算が2次の複雑さになるので、線形で抑えつつ大きくしやすいMambaを利用
  - 但し大きすぎると局所情報を忘却するので、Residual処理、畳み込み、チャネル間の冗長性を抑制するチャネルアテンションを追加したResidual State Space Block (RSSB) を提案
  - ほぼ全ての検証でCNNベース・Transformerベースの手法を凌駕

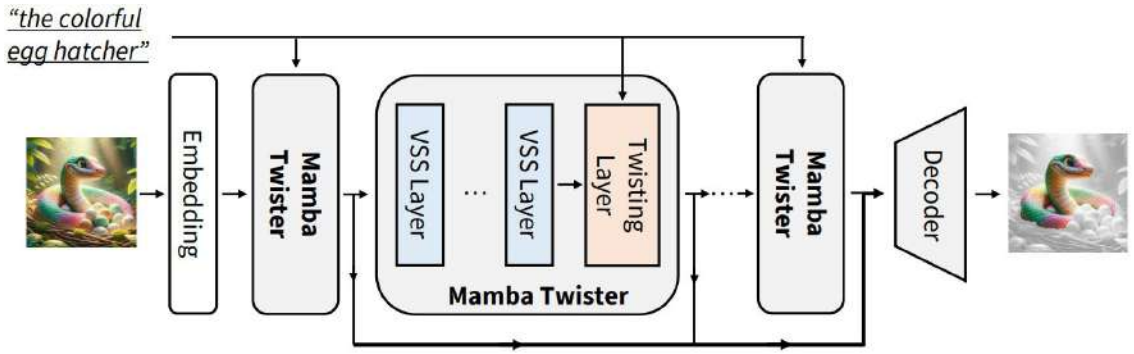


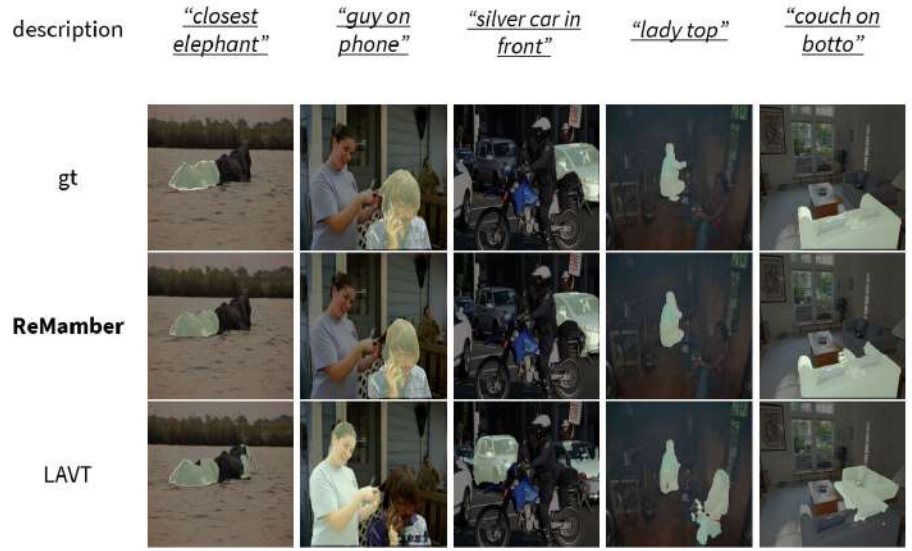


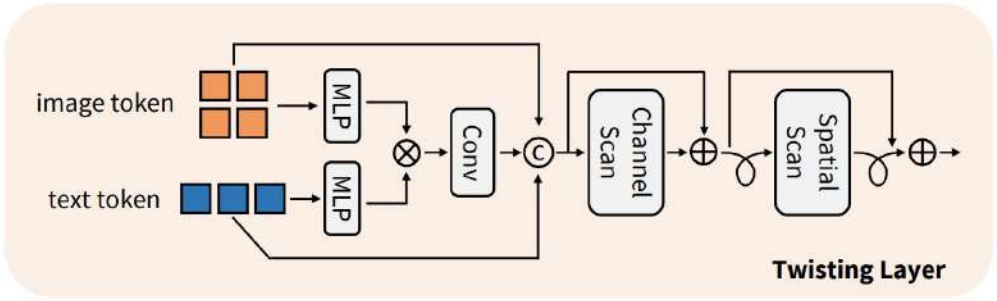
ECCV 2024 の動向・気付き（100/132）

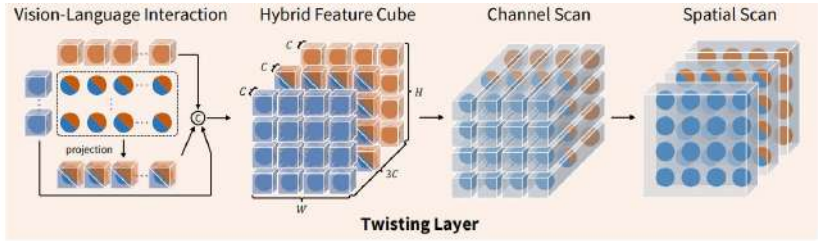
ReMamber: Referring Image Segmentation with Mamba Twister

- 参照（テキストガイド）image segmentationのタスクにMambaを適用
- Image-textのマルチモーダル向けに改良
  - Twisting Layer：image tokenとtext tokenを共有空間にマッピングし、計3つの特徴（token）でfeature cubeを作成、各scanを実施
- 各評価でTransformerベースの手法超え







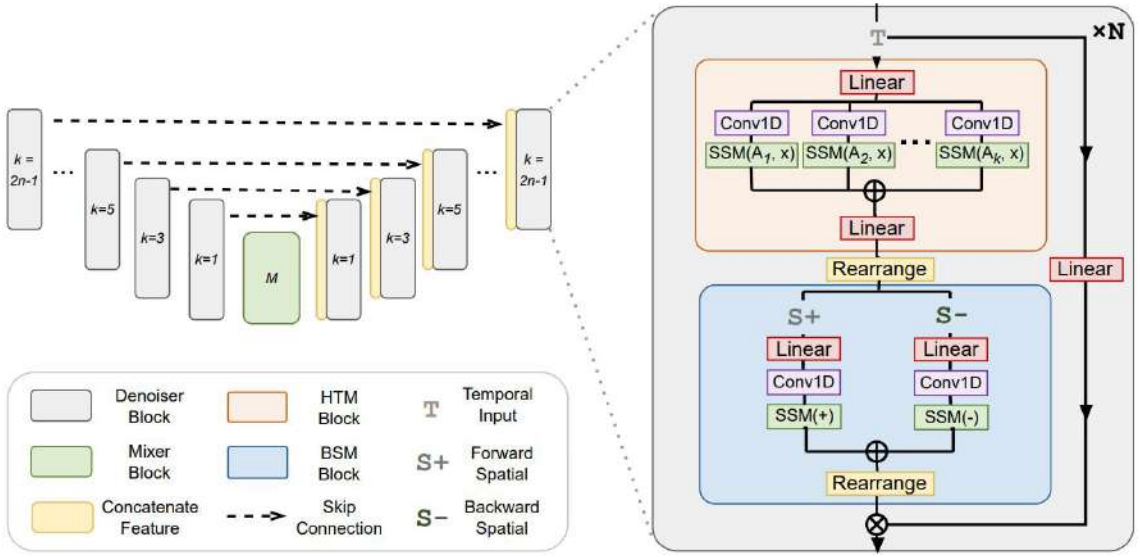


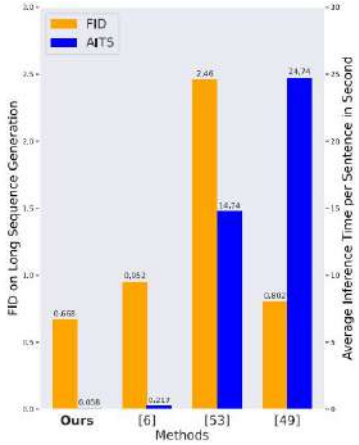
処理のイメージ

ECCV 2024 の動向・気付き（101/132）

Motion Mamba: Efficient and Long Sequence Motion Generation

- 人物動作などのモーション生成に目途とした拡散モデルにMambaを適用
  - Long-range dependencyの把握に優れる
- U-net構造の各ブロックに以下を利用
  - Hierarchical Temporal Mamba (HTM)
    - Linearで潜在空間に落とす方がより運動に関する情報が表現できる
  - Bidirectional Spatial Mamba (BSM)
    - 上位（k=1）ほどスキャン数を多く、下位（k=2n-1）ほどスキャン数を少なくする
- HumanML3Dなどで、FIDで最大50%の削減、推論時間が4倍高速



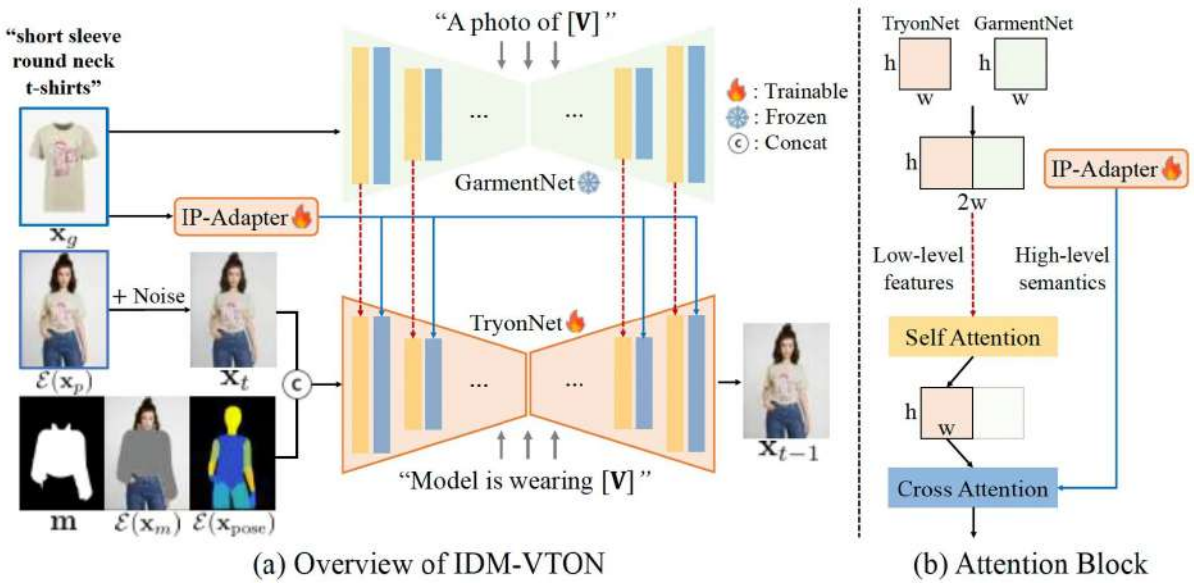


ECCV 2024 の動向・気付き（102/132）

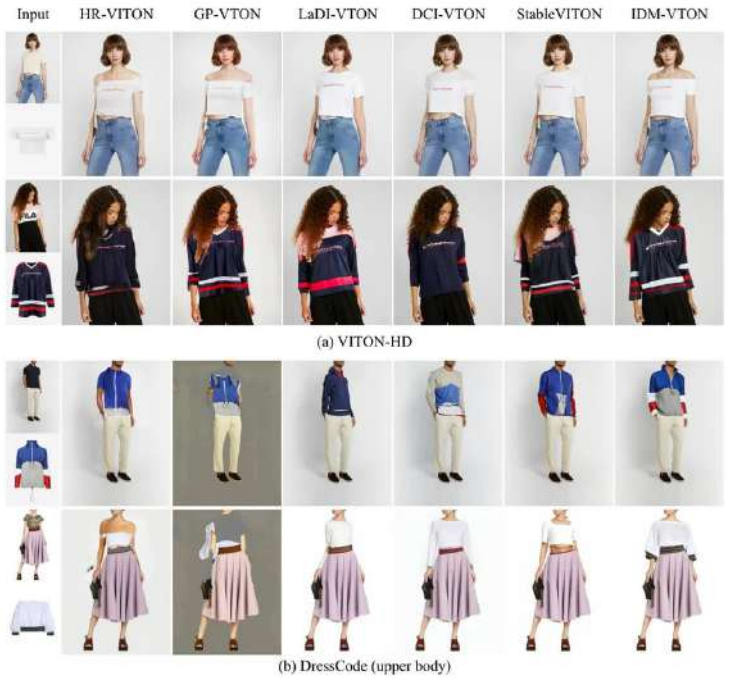
IDM-VTON: Improving Diffusion Models for Authentic Virtual Try-on in the Wild

- 概要：TryonNetとGarmentNetという二つのUNetを用いて，VTONタスクを解くことを提案した．さらに，IP-Adapterを用いて，既存研究よりさらに特徴量をinjectionすることに成功．また，デザイナーさんの方々によって行われた服のannotationを用いることで，服の質感を考慮したより実用的な問題まで解決した．

- ポイント：既存のVTON modelは主に，GarmentNetを学習することで，服の再構成精度を上げることを試みていたが，今回のモデルでは，着衣部分をFine-tuningを行うことで精度が上がることを実証した．



(a) Overview of IDM-VTON　　(b) Attention Block



(a) VITON-HD

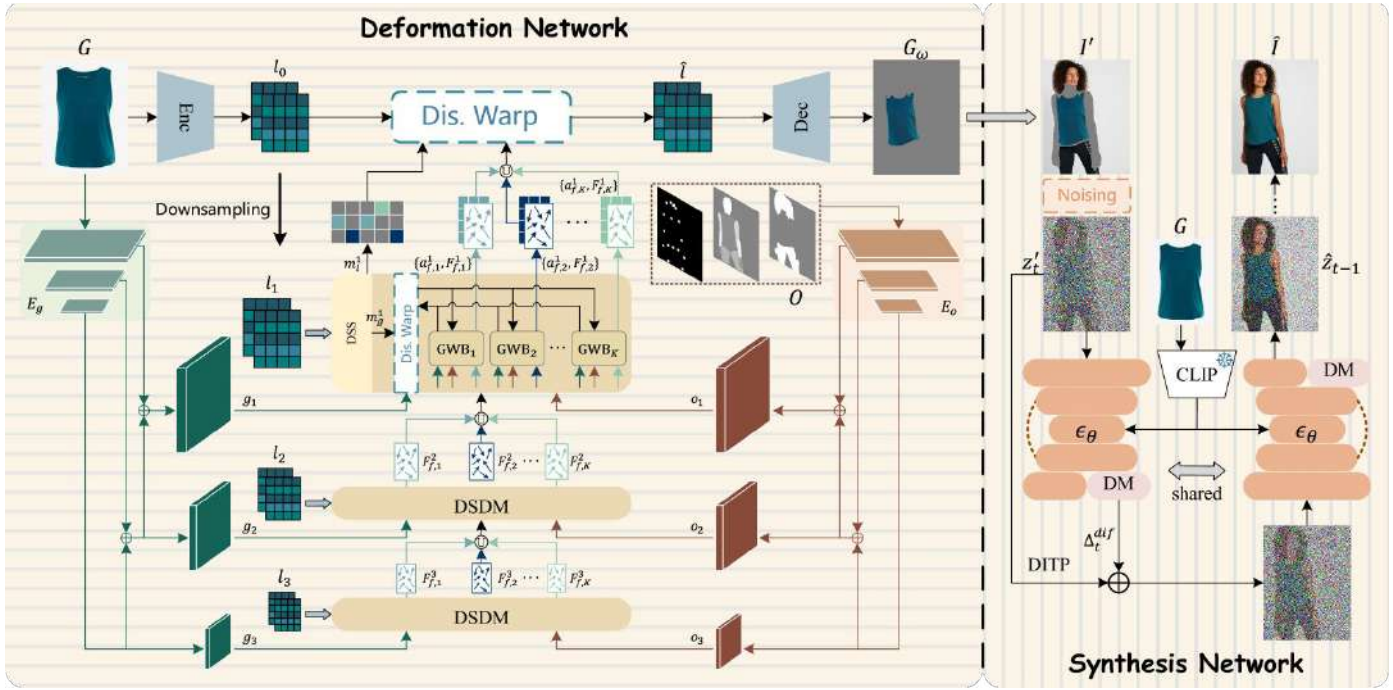
(b) DressCode (upper body)

ECCV 2024 の動向・気付き（103/132）

D4-VTON: Dynamic Semantics Disentangling for Differential Diffusion based Virtual Try-On

- 概要：衣服変形後のsemanticな一貫性の欠如や、annotaion-bsedの衣服parserへの依存を解消し、diffusion modelにおけるノイズ除去やInpaintingの同時処理の課題に対処した論文

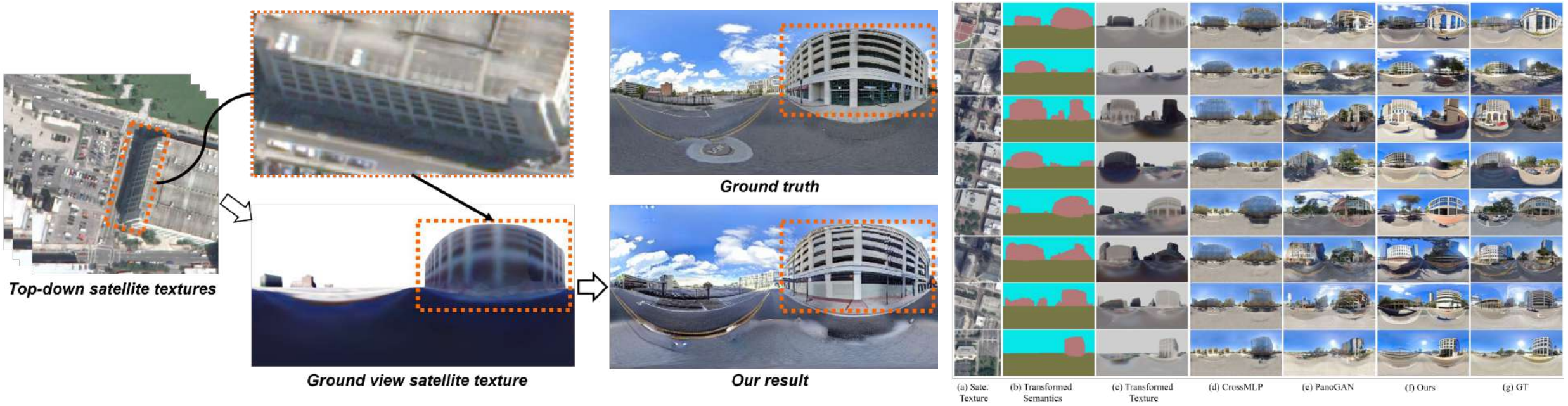
- ポイント：Dynamic Semantics Disentangling Moduleを用いることで，衣服のセマンティック要素の自動抽出を可能にし，局所的な衣服の変形の精度を上げる．さらに，Differential Information Tracking Pathを用いて，inapinting taskとDenosing taskを分離処理し、学習の曖昧さを軽減



ECCV 2024 の動向・気付き（104/N）

Geospecific View Generation – Geometry-Context Aware High-resolution Ground View Inference from Satellite Views

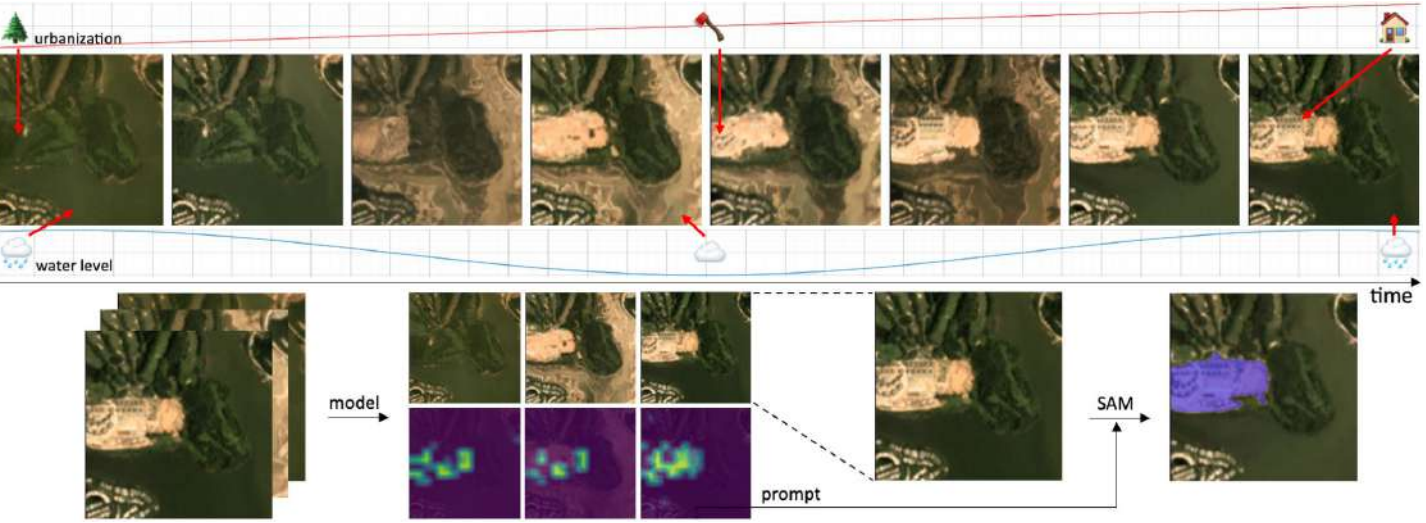
- 衛星画像から地上の360度画像を推定する手法
  - 衛星画像から建物に張り付けたテクスチャでは、地上で肉眼で見た場合との差が大きい
  - 拡散モデルを使って、衛星画像の建物テクスチャ画像をリファインする
  - 隣接する場所間での整合性が取れていないので、将来の課題となっている

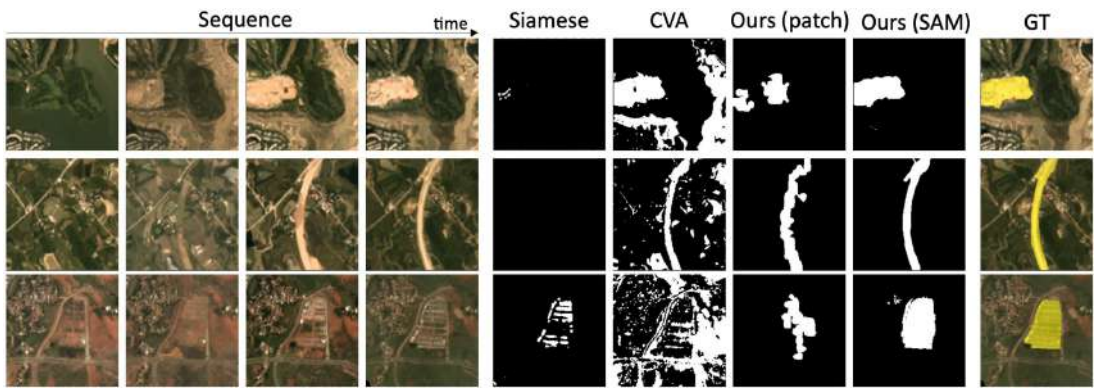


ECCV 2024 の動向・気付き（105/132）

Made to Order: Discovering monotonic temporal changes via self-supervised video ordering

- 衛星画像などの時系列動画像データから季節変化を排除した変化検出する手法
  - 衛星画像のような時系列な動画像から変化検出する場合には、建物の有無などの変化の他に季節的な変動が邪魔になる
  - 季節変動以外の変化を単純な変化とし、単調な時間的変化を時間の経過とともに一方向にのみ（増加または減少）発生する変化として定義すると、自己教師あり学習が可能になる
  - 元々は時系列に並んだ画像を時間的にシャッフルし、それを画像の順序付けという、シンプルな自己教師あり学習を行うと、単調な変化検出が可能になる





ECCV 2024 の動向・気付き(106/132)

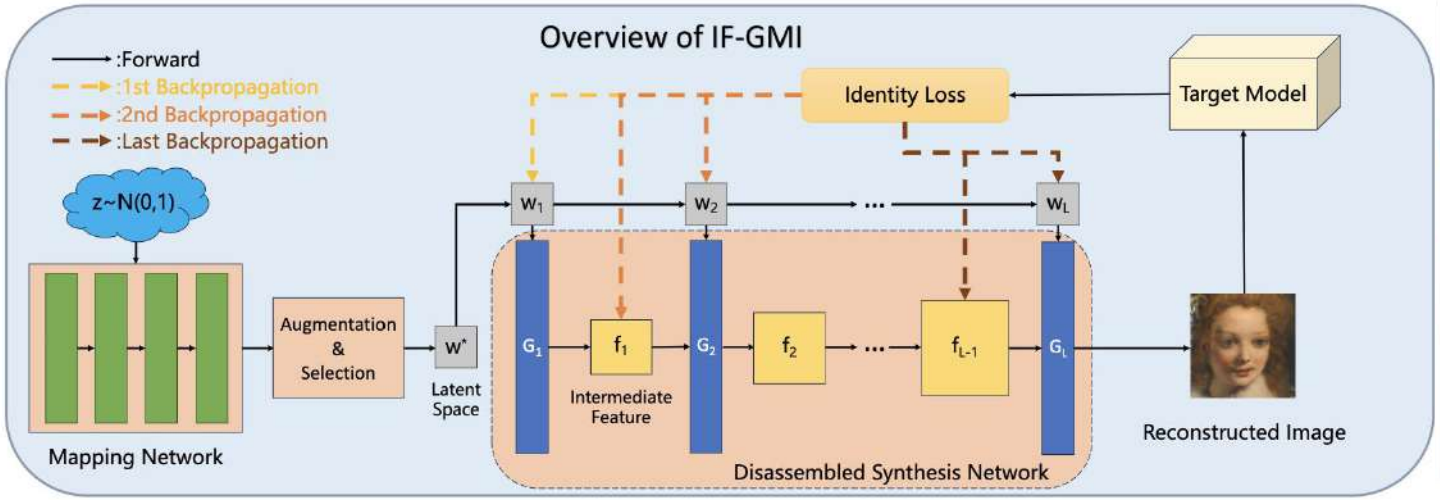
A Closer Look at GAN Priors: Exploiting Intermediate Features for Enhanced Model Inversion Attacks

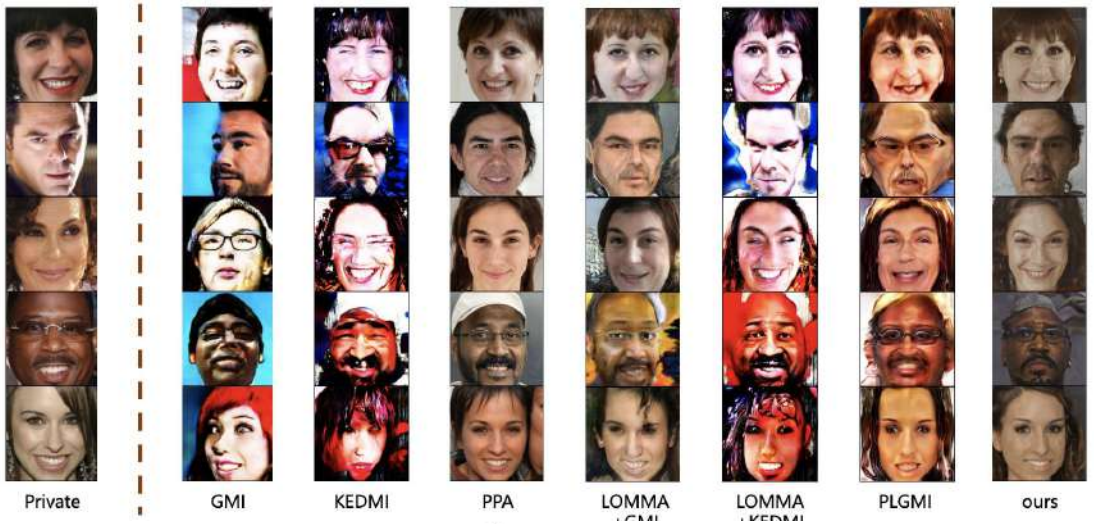
- 概要

  生成モデルを事前学習するデータと攻撃対象のモデルが実際に学習したデータのズレにロバストなModel Inversion Attack手法IF-GMIを提案.

- ポイント

  従来の潜在変数の最適化に加えて，Style-GANの中間特徴量も初期層から最終層にかけて段階的に最適化する.





ECCV 2024 の動向・気付き(107/132)

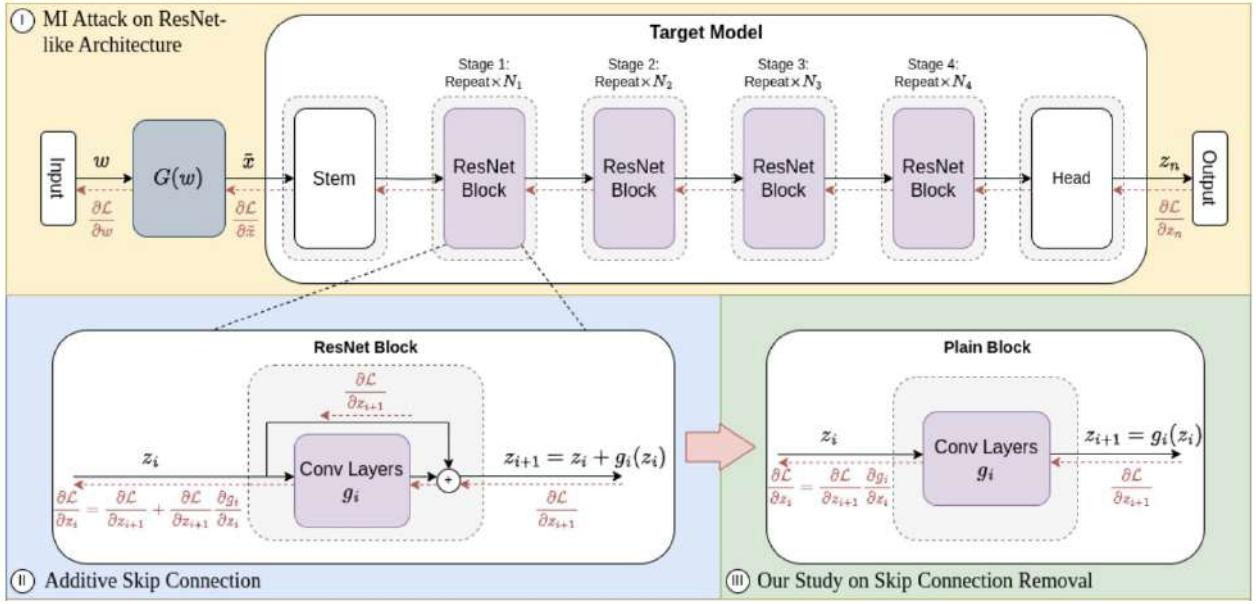
On the Vulnerability of Skip Connections to Model Inversion Attacks

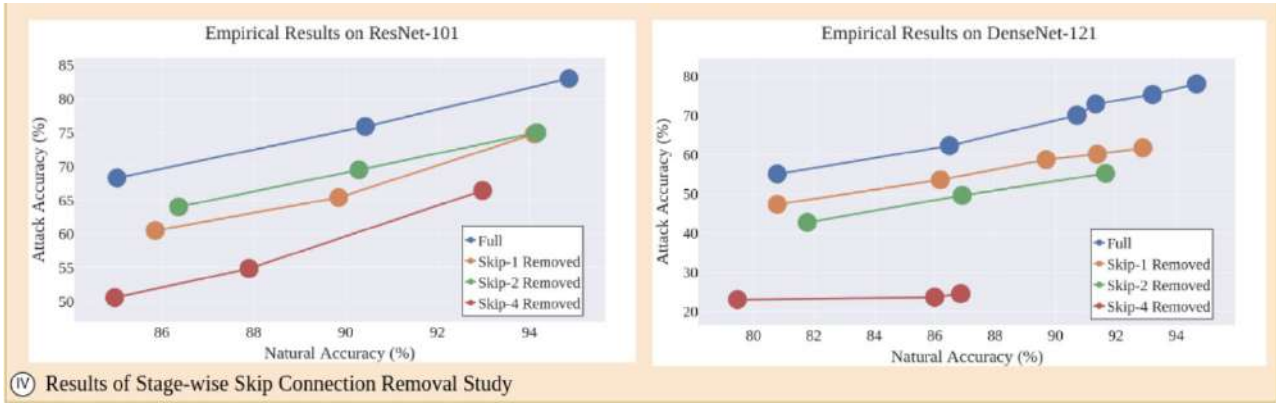
- 概要

  Model Inversion Attack(MIA)の防御手法に関する研究の中でもDNNのアーキテクチャに焦点を当てて分析を行い防御手法を提案した研究.

- ポイント

  最終層へのSkip Connectionを削除するRoLSS, その他のSkip Connectionの加法時に1より小さい値でスケーリングするSSF, 学習初期では全てのSkip Connectionを維持したまま学習し, その後RoLSSを適用する方法を提案している





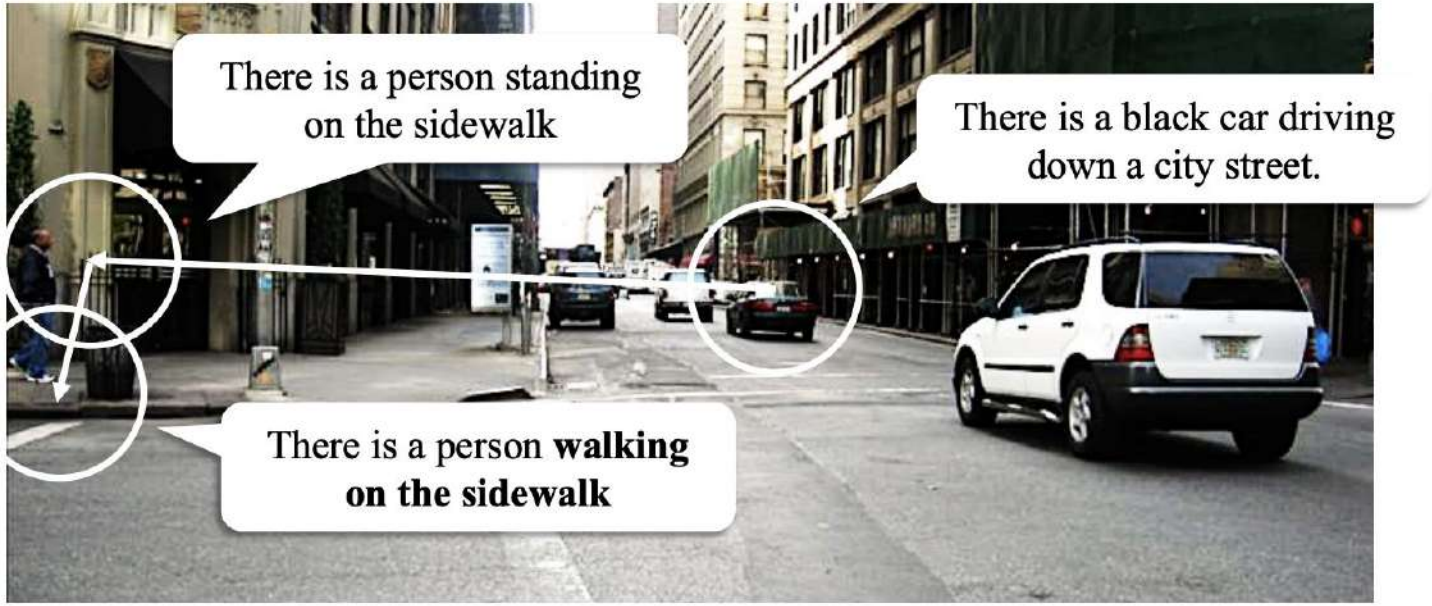
ECCV 2024 の動向・気付き（108/132）

GazeXplain: Learning to Predict Natural Language Explanations of Visual Scanpaths（Oral）

- 概要：Gazenのshiftとそのexplanatioを同時に推定するタスク・データセット・手法の提案。
- ポイント：LLMを利用して、Gazeのshiftルートの詳細解釈を与えた。
- 感想：Gazeの shiftは人間の認識・考えのプロセスを示せる。人間の考えるプロセスの理解、意図の推定やミスの分析など色々なところで使える。また、Gazeのようなデータが比較的に研究が多くない。Gazeのみではなく、シーンを観測する際に、人間の姿勢・ポーズや手の姿勢なども、人がどのように周りの情報を理解しているのに使えそう。

Q: Does the person on the sidewalk appear to be walking?

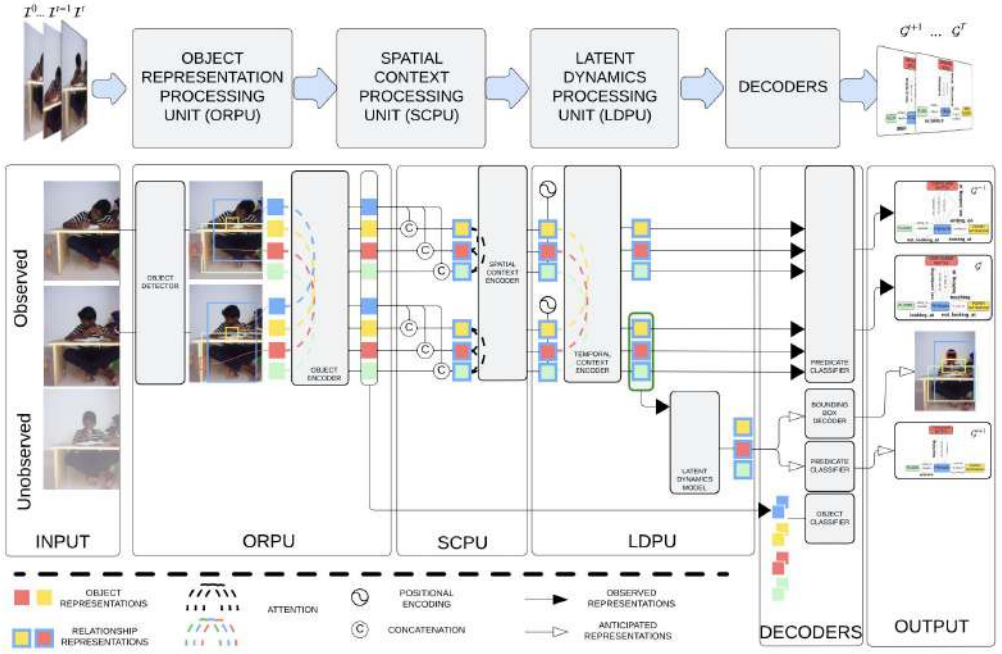
A: Yes



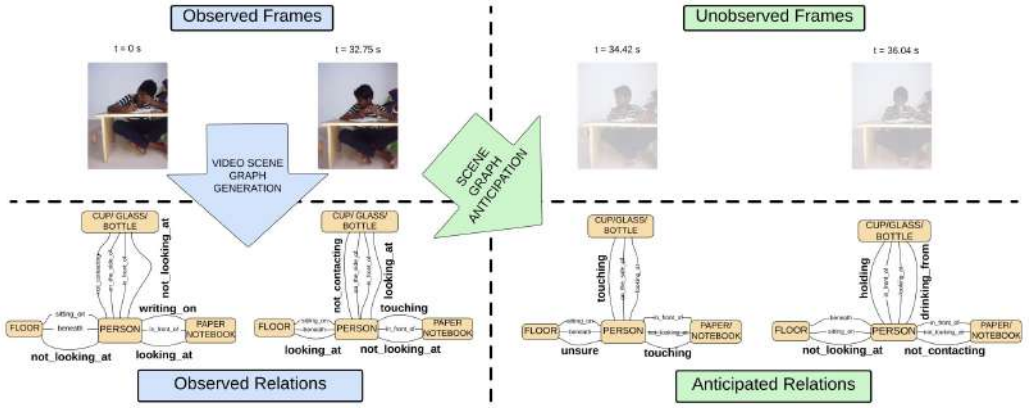
ECCV 2024 の動向・気付き（109/132）

Towards Scene Graph Anticipation（Oral）

- 概要：過去のビデオから未来の物体間関係を推定するタスクを提案。また、NeuralODEとNeuralSDEをベースとした物体関係性のLatent dynamicsを学習する手法を提案。複数のレイヤーを使って段階的に物体間の関係性を推定。
- ポイント：Scene Graph Anticipationタスクの提案。
- 感想：未来の情報を構造化した知識で予測することが面白い、生成系も構造化知識の使用が多くなってきたイメージ。展開として、Out-of-viewのScene Graphの予測や、音声を使った予測もできそう。



提案手法

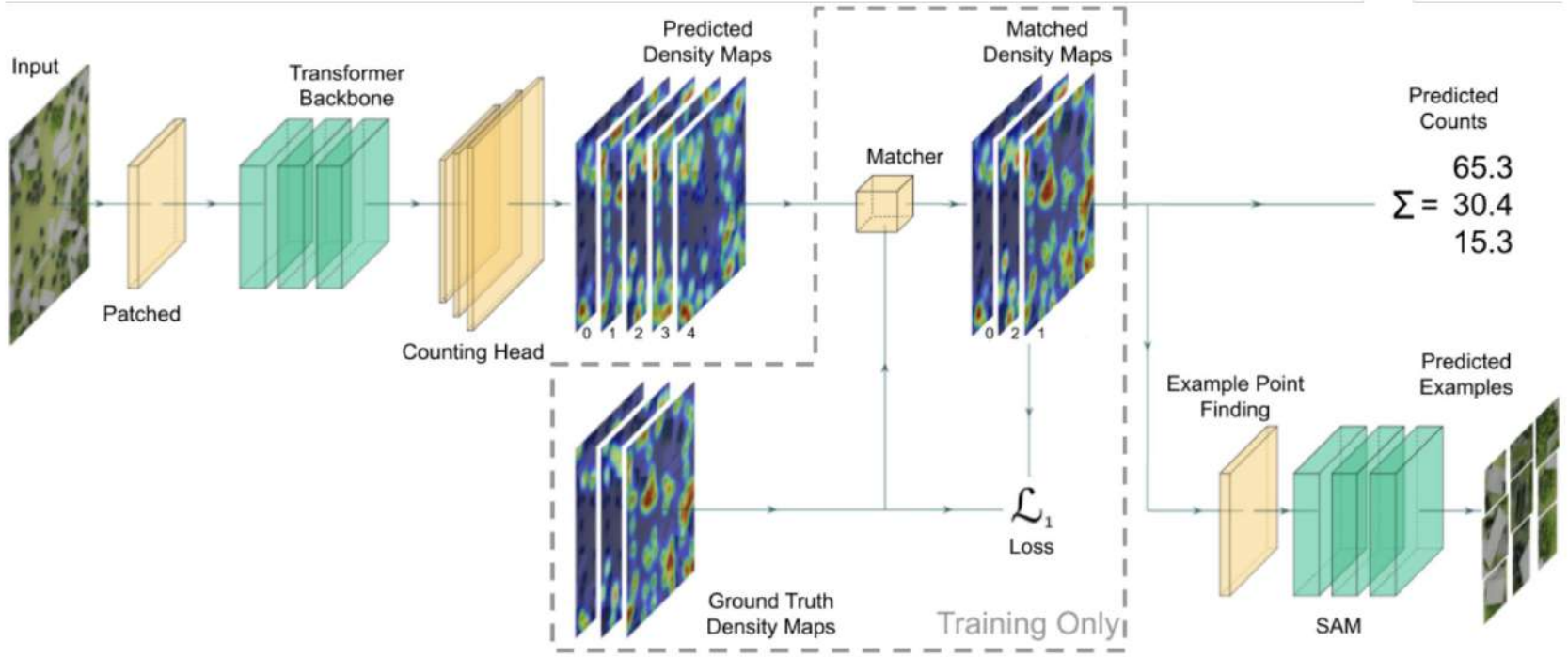


タスク

ECCV 2024 の動向・気付き（110/132）

ABC Easy as 123: A Blind Counter for Exemplar-Free Multi-Class Class-agnostic Counting（Oral）

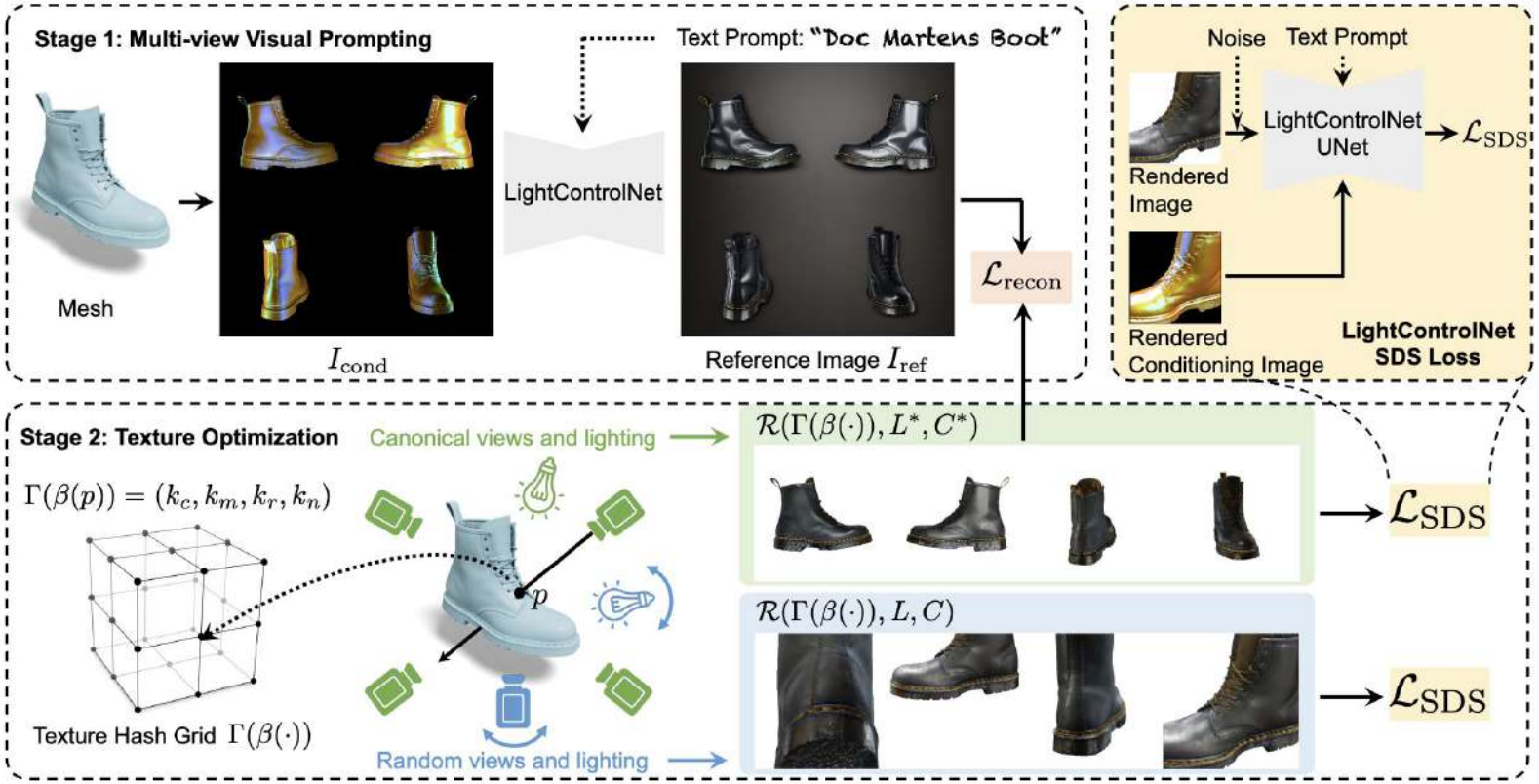
- 概要：Exemplar-Freeで画像から複数のクラスの物体をCountingする手法ABC123を提案。ABC123がまずクラスごとのdensity mapを推定し、その後クラスごとの数をintegrateする。高質なデータセットがないため、大規模Simulation Countingデータセットも提案。大規模データで学習した結果、ABC123がUnknownのクラスのカウンティングでも高い精度。
- ポイント：Countingタスクをシンプルなモデルで高質な解け方。
- 感想：言語と組み合わせていろんな応用ができる。Countingと似たように、画像から自動的に視覚ベースな計算・推理ができるモデルが重要で面白そう。パーツとして現在のMLLMsにつけると応用が広がりそう。



ECCV 2024 の動向・気付き（111/132）

FlashTex: Fast Relightable Mesh Texturing with LightControlNet (Oral)

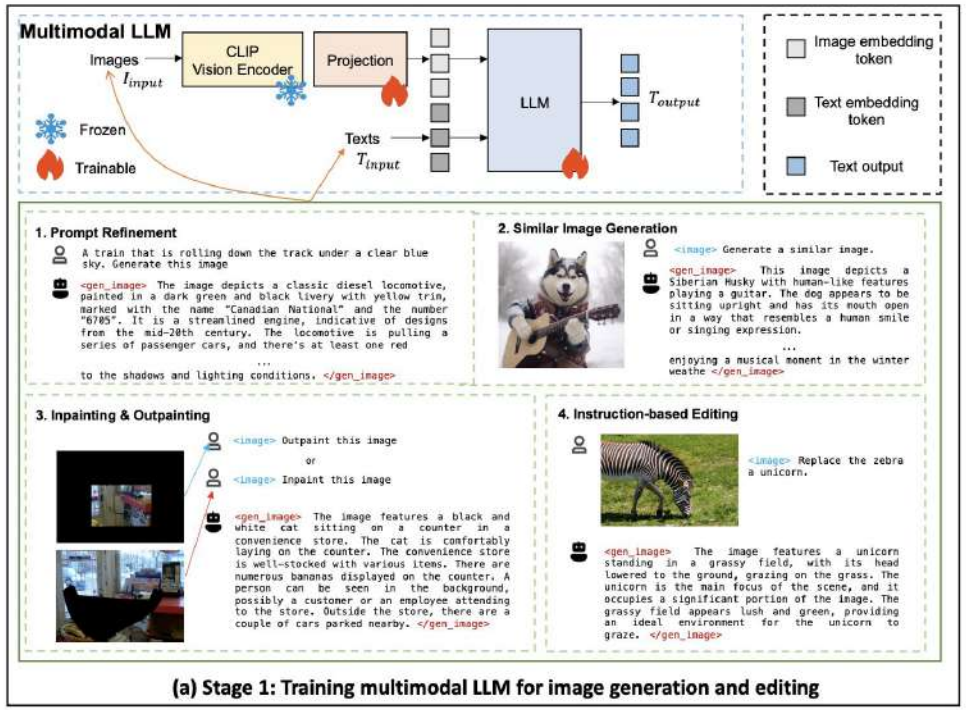
- 概要：テキストから3Dメッシュのテクスチャーマップを生成する手法の提案。特に、surface materialからlightingを独立し、さまざまな環境光の元の3Dモデルをリアリティで生成可能。提案手法がControlNetをベースにしている。まず多視点のConsistencyを考慮した多視点のテクスチャー生成を行う。次に、Score Distillation Samplingを使って、テクスチャーの最適化を行う。
- ポイント：Lightingに着目し高精度の結果を達成したところ。
- 感想：Disentangled的な精密な生成の手法が多くなってきた。いくつかの小さいテクニックを使った印象。



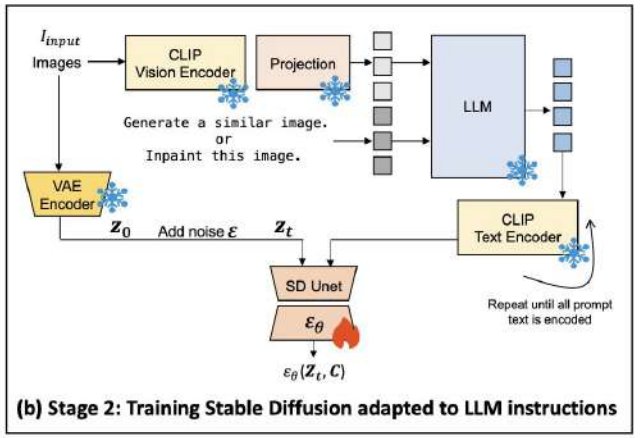
ECCV 2024 の動向・気付き（112/132）

LLMGA: Multimodal Large Language Model based Generation Assistant (Oral)

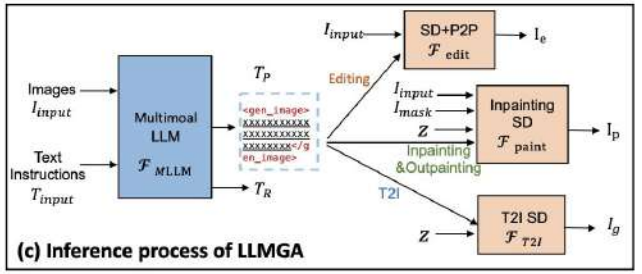
- 概要：画像生成と編集のためのツール的な手法を提案。ユーザと生成モデルの間を介入し、LLMに含まれる知識とReasoning能力を使って、ユーザーの入力を生成モデルが生成しやすいように詳細的な言語Promptを生成しながら、画像を生成する。複数のタスクで提案のモデルが使える（図の左下）。2段階の手法を提案。段階1では、詳細的なテキストと画像のアラインメントを学習。段階2では、生成モデルの再学習を行なって、詳細Promptからの画像生成を可能にした。
- ポイント：生成系のモデルの精密コントロールをツール化。
- 感想：MLLMや生成モデルをより解釈できるようにする論文が多くなってきた。



(a) Stage 1: Training multimodal LLM for image generation and editing



(b) Stage 2: Training Stable Diffusion adapted to LLM instructions

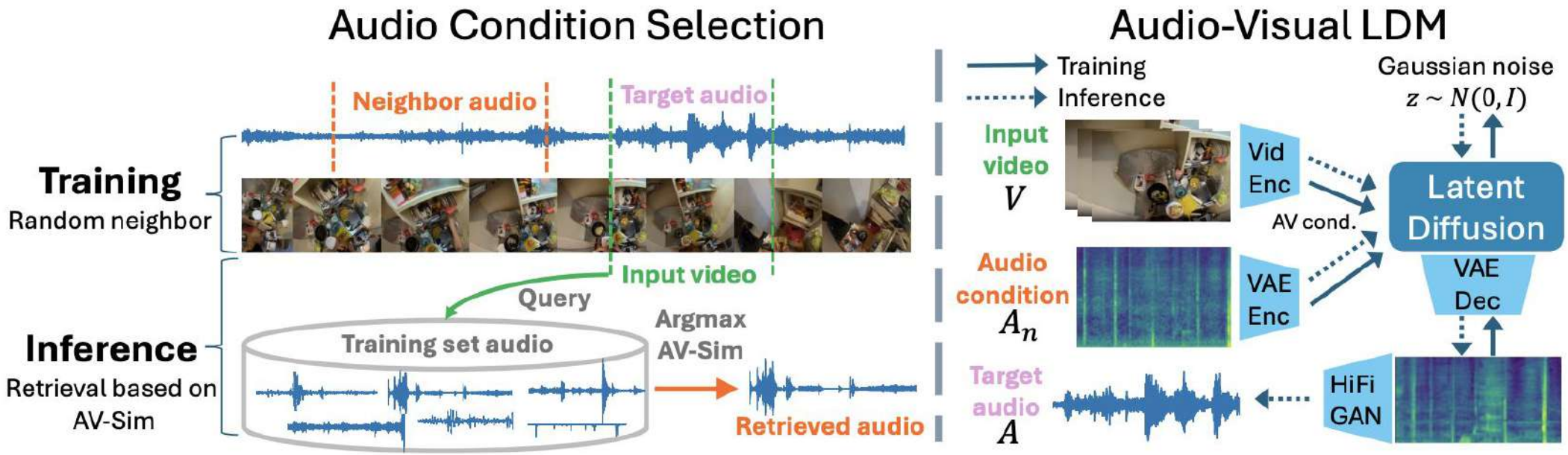


(c) Inference process of LLMGA

ECCV 2024 の動向・気付き（113/132）

Action2Sound: Ambient-Aware Generation of Action Sounds from Egocentric Videos (Oral)

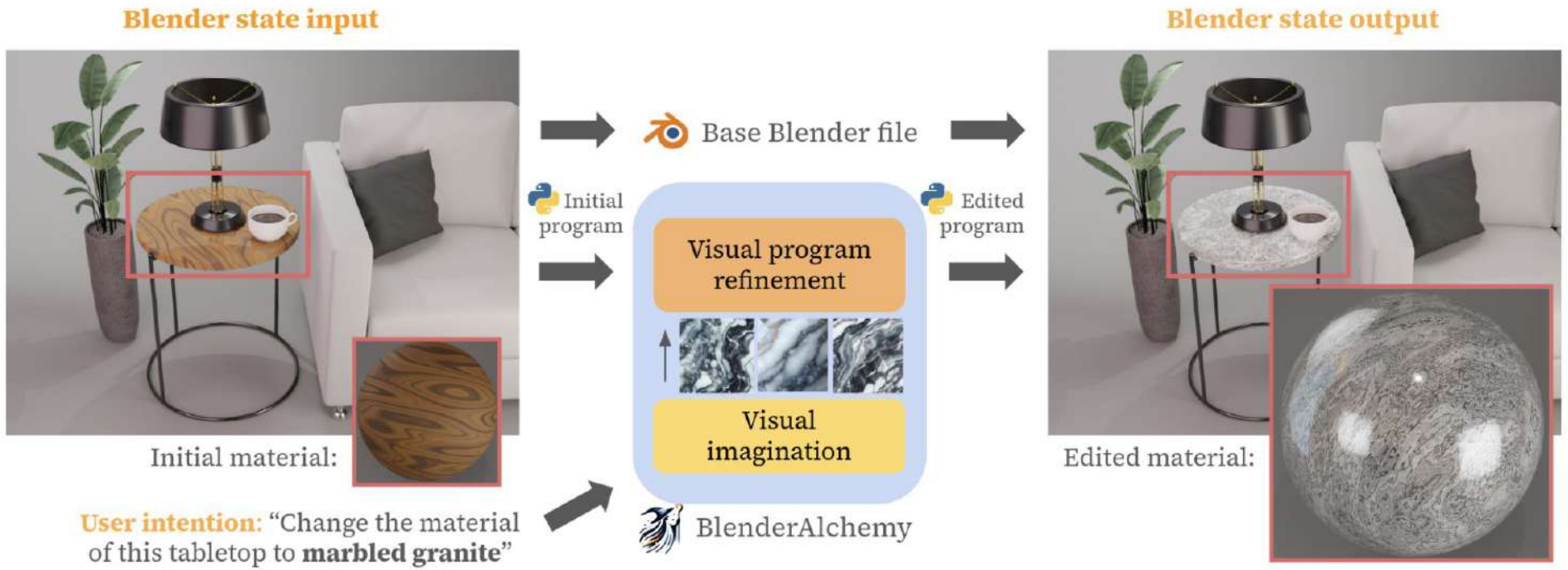
- 概要：Retrieval-augmented generation（RAG）ベースのSilentビデオから音声を生成するDiffusionモデルの提案。音生成する際に同じビデオの他の時間帯の音声もConditionに入力し音をReconstructionして、元のビデオに含まれる背景音の影響を緩和できる。推定段階では、まずビデオから音声をRetrievalする。得られた音声をDiffusionモデルのConditionとする。またモデルの学習するために、Ego4Dデータセットをベースとし、でEgo4D-Soundsを提案。
- ポイント：シンプルで賢い手法でAmbient音の影響を受けずにビデオからの音生成ができた。
- 感想：拡張としては、見えていない部分の音も生成できると面白そう。



ECCV 2024 の動向・気付き（114/132）

BlenderAlchemy: Editing 3D Graphics with Vision-Language Models

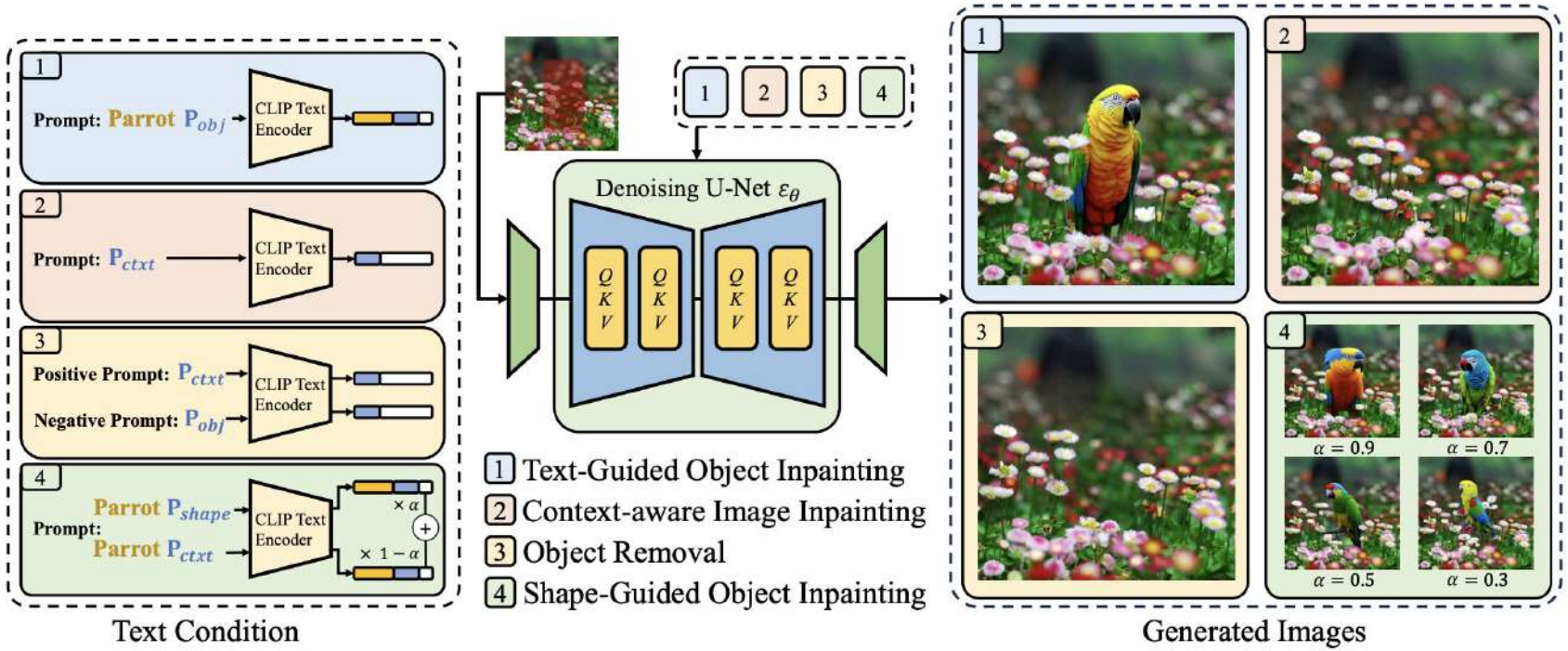
- 概要：テキスト入力とBlenderのコードから、テキストのDemandに応じてコードを編集するタスクと手法を提案。編集後の画像をDiffusionモデルで生成し、編集後の画像が段階的にDiffusionが想像した画像に近づけるように段階的にBlenderコードを最適化。現状、最適化ステップごとにBlenderで画像をrenderinする必要がある。
- ポイント：タスク自体がすぐにBlender使用者に適応できそうで、さまざまな応用まで拡張可能。研究テーマとしても拡張可能な面が多い。
- 感想：コードベース画像認識、分析、編集、生成などが実応用しやすい、解釈性が高い。BlenderAlchemyを拡張していくと動画の編集や複雑な画像の編集、実画像のコントロールなどに応用可能。実世界画像をコードに近似する研究あったら良さそう！



ECCV 2024 の動向・気付き（115/132）

A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting

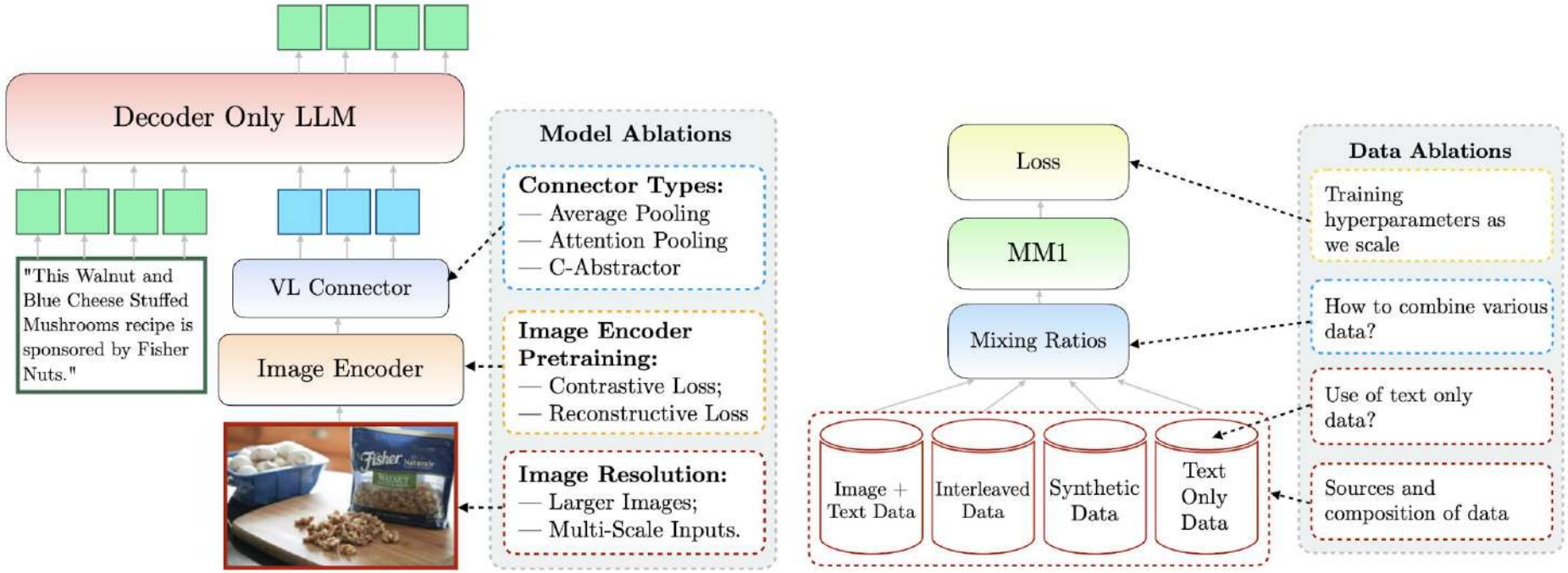
- 概要：複数のInpaitingのタスクをプロンプトで対応できる手法の提案。既存のDiffusion手法に二つの学習可能なタスクPrompt：PobjとPctxtを追加し、テキストとコンテキストに応じて物体の追加、削除、Outpaintingなどができる。
- ポイント：同じモデルで複数のInpaintingタスクを対応できるところ。他のMultimodalタスクにも拡張できる。
- 感想：画像Inpaitingから、一般的なタスクプロンプトをMLLMに学習させることもできそう。Task Promptと比較してInstruction Learningの方が使う時に自由度が高い。



ECCV 2024 の動向・気付き（116/132）

MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

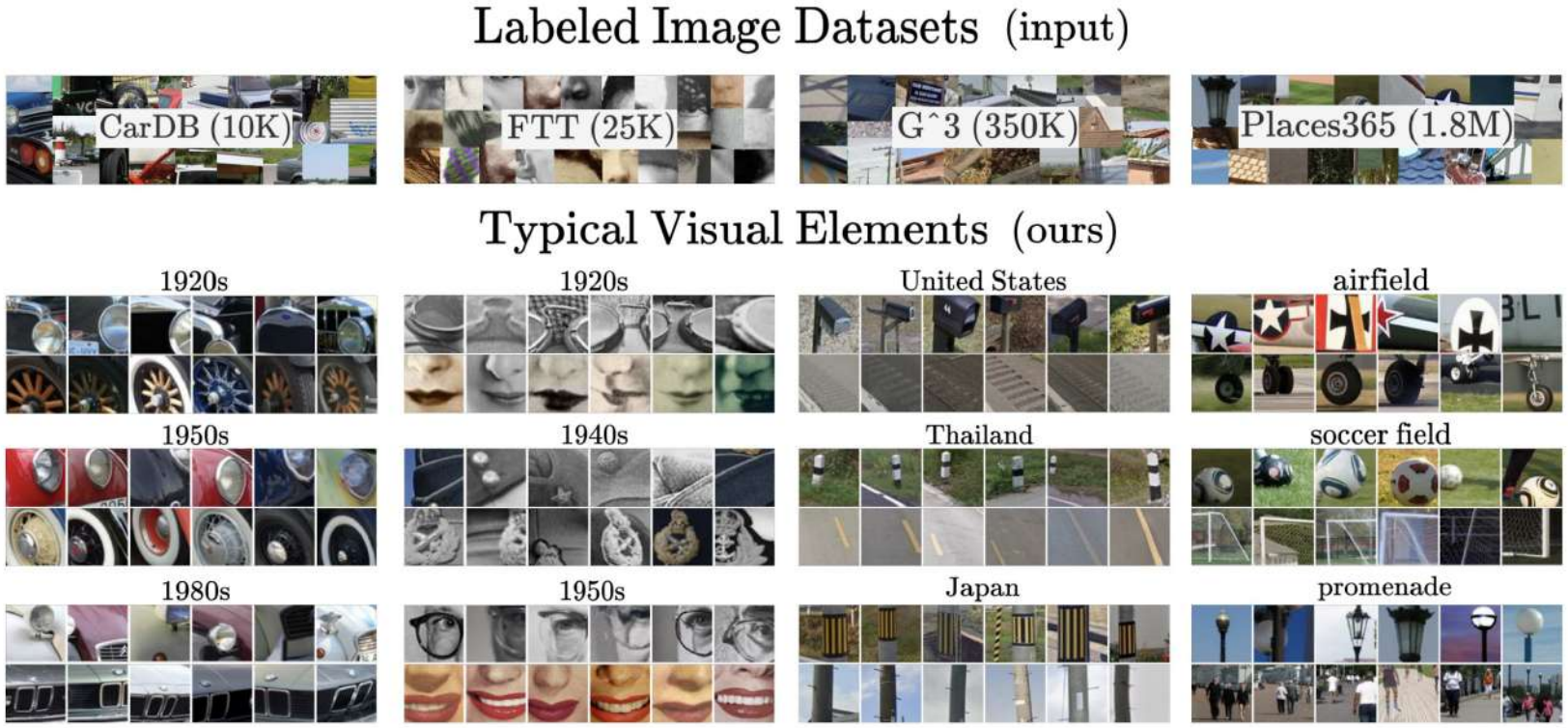
- 概要：Multimodal pretrainingにおいてMLLMのEncoder、V&L connector、学習データの選択などについて網羅的に実験調査をした。最後に、得られた結論によりMLLMのMM1（30B /64B）を提案。MM1がFew shot chain-of-thought, 複数画像のreasoningなども強い性能を得られた。
- ポイント：Multimodal pretrainingのレシピーを公開し、大規模の実験分析でMLLMの構築と学習に知見を示した。例えば、まず画像解像度が最も性能に影響し、その次はエンコーダーのCapacityとロスの選択、最後に、学習データはさまざまなタイプのものをコンバインして用意した方が性能が出る（例：image-caption, interleaved image-text, text-onlyなどが重要）。また、V&L connectorの構造選択が比較に重要ではないことも示唆した。Mixture-of-expertsを使うことで、MM1が様々なベンチマークでSOTAを達成。
- 感想：Appleの研究でかなりのリソースを使った。



ECCV 2024 の動向・気付き（117/132）

Diffusion Models as Data Mining Tools

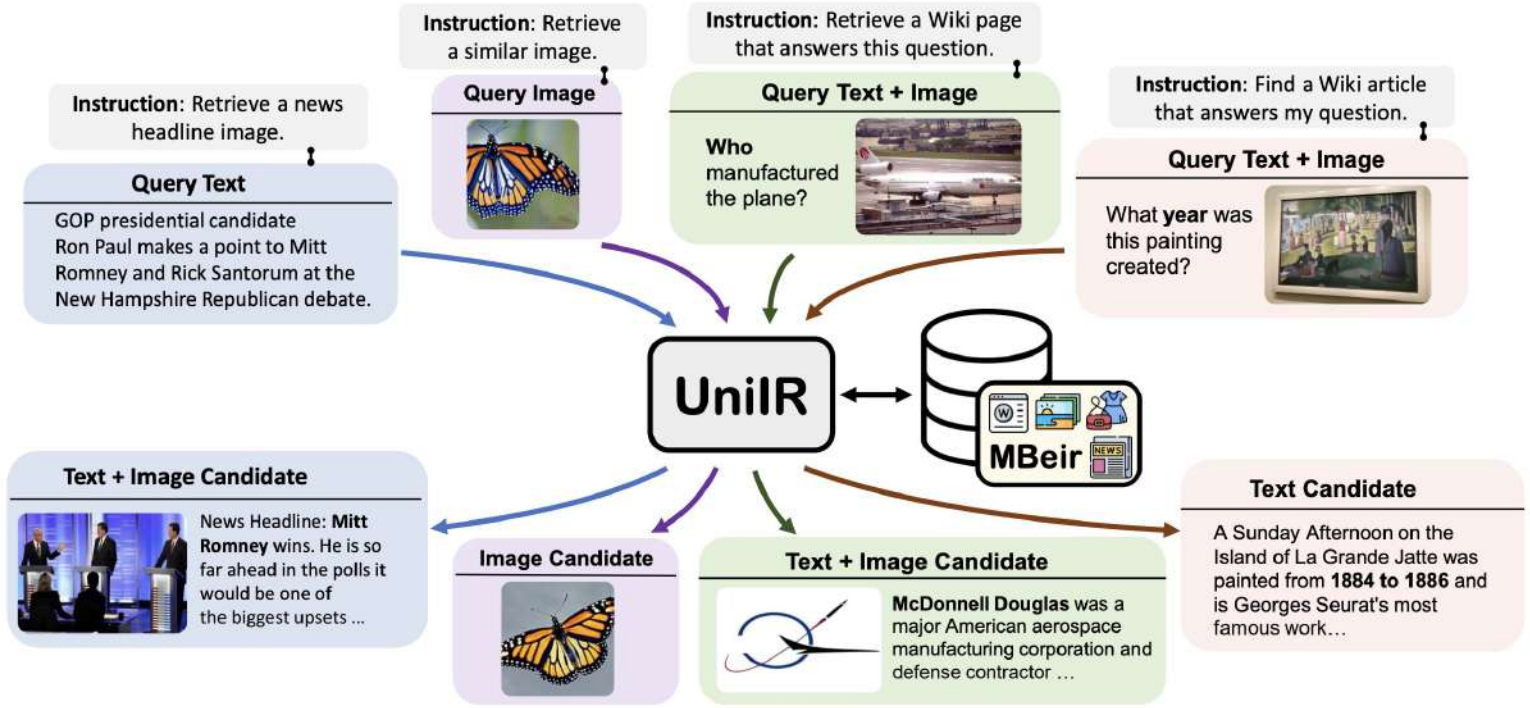
- 概要：学習済みのDiffusion modelをvisual data miningに使う手法を提案。データパターンの分析や、学習データセットの詳細分析などができる。分析したい要素（例：時間、場所）をConditionでターゲットデータセットでDiffusionモデルを学習させることでそのデータセットのその要素についての分析をする。
- ポイント：Diffusion modelをData Miningにするアイディア。Analysis-by-synthesisアプローチは興味深い。
- 感想：Diffusion modelで広い分布を持った詳細な画像を生成できることが知られている。人間がラベリングしたデータセットにはラベリングのバイアスがあると比較して学習済みDiffusion modelではラベリングデータセットのラベルバイアスを一部緩和できることでデータMiningに使いやすい。また、Large modelはGenerationモデル／Generation ＋Recognitionモデルの形式が良さそう。データセットを比較する際に、データセットの比較ではなく生成モデルで便利で条件付けしながら比較できるポイントも面白い。



ECCV 2024 の動向・気付き（118/132）

UniIR: Training and Benchmarking Universal Multimodal Information Retrievers（Oral）

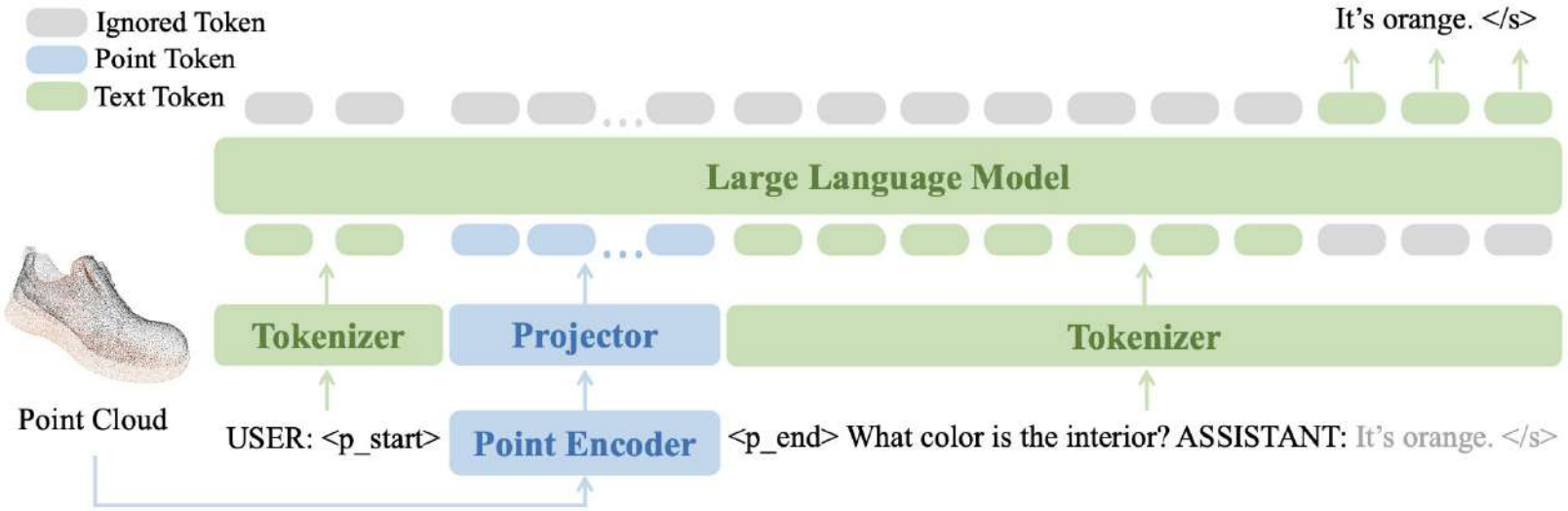
- 概要：単独の画像／テキストから、Interleaved画像・テキストまでの入・出力のRetrievalタスクを対応可能なInstruction-guidedモデルUniIRを提案。CLIP特徴やScore、Fusionの方法などについてAblation studyをした。また、UniIRの大規模Multi-Taskでモデルを学習することの重要性を示した。
- ポイント：多種類のRetrievalタスクを同時に対応可能でもっと自然なUser Inputに近づいた感じ。
- 感想：Interleaved image + textで認識・生成・学習が多くなってきた。UniIRはビデオへも適応可能。Newsデータセットも使った。NewsやWikipediaなど実世界の知識ソースの使用も多くなってきた。Retrievalタスクや、Retrieval augmented generation、データセットMiningなどの研究も増えてきた。



ECCV 2024 の動向・気付き（119/132）

PointLLM: Empowering Large Language Models to Understand Point Clouds（Best Paper Candidate）

- 概要：Point cloud入力を直接対応できるPointLLMを提案。PointLLMの構造がシンプルでStraightforward（左下）。PointLLMの学習のためのSyntheticデータセットも提案。既存のデータセットと比較して提案データセットはPoint cloudの詳細を細かく言語でアノテーションをしている。
- ポイント：Point cloudをLLMにリンクさせて、従来直接扱いにくいPoint cloudでも言語ベースで認識（今後編集なども可能？）と分析を可能にしたところ。
- 感想：PointLLMでも詳細的なアノテーションデータで学習することの重要性を示した。学習データの工夫やモデルの工夫など様々なところからこの研究を展開できる。



PointLLMの構造

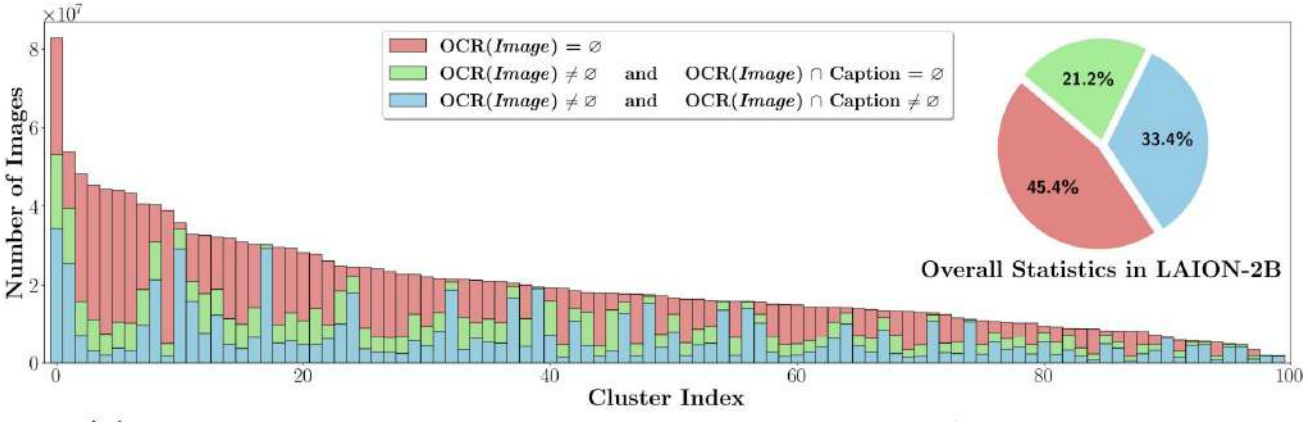
結果例

ECCV 2024 の動向・気付き（120/132）

Parrot Captions Teach CLIP to Spot Text（Oral）

- 概要：CLIPモデルはParrot Caption現象（ 画像コンテキストの認識をせずに、画像中の文字を単純にSpotする）がある。その現象の要因と影響を網羅的に調査した。具体的に、LAION-2Bデータセットで画像中の文字とCaptionの重なり度合いを分析したり、Parrot Captionの度合いをコントロールし、画像と画像中にEmbededしているテキストの類似度を調査した。Embeded CaptionとImage Captionの重なりが、CLIPモデルのParrot Captionを起こす重要な要因と明示化した。
- ポイント：CLIPのParrot Caption問題を網羅的に調査しその要因を示した。そのようなParrot Captionが実際CLIPのEncode能力を大きく影響することも示した。CLIPがMLLMで広く使われているため、分野に対してはこの知見の重要度が高い。
- 感想：画像中の文字とCaptionの重なりが多いデータセット自体は問題なさそうだが、ロスなどでモデルがデータセットに含まれるテキスト以外のテキストも画像から認識できる、画像の意味や文字の意味そして文字と画像の関係性がわかるように、モデルを評価することが重要。

LAION-2Bデータセット例（上部）とCaption（下部）

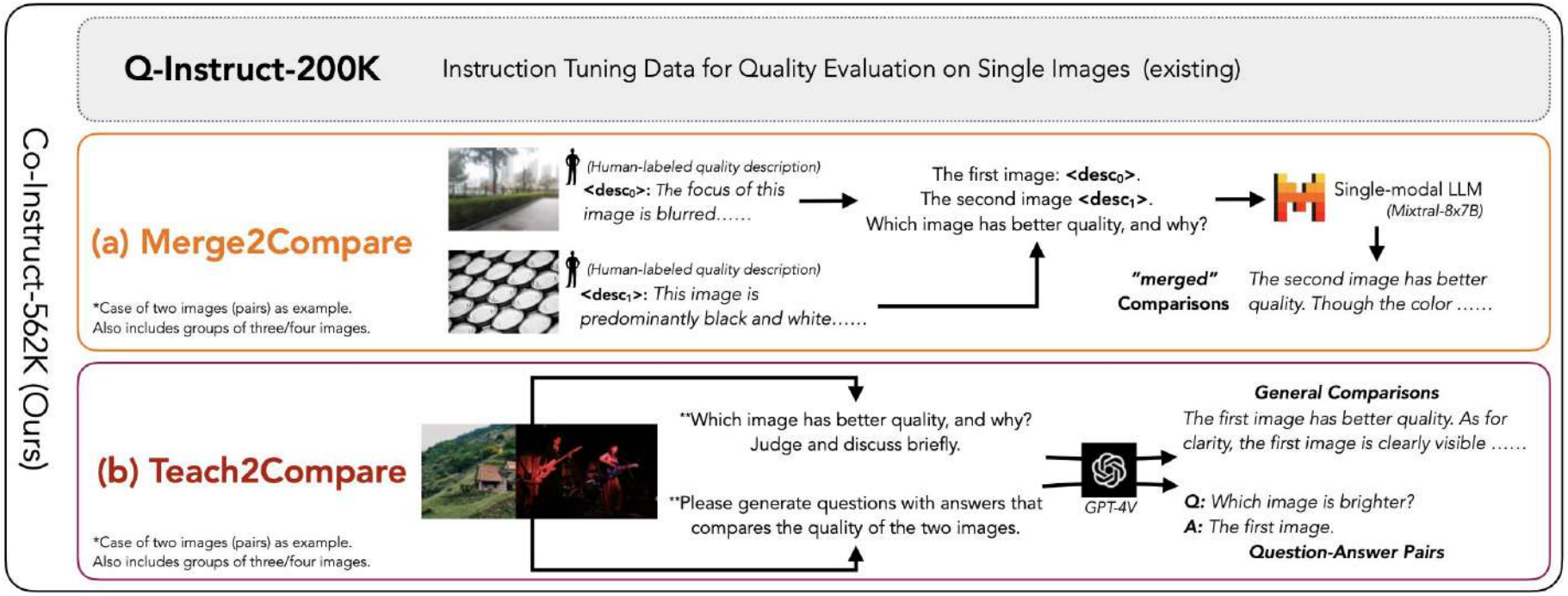


LAION-2Bデータセット例に含まれるOCRベース画像とCaptionの割合

ECCV 2024 の動向・気付き（121/132）

Towards Open-ended Visual Quality Comparison（Oral）

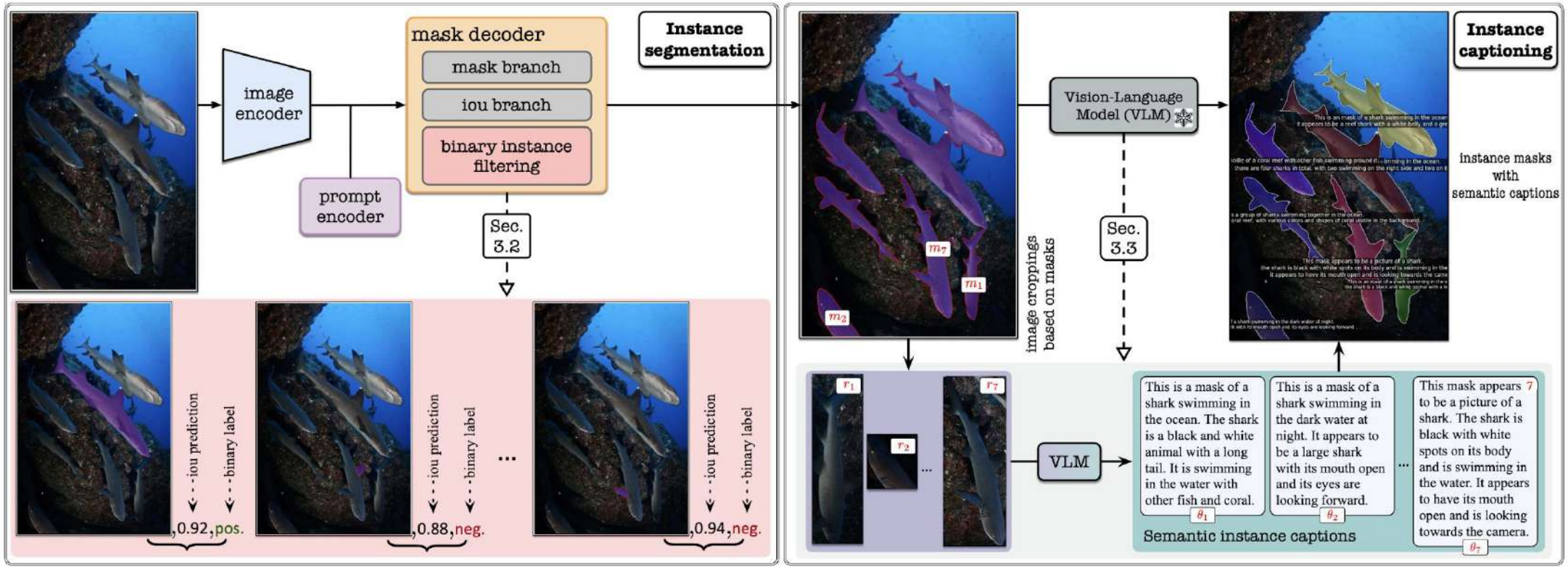
- 概要：従来の1枚の画像からVisual Quality Assessmentを行うタスクをOpen-endedにした。具体的に、複数枚の画像を入力してそれらのQualityをRankingするタスクとデータセットを提案した。また、複数枚の画像をベースにReasoningできるモデルCo-Instructを提案した。さらに、自動構築されたCo-Instruct-562Kを構築し、GPT4Vより高い精度を得られた。
- ポイント：Open-endedでVisual Qualityを評価するタスクの提案。
- 感想：複数枚の画像からの認識やReasoning、もしくはGeneral的なVisual Reasoningがまだ検討すべきな課題が多い。Co-Instructの性能がモデルより、学習データの面が重要。GPT4VなどのClosed sourcedのモデルが色んな論文で評価に使われていて、手法の比較よりは、データセットの性能を示す観点で実験を行う感じが強い。



ECCV 2024 の動向・気付き（122/132）

MarineInst: A Foundation Model for Marine Image Analysis with Instance Visual Description（Oral）

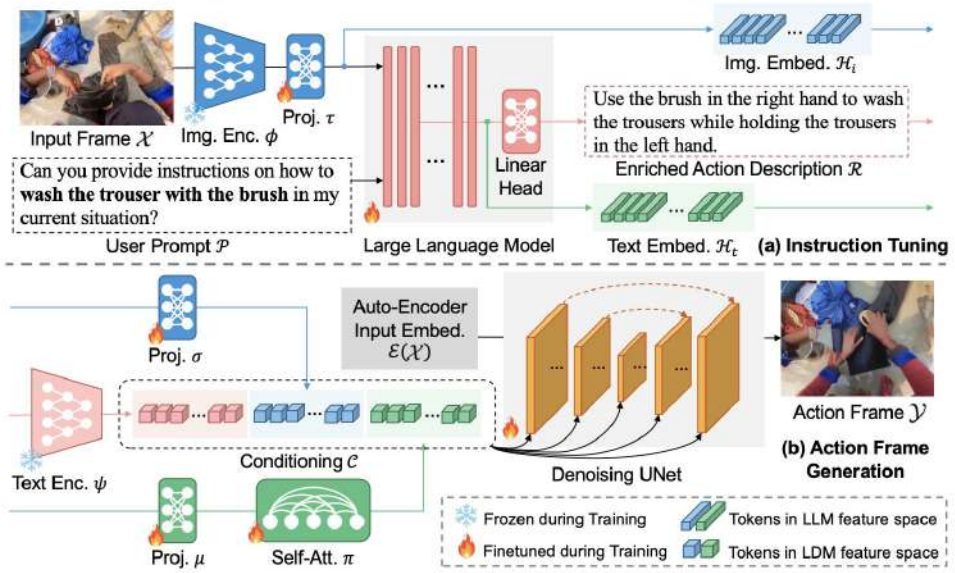
- 概要：Instance Segmentation とCaptionをベースとしたMarine生物認識のためのFoundationモデルMarineInstを提案。Underwater Marine Imagery認識に特化したモデルのデザインも行なった。また、MarineInst20Mデータセットも提案。 VLMを使ってMarineInst20MデータセットにSemantic密なAnnotationを付与した。
- ポイント：Marineの画像理解とLLMをリンクさせた。Instance Segmentation & Captionベースのモデルデザインも Solid。
- 感想：AI for goodやAI特にLLMやFoundation Modelを生物・地理・医学などの研究が多くなってきた。



ECCV 2024 の動向・気付き（123/132）

LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning（Best Paper Candidate）

- 概要：Egocentricの画像とHowToを表すSentence入力から、センテンスが指定するようにEgocentric画像を生成する手法。提案手法は2段階で行なっている。段階1では、LLMで大規模Instruction Tuningのデータで HowToと画像の中の密なSemantic情報アライメントを行う。段階2では、段階1で得られる豊かなText条件、オリジナル画像のEmbeddingなどからDiffusion modelで画像を生成。
- ポイント：まずHowToの動画もしくはフレーム生成が色んなアプリケーションで重要。さらにLLMを利用し、Instructive Videoの豊かな意味情報を画像生成に持ってきた。
- 感想：CVPR2024のGenHowToモデルとモチベーションが似ていて、解き方はGenHowToよりはFlexible。Roboticsアプリケーションや Step Error Detectionなどにも使えそう。



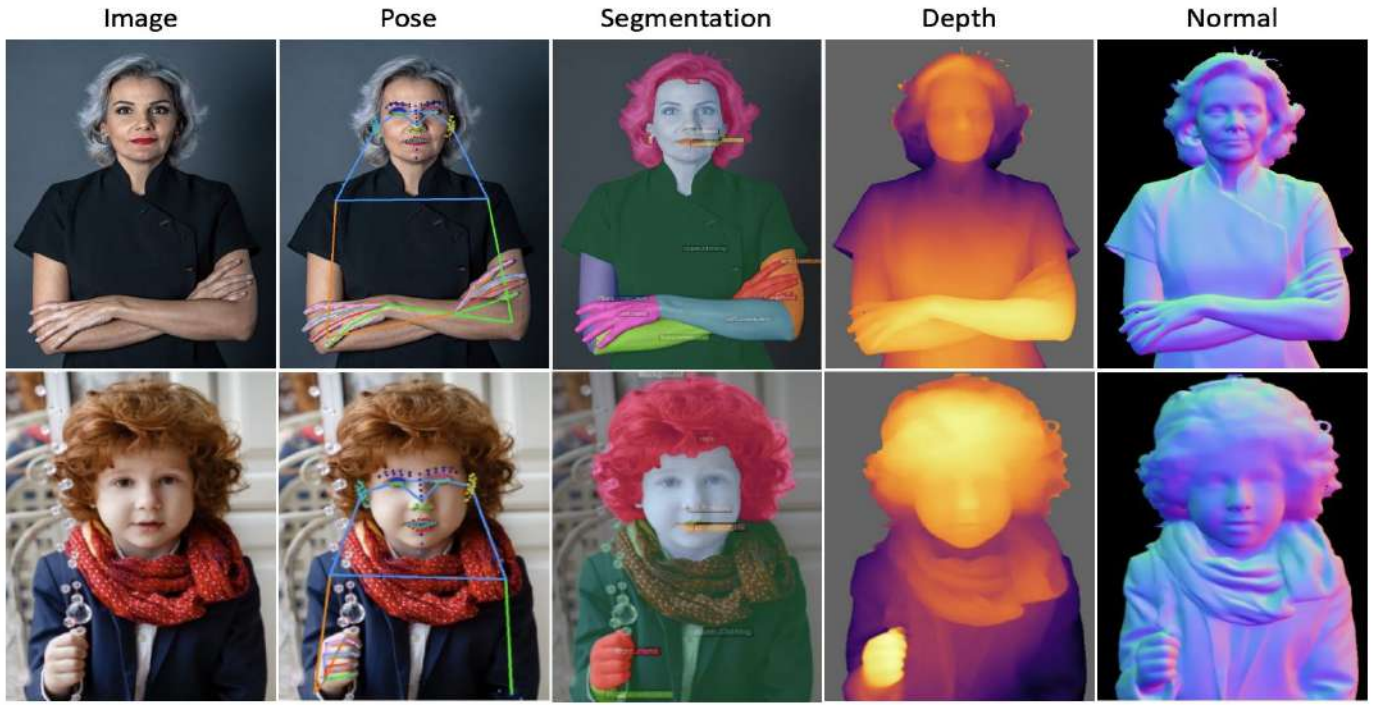
LEGOモデル

結果例

ECCV 2024 の動向・気付き（124/132）

Sapiens: Foundation for Human Vision Models（Best paper candidate）

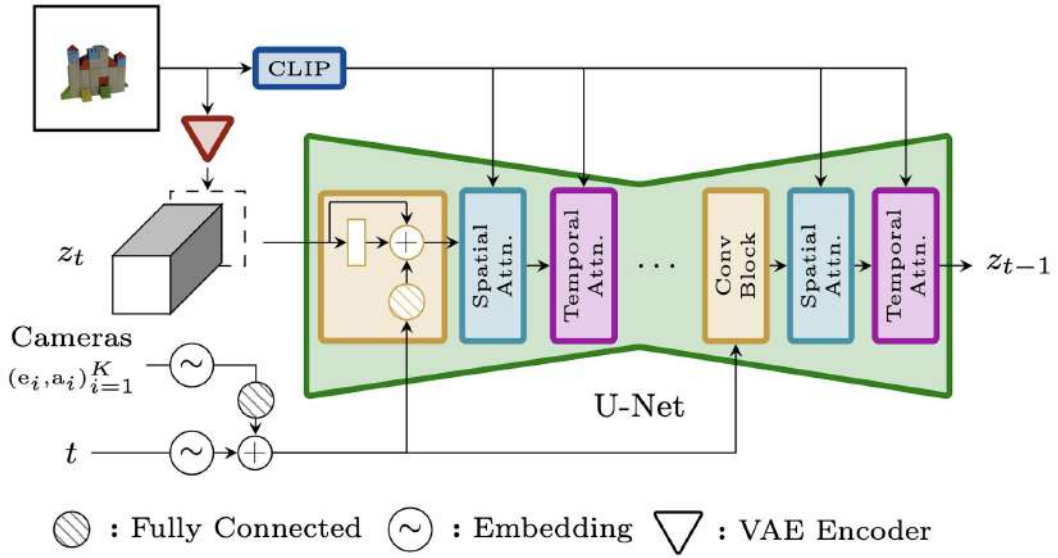
- 概要：4つのタスク（Pose推定、Segmentation、Depth推定、Normal）を高精度（SOTA）でできる人間認識モデルSapiens（複数モデル）を提案。Sapiensはまず300Mの人間写真セット（解像度1024＊1024）でself-superviseで事前学習し（mask-autoencoder (MAE)）、その後4つのタスクで別々でシンプルなモデルでシンプルなFine-tuneをした。Fine-tuneする際に、質の高いデータで行うことが重要。
- ポイント：Reasonableの実験デザインでSOTAなHuman Vision Modelsを提案。
- 感想：実際著者に聞いていたところ、Multi-taskで学習したが、タスク単独学習の方が精度が高かった。高解像度の画像で大規模事前学習したところ、高解像度画像の使用も性能に大きく貢献した。モデルがシンプルのように感じでいて、モデルデザインをしてもまだ大規模学習の方が一気に精度が上がるところにある。



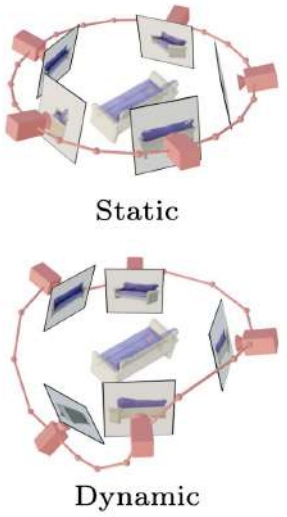
ECCV 2024 の動向・気付き（125/132）

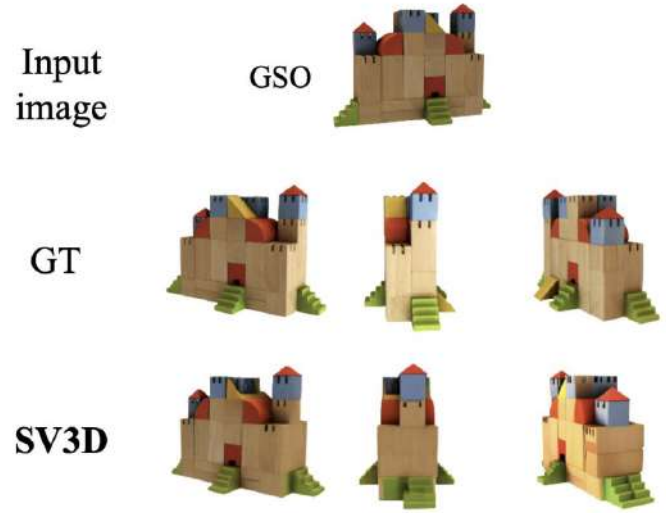
SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion（Oral）

- 概要：高解像度で3D物体のOrbitalビデオを生成するLatent video diffusionモデルSV3Dを提案。既存のImageベースDiffusionモデルを利用する手法と比較し、この研究は既存手法の SVD（Stable Vision diffusion）をNovel view synthesis（NVS）に使用し、 Multiview Consistencyを向上。3D Generationをする際に、まず生成したMultiview画像をガイドで、Score Distillation Samplingロスを導入しNeRFの学習をガイドする。また Coarse-to-fine学習やDisentangled Illuminationなどいくつかのテクニックも使った。
- ポイント：Image2Video diffusionでNVSと 3D Generation をするアイデアがいい。手法がシンプルで性能が高い。
- 感想：SV3Dのように、今回のECCVではVideoデータを使って3D Reconstructionを行う研究が何件ありました。また、Videoベース Diffusionモデルもかなり多くなってきた。3V3Dの後、非剛体運動・動きをする人物のモデルを生成できる3V4Dも提案。



モデル





結果例

ECCV 2024 の動向・気付き（126/132）

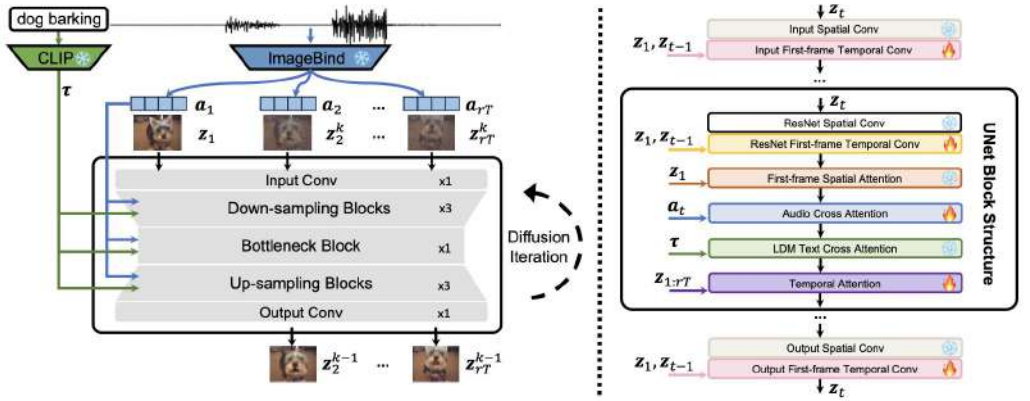
Watch Your Steps: Local Image and Scene Editing by Text Instructions（Oral）

- 概要：Editする際にEditの対象領域に集中してEditする手法を提案。提案手法がInstructPix2Pixをベースにしている。Editの対象領域をLocalizationする際に、2回IP2Pを使う。まず編集対象を含むセンテンスを入れて画像を生成する、次に何もテキストを入れずに画像を生成する。生成した画像の差分をNormalizeしRelevance mapとして使う。多視点からのRelevance mapからRelevance Fieldを計算し、Multiviewから指定編集対象の3D領域を生成する。Relevance mapに関する閾値の設定を実験で調整した。
- ポイント：シンプルなEditの対象領域をPinpointするRelavance mapの提案。明示的に対象領域を編集できることが使いやすい。
- 感想：手法がシンプルで効果が良さそう。

ECCV 2024 の動向・気付き（127/132）

Audio-Synchronized Visual Animation（Oral）

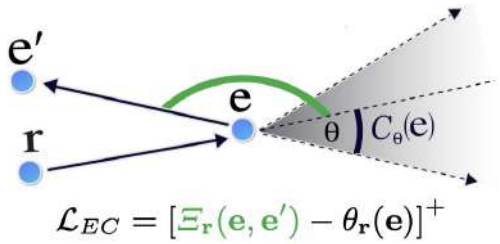
- 概要：Audioと初期画像から、ビデオの続きを生成するタスクASVAを提案。また、VGGSoundをベースに15カテゴリのaudio-visual eventsを集めた。また、ImageBindを使って、音声をエンコードしそれらと入力のテキストをベースに画像編集Difussionモデルを提案した。また、いくつかのAttentionレーヤーを再学習した。
- ポイント：Audio-synchronized video animationタスクの提案。音からビデオ編集・生成できることを示した。
- 感想：この研究が比較的に小さいデータセットで少ないクラス（15クラス）でしか検証していないですが、いろいろ可能性を示してくれた。ASVAで提案しているAudioからビデオ生成、ビデオAnimationする研究がこれから増えそう。研究できる方面が多い！



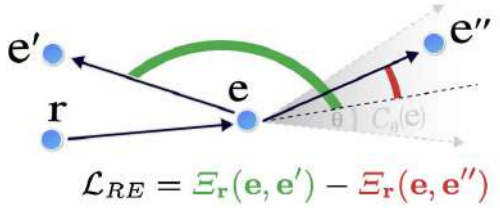
ECCV 2024 の動向・気付き（128/132）

Emergent Visual-Semantic Hierarchies in Image-Text Representations (Oral)

- 概要：既存のFoundation models, VLMsなどがvisual semantic hierarchiesをどれくらい理解できているのかを評価した論文。Radial Embedding（RE）フレームワークを提案した。REモデルでは、visual semantic hierarchies のLayer上の概念の特徴の間の距離と角度で計算されている。REベースのContrastive Learningで、著者が作った評価用データセットでFinetuningした結果、既存手法のVisual Semantic Hierarchiesについての認識能力があがった。
- ポイント：提案のRE構造がシンプル。また、REで再学習した結果がよかった。
- 感想：画像にはHierarchical構造で表示しにくい部分もあるから、Hierarchical Structure以外の構造で現状のVLMsが視覚要素を理解しているのかを検討するのもできる。



$$\mathcal{L}_{EC} = [\Xi_r(e, e') - \theta_r(e)]^+$$



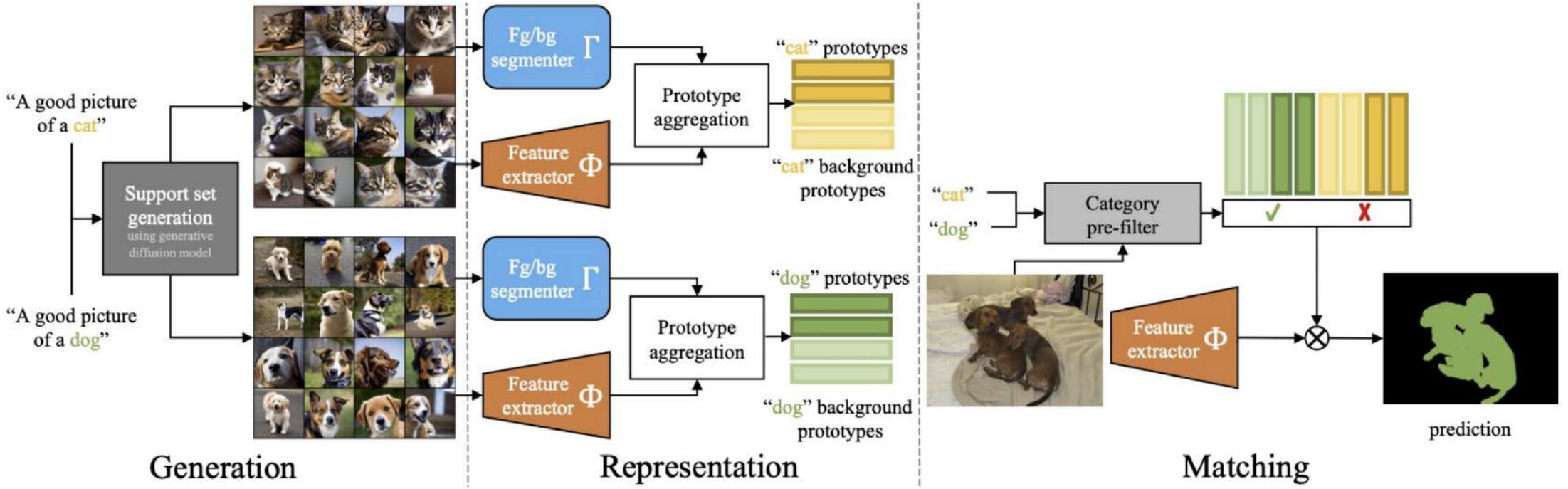
$$\mathcal{L}_{RE} = \Xi_r(e, e') - \Xi_r(e, e'')$$

CLIP	+alignment	CLIP	+alignment
fun	*food animal*	*two*	*cat*
top	*vegetables*	*sleep*	*cats*
…	…	…	…
a close up of a plate of food with broccoli	*A raw piece of broccoli with something growing from it.*	*cats sleeping with a remote*	*Couple of cats sleeping on opposite ends of the couch*
A worm sits on top of a piece of broccoli.	*A worm sits on top of a piece of broccoli.*	*Two cats sleeping with a remote control near each of them.*	*Two cats sleeping with a remote control near each of them.*

ECCV 2024 の動向・気付き（129/132）

Diffusion Models for Open-Vocabulary Segmentation（Oral）

- 概要：Diffusion modelをOpen vocabulary segmentationに適応した手法OVDiffを提案。提案手法が学習せずに高精度でSegmentationができる。Inthewildタスク（例：Cat）をSegmentする際に、まず”a good picture of a cat”を学習済みのDiffusionモデルに入力し、画像セットを生成する。次に、Offtheshelf特徴抽出器で画像特徴を抽出しAggregationによりセット画像のCategory prototypesを生成する。最後に、nearest-neighbor matchingでマスクを生成する。
- ポイント：Diffusion modelをSegmentationタスクへ応用するポイント。また、手法がシンプルで学習がいらない。
- 感想：Diffusion modelではクラス分類だけではなく、もっと詳細的なSegmentationもできる可能性がありそう。Diffusion modelから知識を得て他のタスクへ適応する研究が多くなった。



ECCV 2024 の動向・気付き（130/132）

The Hard Positive Truth about Vision-Language Compositionality

- 前提知識

  合成性：文全体の意味は，文の部分要素の意味の組み合わせから構築される．今回の論文は合成性を獲得していればわかる文についての論文．合成性の理解の例としては，「白いフリスビーを咥えた茶色の犬」と「茶色いフリスビーを咥えた白い犬」の違いが理解できること（白と茶色が入れ替わっている）

  ハードネガティブ：この論文のハードネガティブは，部分的に意味を取り替えたときに意味が **変わる** 文．
  例は，「白いフリスビーを咥えた茶色の犬」と「茶色いフリスビーを咥えた白い犬」

  ハードポジティブ：この論文のハードポジティブは，部分的に意味を取り替えたときに意味が **変わらない** 文．
  例は，「白いフリスビーを咥えた茶色の犬」と「乳白色のフリスビーを咥えた茶色の犬」

- 概要：Contrastive Language-Image Pretraining（CLIP）などの視覚言語モデルは合成性を持っておらず，ハードポジティブでデータオーグメンテーションした手法で合成性の性能改善が提案された．しかしながら，ハードポジティブだけでは改善しないことをこの論文は明らかにし，ハードポジティブも加えたデータオーグメンテーションをして，合成性の性能を改善した．

- ポイント：言語を伴う大規模モデルは人間がしているような理解とは程遠い部分があり，その点について取り組むような本質的な研究．

- 感想：未だに大規模モデルはそれっぽい挙動をする怪物のようなものであると思っており，その点について分析を行なっていくところは非常におもしろいし，本質的だと感じました．

Image i

Existing work

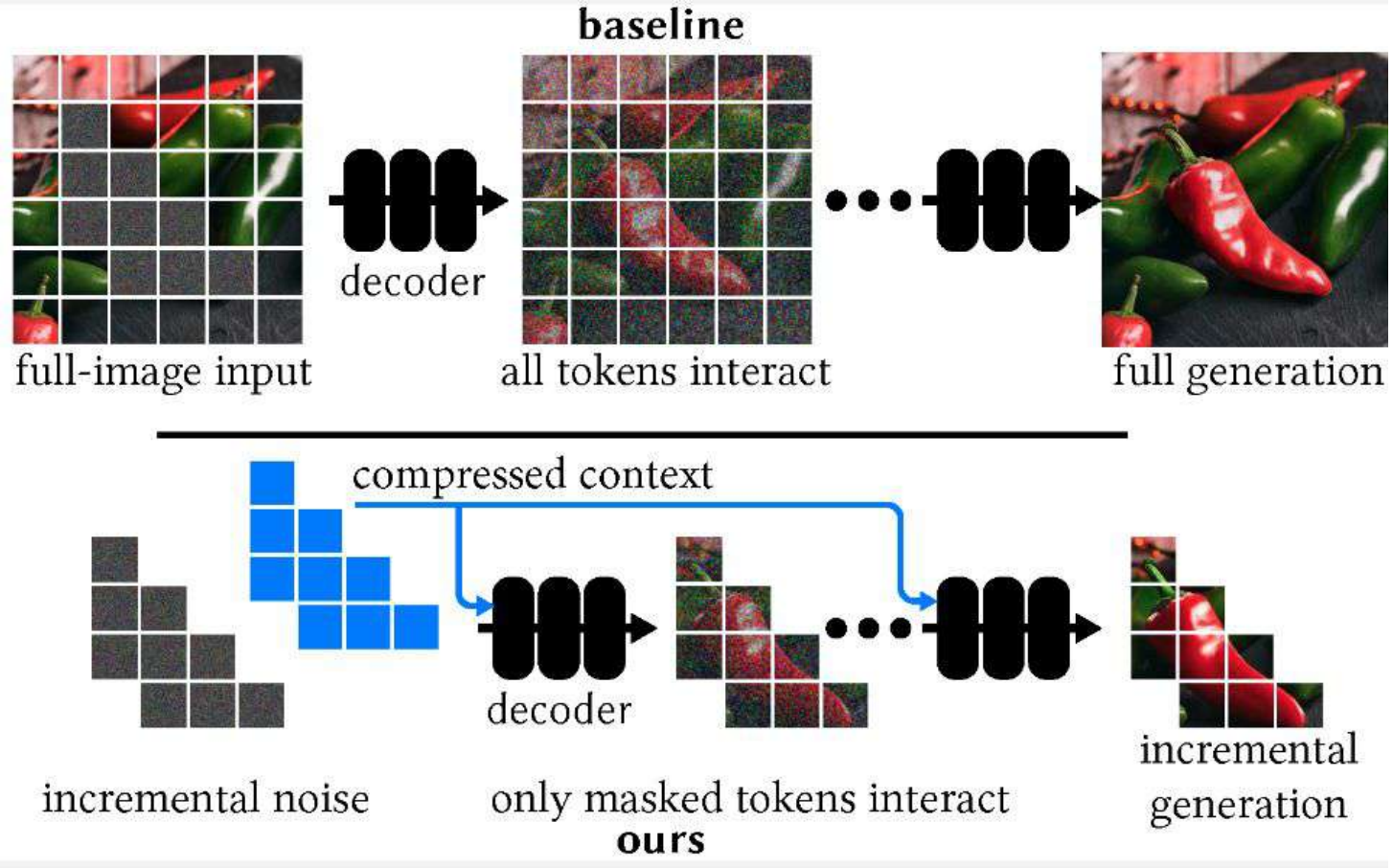
	Captions	CLIP	Hard Negative Finetuned	Ours
Original Caption c	brown grass	0.236	0.152	0.240
Hard Negative c_n	blue grass	0.240	0.143	0.231
Hard Positive c_p	chestnut grass	0.249	0.134	0.241

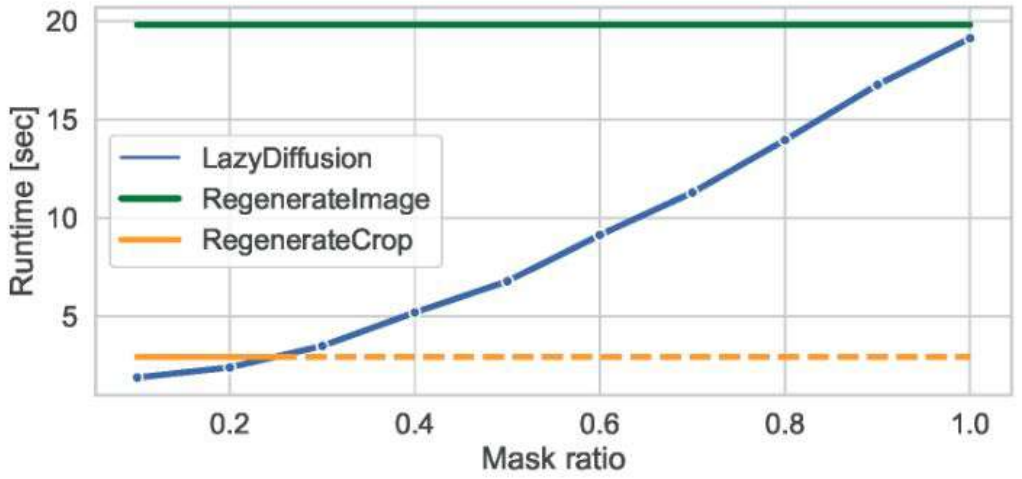
Our work

ECCV 2024 の動向・気付き（131/132）

Lazy Diffusion Transformer for Interactive Image Editing

- 概要：インタラクティブな画像編集タスクにおいて、要編集の部分のみを生成することで、高速化する手法
- ポイント：入力トークンを編集領域の形にして、周りのコンテキスト情報を入力トークンに埋め込む。その後、拡散プロセスを行い、編集領域の結果を出力
- 感想：非常にSimple yet effectiveな手法であり、局所的な画像編集タスクでは高速かつ自然な結果を得ることができる

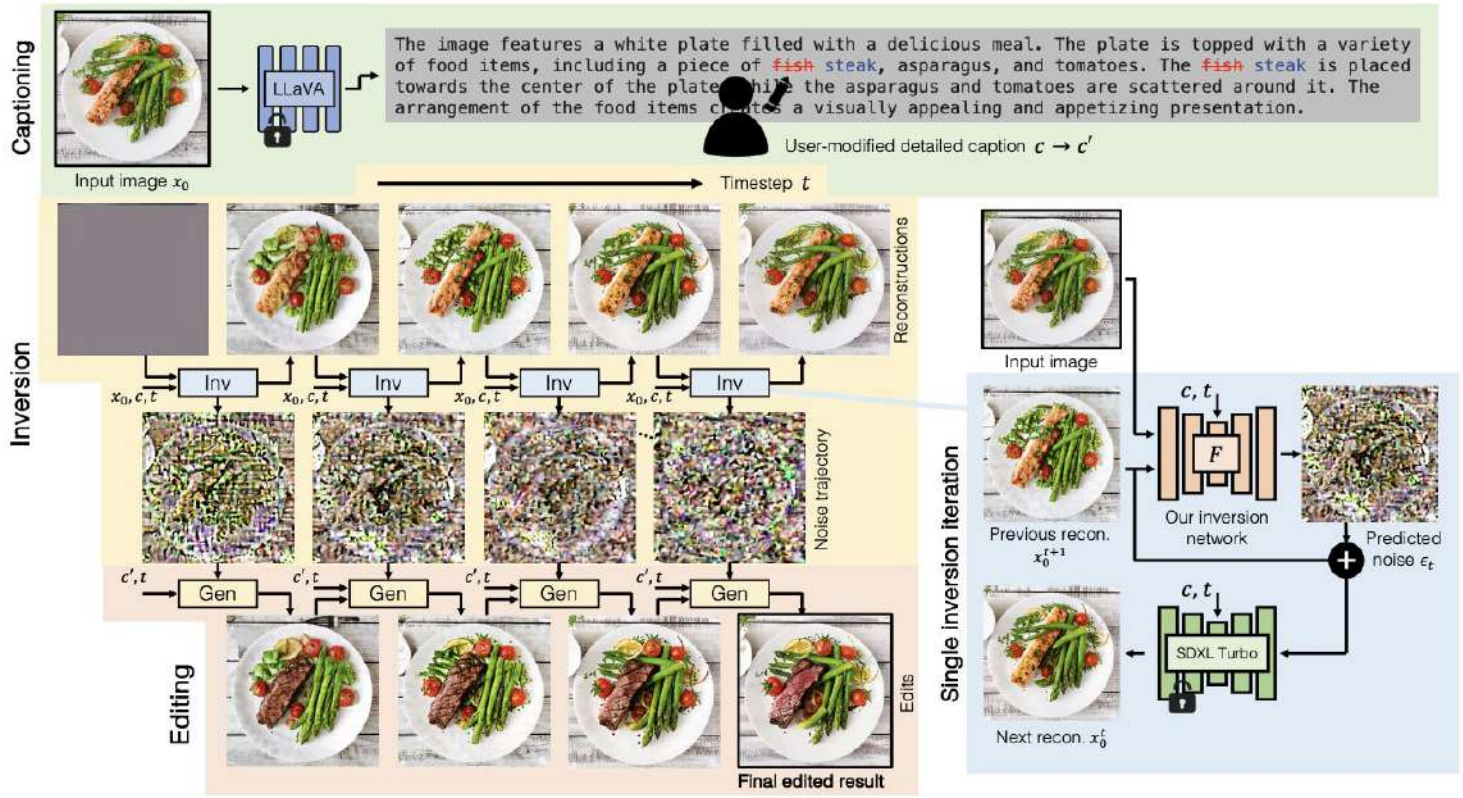




ECCV 2024 の動向・気付き（132/132）

TurboEdit: Instant Text-based Image Editing

- 概要：SDXL Turboをベースに、ユーザーがリアルタイムで試行錯誤しながら画像編集を行うことが可能
- ポイント：実画像を編集するには、image inversionで画像→潜在空間にマッピングする必要がある。この手法では高速かつ正確なInversionネットワークを提案している。また、プロンプト内の一つの属性を変更するだけで、対応する画像の属性のみを変更できる
- 感想：テキストベースの実画像編集タスクにおいて、正確なImage inversion、実行時間、属性間のEntanglementなどを考慮した手法



Oct 4th, 2024 @ Mico Milano