Research Statement

Jianwei Ren

1. Introduction

I am Jianwei Ren, currently a research assistant at IIIS, Tsinghua University. Prior to this, I earned my MEng (2024) from Beijing University of Posts and Telecommunications (BUPT) and my BEng (2021) with a UK **First Class Honours Degree** from the joint program between BUPT and Queen Mary University of London (QMUL).

My academic excellence, reflected in my Master's (91.36/100) and Bachelor's (87.83/100) GPAs, is bolstered by a solid background in mathematics, informatics, and programming, along with rich project experience. I am proficient in PyTorch and MMCV.

From January to July 2024, I served as an intern at Xiaomi EV, during which I secured the **2nd** place in the Mapless Driving track of the CVPR2024 Autonomous Grand Challenge.

My research interests lie at the intersection of artificial intelligence and computer vision, with a particular focus on *generative models, autonomous driving, and multi-modal machine learning*. For more information, please visit my personal website.

2. Research Experience

2.1. Simulation for Scalable Autonomous Driving

This is my current research project at Tsinghua University. It formulates the task as joint distribution modeling of multi-agent states, with the goal of aligning the distribution with real-world driving scenarios. By achieving this alignment, the simulator can freely sample multi-modal trajectories for dynamic evolution. Inspired by recent advancements in generative foundation models like Sora [2] and GameNGen [7], this work adopts diffusion models as the backbone of a stochastic virtual world to underpin agents' behavior logic. This approach promises to substantially reduce the cost of algorithm validation, enhance efficiency, and address safety concerns.

2.2. Autonomous Grand Challenge

During my internship at Xiaomi, I participated in the Mapless Driving track of the Autonomous Grand Challenge. Not only did we achieve the **2nd** place in the CVPR2024, but we also secured the **1st** place in both the CVPR2023 and China 3DV2024 leaderboards for the same challenge. I take great pride in the fact that our team was comprised of just *two members*, with myself shouldering the majority of the responsibilities.

The task of this challenge is to construct a high-definition (HD) map using surrounding multi-view images of autonomous vehicles. Specifically, we are required to detect traffic elements, including lanes, traffic signs, road boundaries, and crosswalks. We also need to model the topological relationships between lanes as well as between lanes and traffic signs.

Taking into account the diversity among elements, we abandon the traditional approach of using a fixed number of points to represent any element indiscriminately. Instead, I propose a joint training framework [5] that allows the network to independently decode three major tasks while sharing the same latent features. Experiments have demonstrated that our framework not only improves training efficiency but also significantly enhances performance. Our method achieves slightly inferior results with only ***1/3*** *the parameters* of the leading model.

This project also endowed me with invaluable experience in hyperparameter tuning. Our final model was trained on 96 A800 GPUs, behind which were hundreds of trials and meticulous calculations to ensure optimal performance.

2.3. Self-Supervised Monocular Depth Estimation

My master's thesis offers two main contributions to the self-supervised monocular depth estimation community: firstly, a representation learning distillation framework, and secondly, an adaptive discrete strategy [6].

This work aims at recovering 3D depth information from a single RGB image, a task complicated by the lack of supervision during training. However, my efforts have enabled neural networks to generate higher-quality depth maps.

My framework employs a new paradigm for depth estimation, inspired by SimCLR [3] and SimSiam [4], which encodes the original frame in parallel and embeds it into a latent space. Recognizing the importance of spatial correspondences for dense prediction tasks, it utilizes pixel-level representation alignment to alleviate the inconsistency [1] between the objectives of perception and reconstruction.

The adaptive discrete strategy, while casting the depth regression into a classification problem, takes an exploratory look at the issue of sample imbalance within the self-supervised environment. It also introduces an adversarial objective function to prevent training collapse caused by the lack of supervision.

Thrilling is that the experience accumulated in this project can readily extend to others, like 3D reconstruction and scene perception. I attribute my rapid mastery of the autonomous driving challenge to the shared knowledge base they offer.

3. Conclusion and Future Work

Having amassed substantial experience in the realms of computer vision and autonomous driving, I am committed to continuing my exploration in the domains of generative models and multi-modal machine learning. I am currently seeking a Ph.D. programs opportunity to contribute more to the community and to facilitate AI's application for human welfare.

If my application is accepted, I can try to apply for the CSC scholarship to support my doctoral studies. Moreover, I plan to proactively collaborate with my prospective supervisor to develop a comprehensive research proposal, which I intend to execute systematically throughout the program.

References

[1] Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024. 1

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1

[5] Guang Li, Jianwei Ren, Quanyun Zhou, Anbin Xiong, and Kuiyuan Yang. Leveraging sd map to assist the openlane topology. 1

[6] Jianwei Ren. Adaptive discrete disparity volume for self-supervised monocular depth estimation. *arXiv preprint arXiv:2404.03190*, 2024. 1

[7] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 1