Research Statement – Jiawei He

jwhe2024@gmail.com - https://jiaweihe.com/

Overview

My research mainly focuses on video-based perception in complex scenes. In recent years, computer vision has made significant advancements with the help of deep neural networks. However, compared with advances in single-image perception, video-based perception is still limited by time and space complexity, and it is difficult to scale up.

Video-based perception can be applied across various domains such as autonomous driving, surveillance, augmented reality, and robotics. The dynamic nature of video data provides a rich source of temporal information, enabling deeper insights into the spatiotemporal characteristics of objects and scenes. This research statement outlines our vision to advance the fields of 3D object detection, multiple object tracking, and generative world modeling within the context of temporal perception in video.

Video data has inherent advantages and challenges that are difficult to leverage. Specifically:

- Objects exhibit temporal continuity in video data, characterized by continuous motion and evolving spatial configurations over time. This temporal coherence forms the cornerstone for understanding object dynamics, facilitating tasks such as object detection and tracking. By exploiting temporal connections, we aim to enhance the robustness and accuracy of these tasks, leveraging the inherent predictability of object behaviors within video sequences.
- Video data often contains redundant information across successive frames, stemming from the spatial and temporal coherence of scenes. Harnessing this redundancy offers opportunities for efficient data representation and inference, enabling streamlined processing pipelines and improved computational efficiency. Our research seeks to leverage redundancy-aware techniques to optimize resource utilization and enhance the scalability of video perception systems.
- The fusion of information from multiple temporal observations presents a unique opportunity to recover detailed geometric attributes of objects and scenes. By integrating observations across consecutive frames, we aim to reconstruct accurate 3D representations of objects, enabling precise localization and volumetric analysis. This approach not only enhances the fidelity of object detection but also facilitates higher-level reasoning tasks such as scene understanding and action recognition.
- Building upon historical observations, we envision a future where video perception systems possess the capability to anticipate and adapt to dynamic environments proactively. By learning from past interactions and experiences, these systems can forecast future trajectories of objects, enabling proactive decision-making and scenario planning. Our research aims to imbue video perception models with predictive capabilities, enabling them to anticipate future events and adapt accordingly in real-time scenarios.

To face the advantages and challenges of video data, my research objective is to develop robust and efficient video-based perception algorithms in computer vision systems, specifically targeting the following areas:

1. **Multiple Object Tracking**: My research focuses on achieving accurate and stable association, which is the main challenge for MOT in crowded and heavily occluded scenes. I propose a novel graph-matching-based data association paradigm in [3] and its extension [4]. In [2], 3D pseudo representations are learned from videos, which help separate objects in the 3D feature space.

2. **3D Object Detection**: My research topic focuses on camera-only 3D object detection. To summarize the research, in terms of methodology, I mainly proposed a series of fully-/weakly-supervised object detectors based on object-centric 3D reconstruction. This geometric constraint can serve as monocular depth solver [6], temporal optimizer [1], and 3D pseudo labeling method [5].

3. **Generative World Models** [7]: As a new research field, [7] pioneeringly proposed a multi-view generative world model applicable to autonomous driving, and provide for the first time an example of downstream end-to-end application of autonomous driving.

Specific Projects

Learnable graph matching for MOT and other data association tasks. Data association is a core problem for many applications, e.g., multiple object tracking, image matching, and point cloud registration. This problem is usually solved by a traditional graph-based optimization or directly learned via deep learning. However, current data association solutions have some defects: they mostly ignore the intra-view context information; besides, they either train deep association models in an end-to-end way and hardly utilize the advantage of optimization-based assignment methods, or only use an off-the-shelf neural network to extract features. I have explored a new paradigm utilizing differentiable graph matching for MOT (CVPR 2021 [3]). Specifically, we model the relationships between tracklets and the intra-frame detections as a general undirected graph. Then the association problem turns into a general graph matching between tracklet graph and detection graph. Furthermore, to make the optimization end-to-end differentiable, we relax the original graph matching into continuous quadratic programming and then incorporate the training into a deep graph network with the help of the implicit function

theorem. However, when scaling up instance numbers, the graph matching method is not efficient enough. To make the matching paradigm more efficient, we focus on the fast solver for graph matching (TPAMI 2024 [4]). We notice that we can design spatiotemporal constraints in the real downstream tasks, then the feasible region is limited and much smaller than the original quadratic programming formulation. Thus we try our best to make our matching paradigm more universal and practical for more data association tasks.

Reconstruction-based 3D object perception. In the early exploration of 3D reconstruction-based perception (ECCV 2022 [6]), we modify PnP algorithm and learn 2D-3D keypoint correspondence as graph matching. So, dense correspondences between 2D and 3D become the geometric constraints to solve monocular depth for the objects. In CVPR 2023 [1], we explore more temporal correspondences as reconstruction constraints. A two-stage object detector is designed, in which temporal correspondence learning is treated as the second-stage temporal learning model. A new featuremetric object bundle adjustment loss is designed for training this module. My research [2] takes advantage of the easy association of objects in 3D space. It uses a pseudo-3D representation of objects in the three-dimensional feature space, and optimizes it jointly with the object association module. This work designs a reconstruction-based pseudo-3D object label generation and 3D object representation learning module. By only learning the 3D representation of objects from monocular videos and supervising them with 2D tracking labels, there is no need for additional annotations from LiDAR or pre-trained depth estimators. This idea was also extended for weakly supervised 3D object detection [5]. In this work, three stages of generalization are developed: from complete to partial, from static to dynamic, and from close to distant.

World models for autonomous driving. In autonomous driving, predicting future events in advance and evaluating the foreseeable risks empowers autonomous vehicles to better plan their actions, enhancing safety and efficiency on the road. In my recent research (CVPR 2024 [7]), the first driving world model was compatible with existing end-to-end planning models. Through a joint spatial-temporal modeling facilitated by view factorization, our model is the first to generate high-fidelity multiview videos in driving scenes. Building on its powerful generation ability, we showcase the potential of applying the world model for safe driving planning for the first time. Particularly, our Drive-WM enables driving into multiple futures based on distinct driving maneuvers, and determines the optimal trajectory according to the image-based rewards. Evaluation on real-world driving datasets verifies that our method could generate high-quality, consistent, and controllable multiview videos, opening up possibilities for real-world simulations and safe planning.

Future Plans

Scaling up the video perception models. Although we have explored the network and the solver for the possibility of scaling up, there is still a large gap in terms of data scale. The available public datasets are insufficient. For example, the nuScenes dataset only consists of 5 hours of videos. And different datasets have varying data collection standards, which makes it challenging to train the model on them together. So creating a bigger dataset with different scenes will be my future work.

Effective and efficient world models. We have explored the world models for autonomous driving scenes. However, we have not yet applied our world model in more scenarios. Besides, the diffusion-based generation is time-consuming. So, in the future, I will design a more efficient world model to be practically applied for more tasks, e.g., robotics, and climate forecasting.

References

[1] Jiawei He, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 3d video object detection with learnable object-centric global optimization. In *CVPR*, 2023.

[2] Jiawei He, Lue Fan, Yuqi Wang, Yuntao Chen, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Tracking objects with 3d representation from videos. *arXiv preprint arXiv:2306.05416*, 2023.

[3] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *CVPR*, 2021.

[4] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: A practical paradigm for data association. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] Jiawei He, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Weakly supervised 3d object detection with multi-stage generalization. *arXiv preprint arXiv:2306.05418*, 2023.

[6] Yingyan Li, Yuntao Chen, Jiawei He, and Zhaoxiang Zhang. Densely constrained depth estimator for monocular 3d object detection. In *ECCV*, 2022.

[7] Yuqi Wang*, Jiawei He*, Lue Fan*, Hongxin Li*, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *CVPR*, 2024.