Statement of Purpose

Jiangxin Sun (sunjx5@mail2.sysu.edu.cn)

As a computer science student, I have always had a natural interest in developing intelligent systems that assist humans. I aspire to develop perceptual systems to achieve human-level recognition capabilities and further build decision-making systems to perform human-like behaviors such as interacting with complex environments. Despite the remarkable progress achieved in scene understanding, rational interactions of intelligent systems in complex environments remain challenging due to the lack of ability to accommodate possible future scenarios and the impact of the interactions performed. Considering the uncertainty and ambiguity embedded in the huge prediction space, it is still very difficult for machines to achieve reasonable forecasting. To overcome this challenge, I worked on future prediction tasks.

I began my research with video-based prediction tasks, specifically future instance segmentation prediction, which forecasts segmentation results of future unobserved frames. To explore the key points of such an emerging task that combines instance segmentation and video prediction, I pursued a research project with **Prof. Wei-Shi Zheng** and **Prof. Jian-Fang Hu**. The mainstream of this task was to insert a prediction block into an instance segmentation model (e.g., Mask R-CNN) and to predict future pyramid segmentation features extracted by FPN. The first weakness we found in existing methods was pyramid segmentation features at each individual level are predicted separately without joint consideration and thus intrinsic relationships among the features at different pyramid levels have been neglected. To solve this problem, we conducted in-depth discussions with **Prof. Wenjun Zeng** and proposed to collaboratively predict pyramid features via exploiting cross-level spatio-temporal contextual information aggregation. Considering that object instances to be segmented could vary greatly in different videos, we developed a novel flexible auto-path aggregation network for selective and adaptive information aggregation. We presented a preliminary work in **ACM Multimedia 2019** [1] and the extended approach was accepted by **TPAMI** in 2021 [2]. The second reflection was that there seems to exist a natural contradiction between learning discriminative segmentation features and learning reliable future prediction. As the resolution increases, the fitting of the high-uncertainty or even unpredictable local details will gradually dominate the learning process of the prediction model, which can hinder the learning of global motion and leads to degradation of prediction accuracy. To address this issue, we proposed an autoencoder to learn the predictive representation of segmentation features and performed prediction on this feature space. We tried to follow human prior knowledge to suppress local details in the encoding process and presented this work in **ICCV 2021** [3]. Then we started a collaboration with **Prof. Jianguo Zhang** and proposed to explicitly model the prediction uncertainty in the designed feature predictor, such that predictive representation can be learned using uncertainty decay. The revised approach is scheduled to be submitted to **TPAMI**.

During my segmentation prediction research, I gradually realized that the ability to anticipate future events is a key factor towards developing intelligent behavior. Also, the characteristics of corresponding fundamental tasks (e.g., instance segmentation) are critical to future predictions. Firstly, intrinsic relationships among pyramid segmentation features require contextual information aggregation during the prediction process. Secondly, discriminative feature representation with rich local details implies the need to reduce uncertainty in forecasting. I guessed such a phenomenon exists for other prediction tasks

as well. To verify this assumption, I continued my research in another prediction task, human motion prediction, which is useful for decision-making systems and human-machine interaction. I joined Huya Inc. in July 2020 as a research intern and was supervised by **Dr. Xintong Han**. Our goal was to predict future unobserved human motion (e.g., 3D mesh) according to observed past RGB frames. We found that actions consisted of common atoms (i.e., sub-actions) and these sub-actions could be shared across different subjects performing the same action. We further constructed an action-specific memory bank to exploit representative sub-actions and retrieved possible motion dynamics for guiding future motion prediction. We published this work in **NeurIPS 2021** [4]. After that, I joined Microsoft Research Asia in August 2021 as a research intern in the intelligent multimedia group and was supervised by **Dr. Chunyu Wang**. We found the assumption of sub-actions was more applicable for human motion prediction tasks in specific situations such as dance generation, which requires predicting future dance choreography conditioned on music pieces and past motion. Similar to my previous work, we constructed a memory bank to learn manifolds representing dance movements. Then we presented bank-constrained manifold projection to reduce the noises in the predicted motions and achieved long-term non-freezing dance generation. We published this work in **NeurIPS 2022** [5].

Although my previous research mainly focused on prediction problems, my primary interest is in developing intelligent systems that can assist humans. On the one hand, I believe the ability to anticipate future events is a key factor towards developing intelligent behavior, since humans naturally take into account possible future situations for making decisions, especially when driving and interacting with others. On the other hand, research on prediction problems also requires in-depth studies of the characteristics of corresponding fundamental tasks (e.g., scene understanding and motion analysis), which gives me alternative perspectives to understand these domains. Therefore, during my Ph.D., I aim to cultivate a solid research sense to discover meaningful yet easily overlooked research problems, especially those related to perceptual/decision-making systems. This ability would also facilitate my career goal of becoming a research professor, leading research to invent intelligent machines that benefit our society.

References

[1] Jiangxin Sun, Jiafeng Xie, Jian-Fang Hu, Zihang Lin, Jian-Huang Lai, Wenjun Zeng, and Wei-Shi Zheng. Predicting future instance segmentation with contextual pyramid convlstms. In *Proceedings of the ACM International Conference on Multimedia*, 2019. 1

[2] Jian-Fang Hu*, Jiangxin Sun*, Zihang Lin, Jian-Huang Lai, Wenjun Zeng, and Wei-Shi Zheng. Apanet: Auto-path aggregation for future instance segmentation prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3386–3403, 2022. 1

[3] Zihang Lin*, Jiangxin Sun*, Jian-Fang Hu, Qizhi Yu, Jiang-Huang Lai, and Wei-Shi Zheng. Predictive feature learning for future segmentation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1

[4] Jiangxin Sun, Zihang Lin, Xintong Han, Jian-Fang Hu, Jia Xu, and Wei-Shi Zheng. Action-guided 3d human motion prediction. In *Advances in Neural Information Processing Systems*, 2021. 2

[5] Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. In *Advances in Neural Information Processing Systems*, 2022. 2