Research Proposal

Geometry Constrained Fusion for Depth-aware Video Panoptic Segmentation

Shuyue Jia, M.Phil. Computer Science, City University of Hong Kong shuyuej@ieee.org

1 Background

Scene understanding with depth-aware panoptic segmentation recently triggered my attention [1]. Video Panoptic Segmentation (VPS) is challenging and crucial for many tasks, such as autonomous driving, robotics, mixed reality, sensor vision, and video editing. The VPS combines semantic and instance segmentation such that all pixels are assigned a class label, and all object instances are uniquely segmented across video frames [2]. It needs an extension to include monocular depth prediction for a complete scene understanding, *i.e.*, Depth-aware VPS (DVPS). Monocular depth estimation aims to predict the spatial position of each 3D point projected to the image plane.

2 Proposed Geometry Constrained Fusion for DVPS

This proposal presents a geometry-constrained fusion approach that aims to improve temporal association for DVPS. It introduces geometry constraints into the joint-learning framework of video segmentation, depth prediction, and optical flow estimation. **In specific, this method aims to enhance segmentation results with temporally consistent instance tracking, consistent depth prediction via multi-task learning, and simple yet effective deep network components.**

2.1 Temporally Consistent Instance Tracking

2.1.1 Research Question and Gap

One of the key research questions is how to achieve the temporally consistent instance association between multiple frames for VPS more effectively. Traditional works aim to build effective and efficient Multi-Object Tracking (MOT) methods to address the instance association problem, such as instance embedding similarity [3,4]. Most methods are deliberately designed from the perspective of high-level correspondence, where the instance-level information between frames is associated [5]. Apart from high-level correspondence, a growing body of research has revealed that motion clues from the low-level pixel perspective are beneficial to the segmentation task [6, 7]. However, the unilateral flow information helps learn a good representation for moving objects, but not for static objects. In addition, flow is not good at distinguishing individual objects, especially when their motions are similar [8, 9] [1]. So far, **the consistency between high-level object association and low-level pixel correspondence has not yet been explored in the literature.** In consequence, to fill the gap, **a temporally consistent instance tracking module is proposed by tracking objects via optical flow, in consistency with the instance embedding association**. The method is designed to achieve a unified representation and processing for the temporal association, which is aimed at promoting the segmentation performance of the VPS task.

2.1.2 Proposed Solution

Firstly, based on the flow information and the tracking results from the track head, the pixel low-level correspondence and instance high-level association have been established. The low-

[1] Some flow preliminary results are on my GitHub through the GMFlowNet model [10].

level correspondence, *i.e.*, optical flow map, is derived from the trained flow network, *e.g.*, GMFlowNet [10]. Secondly, patch-level and pairwise embedding methods are presented to obtain instance flow embedding and instance track embedding [5]. Thirdly, a shared fusion module is introduced to generate the estimated instance embeddings of the next frame by integrating the flow and track information with instance information, respectively. Finally, instance embeddings from the flow and track branches are constrained via a consistency loss. This instance similarity constraint between low-level pixels and high-level instances may alleviate the instance switch problem when processing a video clip.

2.2 Consistent Depth Prediction

2.2.1 Research Question and Gap

Since segmentation, depth, and flow are tightly coupled with the inherent geometry constraints, they should be considered together to benefit from each other and achieve much higher accuracy. **Another key research question for DVPS is the influence of depth on segmentation and vice versa.** **The general idea of this proposal for DVPS is to utilize the previous general representation, such as depth or flow, to help provide hints for later frame's predictions, and explore more possibilities of using this multi-modal information** [2].

2.2.2 Proposed Solution

Since segmentation, depth, and flow tasks are deeply correlated, **separate processing is non-optimal**. Thus, **consistent depth prediction is achieved by joint learning of multi-tasks employing a depth loss** [1,4]. The features of the segmentation network are shared with the depth network through the backbone. In addition, **current depth-segment interaction is limited**. For example, the recent DVPS state-of-the-art (SOTA) method, *i.e.*, PolyphonicFormer, only applies independent processing and a query linking [4]. Thus, the **interaction between segmentation and depth prediction deserves further study**. The **depth positional encoding** injects depth positional hints into the backbone, which may be one of the potential solutions [12].

2.3 Optical Flow Estimation

As for the optical flow estimation module, it could be **fine-tuned together with the segmentation and depth prediction tasks**, depending on performance and efficiency [3]. Through multi-task learning, segmentation, depth prediction, and flow estimation can be solved together. Prior to end-to-end training, the optical flow maps can be generated offline by a trained flow network.

2.4 Backbone Architecture

My previous research on electroencephalogram (EEG) signal processing and neuroscience focused on extracting spatio-temporal features from time-series signals [13–15]. To build a robust and efficient DVPS backbone, an **object-centric learning** framework with **multi-scale spatio-temporal feature extraction** can be employed. In detail, the network extracts spatial features from video frames and then utilizes temporal attention to refine the extracted spatial features. It was also inspired by my non-local dependency modeling research [16]. Last but not least, another

[2] The unified 3D representation for DVPS (first predicts depth from the constructed 3D scenes, then conducts segmentation) is also worth exploring [11].

[3] The Ground Truth of flow maps can be produced by a pre-trained flow model.

consideration is that some highly performant Video Instance Segmentation (VIS) methods are mainly based on the multi-scale spatio-temporal feature extraction scheme [17, 18]. **The backbone design for DVPS may refer to the SOTA VIS methods.**

2.5 Multi-source Data Employment

A robust and effective DVPS system does not merely rely on effective network and module design but also on technical details, such as data pre-processing. For instance, in the SemKITTI-DVPS dataset, different data may have much sparser Ground Truth (GT) or different semantic labels for the content [4]. Thus, **a proper training strategy should be investigated to balance the data of such partial labels, and the semantics of the wrong labels should be corrected and unified.** As a consequence, during the data pre-processing stage, I conducted experiments to generate dense masks from sparse masks through linear interpolation [5].

Another issue is lacking labeled data. GT data for dense prediction tasks, such as depth prediction, image & video segmentation, or optical flow estimation, costs much more resources than sparse prediction tasks. Thus, a data-efficient training scheme may be employed. For example, labeled pseudo-video data can be synthesized by deforming and shifting the instances from labeled images. Self-supervised learning may also be applied to explore the data distribution and inherent geometry constraints.

References

[1] S. Qiao, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, “ViP-DeepLab: Learning visual perception with depth-aware video panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3997–4008, June 2021.

[2] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, “Video panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9859–9868, June 2020.

[3] Y. Zhou, H. Zhang, H. Lee, S. Sun, P. Li, Y. Zhu, B. Yoo, X. Qi, and J.-J. Han, “Slot-VPS: Object-centric representation learning for video panoptic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3093–3103, June 2022.

[4] X. Li, H. Yuan, Y. Yang, L. Zhang, Y. Tong, and D. Tao, “PolyphonicFormer: Unified query learning for depth-aware video panoptic segmentation,” in *European Conference on Computer Vision*, vol. 7, Springer, Oct. 2021.

[5] X. Li and D. Chen, “A survey on deep learning-based panoptic segmentation,” *Digital Signal Processing*, vol. 120, p. 103283, Jan. 2022.

[6] W. Ye, X. Lan, G. Su, H. Bao, Z. Cui, and G. Zhang, “Hybrid tracker with pixel and instance for video panoptic segmentation,” *arXiv preprint arXiv:2203.01217*, 2022.

[7] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, “Conditional object-centric learning from video,” in *International Conference on Learning Representations*, Apr. 2022.

[4] Some labels are on my GitHub. GT can be better visualized through the Computer Vision Toolkit (cvkit).

[5] https://github.com/SuperBruceJia/Samsung_Internship/tree/main/pre-process-dataset

[8] C. Yang, H. Lamdouar, E. Lu, A. Zisserman, and W. Xie, “Self-supervised video object segmentation by motion grouping,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7177–7188, June 2021.

[9] M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu, and P. Luo, “Every frame counts: Joint learning of video segmentation and optical flow,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 10713–10720, Feb. 2020.

[10] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, and D. Metaxas, “Global matching with overlapping attention for optical flow estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17592–17601, June 2022.

[11] Z. Teed and J. Deng, “DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16558–16569, Dec. 2021.

[12] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, “MonoDTR: Monocular 3d object detection with depth-aware transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4012–4021, June 2022.

[13] **S. Jia**, Y. Hou, X. Lun, Z. Hao, Y. Shi, Y. Li, R. Zeng, and J. Lv, “GCNs-Net: A graph convolutional neural network approach for decoding time-resolved eeg motor imagery signals,” *IEEE Transactions on Neural Networks and Learning Systems*, Aug. 2022. DOI: 10.1109/TNNLS.2022.3202569.

[14] Y. Hou, **S. Jia**, X. Lun, S. Zhang, T. Chen, F. Wang, and J. Lv, “Deep feature mining via the attention-based bidirectional long short term memory graph convolutional neural network for human motor imagery recognition,” *Frontiers in Bioengineering and Biotechnology*, vol. 9, Feb. 2022. DOI: 10.3389/fbioe.2021.706229.

[15] Y. Hou, L. Zhou, **S. Jia**, and X. Lun, “A novel approach of decoding eeg four-class motor imagery tasks via scout esi and cnn,” *Journal of Neural Engineering*, vol. 17, no. 1, p. 016048, Feb. 2020. DOI: 10.1088/1741-2552/ab4af6.

[16] **S. Jia**, B. Chen, D. Li, and S. Wang, “No-reference image quality assessment via non-local dependency modeling,” in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, IEEE, Aug. 2022.

[17] J. Wu, S. Yarram, H. Liang, T. Lan, J. Yuan, J. Eledath, and G. Medioni, “Efficient video instance segmentation via tracklet query and proposal,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 959–968, June 2022.

[18] O. Thawakar, S. Narayan, J. Cao, H. Cholakkal, R. M. Anwer, M. H. Khan, S. Khan, M. Felsberg, and F. S. Khan, “Video instance segmentation via multi-scale spatio-temporal split attention transformer,” *European Conference on Computer Vision*, Oct. 2022.