HVPS: A Human Video Panoptic Segmentation Framework

Yizhou Wang, Haotian Zhang, Zhongyu Jiang, Jie Mei, Cheng-Yen Yang, Jiarui Cai, Jenq-Neng Hwang
University of Washington, Seattle, WA, USA
{ywang26, haotiz, zyjiang, jiemei, cycyang, jrcai, hwang}@uw.edu

Kwang-Ju Kim, Pyong-Kun Kim
Electronics and Telecommunications Research Institute (ETRI), South Korea
{kwangju, iros}@etri.re.kr

Abstract

Segmentation and tracking are two crucial tasks in smart city applications. Especially, video panoptic segmentation is a task that manages to assign semantic IDs and tracking IDs to every pixel in a video. In this paper, we propose a robust framework to achieve human tracking and panoptic segmentation simultaneously. First, the human tracking and segmentation are accomplished by a spatio-temporal attention based neural network for object appearance feature extraction and Hungarian tracking algorithm considering short-term retrieval and long-term re-identification (re-ID). Second, we fuse the semantic segmentation from EfficientPS with the human instance segmentation as our panoptic segmentation results. Our proposed framework can achieve 48.6% *of Segmentation and Tracking Quality (STQ), validated on the MOTChallenge-STEP dataset, achieving a new state-of-the-art method.*

1. Introduction

Video panoptic segmentation is aimed to segment and track all pixels in a video. This task is potentially very useful for many smart city applications, e.g., surveillance and security purposes. It is trying to combine the existing panoptic segmentation and multi-object tracking (MOT) tasks together to formulate a new and comprehensive objective.

In this work, we propose a novel framework, named Human Video Panoptic Segmentation (HVPS), to accomplish video panoptic segmentation on MOTChallenge-STEP, which includes a large number of person in the city scenes with both static and moving camera. Our proposed framework can be divided into two steps: 1) human tracking and segmentation; 2) semantic segmentation for background. The human tracking and segmentation part can be addressed by the recent proposed topic, called multi-object tracking and segmentation (MOTS) [11]. MOTS task tries to predict object masks with track ID for each instance in a video. After that, the remaining segmentation is the background, i.e., stuff classes in panoptic segmentation. For MOTS, we first propose a robust embedding feature extractor with spatio-temporal attention to extract reliable and distinguished features for each object. After that, a Hungarian algorithm is utilized to track the objects. The short-term retrieval and long-term re-ID are adopted to refine the tracking performance.

The evaluation results are based on Segmentation and Tracking Quality (STQ) on MOTChallenge-STEP dataset [12], which jointly consider segmentation and tracking quality into one metric. Overall, our proposed HVPS achieves the first place in this challenge with 48.6% STQ.

The contributions of this paper can be summarized as follows:

- Propose a new framework, named Human Video Panoptic Segmentation (HVPS), which is a human-based multi-object tracking and segmentation pipeline with background semantic segmentation.
- Consider both spatial and temporal attention during the feature extraction stage to obtain more robust embedding features for multi-object tracking.
- Conduct short-term retrieval and long-term re-ID mechanisms into a Hungarian algorithm for the tracking step.
- Achieve the state-of-the-art on MOTChallenge-STEP dataset with 48.6% of STQ.

2. Related Works

Multi-Object Tracking and Segmentation. The concept, Multi-Object Tracking and Segmentation (MOTS), was proposed in [11]. The baseline method of MOTS,

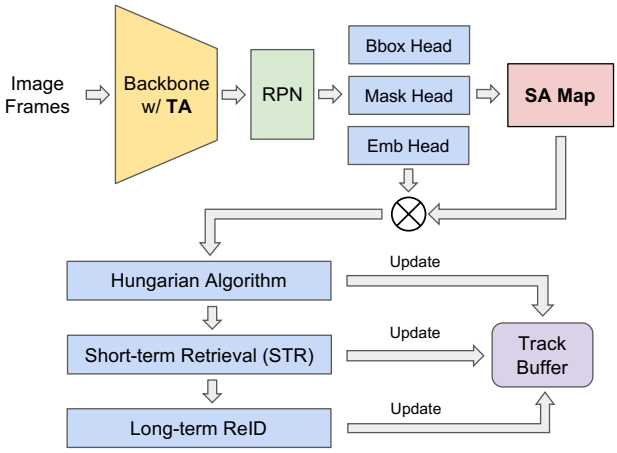

Figure 1. Our human tracking and segmentation framework with spatio-temporal attention.

named Track R-CNN, is a conventional Mask R-CNN [5] with two temporal 3D convolution layers to incorporate the adjacent frames and an additional embedding head to extract instance features for tracking. The embedding head idea is consistent with the aforementioned joint detection and tracking, which is efficient and practical.

Video Panoptic Segmentation. Recently, combining all the above methods with semantic segmentation, the Video Panoptic Segmentation is introduced. Motion-DeepLab [12] firstly proposed this task and released a dataset called KITTI-STEP. Then, furthermore, ViP-DeepLab [9] integrated the depth into this task and proposed Depth-aware Video Panoptic Segmentation, which not only requires tracking and panoptic segmentation results through frames but also requires accurate depth estimation results.

3. The Proposed Framework

Our proposed HVPS framework mainly consists of two parts, i.e., human tracking and segmentation and background segmentation. The human tracking and segmentation are based on a novel feature extractor with spatio-temporal attention, appended by a Hungarian tracking algorithm [2]. Whereas the background segmentation is inferred from EfficientPS [8] and fused with our previous human instance segmentation results. The details of these two parts are described below.

3.1. Human Tracking and Segmentation

In this section, we introduce a video-based multi-object tracking and segmentation (MOTS) framework with spatio-temporal attention, shown in Fig. 1. First, a two-stage CNN-based object detector with cascaded multi-task heads is proposed for separately handling bounding box regression, object classification, mask generation, and instance-aware embedding feature prediction with spatio-temporal attention for video clips.

Temporal Attention. The temporal attention aggregates information from short video clips for the smoothness of prediction. We use the features of three consecutive frames generated by a feature pyramid network (FPN) [7]. The temporal attention (TA) module learns a pixel-wise attention map for each pyramid level and uses them to compute weighted features. First, the features of the three frames are concatenated, and 3D convolutions are applied to achieve feature aggregation. Then, Softmax is performed over each pixel location to generate the TA maps, which are used to create the TA feature, the weighted sum of the original FPN features.

Spatial Attention. The spatial attention performs the interest within each object bounding box and removes ambiguous and irrelevant features. The intuition of the SA module is to highlight the foreground and suppress the background so that more concentrated appearance features can be obtained. To achieve this, the features from the backbone are passed through 2D convolutional layers and a Sigmoid operation to generate a SA map, indicating the probability of objectness or foreground. Then, with the intermediate output of the SA module, several 2D convolution layers are applied to produce the object mask.

Tracking. The overall tracking method is based on [2], considering both short-term retrieval and long-term re-ID. First, the matched tracks are updated with new detections, while the unassigned ones are sent into a short-term retrieval module to be associated with the live tracks without detection in the previous frames. Besides, the bounding box intersection-over-union (IoU) and distance are considered as other constraints during the tracking. Afterward, re-ID could reduce identity switch (IDS) by reconnecting broken-up tracks. Here, the long-term occlusions are recovered by feature-based re-ID [6].

3.2. Semantic Segmentation

EfficientPS [8] is a novel and state-of-the-art semantic as well as panoptic segmentation method and is adopted in our pipeline for semantic segmentation branch since the original instance segmentation results from EfficientPS are not precise enough comparing with our HTS method.

By combining a two-stage object segmentation branch with a semantic segmentation branch, EfficientPS is able to generate the final panoptic segmentation. Furthermore, especially for the semantic branch, Feature Pyramid Network

Table 1. Quantitative results on the MOTChallenge-STEP dataset.

Method	STQ↑	AQ↑	SQ(IoU)↑
siain	31.8	15.4	65.7
EffPS_MM	42.8	26.4	**69.2**
IPL_ETRI (Ours)	**48.6**	**43.3**	54.5

Table 2. Multi-object tracking and segmentation results on the MOTS dataset. (Best in **bold**, second is underlined)

Method	sMOTSA↑	IDF1↑	MOTSA↑	MOTSP↑
TrackRCNN [11]	40.6	42.4	55.2	76.1
SORTS [1]	55.0	57.3	68.3	81.9
GMPHD_MAF [10]	69.4	66.4	83.3	84.2
ReMOTS [13]	**70.4**	**75.0**	**84.4**	84.0
IPL_ETRI (Ours)	69.5	70.3	83.3	**84.2**

(FPN) is utilized for multi-scale feature integration. Thus, EfficientPS provides a solid semantic segmentation result for our pipeline.

During the implementation of EfficientPS, we first use the pre-trained model on KITTI Panoptic Segmentation dataset [4]. The model is then finetuned on the KITTI-STEP dataset and finally finetuned on MOTChallenge-STEP dataset.

4. Experiments

4.1. Dataset and Evaluation Metrics

The MOTChallenge-STEP dataset [12] is extended from MOTChallenge dataset [3]. It consists of 2 training sequences and two testing sequences. The evaluation metric of this challenge is the Segmentation and Tracking Quality (STQ), consisting of two factors, association quality (AQ) and segmentation quality (SQ), that measure the tracking and segmentation quality, respectively.

4.2. Evaluation Results

After evaluating the STQ performance on the MOTChallenge-STEP dataset, we can achieve first place with 48.6% of STQ, 43.3% of AQ, and 54.5% of SQ. The results of the final leaderboard are shown in Table 1.

From the results, we notice that our AQ outperforms other methods by a large margin, which means we do very reliable human tracking and segmentation results. However, our SQ is not very promising. This issue might be due to insufficient training data or strategies.

Besides, we also evaluate our human tracking and segmentation method on the MOTS dataset, shown in Table 2. It shows that our method can also get a great performance on the MOTS task.

5. Conclusion

In this challenge, we proposed HVPS for Human Video Panoptic Segmentation and achieved 48.6% of STQ in the MOTChallenge-STEP dataset, which is the best result in this challenge. However, compared with other methods, the Segmentation Quality (SQ) of our approach is relatively worse. Thus, in the future, temporal information can be added to our framework for superior semantic segmentation results.

References

[1] Martin Ahrnbom, Mikael G Nilsson, and Håkan Ardö. Real-time and online segmentation multi-target tracking with track revival re-identification. In *VISIGRAPP (5: VISAPP)*, pages 777–784, 2021. 3

[2] Jiarui Cai, Yizhou Wang, Haotian Zhang, Hung-Min Hsu, Chengqian Ma, and Jenq-Neng Hwang. Ia-mot: Instance-aware multi-object tracking with motion consistency. *arXiv preprint arXiv:2006.13458*, 2020. 2

[3] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 3

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[6] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *CVPR Workshops*, 2019. 2

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[8] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021. 2

[9] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh 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*, pages 3997–4008, 2021. 2

[10] Young-min Song and Moongu Jeon. Online multi-object tracking and segmentation with gmphd filter and simple affinity fusion. *arXiv preprint arXiv:2009.00100*, 2020. 3

[11] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 7942–7951, 2019. 1, 3

[12] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 1, 2, 3

[13] Fan Yang, Xin Chang, Chenyu Dang, Ziqiang Zheng, Sakriani Sakti, Satoshi Nakamura, and Yang Wu. Remots: Self-supervised refining multi-object tracking and segmentation. *arXiv preprint arXiv:2007.03200*, 2020. 3