Robust Video Panoptic Segmentation and Tracking

Jincheng Lu Yue He Minyue Jiang Meng Xia Wei Zhang Xiao Tan
YingYing Li Hao Sun Errui Ding
{*lujincheng01, heyue04, jiangminyue, xiameng02, zhangwei99, tanxiao01, liyingying05, sunhao10, dingerrui*}*@baidu.com*

Abstract

In this paper, we propose a video panoptic segmentation and tracking framework. Our proposed framework called REPAET, encompasses an ensemble of the segmentation results of leading frameworks such as MaskFormer and HMSA, together with a carefully designed multi-object tracking strategy. We conduct extensive experiments using the newly proposed KITTI-STEP dataset and the evaluation metric named Segmentation and Tracking Quality (STQ). The results show that our proposed method achieves leading results on KITTI-STEP.

1. Introduction

Pixel-wise analysis of video scene understanding is highly demanded in many areas such as autonomous driving. Towards this goal, one new task Segmenting and Tracking Every Pixel (STEP) has been proposed, which contains two data sets KITTI-STEP and MOTChallenge-STEP together with a new evaluation metric, named Segmentation and Tracking Quality (STQ) . Compared with the task of panoptic segmentation, STEP incorporates extensions in video domain and requires to assign a semantic class and identity-preserving track ID to each pixel throughout the video. Compared with the task of multi-object tracing and segmentation (MOTS), STEP not only consider foreground instances but also take the non-countable regions into considerations, such as the sky and crowds. Figure 1 visualizes of the image and the panoptic segmentation on KITTI-STEP. In this paper, we propose a video panoptic segmentation and tracking system and verify through extensive experiments that our proposed method demonstrates leading performance.

2. Methodology

In this section, we introduce the panoptic segmentation and multi-object tracking strategy in our proposed framework, respectively.

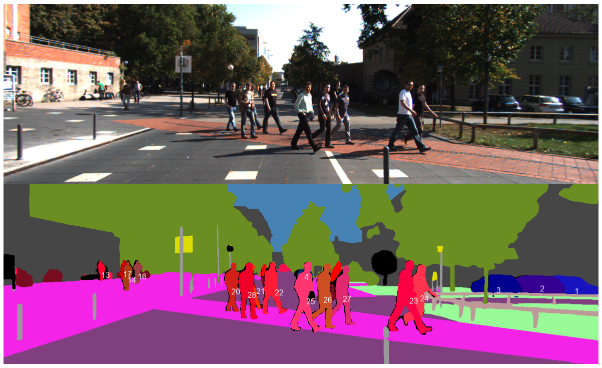

Figure 1. The visualization of the image and the video panoptic segmentation and tracking example on KITTI-STEP.

2.1. Panoptic Segmentation

We adopt the state-of-art segmentation network MaskFormer [2] to solve the panoptic segmentation task. MaskFormer predicts a set of binary masks, each of which is associated with a single global class label. In this way, MaskFormer can offer both instance- and semantic-level segmentation results. In order to boost the performance of our model, we choose the Swin-Large version of Swin-Transformer [5] as our backbone architecture. Furthermore, similar as in Panoptic-DeepLab [1], we also increase the weights for small objects during the training phase of MaskFormer. We verify through experiments that this modification greatly improves the overall performance of MaskFormer, surpassing the results of Panoptic-DeepLab on KITTI-STEP dataset.

To further improve the performance of semantic segmentation, we adopt Hierarchical Multi-Scale Attention (HMSA) [7] which combines multi-scale prediction results. Specifically, we adopt HRNet-OCR [11] as the backbone for HMSA and incorporate dice loss as well as the bootstrapped cross entropy loss instead of cross entropy during training.

Finally, we merge the results of MaskFormer and HMSA to obtain our segmentation results.

2.2. Multi-Object Tracking

Provided with the high-quality panoptic segmentation results, our proposed multi-object tracking method focuses on associating instances throughout the video frames. We first initialize a number of tracklets based on the instances detected in the first frame. Then in the subsequent frame, we associate the estimated instances with the existing tracklets according to their similarity.

To realize a robust association, we propose to measure the similarity of the combination of the appearance feature and the IoU of temporally propagated masks. Specifically, we employ two ReID models using the state-of-the-art ReID framework [3] to extract 2048-dimension appearance features for ”car” and ”person”, respectively. Then, similar to [9], we adopt optical flow method RAFT[8] to warp each predicted mask at frame $t-1$ into frame t and calculate the mask IoU between the masks in consecutive frames. Further, we form the combined similarity distance as the product of cosine distance of the ReID features and the mask IoU distance. Finally, we adopt Hungarian algorithm[4] to perform the association.

To minimize the accumulated errors in long-term tracking, one common practice is that the unmatched predictions are kept for a limited number of frames. However, this often result in broken trajectories when the instance reappears after a long period of occlusion. Thus, we further incorporate post-processing strategies to link these broken tracklets. Specifically, we adopt a greedy algorithm to merge the non-overlaping tracklets based on the similarities of their tracklet-level appearance feature.

3. Experiments

KITTI-STEP Dataset All experiments are conducted on the KITTI-STEP Dataset. This dataset is built on KITTI-MOTS Dataset, which has 21 and 29 sequences for training and testing, respectively. Validation set is further split from the training sequences, making 12 sequences for training and 9 sequences for validation. Readers can refer to [9] for more details.

Evaluation Metrics STQ metric is used for evaluation, which combines Association Quality(AQ) with Segmentation Quality(SQ).

$$STQ = (AQ \times SQ)^{\frac{1}{2}} \tag{1}$$

Implementation details For segmentation models, we first pretrain MaskFormer and HMSA using the cityscapes dataset then finetune the pretrained models using the KITTI-STEP dataset.

For ReID models, we use the minimum enclosing rectangle of instance masks to crop cars and persons from the original images for training CNN models. Two separate CNN models pretrained on VERI-Wild[6] and MSMT17[10] are utilized to extract features for car and person.

Model	SQ
MaskFormer	71.23
HMSA	71.06
Merged	**73.02**

Table 1. Segmentation results on the validation set.

Method	w/ post	$carAQ$
Temporal Propagation		74.37
ReID		73.56
ReID	✓	74.30
Temporal Propagation × ReID		77.31
Temporal Propagation × ReID	✓	**78.90**

Table 2. Comparisons with different strategies for associating instances on the validation set.

Method	STQ	AQ	SQ
UW_IPL/ETRI_AIRL	67.55	71.26	64.04
REPEAT(Ours)	67.13	65.81	68.49
EffPS_MM	62.93	61.49	64.41
siain	57.87	55.16	60.71
HybridTracker	54.99	54.44	55.54
Motion-Deeplab[9]	52.19	45.55	59.81

Table 3. Overall scores on KITTI-STEP test set.

3.1. Ablation Experiments

The ablation study on the KITTI-STEP Dataset mainly shows: (1) the results of using different segmentation methods; (2) the effectiveness of tracking strategies. Here, tracking strategy experiments are conducted only on car instances for simplicity.

As we can see from table 1, by merging two different semantic models' output, SQ is boosted from 71.23 to 73.02, achieving 1.8% improvement.

Beside, Table 2 shows that considering temporal propagation (optical flow) and feature similarity (ReID) simultaneously is able to benefit high quality associations.The post-processing strategy also improves our result.

3.2. Overall Score on KITTI-STEP Dataset

As shown in Table 3, our proposed robust video panoptic segmentation and tracking method shows superiority on SQ score and yields a competitive STQ score.

4. Conclusion

In this paper, we present our solution for the ICCV2021 Workshop 6th Benchmarking Multi-Target Tracking Competitions Video Track. The overall STEP task is divided into panoptic segmentation and instance associating, and both are optimized to provide robust tracking. Extensive experiments have shown our effectiveness and superiority to other methods.

References

[1] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020.

[2] Bowen Cheng, Alexander G Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint arXiv:2107.06278*, 2021.

[3] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

[4] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[6] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019.

[7] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[9] 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.

[10] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person reidentification. *CoRR*, abs/1711.08565, 2017.

[11] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer, 2020.