Tracking by Segmentation: Person-ReID and Optical Flow Based Offline Tracker for the MOTSChallenge 2020

Yang Liu[*]
Technical University of Munich
AI Labs, DiDi
yang14.liu@tum.de

Lyuwei Wang
AI Labs, DiDi
wanglvwei@didiglobal.com

Yuan Zhao
AI Labs, DiDi
zhaoyuanjason@didiglobal.com

Haifeng Shen
AI Labs, DiDi
shenhaifeng@didiglobal.com

Jieping Ye
AI Labs, DiDi
yejieping@didiglobal.com

Abstract

In this work, we follow the classic Tracking-by-Detection approach. We build upon the previous unsupervised video object segmentation method known as UnOVOST[8]. Since the main components (segmentation network, optical flow vectors, ReID embeddings) of this method are highly modularized, we also took our effort on adapting each part of the components to fit the MOTS20 dataset in order to achieve a better tracking result.

1. Introduction

Multiple Object Tracking (MOT) is a task of locating different objects in a video and maintain their identities throughout space and time. Previous works utilize high performance detectors to regress a bounding box around the objects of interest to indicate their location in the image. But the accuracy of the bounding box easily drops when the objects are heavily occluded, thus limiting the performance of tracking algorithms. With the hope of pushing the tracking performance by replacing bounding box with pixel level annotations, the 2020 MOTS challenge [13] added instance segmentation annotations to previous benchmark dataset and published KITTI-MOTS and MOTS20. The tracking algorithm in this paper is mainly aimed for MOTS20 dataset, which contains 4 videos for training and 4 videos for testing.

Before MOTS20 [13] there were several benchmarks targeting at video object segmentation and tracking, such as DAVIS[10][11][1][2], YT-VOS[15] and YT-VIS[16]. Comparing to MOTS20[13] which only have one class: person, the other three benchmark datasets contain more types of objects in different video sequences. Although the classification task in MOTS20[13] regresses to simple distinction of foreground (person) and background (other things), the challenge lies in accurately segmenting occluded person and tracking pedestrians in a longer sequence length with much higher occlusion rate because the scenes are constantly crowded. In this paper, we demonstrate our attempt to solve this task by adapting UnOVOST[8], a highly modularized tracking algorithm based on optical flow and ReID embeddings. We focus on improving the segmentation network to generaete finer proposals, and replace the original ReID module with a specialized person-ReID network for extracting re-identification vectors for each proposals.

2. Method

Our approach is an adaptation of UnOVOST[8], which won both the 2019 DAVIS Challenge and the 2019 YouTube-VIS Challenge. UnOVOST[8] is designed to be modular, this makes it easy to replace its components and thus can be conveniently adapted to tackle different kinds of video segmentation and tracking problems. The framework proposed by UnOVOST[8] follows three stages, as shown in Figure 1. Stage 1 is the proposal generation, the instance segmentation network will process every frame in each video and generate proposals. In stage 2, short tracklets will be generated by connecting corresponding proposals between two consecutive frames. Then short tracklets which belong to the same object are merged using ReID embeddings as visual similarity cues in stage 3. In the following sections, we discuss the adaptations we made for these 3 stages.

[*] Work done during an internship at DiDi AI Labs

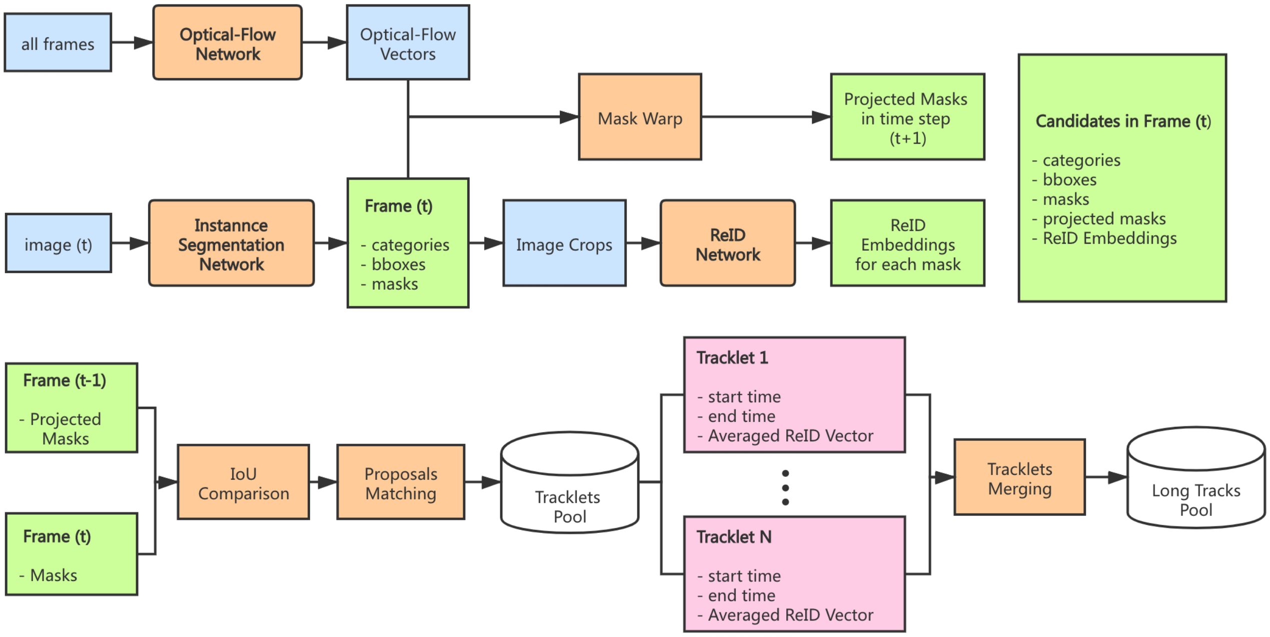

Figure 1. Overview of the method

2.1. Segmentation

The tracking results dependent heavily on the quality of the segmentation. In most of the top-down tracking methods, Mask R-CNN[5] with ResNet[6] as backbone is still the common choice for proposal generation. But lacking of cross channel interaction and the limited size of receptive field makes the feature extraction ability of ResNet suboptimal. Successive variants of ResNet have shown fruitful results in various computer vision tasks. ResNeSt[17] is one of the successful variants and recently achieved state-of-the-art performance on different benchmarks.

We use ResNeSt[17] as the backbone for Cascade Mask R-CNN[3] and perform frame by frame segmentation to generate proposals. We take COCO pretrained weights from [17] and then finetune the network on the training set. The training set contains only 4 videos with 2862 frames in total, and most of the frames within the same sequence have similar image content. To prevent the model from overfitting, we freeze all the layers except the ROI-heads for bounding box regression and mask prediction. This significantly reduces the training parameters and speeds up the finetuning process. Furthermore, we randomly insert COCO data which has the class ”person” into the training set and apply common image augmentation techniques. In order to enhance the model so that it can produce finer segmentation for occluded person, we hand picked key frames with heavy occlusions and repeat them in the dataset. During inference time, we uniformly filter out proposals with confidence lower than 0.5 for all test videos.

2.2. Tracklets Generation

According to UnOVOST[8], tracklets are generated using optical flow vectors. We use PWC-Net[12] with pretrained weights from [7] to process every two consecutive frames to get the corresponding optical flow vectors that describe the motion field from frame $t-1$ to frame t. For every proposal in frame $t-1$, the mask will be warped by the corresponding optical flow vector. The warped mask represents the estimated shape of the object in the next frame. By comparing the IoU score of the masks in current frame t and the warped masks from previous frame $t-1$, we can determine which two masks belong to the same object. Borrowing the idea from UnOVOST[8], the optimal match of proposals from frame $t-1$ and in frame t is determined by the Hungarian algorithm.

2.3. Tracklets Merging

Optical flow based tracking shows remarkable results especially when the objects don’t have large motion in the video. This suits very well for the MOTS20 dataset since the movements of pedestrians are relatively slow and predictable. But tracking based on optical flow only helps to connect tracklets that are present in continuous frames. In the case of heavy occlusions, or when people exit and then re-enter the scene, we use ReID embeddings to compare the visual similarity of different tracklets, and merge them into one longer track if they belong to the same object. Since the targets in MOTS20 dataset are comprised solely of person, we employ a specialized ReID network named MGN[14] for person re-identification. We use pretrained weights from [14] and then finetune on MOTS20 training set. During inference time, we use our finetuned person-ReID network to extract re-identification vector of length 2048 for each proposals cropped from the input images.

Proposals that are merged into one tracklets are assumed to belong to the same object. Sometimes these proposals contain masks that only cover a small part of the human body because of occlusion. The re-identification vectors for these proposals are inevitably to deviate from the the ones that cover the whole body. To compensate the outliers, UnOVOST[8] stacks all the re-identification vectors within one tracklet vertically and take the average value on each dimension. We adapt that idea and add a preprocess step. We consider tracklets that contain decent amount of proposals and calculate a reasonable range of each dimension based on the distribution of the values. Values that deviate too much from this range are ignored before calculating the average.

In the final step, each tracklet has an averaged re-identification vector of dimension $(1, 2048)$ as its representation. We then use the Forest Path Cutting(FPC) algorithm[8] to merge tracklets into long tracks that may span the whole video.

2.4. Post-processing

In the MOTS20 challenge, segmentations are not allowed to overlap with each other. Most of the proposals generated by our finetuned segmentation network don't have an overlap greater then 0.1 (judged by IoU score). During our experiment, we observe that, assigning the overlapped pixels to smaller objects rather then bigger objects has a 0.1 boost for the sMOTSA[13] score. Thus we define a simple criterion to always assign the overlapping pixels to a smaller object.

Because of the confidence score clipping, some proposals are dropped. This has the side effect that, some tracks that could have been connected by only using optical flow vectors, are now seperated into short tracklets. Although cases like this can be easily fixed by using ReID embeddings for merging, we still lose some true positives because of the missing proposals. Therefore, in the post-processing step we also pick out merged tracks and check if there are missing proposals that only span one time step. If so, we fill in the gap by merging the projected mask from previous frame$(t-1)$ and the backward-projected mask from next frame$(t+1)$.

3. Experiments

The initial evaluations were performed on the MOTS20 training set with the help of the provided evaluation tools[1]. Since the building blocks of the tracking algorithm is highly modularized, we first tested the importance of different parts. Given that the training set is small, train/validation split was not performed. We picked two best performing networks on COCO instance segmentation benchmark, BlendMask[4] and Cascade Mask R-CNN[3] with ResNeSt[8] backbone, and trained them with only COCO "person" class for comparison. As mentioned above, all proposals whose confidence score is below 0.5 are clipped. With finer segmentation results, the tracking algorithm gets better at connecting proposals using optical flow and has less chance to mix other objects into the same tracklet. We then compared the tracking performance with two different ReID networks. The ReID network[9] used in PReMVOS[7] is originally designed for objects with diverse classes and appearances. While specialized ReID network MGN[14] manage to reduce the ID-switches by 19 in total, it doesn't provide a significant performance boost on the tracking score. But comparing with the tracking strategy that doesn't use ReID embeddings, we observed that the merging strategy based on ReID embeddings significantly reduce the numbers of ID-switches. Table 1 shows our experiment results on MOTS20 trainset.

We also investigated the trade-off between confidence clipping and the gap filling using optical flow warped masks. By allowing proposals with lower confidence to participate in the tracking process, the tracking performance dropped. One reason is that more false positives were introduced by low confidence proposals. Another reason is that, low confidence proposals usually appear on heavy occluded person. In cases like this, small body parts of the occluded person are usually grouped together with another person. Optical flow based merging strategy only considers the shape of the masks, but ignores the appearance of the person. Therefore, ID-switch happens more often when two people move across each other. Considering cases like this, we use confidence clipping of 0.5 on the test set, and use post-processing step to compensate part of the missing true positives. The performance of our methods on the MOTS20[13] test-set is shown in Table 2

4. Conclusion

In this work, we adapt UnOVOST[8] for the multi-object tracking and segmentation task. We found out that the quality of the pedestrian segmentation, especially a finer segmentation for heavy occluded person, plays an important role for the tracking method to connect

[1] https://github.com/VisualComputingInstitute/mots_tools

Proposals from	ReID Network	sMOSTA score	TP	FP	IDS
BlendMask (ResNet101)	No ReID	60.7	21787	1477	483
BlendMask (ResNet101)	PReMVOS ReID	61.4	21792	1472	294
BlendMask (ResNet101)	MGN	61.5	21800	1470	275
Cascade Mask R-CNN (ResNeSt200)	MGN	64.4	22890	1687	230
Provided segmentations	MGN	64.9	23006	1960	281

Table 1. Tracking results on MOTS20 train-set with different components

Tracker	sMOTSA	IDF1	MOTSA	MOTSP	MODSA	TP	FP	Recall	Precision	ID Sw
ReMOTS	**69.9**	**75.0**	**83.9**	84.0	**85.1**	28270	**819**	87.6	**97.2**	388
PTPM	68.8	68.5	82.6	84.1	83.7	28108	1084	87.1	96.3	368
GMPHD_SAF	68.4	64.9	82.6	83.9	84.4	**28382**	1161	**88.0**	96.1	569
PT	66.8	67.3	79.9	**84.5**	81.1	27215	1059	84.3	96.3	370
ours	66.6	71.8	79.7	84.4	80.7	27114	1067	84.0	96.2	**341**

Table 2. Our results on MOTS20 test-set comparing to the other top five tracking methods

corresponding proposals across different frames accurately. We also replace the general purpose ReID network to a specialized person ReID network to better suit the tracking of pedestrians. Our improvements upon UnOVOST[8] achieved decent result on the 2020 MOTS Challenge[13]. As future work, we proposal to incorporate methods like bi-directional optical flow or integrate visual similarity in the optical flow based merging step. For a better video segmentation result, it is also worthwhile to consider adding temporal information into the network for a more context-aware and temporal coherent segmentation among continuous frames.

Acknowledgements This work was supported by DiDi AI Labs. We would also like to thank Jonathon Luiten and Idil Esen Zulfikar for their inspiring work on UnOVOST[8].

References

[1] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv:1803.00557*, 2018.

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *arXiv preprint arXiv:1906.09756*, 2019.

[4] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[7] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 2018.

[8] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *Proceedings of the IEEE Winter Conference on Applications in Computer Vision*, 2020.

[9] Aljosa Osep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object discovery and detector adaptation from unlabeled video. *CoRR*, abs/1712.08832, 2017.

[10] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[11] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[12] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. 2018.

[13] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.

[14] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *ArXiv e-prints*, Apr. 2018.

[15] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018.

[16] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *CoRR*, abs/1905.04804, 2019.

[17] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Muller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.