Simple re-Identification feature Association for Robust Multi-Object Tracking and Segmentation

Jeongwon Ryu
SI Analytics
34047 Yuseong-gu Daejeon, Korea
wjddnjs7678@gmail.com

Kwangjin Yoon
SI Analytics
34047 Yuseong-gu Daejeon, Korea
yoon28@si-analytics.ai

Abstract

Though tremendous progress have been made in multiple-object tracking and segmentation (MOTS), which jointly perform multiple object tracking (MOT) and instance segmentation, designing a robust tracker for various domains remains an open challenge. In this work, we present a pragmatic approach, the Simple re-Identification feature Association framework (SIA). We designed a re-ID feature extractor based on IBN-Net50-a with batch normalization neck and trained it with pre-defined detection and segmentation results. We also applied various effective tricks for training the network. In the tracking stage, robust appearance feature extraction and simple measurement-to-track associations are applied. Our method achieves 3rd place on the RobMOTS Challenge workshop of CVPR 2021.

1. Introduction

Multi-object tracking (MOT) is considered one of the most challenging problems in computer vision research. Most existing approaches [18, 1, 15, 21] follow the tracking-by-detection paradigm. It isolate the problem of object detection from object tracking. Accordingly, tracking task can focus on the track management, *i.e.*, track initiation, termination, and data association. For object detection, an object detector [5, 23] is used to detect objects frame by frame. For the track management, similar object instances are linked across time where the similarity is measured by either both spatial domain [18, 1] and appearance domain [17] or one of those. Then, incoming instances of current frame that are not associated with previous one initiate new tracks, and a track is terminated if it has been repeatedly missed in previous frames.

In this paper, we present an object tracking and segmentation method, Simple re-Identification feature Association (SIA). Our method adopted the track-by-detection paradigm. Initially, a detector or a segmentor locates object instance from an image, then objects are associated with previous one that is close in the appearance domain. To measure the appearance similarities among object instances, we train a re-identification network [8] whose backbone is IBN-Net50-a [9].

Our contributions are summarized as two folds: (1) we present a new MOTS algorithm, SIA, that robustly tracks object segmentations, (2) With our simple and practical method, we ranked third place on the RobMOTS Challenge.

2. Related Works

2.1. Multi-Object Tracking

Modern MOT method based on the tracking-by-detection paradigm achieve a good performance thanks to the improvement of detection performance [18, 1, 15, 21]. In [18], they used FrRCNN [12] as a detector and associate detection responses by finding the minimum cost of the assignment cost matrix. The cost matrix is computed IoU (intersection over union) between each detection and all predicted bounding boxes from the existing targets. [1] proposed a method similar to [18] except that they used a greedy algorithm to find the minimum association. In [15, 21], they proposed a network that can jointly detect an object and extract a feature of the detected object. The main difference between them is that [15] used YOLOv3 [11] as the detector and [21] used [23]. After they detect and extract object, they applied the similar approaches for tracking task like [18, 1].

2.2. Person Re-Identification

Identifying a person that appeared in past time or different location is called *person re-identification* (person ReID). Recently, much progress has been made in the person reid researches thanks to the deep learning and the advent of challenging datasets. Furthermore, Luo *et al*. [8] thoroughly investigated effective training tricks and design choices for person ReID, and they set a current state-of-the-art baseline for person ReID by combining tricks they found to-

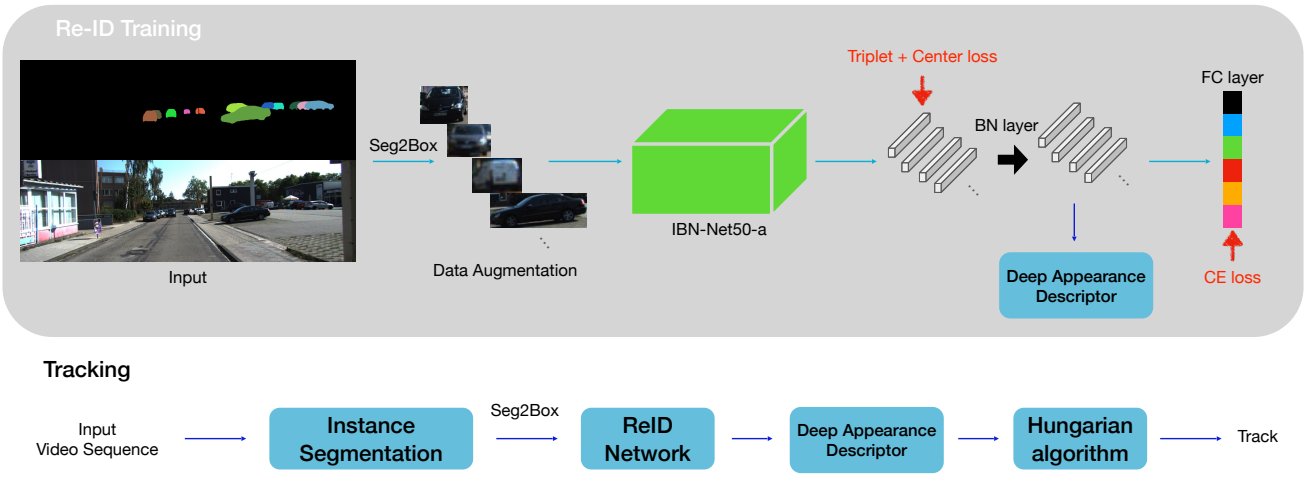

Figure 1. An overview of the proposed simple re-identification feature association approach.

gether. We used ReID network of [8] whose backbone is IBN-Net50-a [9] to extract ReID features.

3. Methodology

The proposed SIA includes two steps: Re-ID feature extraction and online object tracking with acquired Re-ID features. The overall framework is shown in Figure 1.

3.1. Re-ID Feature Extraction

We take object Re-ID works to get the appearance features of the pre-defined bounding box and segmentation mask. IBN-Net50-a [9] pre-trained on ImageNet [4] is used as our backbone model. Then, following a strong baseline [8], add a batch normalization (BN) layer before the classifier fully-connected layer (FC) layer to simultaneously consider cosine distance optimization and Euclidean distance optimization in image pairs. For training, our Re-ID model includes three losses as follow:

$$L = L_{CE} + L_{Triplet} + \beta L_{Center} \quad (1)$$

The cross-entropy loss (L_{CE}) mainly optimizes the cosine distance, while triplet loss ($L_{Triplet}$) focuses on the Euclidean distance. Center loss (L_{Center}) [16], which minimizes intra-class variations while keeping the features of different classes separable, makes up for the drawbacks of the triplet loss. β is the balanced weight of center loss. In our experiments, β is set to be 0.0005. In the inference stage, we extract the 2048-dimension feature after the BN layer as the appearance feature of the observations.

3.2. Online Object Tracking

Objects with 80 categories bring a highly irregular challenge while tracking objects across 8 different benchmarks. Combining with the camera motion, object trajectories shown in those videos are very difficult to predict, and SORT-like trackers based on the Kalman filter cannot achieve satisfactory results.

Therefore, without considering motion information, we acquire the appearance features from the Re-ID model and compute the appearance affinity matrix A_e between all observations and the tracklet of the pool for the incoming frame. The appearance affinity is computed using cosine similarity. Then we solve the linear assignment problem by Hungarian algorithm with cost matrix $C = A_e$.

We adopt the matching cascade proposed in Deep-SORT [18] to solve a series of subproblems. First, we compute the association cost matrix and the matrix of admissible associations. We then iterate over track age n to solve a linear assignment problem for tracks of increasing age. A new tracklet will be initialized for an unmatched observation if its confidence is higher than a pre-defined threshold. After that, we update the set of matches and unmatched observations.

4. Experiments

4.1. Datasets

The data of the RobMOTS Challenge consists of MOTS Challenge [14] KITTI MOTS [14], DAVIS Challenge Unsupervised [2], YouTube-VIS [19], BDD100K MOTS [20], TAO [3], Waymo [13], and OVIS [10] datasets are adopted in our experiments. The pre-computed detections are generated from Mask R-CNN X152 [5] and refined by the refinement net [7].

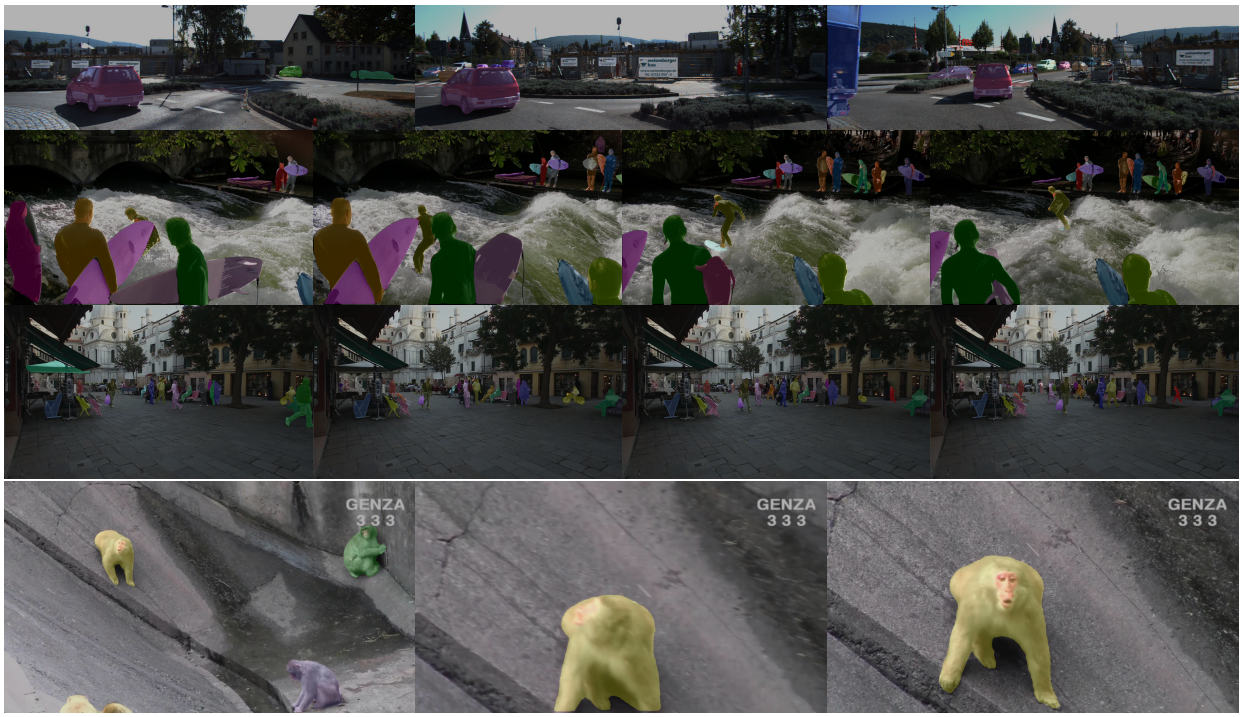

Figure 2. Qualitative results of the proposed SIA. 1st row: KITTI MOTS (moving camera); 2nd row: DAVIS Challenge Un-supervise (moving camera); 3rd row: MOTS Challenge (static camera); 4th row: TAO (moving camera, zoom in).

Rank	Name	Overall	KITTI	BDD	DAVIS	YT-VIS	TAO	MOTS	OVIS	Waymo
1st	RobTrack	61.20%	71.64%	57.86%	56.90%	68.32%	54.99%	61.04%	61.62%	57.21%
2nd	SBT	58.59%	74.01%	53.05%	50.26%	64.41%	51.76%	64.43%	55.61%	55.23%
3rd	SIA	56.87%	70.76%	53.42%	47.42%	62.70%	49.60%	62.18%	54.76%	54.09%
4th	MeNToS	55.52%	69.71%	52.33%	49.60%	64.19%	39.23%	60.15%	55.56%	53.42%
Baseline	STP	54.35%	66.35%	49.35%	48.21%	62.27%	43.76%	60.35%	52.79%	51.75%

Table 1. HOTA metric performance on CVPR 2021 RobMOTS Challenge test set.

4.2. Implementation Details

Data Augmentation. We crop the detected objects from the RobMOTS training set image based on pre-defined bounding boxes and segmentation masks and annotate the ID labels. We resize the patches to 224 x 224 and apply the proposed REA (Random Erasing Augmentation) [22] to solve the occlusion problem and improve the generalization ability of the Re-ID model. More specifically, we delete the rectangular image region, and at every iteration, the region size is randomly defined with upper and lower limits on the region area and aspect ratio.

Training Details. The Re-ID model is trained on an Nvidia A100 GPU following little changes to the default settings of [8]. Here, we omit the other details described in [8].

Testing Details. As max_{age} is a critical hyperparameter, we experimented with the best parameter combination for all datasets. The parameter $max_{age} = 12$ gave the best results in all data sets. As in baseline, detections have been thresholded at a confidence score of 0.5.

4.3. Challenge Results

We report the performance of the SIA to the RobMOTS challenge rating system, with HOTA metrics [6]. It finished third with a 56.87 HOTA final across all eight benchmarks. The quantitative results of the top-4 and baseline methods are shown in Table 1, and some qualitative examples are shown in Figure 2. Specifically, the proposed framework achieves rather low performance in DAVIS but mostly high in other benchmarks, indicating its ability to generalize to various object categories and the potential to handle various Multi-Object Tracking and Segmentation scenarios robustly.

5. Conclusion

In this paper, we propose SIA, a simple framework for robust multi-object tracking and segmentation. We train a Re-ID network based on pre-defined bounding boxes and segmentation masks and perform associations by appearance similarity. The proposed method can track 80 categories of objects using fixed or moving cameras and is flexible in various domains. In addition, it can also be used in real-time scenarios. Our proposed SIA is evaluated across eight benchmarks and achieves 3rd place on the RobMOTS challenge workshop of CVPR 2021.

References

[1] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017*, Lecce, Italy, Aug. 2017. 1

[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 preprint arXiv:1905.00737*, 2019. 2

[3] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[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. 1, 2

[6] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021. 3

[7] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018. 2

[8] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 3

[9] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 1, 2

[10] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 2

[11] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[13] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 2

[14] 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. 2

[15] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2(3):4, 2019. 1

[16] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2

[17] Mikolaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. *arXiv preprint arXiv:2104.13643*, 2021. 1

[18] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2

[19] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2

[20] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[21] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020. 1

[22] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 3

[23] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1