2nd Place Solution for SSLAD Challenge 2022: Multiple Object Tracking

Jinming Su, Junfeng Luo and Xiaolin Wei

Meituan Vision AI Department
{sujinming,luojunfeng,weixiaolin02}@meituan.com

Abstract. In the report, we introduce our solution for ECCV 2022 SSLAD Challenge Track 4: Multiple Object Tracking. In the past few years, multiple object tracking has made great progress with the development of deep learning. However, there are still several problems in the driving scene, including dense objects, large differences in shape and scale, occlusion and blurring. To address these problems, we propose a cascade re-identification tracking framework based on multi-model fusion. In the framework, we propose the motion synthesis of dense objects to alleviate the problem caused by dense objects. Then, we use multi-model fusion to complementarily detect objects with varied shapes and scales. In addition, the cascade re-identification strategy is used to cascade track all objects by using re-identification features and motion features as similarities, which ensures accurate multiple object tracking in various scenes such as occlusion and blurring. Based on this framework, we conducted lots of experiments and finally won second place in the multiple object tracking challenge.

Keywords: Multiple object tracking, autonomous driving, re-identification

1 Introduction

Multiple Object Tracking (MOT) is a fundamental task in computer vision, widely in autonomous driving, video behavior analysis and other applications. To address the MOT problem, many deep learning based models [7, 12, 11] have been proposed in recent years, and constantly improved the state-of-the-art performance on benchmarking datasets [6, 2]. However, many works mainly focus on multi-person tracking. Although the environment is complex and crowded, it can only reflect the effect of person tracking, and there is little research on tracking various objects in the real world.

To understand the temporal association of objects within the videos in the driving scene, BDD100K dataset [10] is released with several tasks including MOT, providing 2,000 videos with about 400K frames with 8 categories of annotations. BDD100K, as a large-scale dataset of the driving scenes, has promoted the development of the MOT field. But, there are still some difficulties in the MOT task. We summarize the main difficulties into four categories: (1) dense objects, *e.g.*, in the first image, (2) large differences in shape and scale, *e.g.*,

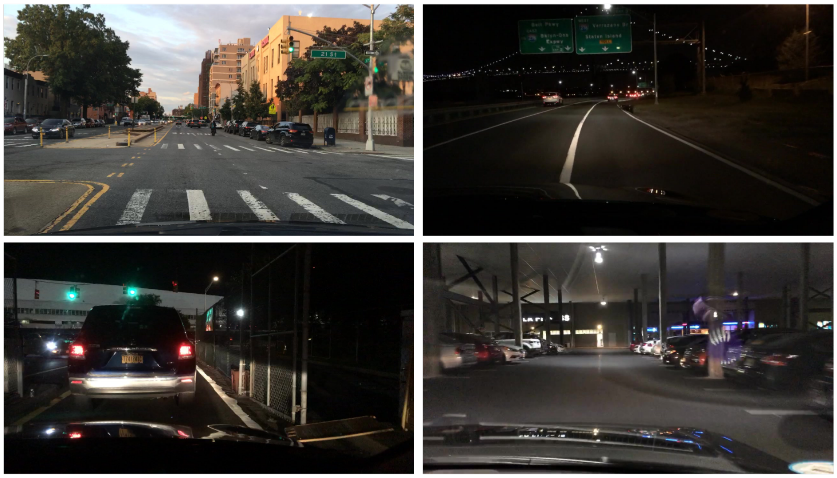

Fig. 1. Representative visualization examples of BDD100K [10].

different illumination and various orientations in the last three images, (3) occlusion, *e.g.*, in the third image, (4) blurring, *e.g.*, in the last image of Fig. 1, which easily lead to missed or false detection and identify switches. Due to these difficulties, MOT remains a challenging task even in the deep learning era.

In order to solve the MOT problem, many methods have made efforts and formed the mainstream tracking-by-detection paradigm, which first detects objects and then associates them over time. These methods can be divided into two categories: single-model methods and separate-model methods based on whether using a single model or separate models to detect objects and extract association features. For single-model methods, FairMOT [12] consists of two homogeneous branches for detecting objects and extracting re-ID features, respectively. The detection branch of FairMOT is implemented in an anchor-free style that estimates object centers and sizes represented as position-aware measurement maps. Similarly, the re-ID branch estimates a re-ID feature for each pixel to characterize the object centered at the pixel. So FairMOT eliminates the unfair disadvantage of the detection branch, effectively learns re-ID features and obtains a trade-off between detection and re-ID. For separate-model methods, deepSORT [9] learns a deep association metric on a large-scale re-identification dataset to track detected objects through longer periods, reducing the number of identity switches. However, these methods may fail in challenging cases of crowded scenes and fast motion, and have difficulty handling occlusion cases.

After the above observation and analysis, we propose a cascade re-identification tracking framework to address the difficulties of MOT, as shown in Fig. 2. In the framework, we first introduce a data augmentation strategy based on the motion synthesis of objects to alleviate the problem caused by dense objects. Next, we use multiple detection models including one-stage and two-stage models with different structures to detect objects with varied shapes and scales, to extract

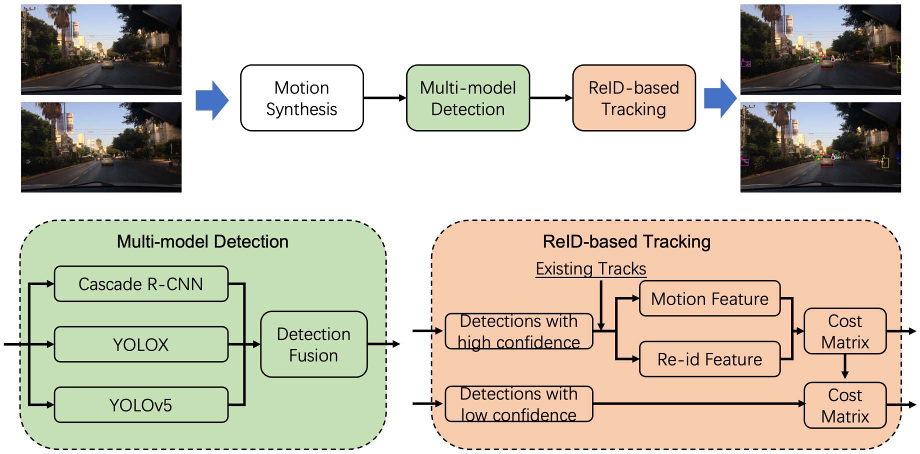

Fig. 2. Framework of our solution.

the information of objects from different aspects so as to achieve complementary promotion. Moreover, we use a cascade re-identification strategy is used to cascade track all objects by using re-identification features and motion features as similarities, ensuring accurate multiple object tracking in various scenes such as occlusion and blurring. Finally, we achieve 46.33% mHOTA on the BDD100K MOT test dataset.

2 Method

To solve the problems (*i.e.*, dense objects, large differences in shape and scale, occlusion and blurring) of MOT, we propose a cascade re-identification tracking framework based on multi-model fusion. In this section, we introduce this framework from three parts as follows.

2.1 Motion Synthesis

In the driving scene, there are usually dense objects, which easily lead to missed detection and identify switches. To alleviate this problem, we propose a motion synthesis method to improve the performance of detection and tracking. In this method, we adopt two ways of data synthesis. In the first way, we extract all the objects in all frames of the training dataset to build an object bank, and then randomly select objects to place them in one of three placing manners including placing randomly, placing along a straight line and densely placing along a radius to obtain the motion of objects. At the same time, the corresponding detection bounding boxes of objects are used as labels of training supervision. In the second way, we extract the information and position of each instance in all videos, make the same position of the instance on new videos, and add their detection bounding boxes into the training supervision. Through these two kinds of motion synthesis, the problems caused by class imbalance and dense objects can be alleviated, and the performance of detection and tracking can be improved.

2.2 Multi-model Detection

In order to detect objects with different shapes and scales, we adopt three different methods to detect targets: two-stage method Cascade R-CNN [1], one-stage methods YOLOX [3] and YOLOv5 [5]. Cascade R-CNN consists of a sequence of detectors trained with increasing IoU thresholds, to be sequentially more selective against close false positives. The detectors are trained stage by stage, leveraging the observation that the output of a detector is a good distribution for training the next higher-quality detector. YOLOX switches the YOLO detector to an anchor-free manner and conducts other advanced detection techniques, *i.e.*, a decoupled head and the leading label assignment strategy SimOTA to achieve state-of-the-art results across a large scale range of models. YOLOv5 is a very popular open source software with various effective modules such as Mosaic data augmentation, adaptive anchor box computing, and so on. We choose these three networks with different structures and configurations to ensure good detection results for objects with different shapes and scales.

In order to fuse the results of different models, we use the weighted boxes fusion method [8] instead of the traditional NMS and soft-NMS, and realize the complementarity of the detection results of different models to obtain better detection results than every single model.

2.3 Cascade ReID-based Tracking

To solve the problem of tracking failure caused by occlusion and blurring, we trained an additional re-id model FastReID [4] to extract re-id similarity as the appearance features. Inspired by ByteTrack [11], we introduce a cascade re-identification strategy to track objects via combing motion and re-id similarity.In the strategy, we first divide the bounding boxes of detection into the high-score boxes and low-score boxes according to the threshold of the detection confidence (the threshold is set at 0.5). Then, the motion features and re-id features are calculated by using the high score box and the previous tracking trajectory, where the motion features are obtained by Kalman filter, and then matched based on the Hungarian algorithm. Next, the low-score box is used for matching with the previous tracking trajectory that didn't match with the high-score box for the first time (for example, the object whose score drops due to severe occlusion in the current frame). Last, for the detection boxes that don't match the previous tracking trajectory and have scores high enough, we will create a new tracking trajectory. For the tracking trajectory that does not match with any detection boxes, we will keep 30 frames. Based on this strategy, we match all the detection boxes with motion and appearance information and ensure the accuracy of various scenes, even occlusion and blurring.

Table 1. Performance on the test leaderboard.

Rank	Team	mHOTA	mMOTA	mIDF1
1	Lenovo_LR_PCIE	49.15	42.97	59.50
2	**bbq**	**46.33**	**38.13**	**55.21**
3		44.36	40.38	52.98
4	CMSQ	42.38	36.15	53.44

3 Experiments

3.1 Dataset

BDD100K is the largest driving video dataset and the dataset splits of the MOT task are 2,000 videos (1400 videos for training, 200 videos for validation and 400 videos for testing) fully annotated 40-second sequences at 5 FPS under different weather conditions, time of the day, and scene types. It needs to track objects of 8 classes and contains cases of large camera motion.

3.2 Metrics

Mean Higher Order Tracking Accuracy (mHOTA, mean of HOTA of the 8 categories) is employed as the primary evaluation metric for ranking. Extra metrics such as mean Multiple Object Tracking Accuracy (mMOTA) and mean ID F1 score (mIDF1) are also employed.

3.3 Implementation Details

All models are trained on BDD100k dataset and pre-trained on ImageNet dataset. For Cascade R-CNN, we adapt ResNet-101 as the backbone and use multi-scale training and AdamW optimizer with an initial learning rate 10^{-4}, the training epoch is set to 36 with the learning rate decayed by a factor of 10:1 at epochs 27 and 33, the image size ranges from (648, 1280) to (720, 1280). For YOLOX-X and YOLOv5-L, we use multi-scale training and SGD optimizer with the initial learning rate is 10^{-3} with 1 epoch warmup and cosine annealing schedule, and the training epoch is set to 50 with weight decay of 5×10^{-4} and momentum of 0.9.

3.4 Performance Analysis

The performance on the test leaderboard is shown in Tab. 1, our approach achieve second place. we also show some of our quantitative results in Figure 3. It can be seen that the proposed solution can accurately track objects in some difficult scenarios including dense objects, large differences in shape and scale, and so on.

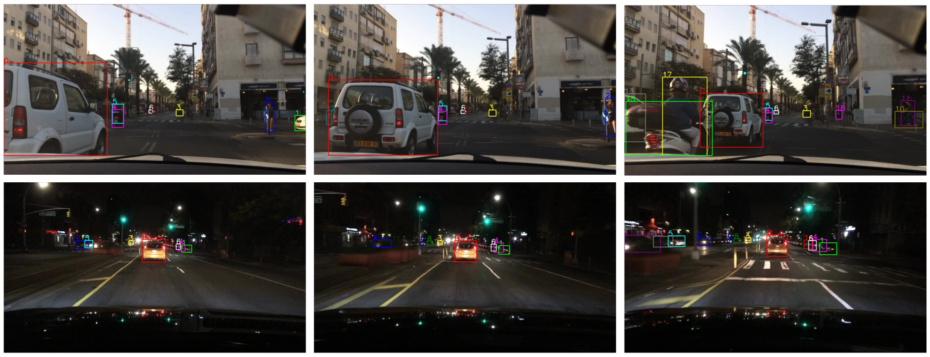

Fig. 3. Representative visual examples from the proposed method.

Table 2. Ablation study on val dataset.

Motion Synthesis	Multi-model Detection			Tracking		mHOTA
	Cascade R-CNN	YOLOX	YOLOv5	motion	re-id	
			✓	✓		41.04
✓			✓	✓		42.59
✓	✓		✓	✓		43.90
✓	✓	✓	✓	✓		44.02
✓	✓	✓	✓	✓	✓	45.72

In order to demonstrate the effectiveness of different components, we conduct several ablation experiments. Quantitative results are shown in Table 2. We boost the performance of the baseline model (yolov5 + SORT) from 41.04% to 45.72% on BDD100K MOT val dataset with the proposed motion synthesis, multi-model detection and reID-based tracking, which shows the effectiveness and compatibility of different components.

4 Conclusion

In this paper, we propose a cascade re-identification tracking framework for multiple object tracking, and make nontrivial improvements and attempts in data, detection and tracking. In the end, we achieved 2nd place on the ECCV 2022 SSLAD Challenge Track 4: Multiple Object Tracking with 46.33% mHOTA.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 6154–6162 (2018)

2. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
3. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
4. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. arXiv preprint arXiv:2006.02631 (2020)
5. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Michael, K., Fang, J., imyhxy, Lorna, Wong, C., Yifu, Z., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, tkianai, yxNONG, Skalski, P., Hogan, A., Strobel, M., Jain, M., Mammana, L., xylieong: ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations. Zenodo (Aug 2022). https://doi.org/10.5281/zenodo.7002879, https://doi.org/10.5281/zenodo.7002879
6. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
7. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 164–173 (2021)
8. Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing pp. 1–6 (2021)
9. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
10. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp. 2636–2645 (2020)
11. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
12. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision (IJCV) **129**, 3069–3087 (2021)