Tracking Multi-objects with Anchor-Free Siamese Network

Bingyu Hui, Jie Feng, Quanhe Yao, Jing Gu, Licheng Jiao

Tracking Multi-objects With Anchor-free Siamese Network

Bingyu Hui[1], Jie Feng[2*], Quanhe Yao[3], Jing Gu[4], Licheng Jiao[5]

[1] Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China
353626620@qq.com

Abstract. Single object tracker based on siamese neural network have become one of the most popular frameworks in this field for its strong discrimination ability and high efficiency. However, when the task switch to multi-object tracking, the development of siamese-network based tracking methods is limited by the huge calculation cost comes from repeatedly feature extract and excessive predefined anchors. To solve these problems, we propose a siamese box adaptive multi-object tracking (SiamBAN-MOT) method with parallel detector and siamese tracker branch, which implementation in an anchor-free manner. Firstly, ResNet-50 is constructed to extract shared features of the current frame. Then, we design a siamese specific feature pyramid networks(S-FPN) to fuse the multi-scale features, which improves detection and tracking performance with the anchor-free architecture. To alleviate the duplicate feature extraction, RoI align is operated to extract all trajectories' template feature and search region feature in a single feature map at once. After that, anchor-free based Siamese tracking network outputs the tracking result of each trajectory according to its template and search region feature. Meanwhile, current frame's detection results are obtained from the detector for the target association. Finally, a simple novel IOU-matching scheme is designed to map the tracking results to the detection results so as to refine the tracking results and suppress the drifts caused by siamese tracking network. Through experimental verification, our method achieves competitive results on MOT17.

Keywords: Multi-object tracking, Siamese Network, Anchor-Free.

1 Introduction

Multi-Object Tracking (MOT) is the problem of detecting object instances and then temporally associating them to form trajectories [1], which is a long-standing problem in the computer vision field. With the development of artificial intelligent society,

This work was supported in part by the National Natural Science Foundation of China under Grant 61871306, Grant 61836009, Grant 62172600, Grant 62077038, by the Innovation Capability Support Program of Shaanxi (Program No. 2021KJXX-08), by the Natural Science Basic Research Program of Shaanxi under Grant No. 2022JC-45 and 2022GY-065, and by the Fundamental Research Funds for the Central Universities under Grant JB211901.

object tracking is widely used in automatic driving, robotics, security monitoring and other fields. Multi-Object Tracking technology has experienced great progress since its emergence, however, this field still faces great challenges due to its stringent performance requirements for precise object detection and fine-grained classification of similar objects.

In earlier work [2,3], MOT has been regarded as a objects-to-trajectories matching problem, most of the earlier works design their networks based on the Tracking-by-Detection (TBD) paradigm, that is, training the detector separately, modeling the motion model to predict each trajectory's position in current frame, and assigning the objects to the trajectories by the Hungarian algorithm or some other data association algorithms. Benefitting from the development of person re-identification, tracking network based on Re-ID has gradually become the mainstream. It is worth mentioning thar Re-ID in object tracking task refers to a column vector of a particular dimension extracted from the object, and the category of the object depends on the data, rather than the characteristics of the person in person re-identification task. Some methods [4,5,6,7,8,9,10] try to solve tracking task by a separate detection model and a separate Re-ID feature extraction network: detection model detects objects firstly and then get the Re-ID features extracted from the detected bounding boxes' corresponding image patch. Finally, links the detection to one of the existing trajectories or creates a new trajectory according to their metrics defined on Re-ID features. In recent years, many methods [11,12,13,14,15] tend to make detection and tracking simultaneously in a common network for its outstanding performance benefited from end-to-end training network, defined as Joint Detection and Tracking (JDT) paradigm. In recent JDT methods, some methods [11,12] design the network structure to extract detection results and Re-ID features at the same time, while others implement tracking [13,14,15] by unique tracking methods.

However, most Re-ID based methods' over-reliance on appearance feature also creates problems hard to solve. One problem is Re-ID feature of similar or occluded object is always indistinguishable at the common natural video scale, such as several people with same clothes or people whose body is blocked by obstacle. Another problem is that continuous information contained in the video is wasted in most Re-ID matching method.

Siamese neural network is widely used in Single-Object-Tracking (SOT) filed for its excellent discrimination of objects and online learning ability. Recently, *Shuai et al* proposed a SiamMOT method [16] for muti-object tracking, which achieved start-of-the-art tracking performance in several public datasets. Nevertheless, due to the anchor-based single object tracking and detection network, during the matching process, each object needs to be re-modeled once, which brings huge time and computing resource cost.

To solve these problems, in this paper, we propose an anchor-free based Siamese network to realize muti-object tracking task, which can achieve both high accuracy and FPS. As show in Fig.1, the proposed network consists of two parallel branch: anchor-free based one-stage detector branch and anchor-free based siamese tracker branch. No predefined anchor is required for the entire network, benefit from the specifically designed S-FPN, the detector directly predict the interest objects' global

heatmap and objects' size information, during the tracking, Siamese-BAM tracker get the template feature of each trajectory and search region feature to regression the tracking result. Finally, after a simple IOU match, tracker matches the objects with the trajectories and generates new trajectories.

The contributions of this work are described as follows:

- We propose an anchor-free based Siamese network for multi-object tracking, which can extract multi-trajectories' template and search region feature at once, enhanced the siamese architecture's inference speed.
- We design a S-FPN module to enhance the tracking and detection performance on anchor-free based network.
- The proposed tracking method can alleviate the object drift problem of siamese tracking network and improve the performance by simple IOU-matching.

2 SiamBAN-MOT

SiamBAN-MOT consists of detection branches and tracking branches. As show in Fig. 1. The detection branch is designed as an anchor-free one-stage detection network, with using ResNet-50[18] as the backbone network to extract the image feature. And we design a S-FPN module to enhance the discrimination of feature. In the tracking branch, a multi-object tracker based on Siamese network is constructed, during the tracking, same as the Siamese RPN++ [19], templates of trajectories are used as convolution kernel to regression tracking results in current frame. Finally, after simple IOU matching, the objects and trajectories are matched. Unmatched objects will be assigned to a new trajectory.

In the next section we will introduce the details of our SiamBAN-MOT, both include the detector (Sec. 2.1) and tracker (Sec. 2.2).

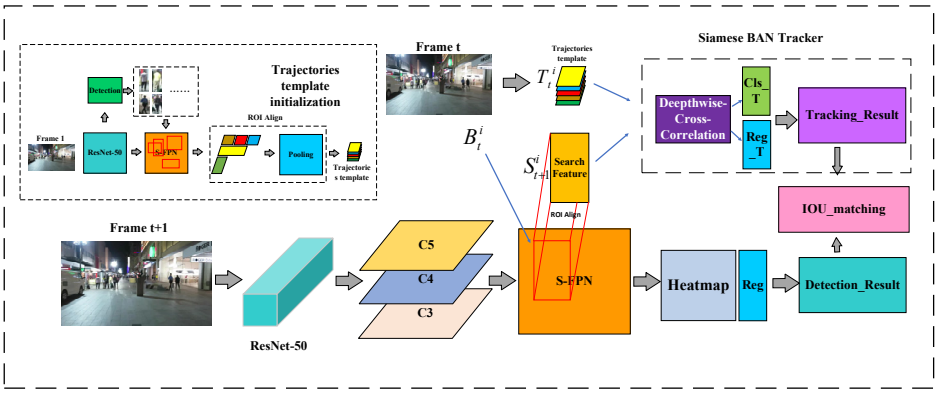

Fig. 1. SiamBAN-MOT network

2.1 Anchor-Free network with Detector and S-FPN

With the excellent feature extraction performance of ResNet-50, we design an anchor-free basic network and used it for detection.

In detail, we extract the features of ResNet-50's convolutional layers 3,4 and 5(denote as C3, C4 and C5), and the stacked features is fed into the S-FPN module for multi-scale fusion (as show in Fig 2). It is important to note that the SFPN is also designed for tracking branch's accurately regression, as template features' requirement of multi-scale enhancement. Then, computing with simple regression head and heatmap head and decoding the heatmap, we can get accurate detection results.

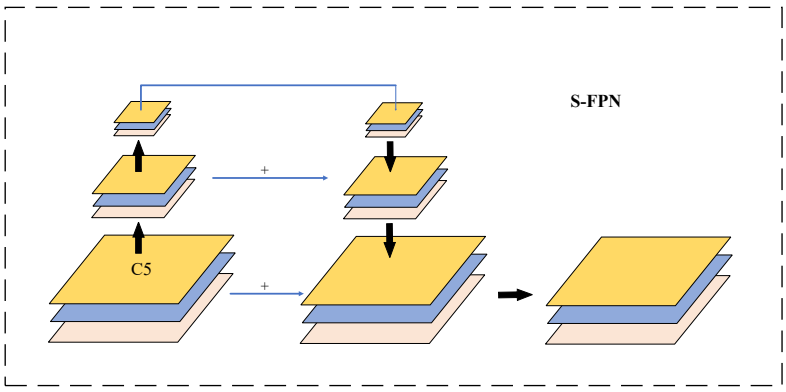

Fig. 2. S-FPN module

In addition, same as Siam-BAN [17], we set the stride to 1 in the conv4 and conv5 blocks in ReNet-50.

2.2 Siamese tracker for multi-object tracking.

The input of tracking branch are template feature (defined as T_t^i , which represents the feature of the i-th trajectory sampled at frame t) and the search region feature (defined as S_{t+1}^i, which represents the feature cropped from a specific region of frame $t+1$ feature , according to the $i\ th$ trajectory's location in frame t). t represents the serial number of the image in the video and i represents the t-th trajectory in trajectories set.

Defining inference result of tracking in frame $t+1$ as $\tilde{B}_{t+1}^i$ and inference result of detection in frame $t+1$ as B_{t+1}^i, the bounding box in ground-truth as $\bar{B}_{t+1}^i$,Formally,

$$\tilde{B}_{t+1}^i = \Gamma(T_t^i, S_{t+1}^i, \theta) \qquad (1)$$

where Γ represent the learnable Siamese tracker with parameters θ.

The strategy for calculating candidate regions of S_{t+1}^i is as follows: set the t-th trajectory's search region in frame $t+1$ according to B_t^i by keeping the same geometric center of B_t^i and expanding region's width and height to α and β times that of B_t^i.

The final accurate tracking result can be obtained by simple IOU matching, which can both realize tracking motion suppression and object matching with existing trajectories.

2.3 Ground-truth and Loss

During the training, we set detector's loss function as:

$$l_{det} = l_{heat} + l_{size} \tag{2}$$

where l_{heat} defined as:

$$l_{heat} = focal\_loss(heatmap, heatmap\_gt) \tag{3}$$

and the l_{size} defined as:

$$l_{size} = l1\_loss(size, size\_gt) \tag{4}$$

the $heatmap\_gt$ is got in the way mentioned in FairMOT [12]. And the $size$ and $size_{gt}$ represent the width and height of B_{t+1}^{i} and $\bar{B}_{t+1}^{i}$.

As for tracker, we use the same strategy in SiamBAN [17] to take positive and negative samples and then we set tracker's loss function as:

$$l_{track} = l_{cls} + l_{reg} \tag{5}$$

Where l_{cls} is the cross-entropy loss and l_{cls} is the IOU loss.

3 Experiment

Our SiamBAN-MOT is implemented in Pytorch on a single RTX 2080Ti 11GB GPU, it achieves the average running speed of 18 FPS on MOT17.

3.1 Datasets and Metrics

MOT17[20] is one of the most popular datasets for multi-object tracking, which consists of both 7 videos in training set and test set, ranging from 7 to 90 seconds long. These videos contain various scene of crowd people indoor or outdoor streets. Following the [16], we only using MOTA (multiple object tracking accuracy), IDF (ID F1 score), and FPS to evaluate the tracker's tracking performance both in accuracy and inference speed.

3.2 Implementation details

Network. We use a fixed ResNet-50 which set the kernel stride to 1 in the conv4 and conv5 blocks in ReNet-50 and keep the feature' size in different blocks by change the padding size. We set (α,β) to (3, 2) to get the search region.

Training. We jointly train the tracker and detection network end to end. SGD with momentum is used as optimizer. Setting learning rate of 0.05 and decrease it by factor 10 after 50% of the iterations, we set the batch size of 16 image pairs and weight decay of 10^{-4}.

Inference. We use a dynamic IOU threshold that is 0.5 while only one object can achieve this threshold, otherwise, match the object with the largest IOU value to the trajectory. We keep a trajectory active 30 frames since it disappears and expand the search region 2 times every 10 frames.

3.3 Experiment results on MOT17

Finally, we compare our SiamBAN-MOT with several excellent models on MOT17 datasets with MOTA, IDF1 and FPS metrics.

Table 1. Results on MOT17 test set with public detection.

Method	MOTA	IDF1	FPS
SST	52.4	49.5	<3.9
CTracker	66.6	57.4	6.8
CenterTrack	67.8	64.7	17.5
SiamMOT	65.9	63.3	16
SiamBAN-MOT	64.8	60.9	18

4 Conclusion

In this report, we present an anchor-free based siamese network for multi-object-tracking. The purpose of this method is to replace the step of data association between frames by template matching, so that the existing object matching methods can be more learnable. However, due to the specific of the siamese-network-based method, it always achieves a low FPS performance in multiple object tracking comparing with Re-id based method. According to the results in Table 1, without using any tricks, our method declined in accuracy, but achieved better FPS performance than other methods. And there is still a great potential for our method because it is not optimized.

References

1. Shuai B, Berneshawi A, Li X, et al. Siammot: Siamese multi-object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12372-12382.
2. Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE international conference on image processing (ICIP). IEEE, 2016: 3464-3468.

3. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.
4. Mahmoudi N, Ahadi SM, Rahmati M (2019) Multi-target tracking using cnn-based features: Cnnmtt. Multimedia Tools and Applications 78(6):7077–7096
5. Zhou Z, Xing J, Zhang M, Hu W (2018) Online multi-target tracking with tensor-based high-order graph matching. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 1809–1814
6. Fang K, Xiang Y, Li X, Savarese S (2018) Recurrent autoregressive networks for online multi-object tracking. In: WACV, IEEE, pp 466–475
7. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B (2016) Simple online and realtime tracking. In: ICIP, IEEE, pp 34643468
8. Wojke N, Bewley A, Paulus D (2017) Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP), IEEE, pp 3645–3649
9. Chen L, Ai H, Zhuang Z, Shang C (2018a) Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6
10. Yu F, Li W, Li Q, Liu Y, Shi X, Yan J (2016) Poi: Multiple object tracking with high performance detection and appearance feature. In: ECCV, Springer, pp 36–42
11. Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]//European Conference on Computer Vision. Springer, Cham, 2020: 107-122.
12. Zhang Y, Wang C, Wang X, et al. Fairmot: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.
13. Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European Conference on Computer Vision. Springer, Cham, 2020: 474-490.
14. Peng J, Wang C, Wan F, et al. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//European conference on computer vision. Springer, Cham, 2020: 145-161.
15. Zheng L, Tang M, Chen Y, et al. Improving multiple object tracking with single object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2453-2462.
16. Shuai B, Berneshawi A, Li X, et al. Siammot: Siamese multi-object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12372-12382.
17. Chen Z, Zhong B, Li G, et al. Siamese box adaptive network for visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6668-6677.
18. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
19. Li B, Wu W, Wang Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4282-4291.
20. Milan A, Leal-Taix´ e L, Reid I, Roth S, Schindler K (2016) Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:160300831