Anchor-Free One-Stage Online Multi-Object Tracking

Zongwei Zhou[1,2], Yangxi Li[3], Jin Gao[1,2],
Junliang Xing[1,2], Liang Li[4], Weiming Hu[5,6]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{zongwei.zhou, jin.gao, jlxing}@nlpr.ia.ac.cn
[3] National Computer network Emergency Response technical Team/Coordination Center of China
[4] The Brain Science Center, Beijing Institute of Basic Medical Sciences
[5] CAS Center for Excellence in Brian Science and Intelligence Technology, National Laboratory of Pattern Recognition, CASIA
[6] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
liyangxi@outlook.com, liang.li.brain@aliyun.com, wmhu@nlpr.ia.ac.cn

Abstract. Current multi-object tracking (MOT) algorithms are dominated by the tracking-by-detection paradigm, which divides MOT into three independent sub-tasks of target detection, appearance embedding, and data association. To improve the efficiency of this tracking paradigm, this paper presents an anchor-free one-stage learning framework to perform target detection and appearance embedding in a unified network, which learns for each point in the feature pyramid of the input image an object detection prediction and a feature representation. Two effective training strategies are proposed to reduce missed detections in dense pedestrian scenes. Moreover, an improved non-maximum suppression procedure is introduced to obtain more accurate box detections and appearance embeddings by taking the box spatial and appearance similarities into account simultaneously. Experiments show that our MOT algorithm achieves real-time tracking speed while obtaining comparable tracking performance to state-of-the-art MOT trackers. Code will be released to facilitate further studies of this problem.

Keywords: Anchor-Free · One-Stage · Multi-Object Tracking

1 Introduction

Multi-Object Tracking (MOT), *a.k.a* Multi-Target Tracking (MTT), is critical in video analysis systems ranging from video surveillance to autonomous driving. The objective of MOT is to determine the trajectories of multiple objects simultaneously by localizing and associating targets with the same identity across multiple frames. It is a very difficult task due to challenging factors like large variations in intra-target appearance and frequent inter-target interactions [13].

[1] student paper

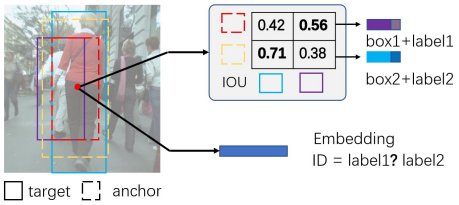

Fig. 1: The label ambiguity of features in an anchor-based MOT tracker.

Tracking-by-detection is the main paradigm for the current multi-object tracking algorithms. It usually includes three steps: object detection in each frame, appearance embedding of each object, and data association across frames. Integrating these steps in one algorithm is usually difficult, especially if real time performance is required. For a MOT framework using a common simple association strategy (*e.g.* Hungarian algorithm), its computing resources are mainly consumed in separated object detection and appearance embedding steps. These two steps can share low-level features to improve the tracking speed. This suggests unifying object detection and appearance embedding in one step.

At present, there are two main schemes for joint detection and embedding learning. One is a two-stage framework similar to Faster-RCNN [17], and the other is a one-stage framework similar to SSD [11]. In the two-stage framework [25], the first stage uses a Region Proposal Network [17] to detect targets, and the second stage uses metric learning supervision to replace classification supervision in Faster-RCNN to learn target embedding. Although it saves some computation by sharing the low-level features, the two-stage design still limits its tracking speed. Moreover, generating a large number of region proposals improves accuracy but reduces efficiency. The solutions in one-stage framework are not well studied yet. The existing methods, such as AJDE [23], learn a joint detector and embedding model based on an anchor-based network, which relies on some predefined proposals named anchor boxes. The framework achieves near real-time tracking speed, but still has two disadvantages. As shown in Fig. 1, according to the Intersection-Over-Union (IOU) values, different anchor boxes (dotted boxes) at the same location are responsible for different targets (solid boxes), but only one feature vector is obtained, making the labels of features ambiguous. The other disadvantage stems from anchor-based structures, such as the manual configuration of hyper-parameters to define anchors and the complex architecture of detection subnets based on the predefined anchors.

To improve the tracking speed and avoid the disadvantages of the anchor-based structure, an anchor-free one-stage network is proposed in this work, where the bounding boxes and their corresponding appearance features are simultaneously extracted from the locations on feature maps directly, rather than predefined anchor boxes. We notice that the idea in [26] is similar to ours, but the method in [26] is more focused on the design of backbone, while our method focuses on the processing of joint object detection and embedding. We name the locations on feature maps as samples in the following. Unlike in general object detection tasks, the targets in multi-target tracking, especially multi-pedestrian

tracking, tend to have similar scales and large occlusions. Thus, general anchor-free detectors (such as FCOS [21]) have a large number of missed detections in MOT due to attention bias and feature selection. Attention bias means that objects with good views tend to draw more attention from the detector making the partially occluded objects being easily missed. The feature selection issue arises because each target is scaled to a single pyramid level. This causes that multiple targets with similar scales may be assigned to same locations, especially if one target occludes another. The embedded features of the targets sharing the same location are ambiguous in that case. Therefore, the proposed model includes two strategies to reduce missed detection while incorporating embedding into the detector. First, the samples used for detection and embedding are re-weighted in the contribution to the network loss based on their distance to the object center. Second, the box regression ranges overlap in adjacent pyramid levels. A multi-task loss is introduced to train the model end-to-end.

Our precise embedding facilitates an improved Non-Maximum Suppression (NMS). The traditional NMS operator only considers Intersection-Over-Union (IOU) values between detections. The appearance information is ignored. As a result, many true targets are suppressed in crowded scenes. The improved NMS suppresses proposals, using both overlaps between the detections and the similarity of the appearances within the detected boxes for reducing over-suppression. The main contributions of this work are in three-fold:

- An anchor-free one-stage joint detection and embedding learning network is presented for online multi-target tracking. The model achieves real-time tracking speed while obtaining state-of-the-art tracking performance.
- Two effective training strategies are proposed to detect targets with similar scales in crowded scenes. The strategies are regression range overlapping and samples re-weighting.
- An improved NMS operator is designed to incorporate both the box spatial and appearance similarity to reduce false negatives in crowded scenes.

We develop a high performance online multi-object tracking system by incorporating the proposed network into a hierarchical data association pipeline. Extensive experimental analyses and evaluations on the MOT benchmark demonstrate the effectiveness and the efficiency of the proposed approach.

2 Related Work

Separate Detection and Embedding for MOT. These methods are dominant in the tracking-by-detection paradigm. Some of these methods build embedding networks upon the detections provided by the MOT benchmark to associate detections across frames, such as DeepSort [24], MOTDT [12], and DAN [19]. Other methods design both detectors and feature extractors to track targets. For example, POI [27] proposes a pedestrian detector based on Faster R-CNN, and Tracktor [1] uses the previous tracking results as proposals to detect the new bounding boxes of the targets for tracking. The single object tracking-based trackers [32] can also be regarded as detectors based on template matching. All these methods need an additional extractor after the detector to handle long-time occlusions. The overall inference time for these methods is approximately

equal to the sum of the times for detection and extraction. This makes real-time operation difficult to achieve.

Joint Detection and Embedding for MOT. These methods reduce the tracking time calculations by combining the detection and the embedding into one step. MOTS [22], STAM [3] and D&T [6] integrate the embedding into a detector in a two-stage network, while AJDE [23] is a one-stage model. In a two-stage model, the detection and the embedding share the low-level features. The embedding is then extracted from the Region-of-Interest (ROI) after the detection. Due to the sequential nature of detection and embedding, the two-stage structure still has a limited tracking speed. Besides, since each target is processed separately in the second stage, the runtime of embedding is proportional to the number of targets. The one-stage model, AJDE, adds an embedding branch to the detection header of the SSD framework to carry out detection and embedding in parallel. This speeds up the tracking while maintaining tracking performance. But it suffers from the anchor-based structures, such as the manual configuration of anchor hyper-parameters and the complex architecture of detection header. Besides, the corresponding relationship between embedding and anchor boxes at the same location is not always one-to-one correspondent (Fig.1). The proposed model is an anchor-free one-stage network, which overcomes the disadvantages of anchor-based structure and further improves the tracking speed.

3 Our Approach

3.1 Anchor-free Joint Detection and Embedding

Network architecture As shown in Figure 2, the network consists of a backbone, a feature pyramid and one prediction header per pyramid level, in a fully convolutional style. The backbone can include commonly used convolutional networks, such as ResNet50 [7]. The feature pyramid is adopted to deal effectively with large scale variations between targets. A pyramid level is represented as P_m where m denotes the level number. The level has $1/s_m$ resolution of the input frame size, where s_m is the stride of down-sampling. A prediction header contains two task-specific subnets, *i.e.* detection and embedding. The embedding subnet has three 3×3 convolutional layers and the output layer extracts a 512-dimensional discriminative feature from each location on the feature map. The detection subnet contains two 3×3 convolutional layers followed by two branches for classification and bounding box regression. The classification branch outputs the probability that each location is a positive sample. The regression branch predicts the distances from each sample to the boundaries of a corresponding target if the sample is positive.

Supervision targets A target in a frame $I \in R^{3 \times W \times H}$ is denoted as $B = (x, y, w, h, c)$ where (x, y) is the center position, w, h are the box width and height respectively. The$c \in Z^k$ is the partially annotated identity label, where -1 indicates a target without an identity label. Given a target, we first assign it to one pyramid level according to its scale. Specifically, the target is assigned to the mth pyramid level P_m if $\max(w, h) \in [a_m, b_m]$, where $[a_m, b_m]$ is the

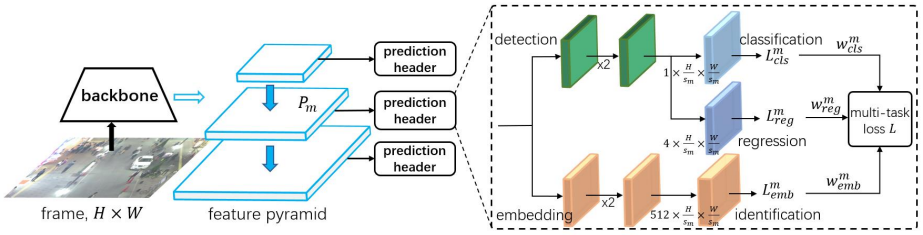

Fig. 2: Architecture of the anchor-free joint detection and embedding model.

predefined regression range of bounding box in P_m. We overlap the predefined regression ranges in adjacent pyramid levels to improve the recall by providing more proposals from different granularity, especially for close and similar-scaled targets. Next we define the positive samples in the mth pyramid level. Each sample p_{mij} with $i = 1, 2, \ldots, W/s_m$ and $j = 1, 2, \ldots, H/s_m$ on P_m has a corresponding image spatial location (X_{mij}, Y_{mij}) where $X_{mij} = s_m(i - 0.5)$ and $Y_{mij} = s_m(j - 0.5)$. The sample is set as positive if its centerness to any one target B_k assigned to P_m is larger than a threshold τ_c. The centerness is the same as that defined in FCOS [21], *i.e.*,

$$\mathrm{CT}(p_{mij}, B_k) = \sqrt{\frac{\min(l^k_{mij}, r^k_{mij})}{\max(l^k_{mij}, r^k_{mij})} \frac{\min(t^k_{mij}, b^k_{mij})}{\max(t^k_{mij}, b^k_{mij})}}, \tag{1}$$

where $(l^k_{mij}, r^k_{mij}, t^k_{mij}, b^k_{mij})$ denotes the distances between (X_{mij}, Y_{mij}) and the left, right, top and bottom boundaries of target B_k.

If the centernesses of a positive sample p_{mij} to multiple targets are all larger than the threshold τ_c, the sample is regarded as ambiguous. The target B_{k^*} with maximal centerness is chosen as the responsible object of the ambiguous sample, where $k^* = \mathrm{argmax}_k\{\mathrm{CT}(p_{mij}, B_k)|k = 1, 2, \ldots, K\}$, and K is the number of targets assigned in mth level. The centerness map on P_m is defined as:

$$M_{mij} = \max_{k=1,2,\ldots,K} \mathrm{CT}(p_{mij}, B_k). \tag{2}$$

Multi-task loss function The loss function of the proposed anchor-free joint detection and embedding model consists of three components for different tasks, the point classification, the box regression, and discriminative feature extraction.

In the detection subnet, we use the IOU loss L_{reg} as in FCOS to regress bounding boxes B_{k^*} from positive samples. For the point classification, the hard-designation of positives and negatives brings more difficulties for training. To reduce the ambiguity of the samples between hard positives and negatives, we apply the centerness map M to re-weight the contributions of ambiguous samples. The focal weight [10] on hard examples are also adopted to combat the extreme class imbalance between positive and negative samples. Let ρ_{mij} be the network's estimated probability indicating whether the sample p_{mij} is positive, and γ be the focusing hyper-parameter. Then, the classification loss in mth pyramid level can be formulated as:

$$L^m_{cls} = -\frac{1}{K} \sum_{i=1}^{W/s_m} \sum_{j=1}^{H/s_m} \alpha_{mij}(1 - \hat{\rho}_{mij})^\gamma \log(\hat{\rho}_{mij}), \tag{3}$$

where

$$\hat{\rho}_{mij} = \begin{cases} \rho_{mij}, & \text{if } M_{mij} > \tau_c \\ 1 - \rho_{mij}, & \text{otherwise} \end{cases}, \quad \alpha_{mij} = \begin{cases} 1, & \text{if } M_{mij} > \tau_c \\ (1 - M_{mij})^{\beta}, & \text{otherwise} \end{cases}. \quad (4)$$

The focusing hyper-parameter γ is experimentally set to 2 as suggested in the Focal Loss [10], and the hyper-parameter β controls the penalty on the ambiguous samples to reduce their contributions to the total loss.

The objective of the embedding subnet is to learn an embedding space where observations of the same target are close to each other, while observations of different targets are far apart. We transform the metric learning problem into the classification problem like many re-identification (ReID) models [30, 20]. Then the cross-entropy loss is used to extract discriminative features. Let $\mathbf{f}_{mij} \in R^{512}$ be the output feature in p_{mij} and c_k be the class label of B_k regressed in p_{mij}. Let $W \in R^{512 \times N}$ be the learnable parameters of the last classifier layer, where N is the number of targets. Then, the embedding loss is defined as follows,

$$L^m_{emb} = - \sum_{ij:M_{mij}>\tau_c} \log \frac{e^{(\mathbf{W}^T \mathbf{f}_{mij})_{c_k}}}{\sum_q e^{(\mathbf{W}^T \mathbf{f}_{mij})_q}}. \quad (5)$$

The automatic learning scheme for loss weights proposed in [8] is adopted to combine these three losses. The total multi-task loss with automatic loss balancing is formulated as,

$$L = \sum_{m,T \in \{cls,reg,emb\}} \frac{1}{e^{w^m_T}} L^m_T + w^m_T, \quad (6)$$

where $w^m_T, T \in \{cls, reg, emb\}$ is the learnable weight parameters.

3.2 Appearance Enhanced NMS (ENMS)

NMS is an integral part of the object detection pipeline. The detected boxes are first sorted according to scores. The box with the highest score is then selected. All the other boxes that have a significant overlap with it are suppressed. This process is applied recursively to the remaining boxes until the final detection result is obtained. Though NMS is efficient in suppressing false positives, it also over-suppresses in dense scenes as it does not take any appearances into account. As shown in Fig. 3, the raw proposals are given in Fig. 3(a) and the detections processed by NMS are given in Fig. 3(b).The arrows in Fig. 3(b) point to targets that are wrongly suppressed by NMS.

Benefiting from the joint model introduced in the last subsection (3.1), which provides detection and embedding simultaneously, we can use the discriminative feature to enhance the NMS operator. Formally, given the raw proposals $\mathcal{B} = \{(B_k, \rho_k, \mathbf{f}_k)k = 1, 2, \ldots, N\}$ and an empty set $\mathcal{B}_f = \emptyset$, where $B_k, \rho_k, \mathbf{f}_k$ denotes the regressed boxes, predicted scores and features respectively, the most reliable proposals $(B_{k^*}, \rho_{k^*}, \mathbf{f}_{k^*}), k^* = \text{argmax}_k \, \rho_k$ are selected firstly. Then get the false proposals of B_{K^*} based on the box overlap and the appearance similarity, *i.e.*,

$$\mathcal{B}_s = \{(B_k, \rho_k, \mathbf{f}_k) | \text{IOU}(B_k, B_{k^*}) > \tau_i \cap \mathbf{f}^T_{k^*} \mathbf{f}_k > \tau_e\}, \quad (7)$$

where τ_i, τ_e are predefined thresholds for IOU and appearance similarity respectively. Update the set $\mathcal{B}_f = \mathcal{B}_f \cup \{B_{k^*}\}$ and apply the above process recursively in $\mathcal{B} = \mathcal{B} \setminus (\mathcal{B}_s \cup \{B_{k^*}\})$ until $\mathcal{B} = \emptyset$. The set $\mathcal{B}_f$ contains the

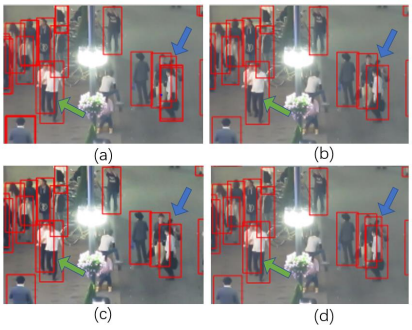

Fig. 3: An exemplar of ENMS. (a),(b),(c),(d) show the detections without NMS, with NMS, with NMS using appearance similarity and with ENMS respectively.

final detections. The detections obtained after suppressing false positives using only appearance similarity are shown in Fig. 3(c), while that obtained using the proposed ENMS are shown in Fig. 3(d). It shows that the ENMS reduces false suppressions in dense scenes.

3.3 Tracking Pipeline

The proposed anchor-free joint detection and embedding model and the appearance enhanced NMS operator are combined with the hierarchical association strategy in MOTDT [12] to form the tracking pipeline in our tracking algorithm.

- Step 1. Given a new frame, obtain the proposals and corresponding features using the proposed anchor-free joint detection and embedding model.
- Step 2. Filter the proposals using the enhanced NMS.
- Step 3. Assign the filtered detections to existing tracklets using feature similarities with a threshold ε_d for the minimum similarity. The similarity is also limited by the distance between the detection and prediction of the tracklet in order to meet the constraint of spatial continuity. That is, the target motion offset in consecutive frames is small. The tracklet feature is online updated as,

$$\mathbf{f}_t = \eta \mathbf{f}_{t-1} + (1-\eta)\mathbf{f}_k, \tag{8}$$

  where η is the momentum cofficient and set as 0.9 as in AJDE [23], $\mathbf{f}_k$ is the feature of associated detection and $\mathbf{f}_t$ denotes the track feature at time t.
- Step 4. Associate the remaining candidates with unassociated tracklets based on the IOU values between candidates and predictions with a threshold ε_{iou}.
- Step 5. Mark any untracked track as lost. Initialize a new trajectory with any unmatched detection with a confidence higher than ε_p. Terminate any trajectory that remains lost for over ε_n successive frames or exits the field of view. Additionally, any new tracks will be deleted if they are lost within the first two frames. This is to suppress false trajectories.
- Step 6. Repeat above steps for the next frame until no more frames arrive.

Table 1: Statistics of the training set.

Dataset	ETH	CP	CT	M16	CS	PRW	Total
#img	2K	3K	27K	5.3K	11K	6K	54.3K
#box	17K	21K	46K	112K	55K	18K	270K
#ID	-	-	0.6K	0.5K	7K	0.5K	8.7K

Table 2: Quantitative analysis of two training strategies.

OR	RW	MOTA	Pre	Rec	IDS	IDF1	mAP	$TFR_{0.1}$
×	×	63.4	80.3	75.7	366	66.3	81.2	88.3
×	√	66.7	85.7	81.1	98	67.1	82.7	89.0
√	×	70.2	**88.9**	81.9	103	66.1	82.2	90.8
√	√	**71.9**	88.4	**83.4**	**78**	**69.8**	**82.8**	**91.7**

4 Experiments

4.1 Experimental Settings

Datasets. Since we transform metric learning into a classification problem, datasets for pedestrian detection, pedestrian ReID and multi-pedestrian tracking are all used to train the anchor-free joint detection and embedding model. The statistics of the training sets are shown in Table 1. ETH dataset [4] and CityPersons (CP) dataset [28] are used for person detection. We mark their targets ID as -1 in training as they have no identity annotations. PRW dataset [29] and CUHK-SYSU (CS) dataset are derived from the ReID task. CalTech (CT) dataset [25] and MOT16 (MT) dataset [15] are collected from the MOT task. The sequences in the ETH dataset that overlap with the MOT16 test set are excluded for fair evaluation. The model is first analyzed on the MOT15 dataset [9] after excluding the sequences appeared in the training, then its performance is compared with the SOTA methods on the MOT16 test set.

Evaluation Metrics. The CLEAR MOT metrics [2] are used to analyze the tracking performance. They include multiple object tracking accuracy (MOTA, ↑), the number of mostly tracked targets (MT, $> 80\%$ covered, ↑), the number of mostly lost targets (ML, $< 20\%$ covered, ↓), false positive (FP, ↓), false negative (FN, ↓) precision (Pre, ↑), recall (Rec, ↑), and identity switches(IDs, ↓). Additionally, ID F1 score (IDF1, ↑) [18] is also employed to measure the identity-preserving ability of trackers. IDF1 denotes the ratio of correctly identified detections over the average number of ground-truth and computed detections. To evaluate the detection accuracy and the appearance embedding, we also use the metrics defined in [23], *i.e.*, the average precision (AP, ↑) at IOU threshold of 0.5 over the Caltech validation set and the true positive rate at false accept rate 0.1 ($TFR_{0.1}$, ↑) over the CUHK-SYSU and PRW validation sets. Here ↑ means higher is better, and ↓ means lower is better.

Implementation Details. We employ DarkNet-53 [16] as the backbone network. The network is trained for 60 epochs with Stochastic Gradient Descent (SGD) optimizer and the batchsize is set as 16. The learning rate is initialized as 10^{-2} and is decreased by 0.1 at the 30th and 50th epoch. The input resolution is 1088×608 if not specified and the data augmentation techniques, such as random rotation, random scale and color jittering, are applied to reduce over-fitting. The predefined regression ranges $[a_m, b_m], m = 1, 2, 3$ are set as $[0, 160]$, $[64, 320]$ and $[256, 608]$ respectively. The parameters τ_i and τ_e used in the improved NMS are set to 0.5 and 0.2 respectively according to the experiment analysis shown in Fig.4. The parameter τ_c used for selecting positive anchor points and the parameter β used in Eq. (5) are analyzed in next subsection. For data association, we set $\varepsilon_d = 0.5, \varepsilon_{iou} = 0.5, \varepsilon_p = 0.6$ and ε_n as the frame rate of the sequence.

4.2 Ablation Study

Analysis of the hyper-parameters. Table 4 analyzes the effects of the hyper-parameters τ_c and β, where τ_c is the centerness threshold used to select positive samples and β is the exponent in Eq.(4) used to penalize ambiguous samples. When $\beta = 1.0$, the higher τ_c obtains a better Pre as the positives are more concentrated, but concentrated positives mean more ambiguity in negatives which decrease the Rec. When $\tau_c = 0.8$, the performances of $\beta > 0$ are all, except Pre, better than the performances with $\beta = 0$. This means that the weighted focal loss is more effective than the original focal loss in MOT. The higher Pre is obtained at $\beta = 0$ because samples around positives are marked as hard negative samples. This enhances the certainty of positives, but also introduces the ambiguity of negatives leading to a lower Rec. The proposed model has the best performance at $\tau_c = 0.8$ and $\beta = 1.0$, so we use these settings when we evaluate the proposed method on MOT benchmark.

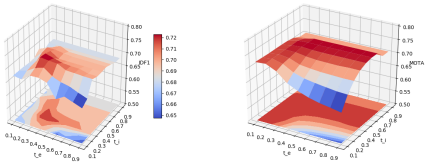

Fig. 4: The analysis of IOU threshold t_i and appearance similarity threshold t_s on the performance of the appearance enhanced NMS on MOT15 train dataset. (a) and (b) are IDF1 and MOTA respectively.

Table 3: Analysis of the appearance enhanced NMS (ENMS). × and √ indicate whether the ENMS module is used.

Method	ENMS	MOTA	Pre	Rec	IDs	IDF1	FPS
AJDE	×	67.3	84.0	85.1	203	66.4	29.4
	√	62.0	78.3	89.6	366	67.4	28.7
AFOS†	×	67.7	87.0	80.4	99	66.2	31.3
	√	68.9	86.8	82.1	106	66.6	30.8
AFOS	×	71.9	88.4	83.4	78	69.8	31.6
	√	72.8	86.3	87.2	80	70.5	30.8

Table 4: Analysis of the hyper parameters on MOT15-train.

τ_c	β	AP	$TFR_{0.1}$	MOTA	Pre	Rec	IDs	IDF1
0.70	1.0	83.3	90.9	70.7	87.2	83.8	96	69.0
0.75	1.0	83.3	89.2	71.2	87.4	**84.2**	109	67.5
0.80	1.0	82.8	**91.7**	**71.9**	88.4	83.4	**78**	**69.8**
0.85	1.0	83.7	90.4	71.7	88.4	83.4	106	68.7
0.80	0.0	82.2	90.9	70.2	**88.9**	81.9	103	66.1
0.80	0.05	**83.9**	91.7	71.8	88.2	82.9	97	66.8
0.80	2.0	82.6	91.5	71.2	87.5	84.0	93	69.8

Table 5: Failure analysis of each subset of MOT16-test.

Sets	Density	FP	FN
MOT16-01	14.2	284 (2.0%)	2675 (5.5%)
MOT16-03	**69.7**	**10928 (75.3%)**	**15068 (31.2%)**
MOT16-06	9.7	651 (4.5%)	4013 (8.3%)
MOT16-07	32.6	967 (6.7%)	5321 (11.0%)
MOT16-08	26.8	667 (4.6%)	8702 (18.0%)
MOT16-12	9.2	420 (2.9%)	2874 (6.0%)
MOT16-14	24.6	594 (4.1%)	9635 (20.0%)
Total	30.8	14511	48288

Table 6: Comparison with the state-of-the-art online MOT trackers under the private detectors on the MOT16 benchmark. In each column of the one-stage and two-stage methods, the best result is in **bold**.

#stage	Tracker	Det	Emb	#box	#id	MOTA	IDF1	MT	ML	FP	FN	IDs	FPS
Two-stage	DeepSORT_2	FRCNN	WRN	429K	1.2K	61.4	62.2	32.8	**18.2**	12852	56668	781	<8.1
	RAR16wVGG	FRCNN	Inception	429K	-	63.0	63.8	**39.9**	22.1	13663	53248	**482**	<1.5
	TAP	FRCNN	MRCNN	429K	-	64.8	**73.5**	38.5	21.6	12980	**50635**	571	<8.2
	CNNMTT	FRCNN	5-layer	429K	0.2K	65.2	62.2	32.4	21.3	6578	55896	946	<6.4
	POI	FRCNN	QAN	429K	16K	**66.1**	65.1	34.0	20.8	**5061**	55194	805	<6.0
One-stage	AJDE_864	Anchor-box	JDE	270K	8.7K	62.1	56.9	34.4	**16.7**	-	-	1608	32.1
	AJDE_1088	Anchor-box	JDE	270K	8.7K	64.4	55.8	**35.4**	20.0	**9172**	54160	1544	25.4
	AFOS_864(ours)	Anchor-free	JDE	270K	8.7K	63.2	59.0	33.6	22.9	13268	52277	1485	**34.3**
	AFOS_1088(ours)	Anchor-free	JDE	270K	8.7K	**64.8**	**63.1**	35.0	22.9	14511	**48288**	**1300**	26.5

Analysis of two training strategies. Two training strategies, *i.e.* overlapping regression ranges (OR) and samples re-weighting (RW), are designed to deal with crowded scenes. The quantitative analyses of the strategies with the enhanced NMS are shown in Table 2. Both Pre and Rec are improved by overlapping regression ranges because targets with similar scales in crowded scenes are assigned to different pyramid levels to reduce their interaction. This also makes it possible to extract more discriminative embedding to reduce the ID switching in MOT. Samples re-weighting further enhances tracking performance by improving Rec, as the contribution of ambiguous samples among hard positives and negatives is reduced. Both strategies improve detection and tracking performance by enhancing the discrimination of targets in crowded scenes.

Analysis of the appearance enhanced NMS. Enhanced NMS (ENMS) introduces feature similarity to conventional NMS to reduce over-suppression. As can be seen from Table 3, AFOS and AFOS†, which represent results with and without overlapping regression ranges respectively, all benefit from ENMS. Although ENMS slightly decreases the Pre values, it improves the recall rates (Rec) by reducing false suppressions. This improves the MOTA. By reducing false suppression, the models also achieve the higher IDF1, which measures the continuity of the trajectory. In addition, the slightly slower speed of the model using ENMS than that of the model using conventional NMS is because ENMS calculates the appearance similarity. For AJDE, we find the performance with ENMS is worse than that with conventional NMS. The reason is that the label ambiguity of embeddings in the training process of AJDE leads to a confusing of the targets in the crowded scenes. This reduces the performance of the ENMS. On the contrary, our model overcomes the label ambiguity, which facilitates ENMS to further improve tracking performance.

4.3 Evaluation on MOT Benchmark

The proposed method is compared with several state-of-the-art trackers under private detectors, such as DeepSORT_2 [24], RAR16wVGG [5], TAP [31], CNNMTT[14], POI [27] and AJDE [23], on the test sets of MOT16. Their configurations and performances are summarized in Table 6. It can be seen from Table 6 that the joint models (AJDE and the proposed AFOS) run at least 3× faster than existing methods while achieving comparable overall tracking accuracy, *e.g.*, as measuted by the MOTA metric.

Compared with AJDE, the proposed method AFOS obtains better IDs and IDF1 as it extracts more discriminative features and avoids the label ambiguity. With the enhanced NMS, AFOS also suppress more false negatives. Note we didn't compare APOS with the performance of AJDE with ENMS because the feature ambiguous in AJDE reduces the performance of ENMS which analyzed in sec. 4.2. As AFOS is an anchor-free model while AJDE is an anchor-based model, AFOS is faster than AJDE. AFOS reaches a real-time speed, *i.e.*, 26.5 FPS for images of size 1088×608. When the image resolution is down-sampled to 864×480, the speed of AFOS can be further increased to 34.3 FPS with only a minor performance drop ($\Delta = -1.6\%$ MOTA). All the experiments are performed on an NVIDIA Tesla V100 GPU.

Analysis of tracking failures. One may notice that AFOS has a much better FN but a worse FP compared to other methods. We analyze the performance of each subset in Table 5 and find that the FP and FN mainly come from MOT16-03 (75.3% and 31.2% respectively). This is because the targets in MOT16-03 are densely distributed with severe occlusions. Many targets are assigned to the same pyramid level, making them difficult to distinguish.

5 Conclusion and Future Work

In this paper, we have proposed a new MOT tracker named AFOS, which allows target detection and appearance embedding to be learned in an anchor-free joint model. AFOS achieves real-time tracking speed with a tracking performance comparable to that of state-of-the-art MOT trackers. Moreover, in order to benefit from the anchor-free joint detection and embedding model, we introduce an appearance enhanced NMS, which combines the appearance similarity with the conventional NMS to prevent over-suppression. We analyze the tracking failures in the proposed model, and plan to perform occlusion model in AFOS to further improve its performance in densely crowded scenes in future work.

Acknowledgements

This work is supported by the national key R&D program of China (No.2018AAA0102802, No.2018AAA0102803, No.2018AAA0102800), the NSFC-general technology collaborative Fund for basic research (Grant No. U1636218), the NSFC (Grant No. 61751212, 61721004), Beijing Natural Science Foundation (Grant No. L172051), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the NNSF of Guangdong (No. 2018B030311046).

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. arXiv preprint arXiv:1903.05625 (2019)
2. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the CLEAR MOT metrics. JIVP **1**, 1–10 (2008)
3. Chu, Q., Ouyang, W., Li, H., Wang, X., Liu, B., Yu, N.: Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In: ICCV. pp. 4836–4845 (2017)
4. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR. pp. 1–8 (2008)
5. Fang, K., Xiang, Y., Li, X., Savarese, S.: Recurrent autoregressive networks for online multi-object tracking. In: WACV. pp. 466–475 (2018)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: ICCV. pp. 3038–3046 (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR. pp. 7482–7491 (2018)
9. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)

10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: ECCV. pp. 21–37 (2016)
12. Long, C., Haizhou, A., Zijie, Z., Chong, S.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: ICME. pp. 1–6 (2018)
13. Luo, W., Xing, J., Zhang, X., Zhao, X., Kim, T.K.: Multiple object tracking: A literature review. arXiv preprint arXiv:1409.7618 (2014)
14. Mahmoudi, N., Ahadi, S.M., Rahmati, M.: Multi-target tracking using cnn-based features: CNNMTT. Multimed. Tools. Appl. **78**(6), 7077–7096 (2019)
15. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. pp. 779–788 (2016)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
18. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCV. pp. 17–35 (2016)
19. Sun, S., Akhtar, N., Song, H., Mian, A.S., Shah, M.: Deep affinity network for multiple object tracking. arXiv preprint arXiv:1810.11780 (2019)
20. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. pp. 480–496 (2018)
21. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355 (2019)
22. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: Multi-object tracking and segmentation. In: CVPR. pp. 7942–7951 (2019)
23. Wang, Z., Zheng, L., Liu, Y., Wang, S.: Towards real-time multi-object tracking. arXiv preprint arXiv:1909.12605 (2019)
24. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP. pp. 3645–3649 (2017)
25. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. pp. 3415–3424 (2017)
26. Yifu, Z., Chunyu, W., Xinggang, W., Wenjun, Z., Liu, W.: A simple baseline for multi-object tracking. arXiv preprint arXiv:2004.01888v2 (2020)
27. Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: POI: Multiple object tracking with high performance detection and appearance feature. In: ECCV. pp. 36–42 (2016)
28. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR. pp. 3213–3221 (2017)
29. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR. pp. 1367–1376 (2017)
30. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM TOMM **14**(1), 1–20 (2017)
31. Zhou, Z., Xing, J., Zhang, M., Hu, W.: Online multi-target tracking with tensor-based high-order graph matching. In: ICPR. pp. 1809–1814 (2018)
32. Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., Yang, M.H.: Online multi-object tracking with dual matching attention networks. In: ECCV. pp. 366–382 (2018)