Pattern Recognition 107 (2020) 107512

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Distractor-aware discrimination learning for online multiple object tracking

Zongwei Zhou[a,b], Wenhan Luo[c], Qiang Wang[a,b], Junliang Xing[a,b,*], Weiming Hu[d,e]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[c] Tencent AI Lab, China
[d] CAS Center for Excellence in Brian Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[e] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190

ARTICLE INFO

Article history:
Received 12 August 2019
Revised 14 April 2020
Accepted 22 June 2020
Available online 24 June 2020

Keywords:
Multi-object tracking
Distractor-aware discrimination learning
Relational attention learning

ABSTRACT

Online multi-object tracking needs to overcome the intrinsic detector deficiencies, e.g., missing detections, false alarms, and inaccurate detection responses, to grow multiple object trajectories without using future information. Various distractions exist during this growing process like background clutters, similar targets, and occlusions, which present a great challenge. We in this work propose a method for learning a distractor-aware discriminative model that can handle continuous missed and inaccurate detection problems due to the occlusion or the motion blur. To deal with target appearance variations, a relational attention learning mechanism is proposed to capture the distinctive target appearances by selectively aggregating features from history states with weights extracted from their appearance topological relationship. Based on the discrimination model, a multi-stage tracking pipeline is designed for automatic trajectory initialization,propagation, and termination. Extensive experimental analyses and comparisons demonstrate its state-of-the-art performance on widely used challenging MOT16 and MOT17 benchmarks. The source code of this work is released to facilitate further studies on the multi-object tracking problem.[1]

1. Introduction

Multi-Object Tracking (MOT), *a.k.a* Multi-Target Tracking (MTT), is an important problem in computer vision with many practical applications such as video surveillance, autonomous driving and human-computer interaction [1]. The goal of multi-object tracking is to determine the trajectories of multiple objects simultaneously by localizing and associating targets with the same identity across multiple frames. It remains a very challenging problem due to factors like target appearance variations, irregular object motions, partial and full object occlusions [2].

A MOT algorithm often relies heavily on object detector to automatically initialize, propagate and terminate object trajectories. The dominant tracking-by-detection strategy [3,4] applies an object detector at each frame first and then associates detection responses across frames to generate the object trajectories. Benefited from the recent advances in deep detection models [5], the object detection performance has been significantly improved. However, the detection results of existing models are far from perfection. As shown in Fig. 1[2], missing detection, false alarm, and inaccurate detection response still occur frequently even with the state-of-the-art detection models. A MOT algorithm thus needs to overcome these intrinsic detector deficiencies to track targets under challenging situations like large pose variations, severe object occlusions, and complex target interactions.

To handle these issues, global association based methods [6–8] generate trajectories in a batch mode by solving a global optimization problem. Those methods utilize the information from both the past and future simultaneously to suppress detection noises occurred in the current frame and to smooth object trajectories across multiple frames. Though tracking in a batch mode typically achieves better performance, it is non-causal and not applicable in online scenarios where a target identity must be determined at the current time step. Without future information available, it

* corresponding author at: University of Chinese Academy of Sciences, Beijing, China.
E-mail address: jlxing@nlpr.ia.ac.cn (J. Xing).

[1] Implementation code link: https://github.com/ZongweiZhou1/DDLTracker

[2] Data from https://motchallenge.net/results/MOT17Det/

https://doi.org/10.1016/j.patcog.2020.107512
0031-3203/

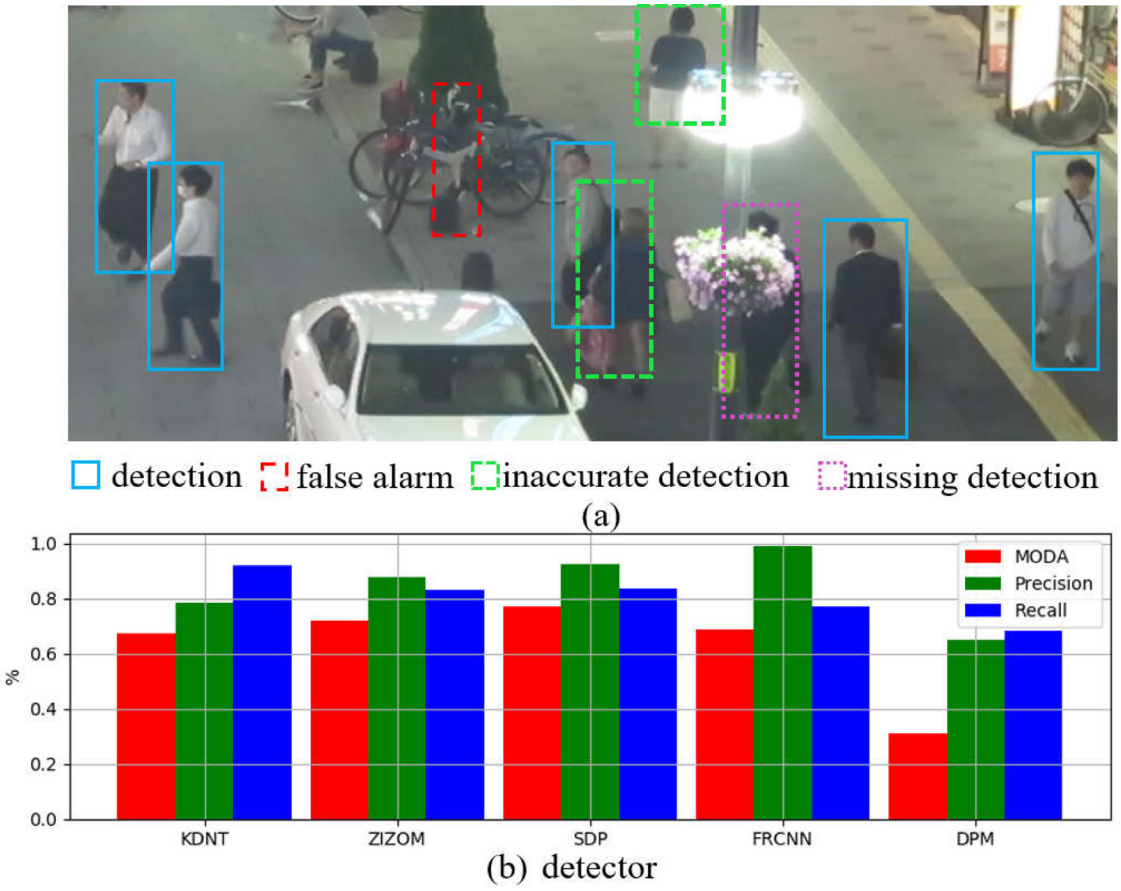

Fig. 1. Exemplary detection failures in MOT challenge. In (a), the cases of false alarm, inaccurate detection and missing detection are demonstrated respectively. In (b), these three kinds of detection failures of different state-of-the-art detectors[2] are evaluated using MODA (Multiple Object Detection Accuracy), Precision and Recall. Best viewed in color.

is more challenging for online multi-object tracking algorithms to grow target trajectories when it is continuously miss-detected or occluded after several frames.

Facing these challenges, this work proposes an online MOT algorithm which learns a unified and multi-functional discrimination model to distinguish the target from both distracting backgrounds and other neighboring or overlapping targets. Inspired by the recent Siamese structure [9], the discrimination model takes two image samples as input and outputs a discrimination confidence value as well as a similarity. This discrimination model is firstly learned offline by distinguishing generic object samples from other targets and background distractions with a distractor-aware loss function. To handle the target appearance variations caused by factors like pose variations, object occlusions and target interactions, the discrimination model is further enhanced by a relational attention procedure, which introduces a lightweight self-attention mechanism by capturing the trajectory feature globally from the history states stored in a temporal window and aggregating them via the weighted fusion learning.

By incorporating the object detection responses and the proposed discrimination model, a multi-stage tracking pipeline is designed for automatic trajectory initialization, propogation, and termination. The discrimination model builds for each initialized target a dedicated appearance model, which is efficiently updated online to preserve its discrimination ability. This dedicated appearance model serves not only as a single object tracker to grow the target trajectory in the scenario that the target is isolated from other targets with inaccurate detection response or even miss-detected but also as a discriminator to distinguish against distractions from the backgrounds and other neighboring or occluding targets. Its predictions are used to replace missing detection responses and refine inaccurate responses, and its confidence scores prevent tracking from drifting in the long term. To summarize, this work incorporates the merits of single object tracker and offline object detector and overcomes their deficiencies to present a new online MOT algorithm with distractor-aware discrimination learning. Its main contributions are threefold.

- A distractor-aware discrimination learning model is proposed to facilitate online multi-object tracking to better differentiate one target from other targets and semantic backgrounds in the scenes.
- A relational attention learning mechanism is introduced to handle appearance variations of targets caused by large pose variations, object occlusions, and target interactions.
- A multi-stage tracking strategy is established within a temporal sliding window which leverages the object detection responses and tracker predictions to deal with trajectory drifting.

Based on the above technical contributions, this study has developed an effective online MOT system. Extensive experimental analyses and evaluations on the widely used challenging MOT16 and MOT17 benchmarks demonstrate the effectiveness of the proposed approach. To facilitate further studies on the online multi-object tracking problem, we will release the source code and trained models of the proposed MOT approach.

2. Related work

Multi-Object Tracking. Tracking-by-detection is becoming the most popular strategy for multi-target tracking with the develop-

ment of object detection methods. The main idea is that trajectories are generated by associating the detected object hypotheses produced by an off-the-shelf object detector. Many methods tackle the task in a batch mode by formulating tracking as a global optimization problem, such as multicut [7], continuous-discrete energy minimization [10], to name a few. These approaches utilize information from both the past and future frames together to handle detection failures. However, tracking in the batch mode is not suitable for time-critical applications in the real world. In contrast, online MOT methods rely only on the information up to the current frame to estimate trajectories. These methods can be divided into two categories: probabilistic inference [11] and deterministic optimization [12]. Such online tracking methods are more sensitive to noisy detections, and detection failures seriously affect the tracking performance. In this work, we integrate merits of object detection and single object tracker to deal with detection failures. Single object tracker can refine the detection and compensate for the missing detection, and confident detection can remedy the tracker drifting.

Object Detection in MOT. Object detection, especially pedestrian detection, receives considerable interests in MOT as it is the first and a critical step for tracking-by-detection methods. Traditional pedestrian detectors, such as ACF [13] and DPM [14], exploit various filters on hand-craft features with sliding window strategy to localize objects. Recently, object detection is dominated by the CNN-based methods [5]. These methods use deep features rather than hand-craft features to classify and localize each target simultaneously. CNN-based detectors outperform significantly traditional detectors both on speed and accuracy. However, even the state-of-the-art CNN-based detectors still inevitably encounter detection failures in practice, especially in crowd scene for MOT, as targets interact with others frequently and the environment sometimes is extremely cluttered. Detection failure is one of the most challenging problem for tracking-by-detection methods. Our work focuses on applying a single object tracking method to handle detection failures including false alarm, missing detection and inaccurate localization.

Single Object Tracker in MOT. Thanks to the significant progress in Single Object Tracking (SOT) field in recent years, single object trackers have been introduced into MOT task in several previous works. Compared with single target tracking, multi-target tracking has several difficulties. First, the number of targets is uncertain, and the start and end points of the trajectory are uncertain. Second, serious occlusions occur between the targets. Finally, there may be strong similarities between the targets. Therefore, single-target trackers cannot be directly applied to multi-target tracking tasks. Xiang et al. utilizes Markov Decision Process (MDP) [15] to track targets in tracked state with optical flow based on the TLD tracker [16]. STAM [17] exploits a spatial-temporal attention mechanism to handle drift issues via regarding all the detections as SOT proposals. DMAN [18] directly applies the ECO tracker [19] from SOT with a cost-sensitive loss and designed a spatial-temporal network for data association when SOT tracker is considered losing the target. However, all these methods are combined with online-updating SOT tracker which is slow in speed and costs a lot of memory. To make matters worse, there are not enough samples to update each tracker, causing the trajectory to drift gradually.

In this work, we propose an online MOT algorithm based on the offline training siamese SOT tracker, SiamRPN [20]. Siamese network-based tracking method [9] contains two CNN branches: one for the template target and the other for the search region, the two branches share the same architecture and parameters. During tracking, the two branches are fed into the cross-correlation layer for sliding window evaluation. Li et al. [20] fuse a Siamese network and Region Proposed Network (RPN) detection method to formulate tracking task as a one-shot detection problem and get the top tracking performance with a high speed. To enhance the robustness and accuracy of existing Siamese-based trackers, Zhang et al. [21] propose new residual modules to eliminate the negative impact of padding. In SiamRPN++ [22], a new architecture is proposed to perform layer-wise and depth-wise aggregations, which reduces the model size and further increase the speed. There are three main changes when we tailor the SiamRPN tracker for MOT in this paper. Firstly, a bi-direction correlation-based tracking structure is exploited in each candidate associate pair to reduce the potential for tracking drift. Secondly, a distractor-aware discriminative loss function is proposed to handle distractors. Finally, a relationship attention mechanism is combined to alleviate the occlusion problem.

3. Proposed online MOT algorithm

As a state-of-the-art single object tracking method, SiamRPN can grow trajectory with bounding box regression from region proposals. However, it does not perform well in the cases when some trajectories are close and interfere with each other or a trajectory is continuously occluded after several frames. Based on these considerations, a distractor-aware discrimination learning model integrating siamese structure is proposed to compensate missing detection, smooth inaccurate detection, and discriminate distractors simultaneously.

3.1. Distractor-aware discrimination learning

The schematics of the proposed discrimination model is shown in Fig. 2. It takes two image samples as input and outputs a discrimination confidence as well as a similarity map. The discrimination confidence is used to discriminate confusing targets while the similarity map benefits to reflect the samples' spatial relationship. The feature extractor f_θ is a modified ResNet-50 structure within which the first four stages are retained and the output stride is reduced to 8 to obtain a higher spatial resolution. The blocks h_ϕ and h_w in the Refining Module (RM) are both two 1×1 convolutional layers with $\{256, 2k\}$ and $\{256, 4k\}$ channels respectively, with k being the number of proposals. The cross correlation operation in RM not only refines the candidate bounding box, but also provides the ROI to extract features for binary classification, *i.e.*, whether the association is correct or not.

In order to explain the loss function more clearly, we first introduce the meaning of the notations used in the loss function. The Discrimination Module (DM, c.f. Fig. 2) outputs the predicted cofidence score c. The ground-truth c^* represents whether the pair of image samples belong to the same target. The output $p = [p_x, p_y, p_w, p_h]$ (c.f. Fig. 2) of S and $p^* = [p_x^*, p_y^*, p_w^*, p_h^*]$ denote the predicted probability and the ground-truth label that the corresponding anchor is responsible for refining the target position. The output t of B is a vector representing 4 parameterized coordinates of the bounding box predicted by each positive anchor while t^* is that of associated ground-truth. The parameterization method is the same as RPN [23]:

$$p_x = \frac{g_x - a_x}{a_w}, \quad p_y = \frac{g_y - a_y}{a_h}$$
$$p_w = \log\frac{g_w}{a_w}, \quad p_h = \log\frac{g_h}{a_h}, \tag{1}$$

where g_x, g_y, g_w, g_h represent the center position and size of the ground truth bounding box while a_x, a_y, a_w, a_h denote that of an anchor.

The loss function of our model consists of three components for different tasks, the box classification $\mathcal{L}_{bc}$, the box regression $\mathcal{L}_{br}$

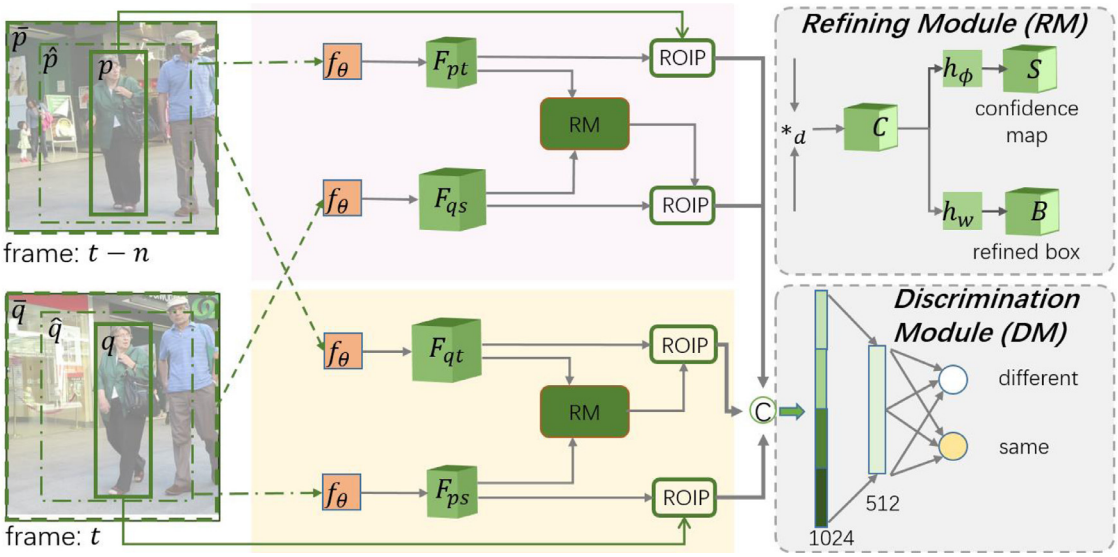

Fig. 2. Schematic of the proposed discrimination model. A pair of Region-of-Interests (RoIs) p, q with their wrappers $\hat{p}, \hat{q}$ and candidate search regions $\bar{p}, \bar{q}$ are input to the network. The template-candidate tuples $(\hat{p}, \bar{q}), (\hat{q}, \bar{p})$ are processed by a Siamese network respectively, where the feature maps of the template $F_{pt}, F_{qt} \in \mathbb{R}^{15\times15\times256}$ are used to provide a refined RoI (q_r, p_r) for the candidate search feature maps $F_{qs}, F_{ps} \in \mathbb{R}^{31\times31\times256}$ with the Refining Module (RM). Compact features extracted from the feature maps with the help of RoIs (p, q_r, q, p_r) are concatenated to form a 1024-dimensional feature vector, which is further exploited in the Discrimination Module (DM) to discriminate whether p and q are of the same identity. ⓒ denotes the concatenation operator. ROIP means ROI Pooling. In RM, $*_d$ denotes depth-wise cross correlation. The cross correlation maps $C \in \mathbb{R}^{17\times17\times256}$ are fed to two convolution branches to generate a confidence map $S \in \mathbb{R}^{17\times17\times2k}$ and to refine bounding boxes $B \in \mathbb{R}^{17\times17\times4k}$. The bounding box with the highest confidence score is selected as the refined target.

and the association classification $\mathcal{L}_{ac}$ as follows:

$$\mathcal{L} = \mathcal{L}_{bc}(p, p^*) + \lambda_1 \cdot \mathcal{L}_{br}(t, t^*) + \lambda_2 \cdot \mathcal{L}_{ac}(c, c^*), \tag{2}$$

where λ_1 and λ_2 are balance parameters. Although the box regression $\mathcal{L}_{br}$ and box classification $\mathcal{L}_{bc}$ are motivated by RPN, there are some differences when it comes to MOT. In RPN, the anchor which has the highest IoU overlap with a ground-truth box or an IoU overlap higher than 0.7 with any ground-truth box is selected as a positive sample, while the negative samples are the anchors whose IoU values are lower than 0.3. The model is prone to just discriminate foreground from the non-semantic background as the training procedure is dominated by easy negative samples, which is very disadvantageous for multi-target tracking task that needs to distinguish between different foreground targets.

Hence, in the proposed distract-aware loss, for the box classification, more hard negative samples, such as other confusing targets and ignored anchors (IoU $\in$ [0.3, 0.7]) are punished, for the box regression, the variation of proposals associated with the same target is also minimized to suppress the occurrence of diffusion box. The new losses can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{bc}(p, p^*) &= (1-\alpha)\mathcal{L}^e_{bc}(p, p^*) + \alpha\mathcal{L}^h_{bc}(p, p^*) \\ \mathcal{L}_{br}(t, t^*) &= (1-\beta)\mathcal{L}^s_{br}(t, t^*) + \beta\mathcal{L}_{cs}(t). \end{aligned} \tag{3}$$

Both $\mathcal{L}^e_{bc}$ and $\mathcal{L}^h_{bc}$ are binary cross entropy losses, where the former acts on positive foreground and easy negative samples like RPN and the latter acts on the hard-negative samples. The hard-negative samples are selected from other foreground with different identity or ignore anchors who have higher responses, to enforce the model to extract more discriminative features. $\mathcal{L}^s_{br}$ is the smooth L_1 loss on all positive samples for box regression and $\mathcal{L}_{cs}(t) = \mathbb{E}\|t - \bar{t}\|_1$ intends to ensure all the positive proposals are close and compact, where $\bar{t}$ means the expectation of the predicted parameterized coordinates and $\mathbb{E}$ is the expectation operator. The close and compact constraint not only benefits to the subsequent Non-Maximum Suppress (NMS) operation, but also helps to make the features more discriminative. The α and β are parameters to balance different components. These two improvements are important for solving the frequent ID switches in MOT tasks.

Furthermore, an association classification loss $\mathcal{L}_{ac}$, the cross-entropy loss over two classes, is also adopted in Eq. (2) to distinguish whether the pair of image samples belong to the same target. Given a new frame, we name the trajectory where at most one detection overlapping with its prediction as isolated trajectory and competitive trajectory otherwise. In training, only competitive trajectories are collected to generate positive and negative samples to train the association classifier.

3.2. Relational attention learning

The target appearance often varies from frame to frame due to factors like object occlusions, pose variations, and target interactions. To handle these variations, history observations are commonly used to character trajectory feature. The most common practice is to normalize the history features with the weights encoding the similarities between them and the candidate. However, the relationship among history observations is typically overlooked. The relationship can weaken the influence from outliers to update the trajectory feature more robustly. To this end, a lightweight self-attention mechanism is introduced in our model to capture the relationship among history observations. Instead of acquiring the importance of history sample through its similarity with current sample, we learn the dependence of the trajectory on each sample online in a self-attention manner, thereby suppressing the negative effects of noise points.

Specifically, as illustrated in Fig. 3, for each trajectory, given N history observations with feature maps $\mathbf{F}_1$, a spatial Gaussian weight is first applied at each channel to reduce the effect of surroundings. The features are further compacted with a 1×1 convolution layer. A global max-pooling operator is followed to abstract invariant features $\mathbf{P} \in \mathbb{R}^{N\times C'}$. A relation matrix is calculated by multiplying $\mathbf{P}$ with its transpose. The row-normalized relation matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ is obtained as

$$\mathbf{D}_{ij} = \frac{\exp(\mathbf{P}_i \cdot \mathbf{P}_j^T)}{\sum_{k=1}^{N} \exp(\mathbf{P}_i \cdot \mathbf{P}_k^T)}, \tag{4}$$

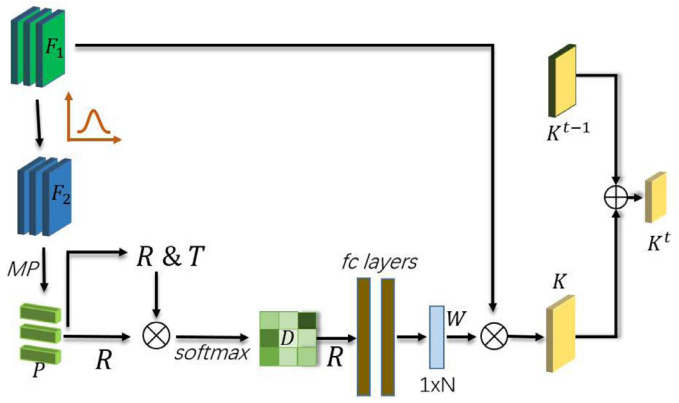

Fig. 3. The details of the Relational Attention Module. *MP* is the global max-pooling operator. *R, T* denote reshape and transpose operators respectively. $\otimes$ represents the matrix multiplication and $\oplus$ means element-wise addition.

where $\mathbf{D}_{ij}$ indicates the jth observation's impact on the ith observation. The relation map $\mathbf{D}$ is then reshaped as a vector and fed into two fully-connected layers ($N^2 \times N^2$, $N^2 \times N$) followed by a softmax layer to obtain the attention score $w \in \mathbb{R}^N$ of each observation.

The final output (trajectory kernel) $\mathbf{K} \in \mathbb{R}^{N\times C\times H\times W}$ is obtained by

$$\mathbf{K} = \sum_{i=1}^{N} w_i \mathbf{F}_{1i}\,. \tag{5}$$

To ensure that the trajectory kernel updates smoothly, a momentum term is used in the update process as $\mathbf{K}^t = \eta\mathbf{K}^{t-1} + (1-\eta)\mathbf{K}$, where $\eta = 0.95$ is the momentum coefficient.

In training, before the unified end-to-end training, relational attention module is first pretrained using specified samples generated from competitive trajectories. Feature maps of $N-k$ observation set $\mathcal{S}_{N-k}$ in the same trajectory are extracted using f_θ, while the other k feature maps are extracted from observations set $\mathcal{S}_k$ of other trajectories. k is a random integer ranging from 0 to $0.3 \times N$. The label of observation o is $\frac{1}{N-k}$ if $o \in \mathcal{S}_{N-K}$, and 0 otherwise.

3.3. Multi-stage tracking

The proposed discrimination model can distinguish the target from both distracting background and other neighboring or overlapping targets, which is essential to grow trajectory. And trajectory propogation is a critical step in multi-target tracking. Benefited from the discrimination model, a multi-stage tracking pipeline (shown in Fig. 4) is designed in this work to track multiple targets in an online mode.

Considering that isolated trajectory and competitive trajectory face large differences in growing, we adopt different tracking strategies for isolated and competitive trajectories. In the *first* stage, each alive trajectory takes its current bounding box as candidate region and refines the bounding box using RM branch. For the isolated trajectory, the refined bounding box is appended as new observation if the trajectory's confidence (as Eq. (6)) is larger than a threshold τ_p.

$$S_{T_k} = \begin{cases} \frac{\sum_i^{n_p} S_i}{n_p} \cdot (2 - \exp(\epsilon\sqrt{n_p})), & \text{if } n_p > 0 \\ 1, & \text{else} \end{cases}, \tag{6}$$

where n_p is the time of continuous tracking in the first stage and S_i denotes the refining confidence in the ith growth. ϵ is a balance parameter. Empirically, the ϵ is related to the allowed maximum number $N_{\max}$ of consecutive failed matches, $\epsilon \approx \log(2)/\sqrt{N_{\max}}$. $\epsilon = 0.1$ in all our experiments.

In the *second* stage, for competitive trajectories, their refining bounding box and overlapped detections after NMS are collected as candidates. The similarities between trajectory and the candidates are calculated using the association classifier branch of the discrimination model. Then the Hungarian algorithm is applied at the association similarity matrix to grow competitive trajectory. In the *last* stage, the remaining detections are further assigned to the untracked trajectories based on IoU between detections and tracker predictions with a threshold τ_{iou}.

After data association, each untracked trajectory is considered as lost in the current frame and a new trajectory is initialized for each unmatched detection with a high response confidence. To alleviate the influence of false detection, any new trajectory will be deleted once it is lost in any of the first τ_i frames. The trajectory will be terminated if it keeps lost for over τ_t successive frames or exits the field of view. For the trajectory kernel update, N history observations are selected as follows,

$$o_i = \arg\max_{t-i\tau_t < j \le t-(i-1)\tau_t} Q_{o_j}, i = 1, \ldots, N\,, \tag{7}$$

where Q_{o_j} is the detection confidence of o_j.

4. Experiments

In this section, we first introduce the experiment settings including datasets, evaluation metrics and the implementation details in Section 4.1. The proposed distractor-aware loss, the relational attention module and the multi-stage tracking strategy are then analyzed respectively in Section 4.2. Finally, in Section 4.3, our proposed online MOT algorithm is compared with the state-of-the-art methods on the public MOT benchmarks.

4.1. Experiment settings

Datasets. We evaluate our online MOT algorithm on the publicly available MOT16 and MOT17 benchmark datasets [24]. The MOT16 dataset consists of 14 video sequences, 7 for training and 7 for testing respectively, and provides public detections derived from DPM [14]. The MOT17 dataset shares the same video sequences with MOT16 and provides another two sets of public detections (by Faster R-CNN [23] and SDP [25]) for more comprehensive evaluation. We use the training sequences in MOT16 benchmark for model training and investigation. Specifically, two sequences, MOT16-09 and MOT16-10, are selected for validation and the remaining ones are used for training. Public detections are used in all experiments for fair comparison.

Evaluation Metrics. We adopt the widely used CLEAR MOT metrics [26,27] to measure the performance of the proposed online MOT algorithm. These metrics include Multiple Object Tracking Accuracy (MOTA ↑), Mostly Tracked targets (MT ↑, the ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span), Mostly Lost targets (ML ↓, the ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span), the number of False Negatives (FN ↓), the number of False Positive (FP ↓), the number of ID Switches (IDS ↓) and the number of Fragments (Frag ↓). Additionally, ID F1 score [28] (IDF1 ↑), which denotes the ratio of correctly identified detections over the average number of ground-truth and computed detections, is also employed to measure the identity-preserving ability of trackers. Here ↑ denotes that higher scores indicate better performance, and ↓ denotes lower scores indicate better performance. Metrics of ACC, EFI, EFP and IoU are used in ablation study, which will be illustrated accordingly.

Implementation Details. The proposed algorithm is implemented with PyTorch. During the further training phase of the discrimination model, the feature extraction layers f_θ are fixed and

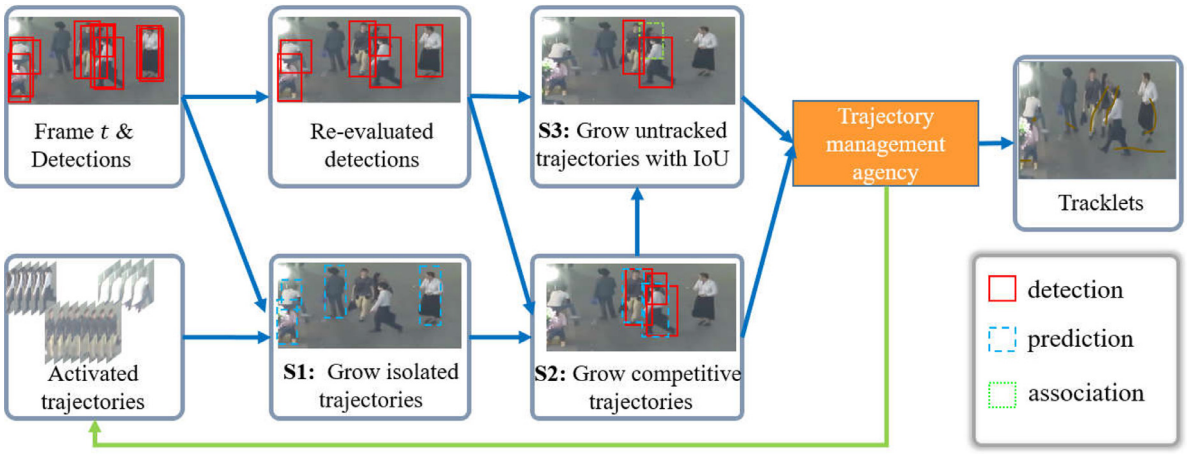

Fig. 4. Pipeline of the proposed MOT algorithm. For each frame, targets are tracked in three stages, *i.e.*, growing isolated trajectories (S1), growing competitive trajectories (S2) and associating untracked trajectories with unassigned detections using IoU (S3). In S1, the RM branch is used to locate new locations which helps to suppress misses or inaccurate detections. The DM branch is used in S2. Facing the competitive trajectories, the bi-directory track in DM provides stronger distinguishing information, which helps to suppress ID switches. Trajectory initialization, propogation and termination are handled in a management agency at each time step. For each tracked track, relational attention is used in the management agency to online update its template to adapt to apparent changes.

the remaining parameters are fine-tuned on MOT training dataset. The ratios and scales of the anchors used in the RM are set as {2, 3} and {7, 9} respectively. The learnable weights of relational attention module are initialized with `xavier` initialization [29]. Any two observations which have the same identity and with temporal distance less then 50 frames are paired as a sample. The network is trained for 25 epochs with the Stochastic gradient descent (SGD) optimizer. Learning rates for region proposal and relational attention module are initialized as $1e-4$ and $1e-2$ respectively. As training proceeds, they are reduced to their 1/10 quantity at the 10th and 18th epoch.

For training data preparation, there are 82,805 samples in each epoch for training and 20,000 samples for validation. Any two observations with the same identity and with temporal distance less than 50 frames are paired as a sample. Data augmentation, such as color jitter, image horizontal flip, and random displacement noise of search region is used in training.

For the parameters setting, in training, the balance parameters in Eq. (2) are experimentally set as $\lambda_1 = 1, \lambda_2 = 2$ in all the evaluations and the parameters α, β are both set as 0.4 when the proposed model is evaluated on MOT benchmark. In the data association, the thresholds for alive trajectory is set as $\tau_p = 0.8$. The balance parameter ϵ in Eq. (6) is set as 0.05 and $N = 8$ historical states are collected to extract the trajectory feature. We set $\tau_{iou} = 0.4$ to suppress the overlapped detections. The trajectory initialization threshold τ_i is set as 2 and the termination threshold τ_t is set as 30. Multi-parameter is a general problem of MOT approaches [7,17,18,30–33]. Most of the common parameters in our method are set as other approaches and without further tuning, the regularizers in loss function will be further analyzed in Section 4.2.

4.2. Ablation study

The Advantage of Distractor-aware Loss. As shown in Eq. (3), we name $\mathcal{L}^h_{bc}$ and $\mathcal{L}_{cs}$ as repulse and consistency terms. The $\mathcal{L}^h_{bc}$ pushes the hard negative samples away from the positive samples and the $\mathcal{L}_{cs}$ concentrates the predictions of the positive samples more concentrated. To better evaluate the effects of these two loss components, we report the performance regarding $\mathcal{L}^h_{bc}$ and $\mathcal{L}_{cs}$ by varying the balance parameters α and β in Table 1.

In the evaluation, given a template and a search area, it is deemed as a correct prediction if the IoU between the ground-truth and the regression is greater than 0.7. We use ACC to denote the ratio of correct predictions. The metric of EFI denotes the ratio of error prediction from ignored anchors in all error predictions, which means some negative samples from ignored anchors are mis-classified as positive. The metric of EFP denotes the ratio of error prediction from positive anchors in all error predictions, which means the classification is right while the regression is inaccurate severely. $\alpha = 0$ means only the original RPN loss is used while $\alpha = 1$ means the samples from the ignored anchors rather than the negative anchors are exploited in training. We evaluate the impact of different β on performance at $\alpha = 0.4$ and the performance of different α at $\beta = 0.4$ in Table 1. It can be concluded that, the best ACC is achieved when $\alpha = 0.4$ and the best EFI is obtained when $\alpha = 1.0$. This demonstrates that the repulse loss term is more effective to distinguish negative samples from ignored anchors. Further more, the ACC is further improved and the EFP is reduced when the consistency loss is used. However, when β is greater than 0.4, the performance is degraded. The reason behind this case is that the loss function pays more attention to regression consistency rather than accuracy. In the extreme case, $\beta = 1$ for example, where the regression accuracy is almost totally ignored, the ACC degrades to nearly 0. Note that EFI and EFP are not necessary to be consistent with ACC because the multiple loss terms influence each other.

We further analyze the proposed loss using MOT metrics along with metrics of ACC and IoU which means the average overlap between correct regressions and targets. To better investigate the influence regarding only $\mathcal{L}^h_{bc}$ and $\mathcal{L}_{cs}$ in Eq. (3), we exclude the influence from the relational attention module by removing this module when we construct baseline variants. Specifically we compare four variants. The first is a plain one with neither $\mathcal{L}^h_{bc}$ nor $\mathcal{L}_{cs}$, *i.e.*, the box regression loss $\mathcal{L}_{br}$ and box classification loss $\mathcal{L}_{bc}$ in Eq. (2) are the same as in RPN. The second and the third ones are counterparts with either $\mathcal{L}^h_{bc}$ or $\mathcal{L}_{cs}$. The fourth one is a variant with both $\mathcal{L}^h_{bc}$ and $\mathcal{L}_{cs}$. A tick mark in Table 2 indicates the corresponding loss term is included in the counterpart. Results in Table 2 suggest that both $\mathcal{L}^h_{bc}$ and $\mathcal{L}_{cs}$ contribute to improve the model. For example, the MOTA, IoU, IDF1, and ACC values increase with varying degrees. The repulse term $\mathcal{L}^h_{bc}$ is especially more effective, as the results indicate.

The Advantage of Relational Attention (RA). Fig. 5 shows the visualization results of the self-attention mechanism. (a) demonstrates eight stored historical states and (b) is the search area of the target in the current frame. The attention weights obtained re-

Table 1
Analysis of the proposed distractor-aware loss with different values of α and β.

Parameter	Metric	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\alpha@(\beta = 0.4)$	ACC(%)	84.20	85.32	85.50	86.43	**86.52**	86.09	85.93	85.83	85.46	85.10	82.19
	EFI(%)	13.71	11.72	11.31	10.54	10.76	10.99	9.96	11.08	9.90	8.86	**7.69**
$\beta@(\alpha = 0.4)$	ACC(%)	86.52	87.12	87.08	87.04	**87.32**	86.29	85.62	85.57	85.72	84.02	0.03
	EFP(%)	10.76	**7.61**	9.83	11.12	10.56	11.45	12.03	16.08	17.99	15.71	86.53

Table 2
Ablation study results on the validation set in terms of different configurations of loss terms and the Relational Attention module (RA).

Method	$\mathcal{L}_{ac}^{h}$	$\mathcal{L}_{cs}$	RA	ACC(%)	IoU	MOTA(%)	IDF1(%)	IDS	Frag
ablation models				84.24	0.7138	45.8	47.2	158	353
	√			87.26	0.7154	46.7	49.9	147	322
		√		87.15	0.7092	45.9	47.4	164	337
	√	√		87.48	0.7169	47.4	51.3	137	305
			√	88.51	0.7267	47.5	52.1	102	**297**
final model	√	√	√	**90.90**	**0.7456**	**48.9**	**54.6**	**91**	299

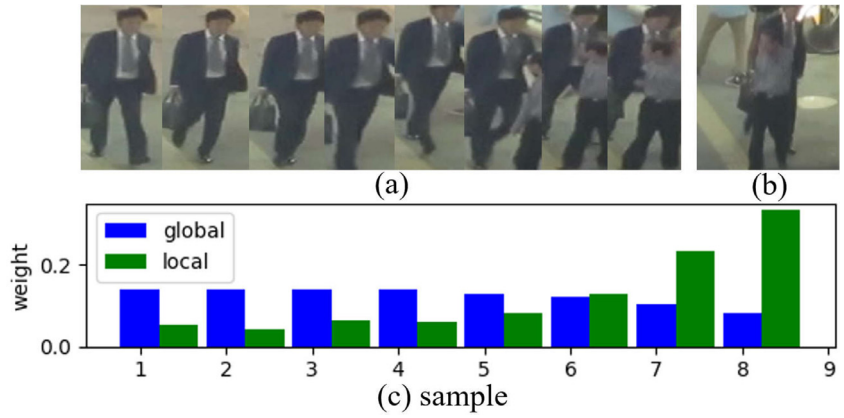

Fig. 5. Visualization of relational attention module. (a) presents eight historical states of an object. (b) shows searching area of the target. (c) compares the attention weights between the historical states and the search candidate by the local mode (without the relational attention module) and the global mode (with the relational attention module). The attention weights in global mode are more consistent than those in local mode.

spectively in global and local modes are compared in (c). Global mode means the attention weights are obtained by the relational attention module, while local mode obtains the attention weights by normalizing the similarities between each historical state and the candidate. It is not difficult to find from (c) that, because of error associations, the weights of the 7th and 8th historical states obtained by the local mode are extremely higher than other values. This results in that the cues of these two observations are dominated when aggregating observations with these weights to character the trajectory. Therefore, the candidate in (b) will be assigned to the trajectory leading to identity switch in this case.

Different from the weights obtained in local mode, weights achieved by self-attention mechanism encode the relationships among all the observations to evaluate the importance of each observation to the trajectory more robustly as shown in (c). The weighted average of the historical samples is used as the current feature of the trajectory, to measure the matching degree between the current target and the trajectory. Thus, the relational attention module can suppress casual mismatches to better collect historical information globally. For the example in Fig. 5, the black suit man will be considered temporarily lost in (b) as the similarity between the target and the trajectory feature after fusion is small, thus avoiding the exchange of track ID with the light shirt man.

Quantitative results in Table 2 also validate that the proposed relational attention module is effective. Regardless of whether distractor-aware loss terms are used, the relational attention module improves model performace. In particular, the better values of IDF1 and IDS demonstrate the benefits from the relational attention module in reducing identity switch in MOT tracking.

Exemplar tracking results are shown in Fig. 6. The detection of trajectory #2 is missing but the proposed model can track it successfully with the RM. Trajectory #3 and #5 are occluded by trajectory #2 in frame #36, but the trajectories' identities are preserved with the help of association classifier branch.

The Advantage of Multi-stage Strategy. In addition to the classification branch and the relational attention learning, the multi-stage strategy also plays an important part in out multi-target tracking model. To analyze the impact of each step, the ablation experiments have been conducted on MOT16 training set as shown in Table. 3, where DAL is the abbreviation of Distractor-Aware Learning.

Though a distractor-aware siamese region proposal network (DaSiamRPN) is proposed in [34] for single object tracking, we select SiamRPN [20] tracker as the baseline. Compared with SiamRPN, the DaSiamRPN has two main improvements. On one hand, the data in the detection task is introduced to expand the diversity of positive samples and improve the generalization ability of the model. On the other hand, different instances are used to build hard negative samples to improve the discriminability of the model. However, these two points are not helpful for MOT tasks. First of all, there is only one main category in the MOT bench-

Fig. 6. Exemplar tracking results. The thin yellow boxes indicate the detection results, while the other boxes indicate the tracking results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Analysis on multi-stage strategy.

Method	MOTA(%)	MOTP (%)	IDs	IDF1(%)	Frag	FPS
SiamRPN (naive)	28.9	44.4	2864	19.6	1935	10.4
+ DAL	32.2	70.4	528	45.8	1027	4.7
+ DAL + (S1,2)	44.1	72.3	511	47.1	785	4.3
+ DAL + RA + (S1,2)	48.2	74.5	122	52.9	358	2.9
+DAL +RA +(S1,2,3) (final model)	48.9	75.9	91	54.6	299	2.7

mark, that is, pedestrian. Secondly, there are a large number of crowded scenes in MOT, and how to distinguish different instances in crowded scenes is the key to MOT, not to distinguish instances from different scenes. Thus, SiamRPN is more appropriate to be our baseline. The main different of applying distractor-aware siamese networks in MOT and SOT is that it aims at different problems. In SOT, the distractor-aware mechanism is used to increase the discrimination of response score to sensitive to the target disappearance, while in MOT, the distractor-aware mechanism is used to distinguish different instances in crowded scenes. Specifically, for each new frame, detections are first suppressed by predictions, and then each detection whose confidence larger than 0.6 is considered to be the starting point of a trajectory to establish a specific SiamRPN tracker. It is straight forward to find from Table. 3 that naive SiamRPN tracker has a very poor performance on MOT task. This is due to the poor ability of the proposed feature to discriminate between foregrounds, resulting in trajectory prone to drift and frequent track id switch. After adding DAL, the extracted features are better for the identification of the foreground, so tracking performance is improved. (S1,2) means the first and second stages are used in the method, *i.e.*the trajectories are divided into two categories, isolated and competitive, for tracking, and the detection is not only used to create new trajectory, but also used to correct the trajectory where the tracking drift occurs. We can find the first two stages are benefit to significant improve MOTA and MOTP. To further utilize the historical information handling the frequent ID switch, RA is combined. It can be find the historical information is important for reducing the ID switch, IDs has dropped from 511 to 122, and IDF1 has increased by nearly 4 percent. Our final proposed method contains the single object tracker with distractor-aware discrimination learning, the RA and the multi-stage tracking strategy.

By analyzing and comparing the experimental results in Table. 3, we get three conclusions. Firstly, naive SiamRPN alone are not sufficient for robust application in multi-object scenarios with many distractors, and discriminative features are necessary. Secondly, RA is important for reducing ID switch. Lastly, the multi-stage tracking strategy can well integrate the functions of each module (*e.g.* DAL, RA, Detections) to achieve better tracking performance.

In addition, we compare model speeds in the last column of Table 3. It can be found that although the speed of SiamRPN can reach higher than 100 FPS in single target tracking, the speed is greatly reduced in MOT task. This is mainly because each target in MOT needs to create a SiamRPN tracker separately. The DAL module reduces speed further because the bi-directory tracking strategy is used to make better use the sequence information in the tracking process. The RA module needs to extract features and build the relationship topology map, which also brings time consumption. In the future work, we will process targets in one frame at the same time instead of processing each target individually to reduce redundant operations and improve the tracking speed of the model.

4.3. Evaluation on MOT benchmarks

The proposed approach is compared with several state-of-the-art MOT methods on the test sets of both MOT16 and MOT17 benchmarks. Quantitative comparison results are presented in Tables 4 and 5, respectively.

For MOT16 dataset, our method achieves the best performance in terms of MOTA and ML metrics and comparable results in terms of IDF1 and FN values against the state-of-the-art online MOT methods. As the most comprehensive metric for MOT, the MOTA value obtained using our approach is even comparable with the performance of state-of-the-art offline methods (*e.g.*, [35]), which demonstrates the effectiveness of the proposed method. In further analysis, we find our approach obtains a comparable IDF1 value but a higher IDS value. As the IDS is the total number of identity switches while IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections, it proves the relational attention module is capable of suppressing casual associations. The under-performance of MT and Frag is mainly due to the adopted naive zero-order motion model where the results in previous frame are directly used as current candidates. Candidates with large deviations lead to a higher Frag and a lower MT. When a more complicated motion mode (5-order) is adopted, the MT increases from 14.0 to 15.2 and Frag drops from 1886 to 1780. More studies on the motion model will be our future work.

Table 4
Evaluation results on MOT16. The best two results regarding each metric are marked by *italic* and **bold** respectively.

Mode	Method	MOTA(%)	IDF1(%)	MT(%)	ML(%)	FP	FN	IDS	Frag
Offline	LMP [7]	*48.8*	51.3	18.2	**40.1**	6654	**86,245**	481	**595**
	GCRA [35]	**48.2**	48.6	12.9	41.1	**5104**	88,586	851	1117
	FWT [36]	47.8	44.3	18.1	*38.2*	8886	*85,487*	852	1534
	NLLMPa [37]	47.6	47.3	17.0	40.4	5844	89,093	629	768
	ASTT [38]	47.2	44.3	16.3	41.6	*4680*	90,877	633	814
	MCjoint [39]	47.1	**52.3**	*20.4*	46.9	6703	89,368	**370**	598
	NOMT [30]	46.4	*53.3*	**18.3**	41.4	9753	87,565	*359*	*504*
Online	Ours	*48.5*	**52.8**	14.0	*37.2*	7525	**85,657**	782	1886
	MOTDT [31]	**47.6**	50.9	**15.2**	**38.3**	9253	*85,431*	792	1858
	AMIR [40]	47.2	46.3	14.0	41.6	*2681*	92,856	774	1675
	DMMOT [18]	46.1	*54.8*	*17.4*	42.7	7909	89,874	**532**	1616
	STAM16 [17]	46.0	50.0	14.6	43.6	6895	91,117	*473*	**1422**
	DCCRF16 [32]	44.8	39.7	14.1	42.3	**5613**	94,133	968	*1378*

Table 5
Evaluation results on MOT17. The best two results regarding each metric are marked by *italic* and **bold** respectively.

Mode	Method	MOTA(%)	IDF1(%)	MT(%)	ML(%)	FP	FN	IDS	Frag
Offline	FWT [36]	*51.3*	**47.6**	**21.4**	*35.2*	24,101	*247,195*	2985	6611
	MHT_DAM [33]	**50.7**	47.2	20.8	36.9	**22,875**	252,889	**2314**	*2865*
	EDMT17 [41]	50.0	*51.3*	*21.6*	**36.3**	32,279	**247,297**	*2264*	**3260**
	IOU17 [42]	45.5	39.4	14.7	40.5	*19,993*	281,643	5988	7404
Online	Ours	*51.4*	**53.7**	16.5	*34.9*	**21,042**	**251,873**	2319	5527
	MOTDT17 [31]	**50.9**	52.7	**17.5**	**35.7**	24,069	*250,768*	2474	**5317**
	HAM_SADF17 [43]	48.3	51.1	17.1	41.7	*20,967*	269,038	*1871*	*3020*
	DMAN [18]	48.2	*55.7*	*19.3*	42.7	26,218	263,608	**2194**	5378

Similarly, for MOT17, Table 5 shows that the proposed approach outperforms the other state-of-the-art online MOT trackers regarding MOTA and ML metrics and achieves the comparable performance in terms of IDF1, FP and FN.

The overall tracking speed of the proposed approach on MOT16 and MOT17 testing sequences is about 2.5 and 2.2 fps using the 2.2 GHz CPU and a TITAN X GPU without dedicated optimization of the code. There are two main reasons for the slow tracking speed. First, a single target tracking is created for each target, and there is a lot of computation redundancy between the trackers. Secondly, in order to make better use the sequence information in the tracking process, the use of bi-directory tracking in the discrimination module has slowed down the speed even more. How to speed up the tracking speed will be one of our future research directions.

5. Conclusion

In this work, we have proposed an online multi-target tracking method which learns a distractor-aware discrimination model to grow each target either when it is continuously miss-detected or occluded after several frames. To handle the appearance variations, a lightweight self-attention module has also been designed to capture the distinctive target appearances by selectively aggregating features from history states with weights extracted from their appearance topological relationship. With the discrimination model, a multi-stage tracking strategy is further designed for multi-target tracking. Experimental results on public MOT16 and MOT17 benchmark datasets verify the effectiveness of the proposed method.

Although the effectiveness of the proposed method has been verified on the MOT benchmark, at least two aspects which can be explored in the future to further improve the performance. Firstly, there is still room for improvement in the MT and the Frag performance. More complex and accurate motion models will further be studied to enhance the model. Secondly, the proposed model locates the current position of each target with the Siamese structure. In essence, each target has experienced the operations such as cropping, wrapping, resizing, feature extraction and cross correlation, resulting in a nearly proportional relationship between the tracking time and the number of targets. In addition, we will process multiple targets in one frame at the same time instead of processing each target individually to reduce redundant operations and improve the tracking speed of the model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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 Natural Science Foundation of China (Grant No. 61672519, 61751212, 61721004), Beijing Natural Science Foundation (Grant No. L172051), the Key Research Program of Frontier Sciences, CAS, Grant No. QYZDJ-SSW-JSC040, and the National Natural Science Foundation of Guangdong (No. 2018B030311046).

References

[1] X. Yan, I. Kakadiaris, A. Shah, Modeling local behavior for predicting social interactions towards human tracking, PR 47 (4) (2014) 1626–1641.
[2] W. Luo, J. Xing, X. Zhang, X. Zhao, T.-K. Kim, Multiple object tracking: a literature review, arXiv:1409.7618 (2014).
[3] H. Wu, Y. Hu, K. Wang, H. Li, L. Nie, H. Cheng, Instance-aware representation learning and association for online multi-person tracking, PR 94 (2019) 25–34.
[4] K. Du Yong, V. Ba-Ngu, J. Moongu, A labeled random finite set online multi-object tracker for video data, PR 90 (2019) 377–389.
[5] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: a survey, arXiv:1809.02165 (2019).
[6] S. Zhang, J. Wang, Z. Wang, Y. Gong, Y. Liu, Multi-target tracking by learning local-to-global trajectory models, PR 48 (2015) 580–590.
[7] S. Tang, M. Andriluka, B. Andres, B. Schiele, Multiple people tracking by lifted multicut and person re-identification, in: CVPR, 2017, pp. 3539–3548.

[8] X. Zhou, Y. Li, B. He, Game-theoretical occlusion handling for multi-target visual tracking, PR 46 (2013) 2670–2684.
[9] B. Luca, V. Jack, F.H. Joao, V. Andrea, H.S.T. Philip, Fully-convolutional siamese networks for object tracking, in: ECCV, 2016, pp. 850–865.
[10] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE TPAMI 36 (1) (2014) 58–72.
[11] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L. Van Gool, Robust tracking-by-detection using a detector confidence particle filter, in: CVPR, 2009, pp. 1515–1522.
[12] W. Brendel, M. Amer, S. Todorovic, Multiobject tracking as maximum weight independent set, in: CVPR, 2011, pp. 1273–1280.
[13] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE TPAMI 36 (8) (2014) 1532–1545.
[14] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE TPAMI 32 (9) (2010) 1627–1645.
[15] Y. Xiang, A. Alexandre, S. Silvio, Learning to track: Online multi-object tracking by decision making, in: ICCV, 2015, pp. 4705–4713.
[16] Z. Kalal, M. Krystian, M. Jiri, Tracking-learning-detection, IEEE TPAMI 34 (7) (2011) 1409–1422.
[17] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, N. Yu, Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism, in: ICCV, 2017, pp. 4836–4845.
[18] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, M.-H. Yang, Online multi-object tracking with dual matching attention networks, in: ECCV, 2018, pp. 366–382.
[19] M. Danelljan, G. Bhat, F. Khan, M. Felsberg, ECO: Efficient convolution operators for tracking, in: CVPR, 2017, pp. 6638–6646.
[20] B. Li, J. Yan, W. Wu, Z. Zhu, X. Hu, High performance visual tracking with siamese region proposal network, in: CVPR, 2018, pp. 8971–8980.
[21] Z. Zhang, H. Peng, Deeper and wider siamese networks for real-time visual tracking, in: CVPR, 2019, pp. 4591–4600.
[22] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, J. Yan, SiamRPN++: evolution of siamese visual tracking with very deep networks, in: CVPR, 2019, pp. 4282–4291.
[23] S. Ren, K. He, B.G. Ross, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, TPAMI 39 (6) (2017) 1137–1149.
[24] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, MOT16: A benchmark for multi-object tracking, arXiv:1603.00831 (2016).
[25] F. Yang, W. Choi, Y. Lin, Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers, in: CVPR, 2016, pp. 2129–2137.
[26] B. Keni, S. Rainer, Evaluating multiple object tracking performance: the clear MOT metrics, EURASIP JIVP 1 (2008) 1–8.
[27] Y. Li, C. Huang, R. Nevatia, Learning to associate: Hybrid boosted multi-target tracker for crowded scene, in: CVPR, 2009, pp. 2953–2960.
[28] Y. Ban, S. Ba, X. Alameda-Pineda, R. Horaud, Tracking multiple persons based on a variational bayesian model, in: ECCV Workshops, 2016, pp. 1–8.
[29] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: AISTATS, 2010, pp. 249–256.
[30] W. Choi, Near-online multi-target tracking with aggregated local flow descriptor, in: ICCV, 2015, pp. 3029–3037.
[31] L. Chen, H. Ai, Z. Zhuang, C. Shang, Real-time multiple people tracking with deeply learned candidate selection and person re-identification, in: ICME, 2018, pp. 1–6.
[32] H. Zhou, W. Ouyang, J. Cheng, X. Wang, H. Li, Deep continuous conditional random fields with asymmetric inter-object constraints for online multi-object tracking, IEEE TCSVT 29 (4) (2018) 1011–1022.
[33] C. Kim, F. Li, A. Ciptadi, J.M. Rehg, Multiple hypothesis tracking revisited, in: ICCV, 2015, pp. 4696–4704.
[34] Z. Zhu, Q. Wang, L. Bo, W. Wu, J. Yan, W. Hu, Distractor-aware siamese networks for visual object tracking, in: European Conference on Computer Vision, 2018.
[35] C. Ma, C. Yang, F. Yang, Y. Zhuang, Z. Zhang, H. Jia, X. Xie, Trajectory factory: tracklet cleaving and re-connection by deep siamese Bi-GRU for multiple object tracking, in: ICME, 2018, pp. 1–6.
[36] R. Henschel, Leal-Taixé, D. Cremers, B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, in: CVPR Workshop, 2018, pp. 1428–1437.
[37] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, B. Andres, Joint graph decomposition and node labeling: problem, algorithms, applications, in: CVPR, 2017, pp. 6012–6020.
[38] Y. Tao, Adaptive spatio-temporal model based multiple object tracking considering a moving camera, in: ICUV, 2018, pp. 1–6.
[39] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, B. Schiele, A multi-cut formulation for joint segmentation and tracking of multiple objects, arXiv:1607.06317 (2016).
[40] A. Sadeghian, A. Alahi, S. Savarese, Tracking the untrackable: learning to track multiple cues with long-term dependencies, in: ICCV, 2017, pp. 300–311.
[41] J. Chen, H. Sheng, Y. Zhang, Z. Xiong, Enhancing detection model for multiple hypothesis tracking, in: CVPR Workshops, 2017, pp. 2143–2152.
[42] E. Bochinski, V. Eiselein, T. Sikora, High-speed tracking-by-detection without using image information, in: IEEE AVSS, 2017, pp. 1–6.
[43] Y.-c. Yoon, A. Boragule, K. Yoon, M. Jeon, Online multi-object tracking with historical appearance matching and scene adaptive detection filtering, AVSS (2018) 1–8.

Zongwei Zhou received the B.S. degree in electronic information and science from China University of Mining and Technology, Xuzhou, in 2013, and M.S. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, in 2016. Currently, he is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include Multi-object tracking and deep learning.

Wenhan Luo is currently working as a senior researcher in the Tencent AI Lab, China. His research interests include several topics in computer vision and machine learning, such as motion analysis (especially object tracking), image/video quality restoration, reinforcement learning. Before joining Tencent, he received the Ph.D. degree from Imperial College London, UK, 2016, M.E. degree from Institute of Automation, Chinese Academy of Sciences, China, 2012 and B.E. degree from Huazhong University of Science and Technology, China, 2009.

Qiang Wang received the B.S. degree from University of Science and Technology Beijing, China in 2015. Currently, he is a Ph.D. student in Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include visual object tracking and target segmentation.

Junliang Xing received the B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Xi'an, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an associate professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Dr. Xing was the recipient of Google Ph.D. Fellowship 2011, the Excellent Student Scholarships at Xi'an Jiaotong University from 2004 to 2007 and at Tsinghua University from 2009 to 2011. He has published more than 40 papers on international journals and conferences. His current research interests mainly focus on computer vision problems related to faces and humans.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Now he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests include visual surveillance and filtering of Internet objectionable information.