Neurocomputing 524 (2023) 17–30

Contents lists available at ScienceDirect

Neurocomputing

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

Center-point-pair detection and context-aware re-identification for end-to-end multi-object tracking

Xin Zhang, Yunan Ling, Yuanzhe Yang, Chengxiang Chu, Zhong Zhou *

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

ARTICLE INFO

Article history:
Received 30 May 2022
Revised 19 October 2022
Accepted 27 November 2022
Available online 17 December 2022

Keywords:
Multi-object tracking
Anchor-free detection
Person Re-Id
End-to-end

ABSTRACT

Online multi-object tracking aims at generating the trajectories for multiple objects in the surveillance scene. It remains a challenging problem in crowded scenes because objects often gather together and occlude in tracking frames. The main impact of crowd occlusions is that it severely harms the performance of the detector and significantly increases the difficulty in extracting object features. In this paper, we propose an end-to-end tracking framework that alleviates such issues and estimates more accurate trajectories. Firstly, We design a Center-Point-Pair detection branch for object detection, which learns the correlations between the object head and the body to simultaneously predict the head and body regions to alleviate unreliable detection in tracking scenes. Secondly, we introduce the context information around the object to the tracker, inspired by the human search pattern. We propose a Context-Aware Re-Identification branch that includes the Previous-Frame Guided Spatial-Attention Model and the Previous-Frame Guided Channel-Attention Model to extract more discriminative object features. Thirdly, to harness the power of deep features for data association in generating reliable trajectories, we propose the Similarity Cluster Trajectory Management method that expands affinity descriptor and adopts the minority obeying the majority principle to association trajectories and detections. The experiments on diverse and challenging MOT datasets show that our tracking framework achieves superior results compared to other state-of-the-arts offline and online multi-object tracking methods.

1. Introduction

Multi-Object Tracking (MOT) aims to conserve object identifications under appearance and motion variation with time and generate the trajectory information for objects in scenes. Additionally, pedestrians are an essential class of tracking objects in research and the tracking algorithms that can accurately estimate pedestrians movements are desirable in broad applications, such as smart city [1,2], autonomous driving [3,4] and video analysis [5,6]. However, tracking performance still needs to be improved due to frequent occlusion of tracking objects by obstacles or other pedestrians, similar appearance features of tracking objects.

The widely used tracking-by-detection frameworks currently employ two separate networks to estimate object tracklets [7–10]. Specifically, the detection model firstly detects tracking objects in each frame. Then the re-identification model extracts appearance features for each object. Finally, the trajectory management method updates trajectory information with similarity measurement. However, those frameworks cannot perform end-to-end training and inference. The real-time and complexity of those tracking frameworks are not satisfactory, especially when there are many objects in the tracking video.

With the in-depth study of multi-task learning networks, single-shot trackers which detect objects and extract object features by applying an end-to-end network have attracted more attention [11,12]. These methods employ standard backbone features for object detection and feature extraction, which can effectively save computation and improve the real-time performance of the tracker. Nonetheless, compared with the tracking-by-detection framework, the tracking performance of those methods is not satisfactory, especially the performance of MOTA and IDS [13]. Therefore, we have analysed the profound reason behind such issues and propose a new end-to-end online multi-object tracker. In particular, we propose three new models in this paper.

In tracking scenes, objects, especially pedestrians and street crowds, usually gather together and occlude each other. This phenomenon seriously affects the detector's performance and significantly increases the difficulty in locating tracking objects, as shown in Fig.1(a). First, in crowd scenes, the object area will

* Corresponding author.
E-mail address: zz@buaa.edu.cn (Z. Zhou).

https://doi.org/10.1016/j.neucom.2022.11.094
0925-2312/

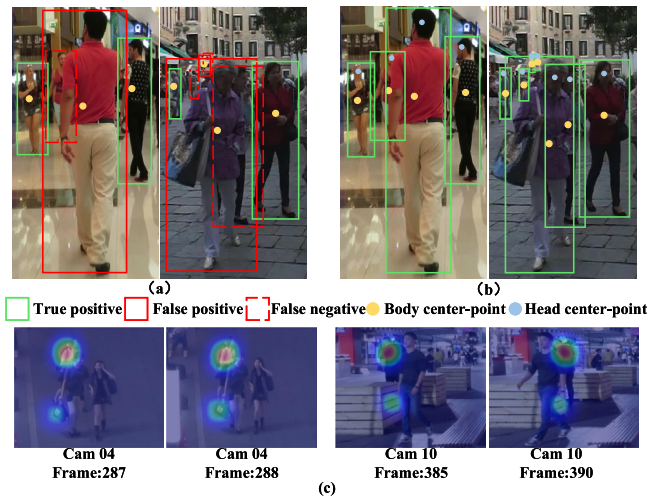


Fig. 1. (a) In crowded multi-object tracking scenes, standard detection methods may generate unreliable results. The predicted boxes in crowded scenes with occlusion using generic detection methods are often inaccurate or covering several mutually occluded individuals. (b) Object locating with head position information guiding. The head position information can be used as robust information to guide the detector to estimate accurate object bounding boxes. (c) Inspired by the human search pattern of humans, the context information around the object can be introduced to guide the network to extract more discriminative features.

be primarily occluded, undoubtedly resulting in the inability to accurately detect each object's bounding box. We have observed that the head region is visible clearly, and the head bounding box regression has few errors caused by the object body occluded. We believe that the head region can be used as robust contextual information to guide the detector to predict the bounding box of the entire object, as shown in Fig.1(b). Second, the anchor-free based detection methods locate the object by regressing object bounding box center point and bottom-up information [14], which can effectively alleviate the above issue. Therefore, we propose a Center-Point-Pair detection branch to capture object head location information and bottom-up information to alleviate unreliable detection.

On the other hand, the occlusion between pedestrians and partial loss in the tracking scene will impact the discriminative ability of the object appearance feature, as shown in Fig.1(c). To alleviate such issues, many researchers focused on designing a more complex neural network to extract the object appearance features to improve tracking performance [15,16]. However, we observed that those models only used the historical and spatial information of the object itself for feature extracting. In contrast, the context information of the object in the scene and inter-frame was not effectively utilized. As humans, when we estimate the similarity between a person and a trajectory in a tracking scene, we would not only look at each individual, but we would also search for peculiar information in the background. For example, the human can learn the information of hidden relationships between pedestrians and surrounding specific objects as additional cues to extract more discriminative appearance features, as shown in Fig.1(c). Motivated by this pattern, we propose a Context-Aware Re-Identification branch to introduce the previous-frame and current-frame context information around the object to guide the tracker to extract object features.

In addition, we propose a new trajectory management method in this paper. Specifically, to fully use the temporal and spatial features of candidates and trajectories. First, we build a new data association model by expanding the affinity descriptor. Then, according to the affinity matrix, we apply the Non-negative Matrix Factor (NMF) algorithm to cluster candidates and trajectories, where each clustering result corresponds to a data association result. Last, we update object trajectories to complete tracking.

In summary, the main contributions of this paper can be summarized as follows:

1. A Center-Point-Pair detection branch is proposed to detect pedestrians in crowded tracking scenes. The branch simultaneously regresses the bounding box of the object head and body. The correlation between different object regions is learned and extracts more distinctive features to alleviate the unreliable detection in tracking.
2. We design a Context-Aware Re-Identification branch to extract more discriminative object features. Inspired by the human search pattern, the branch guides the tracker to exploit the context information around the object from the previous and current frames. The tracker learns the detailed feature of the object and its surroundings to improve the discriminative of object features.
3. In terms of trajectory management, we propose a new data association method that can more accurately describe the similarity between trajectories and detections by considering the historical features contained in trajectories. Besides, we use the NMF algorithm to group trajectory and detection as trajectory association results, which effectively keeps the stability of trajectories during the tracking.
4. Our tracker dramatically enhances tracking performance in high-density crowds and complex scenes. Experiments with challenging public datasets show distinct performance improvement over other state-of-the-arts offline and online tracking frameworks.

2. Related works

Given detection by the detector at each frame, the tracking framework locally associates detections frame-by-frame to generate object trajectories in general. In recent years there has been an explosion of technological progress in MOT driven primarily by object detection strategy [17,18]. Additionally, some recent works have directly used the dense detection output, before the non-maximum suppression, as the input to their tracker [19,20]. Although these methods alleviate the unreliable detection results, they still use one kind of detection information. Hence these methods cannot effectively alleviate the issue of missing detection or multiple objects in the same bounding box. Several works use other category location information to determine the coordinates of the tracking candidates [21–23]. In [24,25], researchers introduced the head detection results as an auxiliary detection result input to improve the final detection result of the tracking-by-detection framework, which can handle occlusion up to a scale. [23,26] further introduced the position results of human keypoints through the pose estimation method to generate higher quality detection results. But the above methods need to design additional detection result screening strategies to generate tracking bounding boxes. Therefore, we propose the Center-Point-Pair detection branch to introduce object head location information. The motivation is that the head region is visible easily in crowds and thus can be a piece of powerful guidance information to direct the detection branch to estimate the bounding box of the tracking object. Our detection branch differs from the above methods. First, the proposed tracker is an end-to-end tracking framework that directly introduces object head information in the inference stage to generate tracking bounding box. Therefore, we do not need to design additional bounding box generation strategies based on multiple detection results. Second, we adopt the anchor-free detection framework, which is more suitable for crowd tracking scenes.

To obtain a more effective appearance model and accurately distinguish different objects, researchers often introduce re-identification networks into the Multi-Object Tracking field [10,19,27]. These methods have shown that discriminative generic features can be trained using the deep learning module. Yin et al. in [28] designed a Siamese network that unifies object motion and affinity model to learn a compact local feature of objects. Zhu et al. in [16] proposed the Dual Matching Attention Networks (DMAN), which introduced spatial and temporal attention mechanisms to extract robust features against appearance variations and cluttered backgrounds. These methods improve the discriminative ability of extracted features by focusing on key regions in the detection images. Consequently, these works ignored the context information around the object, thereby causing inaccurate object representation. However, in the person search field, researchers introduce the query image to guide the network to learn context information around the person, improving search accuracy [29,30]. Therefore, in this paper, we exploit the context information around the object with the Context-Aware Re-Identification branch. This branch includes the Previous-Frame Guided Spatial-Attention Model and the Previous-Frame Guided Channel-Attention Model. Our re-id branch differs from the above methods. First, guided by the previous-frame image, our model can not only focus on the area of the object itself in the image but also exploit the context information around the object. Second, our tracker can learn the hidden context information according to the inter-frame dependencies by introducing the previous-frame image.

For generating object trajectory, employing the Hungarian algorithm to data association based on affinity matrix are the generic methods employed by several multiple object trackers [10,16,31,17]. In global association methods, trackers build accurate and stable trajectories with all detections for a sequence in general under tracking scenes [32,33]. In [34], researchers designed a hierarchical association framework to gradually produce longer tracklets at each level. Some works solve a global data association problem using a min-cost flow algorithm in a network flow [35,36]. In addition, some researchers try to estimate the object trajectories by optimizing the energy function [33]. Yang et al. in [37] design an online-learned CRF model and link tracklet by minimizing an energy function. Inspired by the recent advances in deep learning, several trackers use deep networks for data association [38,39]. In [40], Ma et al. employ GNN to compute the final affinity matrix to improve the quality of object trajectory. In [41], Wang et al. designed a Recurrent Tracking Unit based on Memory Network that extracts long-term information to score potential trajectories for tracking. In this paper, we propose the Similarity-Cluster Trajectory Management method that expands the affinity descriptor and use the NMF algorithm to perform data association. Our trajectory management method can be easily integrated into other tracking frameworks and improve the tracking performance with a small overhead.

The end-to-end multi-object tracking methods [22,42,43] complete the detection and feature extraction process in one model and have attracted more attention because of their efficiency and simplicity. In [22], Voigtlaender et al. proposed Track-RCNN tracker that introduced a Re-ID head on top of Mask R-CNN [44] to make the model regress the object bounding boxes and extracted object feature. Wang et al. [12] designed the JDE tracker built on top of YOLO v3 [45] to achieve a near real-time tracking rate. Zhang et al. [11] proposed FairMOT methods based on an anchor-free detection framework. However, these methods ignored the detection and feature extraction issues in the MOT that affects the tracking results. Unlike the previous trackers, we deeply analyse the reasons behind the issues and propose our end-to-end tracking method, which achieves a good performance in multi-object tracking.

3. Proposed method

The overall framework of our tracking method is depicted in Fig. 2. Different from the typical tracking-by-detection framework, our method is mainly divided into two steps. First, we obtain pedestrian detection results and object appearance embedding simultaneously with the proposed multi-task network. Then, we apply the designed Nonnegative Matrix Factor-based trajectory association method to manage trajectory. In this section, we introduce our precise tracking method in three components. First, a Center-Point-Pair detection branch is applied to work out more reliable detection results in Section 3.1. Then the Context-Aware Re-Identification branch for extracting more robust object features is shown in Section 3.2. Finally, the specific process of Nonnegative Matrix Factor-based trajectory association to improve the rationality and stability for tracking is elaborated in Section 3.3.

We adopt ResNet-50 [46] as the backbone in order to extract valuable features. We apply a new Deep Layer Aggregation (DLA) to fuse multi-size features, as shown in Fig. 2. Compared with the original DLA [47], it not only has more skip connections between low-resolution features and high-resolution features but also can perform more inferences at each resolution. Moreover, we adopt deformable convolution to replace the original convolution for up-sampling in DLA that can dynamically adjust the receptive field of the network to excavate more hidden features. The input of the multi-task network is a tracking image of $3 * H * W$ size, and the output of the backbone network is a feature map of $C * H_f * W_f$ size. Then we predict bounding boxes of pedestrians and generate the object Re-id feature with different branches of the multi-task network.

3.1. The center-point-pair detection branch

3.1.1. Detection branch architecture

The architecture of our designed Center-Point-Pair Detection Branch is shown in Fig. 3. The detection branch includes three sub-branches, which are the body detection sub-branch, the head sub-branch and the center-point-pair detection sub-branch. Different from other detection branches in MOT, our method predicts two bounding boxes for each tracking object in the tracking frame which specify object head and full-body bounding box, respectively. Meanwhile, the three parallel heads are appended to each detection branch to estimate heatmaps, object center offsets and bounding box sizes, respectively, to generate the bounding box.

When detecting objects in crowd scenes, the humans tend to determine the number and location of person heads first, which are the most distinct region and can be easily observed. After that, one can locate objects based on the hidden connections of the object's head and body. Therefore, we design the detection branch to extract the object's head, and body features simultaneously to learn the same behavioural patterns as humans. The detection branch is trained using three sub-branches. The body detection sub-branch is used to learn to detect object body and judge whether there is a pedestrian. The head detection sub-branch is used to learn to detection object head and help the network predict the number of objects. The center-point-pair sub-branch is used to learn the hidden connection between the object body and head. Last, we only use center-point-pair branch to estimate object body and head bounding box simultaneously to generate the final detection results.Thereby accurately regressing the bounding box of each pedestrian in the crowd area.

Specifically, the Center-Point-Pair detection branch takes backbone network feature F_D as input. The backbone feature map F_D will first be fed into two different convolution layers to extract primary head feature PF_H and primary body feature PF_B, which are

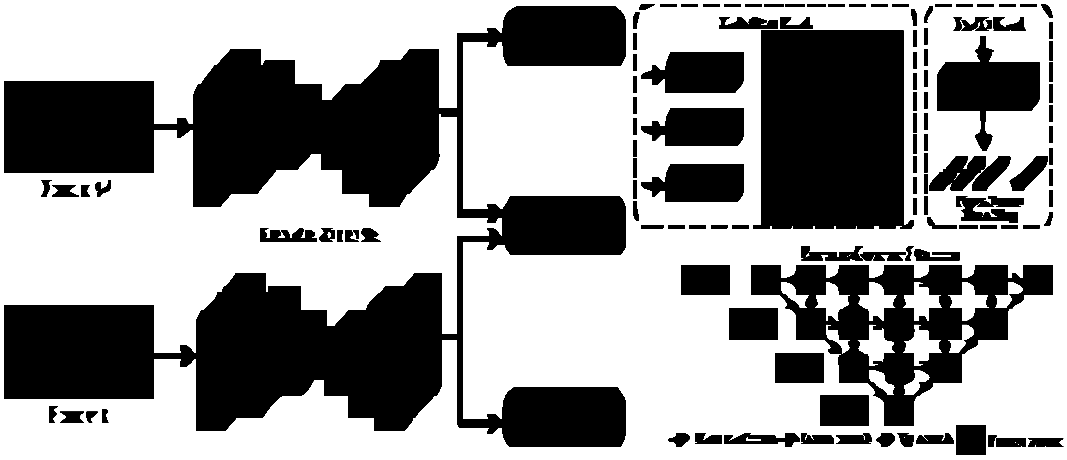


Fig. 2. Overview of our proposed approach for multi-object tracking. The tracker takes the tracking frame as input, and fed it to the encoder-decoder network to extract the backbone feature. Then we apply the multi-task network structure to perform object detection and object feature extraction, respectively. The anchor-free based detection head introduces the object head position information to simultaneously predict the object head bounding box and body bounding box. The Re-Identification head takes the current frame feature and previous frame feature as input and guides the tracker to extract the context information around the object to generate more discriminative features.

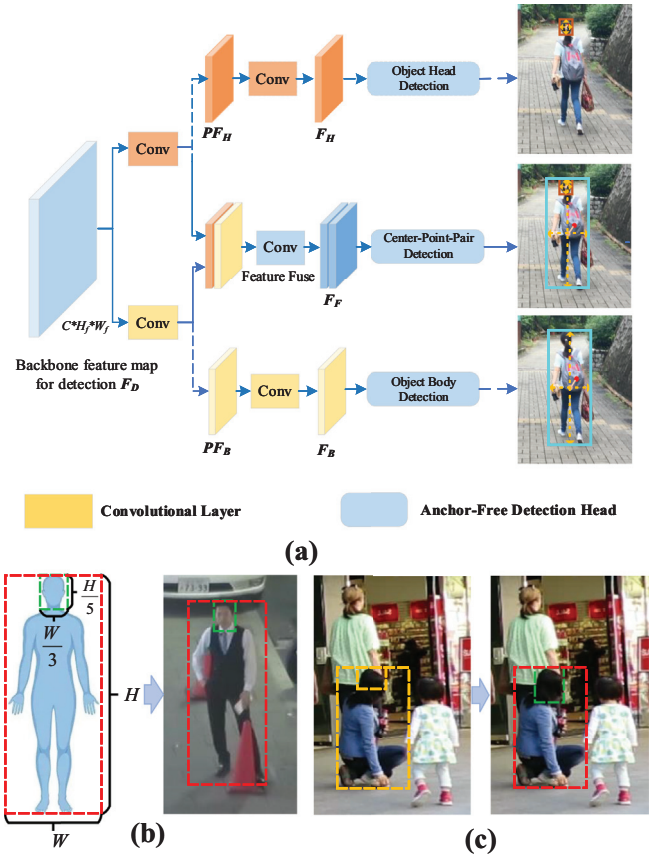


Fig. 3. (a) The structure of our designed Center-Point-Pair Detection branch. which is consisted of object head detection sub-branch, object body detection sub-branch and center-point-pair detection sub-branch. The detection branch will simultaneously localize the head region and the full-body to generate object bounding boxes. Detection sub-branches with dotted line only work during training. (b) The object head bounding box, which is generated based on the prior research statistics. (c) We manually label the heavily offset object head bounding box.

specifically extracted to characterize the head and body of an object. After that, the obtained primary feature PF_H and PF_B will be further propagated forward to generate the head feature map F_H and body feature map F_B to predict the bounding box of the head and body. In addition, we will fuse the two primary feature maps with a feature fusion module. Finally, based on fused feature map F_F, the center-point-pair sub-branch will regress the center-point-pair to predict two bounding boxes representing the regions of the object's head and body.

3.1.2. Center-point-pair detection

Unlike other anchor free based detection branches, our detection branch will regress two types of bounding boxes, one for the object head and the other for the object's full-body. We use two different center points, different bounding box sizes and offsets to regress the head and body region. However, how to associate the detection results of two types of bounding boxes from an object individual is still a problem. We propose a feature fusion module to make the network learn the correlation between the detections of the head and body. Meanwhile, we embed this correlation directly into the center-point-pair, so that the detection branch can use the correlation to improve detection results.

First, in the heatmap head, we estimate the position of the object's head and body centers. In particular, the dimension of the heatmap is $2 * H * W$ since two types center points need to be estimated, object body center and object head center. When a location in the heatmap overlaps with the ground-truth object center point, the response at that location is expected to be one. In this way, two types of object center points can be predicted, denoted as N centre-point pairs $\left\{\left(C_H^i, C_B^i\right)\right\}_{i=1\cdots N}$, where $C_H^i = \left(C_{H_x}^i, C_{H_y}^i\right)$ specifies the center point coordinates of i-th object head bounding box, and $C_B^i = \left(C_{B_x}^i, C_{B_y}^i\right)$ specifies the center point coordinates of i-th object body bounding box.

In the training process, for each ground truth bounding box $Gb_i = (x_1^i, x_2^i, y_1^i, y_2^i)$, its center-point can be represented as $\left(C_x^i, C_y^i\right)$, where $C_x^i = \frac{x_1^i + x_2^i}{2}$ and $C_y^i = \frac{y_1^i + y_2^i}{2}$, respectively. Accordingly, the center point location in the feature map can be computed by $\left(\widetilde{C}_x^i, \widetilde{C}_y^i\right) = \left(\frac{C_x^i}{n}, \frac{C_y^i}{n}\right)$, where n is stride. In the heatmap, the response at the point (x, y) is obtained as $M_{xy} = \sum_{i=1}^{N} \exp \frac{-\left(x-\widetilde{C}_x^i\right)^2 + \left(y-\widetilde{C}_y^i\right)^2}{2\sigma_c^2}$, where N is the number of tracked objects in the tracking frame and σ_c is the standard deviation. Therefore, we define the heatmap ground truth of center-point-pair sub-branch as a heatmap-pair

(M_B, M_H) where M_B is object body detection heatmap and M_H is object head detection heatmap to train the proposed detection branch to generate object center-point pair.

Then, the network needs to regress the offset and bounding box size to generate object bounding boxes. For the offset head, since there are multiple down-sampling and up-sampling operations in the backbone network, the center-point of the object will be offset. Therefore, the offset head needs to predict center offset value to alleviate the impact of down-sampling and up-sampling in the backbone network to make object location more precise. The offset ground truth can be generated by $O^i = \left(\frac{C_x^i}{n}, \frac{C_y^i}{n}\right) - \left(\left\lfloor\frac{C_x^i}{n}\right\rfloor, \left\lfloor\frac{C_y^i}{n}\right\rfloor\right)$. The box size head is used to estimated the width and height of the object bounding boxes. We compute the ground truth of bounding box size by $S^i = (x_2^i - x_1^i, y_2^i - y_1^i)$.

3.1.3. Detection loss

We employ the original anchor free based detection losses to train the object head detection sub-branch, and object body detection sub-branch, including (1) heat map prediction loss L_{heat} that encourages correct heat map. (2) bounding box prediction loss L_{bbox} that ensures correct estimated bounding box size from feature map. For training data, we use MOT 17 training data to train the body detection sub-branch and apply generated head detection training dataset to train the head detection sub-branch. We generated the preliminary head bounding boxes based on the prior research statistics. Specifically, the height of the head bounding box is one-fifth of the height of the body bounding box, and the width of the head bounding box is one-third of the width of the body bounding box [48–50]. Afterward, we manually screen the preliminary head bounding boxes. Last, we manually label the heavily offset head bounding box to generate the object head's final ground truth bounding box, as shown in Fig.3(b) and Fig.3 (c). Concrete definitions of these losses are provided below:

$$L_{heat} = -\frac{1}{N}\sum_{xy}\begin{cases}\left(1-\widehat{M}_{xy}\right)^{\alpha} log\left(\widehat{M}_{xy}\right) & \widehat{M}_{xy}=1\\ \left(1-M_{xy}\right)^{\beta}\left(\widehat{M}_{xy}\right)^{\alpha} log\left(1-\widehat{M}_{xy}\right) & otherwise\end{cases} \tag{1}$$

$$L_O = \sum_{i=1}^{N}\left\|O^i - \widehat{O}^i\right\|_1 \tag{2}$$

$$L_S = \sum_{i=1}^{N}\left\|S^i - \widehat{S}^i\right\|_1 \tag{3}$$

$$L_{body} = L_{heat}^b + L_O^b + L_S^b \tag{4}$$

$$L_{head} = L_{heat}^h + L_O^h + L_S^h \tag{5}$$

In the above equations, $\widehat{M}$ is the estimated heatmap, α, β are the predetermined parameters, $\widehat{O}$ is the generated offset and $\widehat{S}$ is predicted bounding box size. L_{body} and L_{head} is the loss function for object body detection and object head detection respectively.

Consistent with the denotation of center-point-pair, the ground truth of our task should also be a set of heat map pairs (M_B, M_H), offset pairs $\left\{\left(O_B^i, O_H^i\right)\right\}_{i=1\cdots N}$ and bounding box size pairs $\left\{\left(S_B^i, S_H^i\right)\right\}_{i=1\cdots N}$, which are generated by the corresponding combination of the MOT17 training data and head detection training data. In center-point-pair detection sub-branch, the heat map pair $\left(\widehat{M}_B, \widehat{M}_H\right)$, the offset pair $\left(\widehat{O}_B, \widehat{O}_H\right)$ and the size pair $\left(\widehat{S}_B, \widehat{S}_H\right)$ can be generated by the corresponding combination of the MOT17 training data and head detection training data. So, we define its loss function as

$$\begin{aligned} L_{cpp} = \; & L_{heat}\left((M_B, M_H), \left(\widehat{M}_B, \widehat{M}_H\right)\right) \\ & +\lambda_B f\left(x_{ij}\right)\left(L_O\left(O_{Bj}, \widehat{O}_{Bi}\right) + L_S\left(S_{Bj}, \widehat{S}_{Bi}\right)\right) \\ & +\lambda_H f\left(x_{ij}\right)\left(L_O\left(O_{Hj}, \widehat{O}_{Hi}\right) + L_S\left(S_{Hj}, \widehat{S}_{Hi}\right)\right) \end{aligned} \tag{6}$$

where $f(x_{ij}) = \{0, 1\}$ is an indicator for matching the i-th center-point-pair to the j-th ground truth of a tracking object instance, λ_B and λ_H are set to 1 as hyper-parameters to balance the three losses.

3.2. Context-aware re-identification branch

In the end-to-end multi-object tracking network, the Re-Identification branch is used to extract the object appearance features. The robust and discriminative appearance feature is crucial in the tracking process to accurately generate the object tracklet. Therefore, we introduced the context information around the object according to the pattern of the human and proposed the Context-Aware Re-Identification branch. The architecture of the Context-Aware Re-Identification branch is shown in Fig. 4., including the Previous-Frame Guided Spatial-Attention Model and the Previous-Frame Guided Channel-Attention Model. We apply a convolution layer with 256 kernels at the end of the network to generate the resulting feature map as $F \in R^{256 * H_f * W_f}$. The Re-ID feature $F_{x,y} \in R^{256}$ of object centered at (x, y) can be extracted from the feature map. Note that the object center point is optimized by offset.

3.2.1. Previous-frame guided spatial-attention model

The Previous-Frame Guided Spatial-Attention Model aims to guide the network to learn the context information around the object through the spatial attention mechanism and the previous tracking frame. Specifically, after getting the current-frame backbone feature map, we use a series of $(1 * 1 * c)$ convolution layers to generate the current-frame feature map F_t and current-frame value feature map F_t^v, respectively. We perform the same operation on the previous-frame backbone feature map to obtain the previous-frame feature map F_{t-1} and previous-frame value fea-

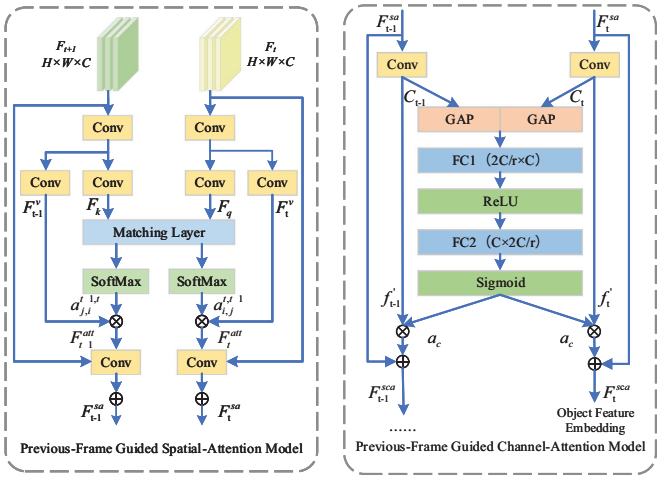


Fig. 4. Our proposed Context-Aware Re-Identification branch. It takes the previous-frame feature and the current-frame feature as input. Then it adopts the Previous-Frame Guided Spatial-Attention Model and Previous-Frame Guided Channel-Attention Model to guide network learning context features around the object according to intra-frame and inter-frame dependencies.

ture map F_{t-1}^{v}. Then, we transform the F_t and F_{t-1} into two feature spaces through a hidden layer. We regard the mapped current-frame feature map F_t as query feature map F_q, and the mapped previous-frame feature map F_{t-1} as key feature map F_k to calculate the spatial attention map of the current frame, as the following formula:

$$a_{i,j}^{t,t-1} = \frac{exp(s_{i,j})}{\sum_{i=1}^{N} exp(s_{i,j})}, s_{i,j} = F_q(x_i)^T F_k(x_j) \tag{7}$$

where $a_{i,j}^{t,t-1}$ is the attention value of the j-th position of key feature map for the i-th position of query feature map, $s_{i,j}$ indicates the similarity matrix. Then we multiply $a_{i,j}^{t,t-1}$ with current frame value feature map F_t^v to get the spatial-attention masked current frame feature map F_t^{att} that weight by the spatial-attention map, where:

$$F_t^{att} = \sum_{i=1}^{N} a_{i,j}^{t,t-1} F_t^v \tag{8}$$

Meanwhile, to prevent the fluctuation of spatial attention training in the early training stage from affecting the network performance. We add the feature map F_t^{att} and F_t. Therefore the final spatial-attention feature map of current frame F_t^{sa} is given by:

$$F_t^{sa} = \theta F_t^{att} + F_t \tag{9}$$

where θ is a learnable scalar. For the previous-frame feature map F_{t-1}, we transpose the similarity matrix in Eq. 7 to calculate new attention map $a_{j,i}^{t-1,t}$. Then we perform the same operation to generate the final spatial-attention feature map of the previous frame F_{t-1}^{sa}.

3.2.2. Previous-frame guided channel-attention model

Many kinds of physical information are included in the channel of the feature map, and some unique information is beneficial to improve the distinguishability of object features. Therefore, after obtaining the previous-frame guided spatial-attention feature map, we propose the Previous-frame Guided Channel-Attention Model to enhance the corresponding channel information between two frames, which structure is shown in Fig. 4(b). The Previous-Frame Guided Channel-Attention Model is inspired by the SE-Net [51]. The main difference is that our model uses both the previous and current frames to calculate a weight vector and re-weight the feature map per channel.

In the Previous-Frame Guided Channel-Attention Model, the two operations are mainly performed on the channel of the feature map, namely, squeeze and excitation. In squeeze operation, we employ global average pooling to condense the channel-descriptors C_t and C_{t-1} of the current frame and previous frame, respectively.

In excitation operation, we concatenate the obtained channel descriptors C_t and C_{t-1} to generate two-frame channel descriptors $[C_t, C_{t-1}] \in R^{2C}$ with a non-linear bottleneck. We use the first fully-connected layer FC_1 to obtain the $2C/r$ channel descriptors, where r is multiple of dimensionality reduction. Then, we employ *ReLU* function and the second fully-connected layer FC_2 to re-expands the channel descriptors to C. Last, the channel attention map is calculated by *Sigmoid* activation being as follows:

$$a_c = \phi(f_2(\delta(f_1([C_t, C_{t-1}])))) \tag{10}$$

whereby, f_1 indicates the first fully-connected layer for dimensionality-reduction, f_2 is the second fully-connected layer for dimensionality-expansion and ϕ is the *Sigmoid* activation. We refer to δ as the *ReLU* function to model nonlinear connection between channels. The outputs of Previous-Frame Guided Channel-Attention Model feature map F_t^{sca} and F_{t-1}^{sca} for the respective current frame and previous frame are:

$$F_t^{sca} = \lambda a_c \otimes f_t' + F_t^{sa} \tag{11}$$

$$F_{t-1}^{sca} = \lambda a_c \otimes f_{t-1}' + F_{t-1}^{sa} \tag{12}$$

where f_t' and f_{t-1}' are the residual outputs of feature map F_t^{sa} and F_{t-1}^{sa} respectively, $\otimes$ is channel-wise multiplication, λ is the weight parameter to balance the two types of attention feature map. In this way, the proposed Previous-Frame Guided Channel-Attention Model re-calibrates channel weights to take into account inter-frame channel similarities and dependencies in multi-object tracking.

3.2.3. Re-identification loss

The objective feature extraction branch can be formulated as a classification task or a verification task for the training process. Ideally, we want to extract the discriminative feature to the object while increasing the distance between different object features. Therefore, we apply a convolution layer on feature map F_t^{sca} to extract appearance features for each object. We denote the resulting feature map as $F_t^{reid} \in R^{256 * H * W}$, and the object Re-ID feature $F_{object}^{x,y} \in R^{256}$ of an object-centred at *(x, y)* can be extracted.

To learn powerful discriminative features for object appearance representations in the classification task, we first use identity loss to classify different object identities with different features. Specifically, we utilize the ground-truth identity to generate the one-hot label $GL_i(m)$. Then we compute the identity loss as:

$$L_{id} = -\sum_{i=1}^{N}\sum_{m=1}^{M} GL_i(M) log(p(m)) \tag{13}$$

Where M is the number of classes, $p(m)$ is the label distribution vector mapped from extracted Re-ID feature F_{reid}. Meanwhile, we employ contrastive loss to increase the distance of different objects to optimise the verification task. Since the same object appears only once in a frame, the contrastive loss is given by:

$$L_{con} = \left(\rho - d_{f_i f_j}\right) \tag{14}$$

Where $d_{f_i f_j}$ is the distance between two different object features, we utilize both identity loss and contrastive loss to combine these two complementary tasks. Thus, the total loss for the Re-ID branch can be calculated as the sum of these two kinds of losses:

$$L_{reid} = \lambda_i L_{id} + \lambda_c L_{con} \tag{15}$$

Where λ_i and λ_c are the coefficients of each term individually.

3.3. Similarity-cluster trajectory management method

In object tracking, we generate the bounding box of the object based on the estimated result by the proposed network. Meanwhile, we also get the object feature embeddings at the estimated object centers. Then, we calculate the affinity score to associate detections and trajectories. However, some objects with occlusion are significantly different from others, so affinity scores between those detections and trajectories cause association errors. Therefore, we propose Similarity-Cluster Trajectory Management Method to associate detections and trajectories.

We design new affinity descriptors to describe the similarity between trajectories and detections more comprehensively. It is assumed that M trajectories and N detections need to be associated. As shown in Fig. 5, we divide the trajectory into k sub-trajectories in time order and measure the affinity scores between

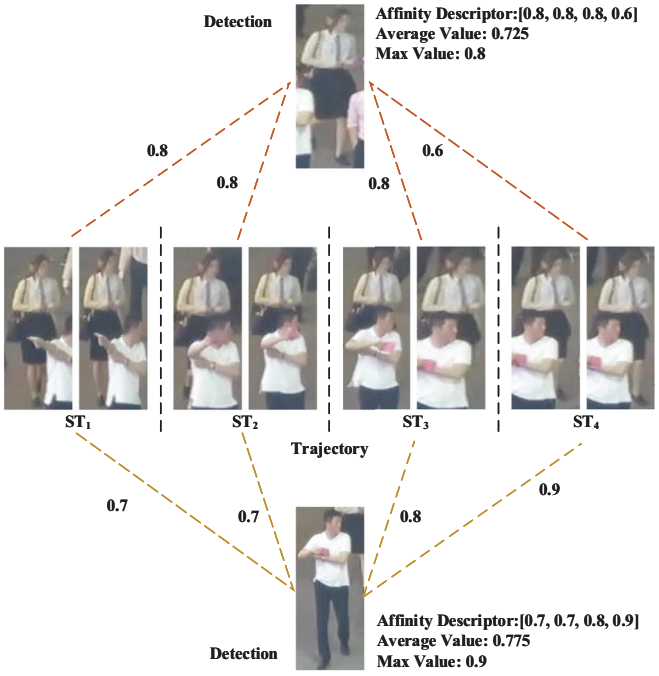


Fig. 5. The illustration of how to build the new affinity descriptor and the rationality of employing the minority obeying the majority principle to associate trajectories and detections.

the sub-trajectories and detections, respectively, to obtain a new k-dimensional affinity descriptor.

Moreover, we adopt the minority obeying the majority principle to decide whether the trajectory and detection match, as shown in Fig. 5. Specifically, after obtaining the k-dimensional affinity descriptors S_1 and S_2 between trajectories and detection. Taking the maximum similarity value in affinity descriptors or the average similarity value of affinity descriptors to determine whether the trajectory and detection are associated will be disturbed by the noise, resulting in the issue of ID switch. However, the minority obeying the majority principle can effectively alleviate this issue to generate correct trajectories.

We employ the similarity-cluster algorithm to implement the above trajectory management idea. Supposing there are *2* trajectories and *3* detections that need to be associated. We divide the trajectory into *2* sub-trajectories. First, we adopt Mahalanobis distance to compute its original affinity matrix, as shown in Fig. 6(a). Further, we use Kalman Filter to predict the bounding box of trajectories in the next frame, since trajectories and candidate detections cannot be associated with themselves. Therefore, the final affinity matrix as shown in Fig. 6(b). Then, we adopt non-negative matrix factorization for clustering to generate association results that conform the minority obeying the majority. The objective optimization function is defined in Eq. 16:

$$MinJ = \left\| W - HH^t \right\|^2, (H \geqslant 0) \tag{16}$$

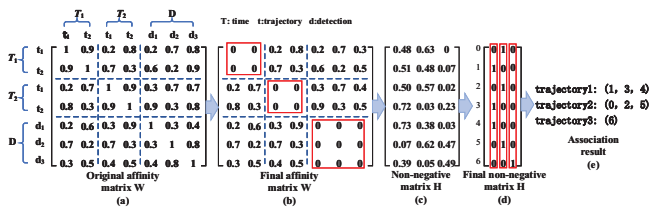


Fig. 6. The process of data association utilizing the similarity clustering algorithm.

Where W is the affinity score matrix, H is a non-negative matrix, and H^t is the transpose matrix of H. When the value of the optimization function 16 is less than the threshold T_{cost} or reaches the maximum limit of iterations, we output H as the decomposition result, as shown in Fig. 6(c). The size of H is $(kM + N) * d$, and d represents the number of categories. For each row in H, we set the maximum value position to *1* and the rest position is set to *0* to obtain the final non-negative matrix H, as shown in Fig. 6(d). Then, we take out the index that the value is *1* in each column of H and add it to the corresponding class to generate the clustering result. Last, in the clustering result, each category is the association result between the trajectories and detections, as shown in Fig. 6(e).

For unmatched detections and trajectories, we try to match them according to the *IoU* between predicting and detecting bounding boxes, with the threshold T_{IoU}. Last, unassociated trajectories are terminated if they are not associated for T_{term} frames. Meanwhile, we initialize the trajectory for unassociated detections, which are not associated with any trajectory in any of the first T_{init} frames.

4. Experiments

In our evaluation, we focus on tracking humans due to its importance. We first verify the effectiveness of the proposed models by applying them for ablative studies. Then, we analyze the performance improvement in MOT by the proposed tracking framework in detail.

4.1. Implementation details

We adopt ResNet-50 [46] as our backbone network and employ the proposed variant of DLA to fuse multi-layer features. The model parameters pre-trained on the COCO dataset are used to initialize our model. Training is performed on two NVIDIA RTX *2080* Ti GPUs. Hyper-parameters of the proposed tracker are optimized with the help of the MOT17 dataset [52], for which we specify a validation set to tune our model. Specifically, we utilize the Cam-02, Cam-04, Cam-05 and Cam-09 in the MOT 17 training data as the training set. The Cam-10, Cam-11 and Cam-13 are used as the validation set for the preliminary optimization of the tracker hyper-parameters. Our network has an input frame size of *1920*1080*, and the feature map resolution is *480*270*. We train our tracker with the Adam optimizer for *100* epochs. We start the learning process with e^{-4} as the learning rate, which is decreased to *1/10th* of the previous value at epochs *50* and *80*. The batch size is set to *6*. We use standard data augmentation techniques, including rotation, scaling and color jittering. Furthermore, we set the threshold $T_{IoU} = 0.7$ for associating the unmatched detection and unmatched trajectories. For trajectory management, we set threshold $T_{init} = 3$ for trajectory initialization and $T_{term} = 10$ for trajectory termination.

4.2. Datasets and metrics

There are four training datasets introduced as follows: We mainly use the MOT17 dataset [52] for tracker training. Moreover, the CalTech [53], CUHK-SYSC [54] and PRW [55] datasets provide both bounding box and identity that allows us to train our tracker. We utilize four datasets for training to enable the proposed tracker to achieve the best performance. However, to fairly compare with different methods in other experiments, we only use the MOT17 dataset to train the proposed tracking framework. We extensively evaluate a variety of factors of our approach on the test sets of three benchmarks: MOT15, MOT16 and MOT17.

In order to measure the performance of tracking results, we adopt the common CLEAR MOT [56] consisting of multiple metrics, which includes Multiple Object Tracking Accuracy (MOTA), ID F1 score (IDF1, the ratio of correct detections over the average number of ground-truth and computed detections), MT (Mostly Tracked objects, the percentage of ground-truth trajectories covered by a track hypothesis for *80%* of their life or more), ML (Mostly Lost objects, the percentage of ground-truth trajectories covered by a track hypothesis for *20%* of their life or less), the number of False Negatives (FN), the number of False Positives (FP), the number of ID Switches (IDS), the number of trajectory fragmentations (Frag). Note that MOTA and IDF1 are considered to be the most important evaluation metrics.

4.3. Ablation studies

In this section, we conduct a number of ablation experiments to study the contribution of each module in our tracker, including Center-Point-Pair Detection branch, Context-Aware Re-Identification branch and Similarity-Cluster Trajectory Management Method. In addition, we set up the experiment to illustrate the impact of the training datasets on tracker performance.

4.3.1. Detection in MOT

In order to verify the effectiveness and contribution of the designed detection branch, we evaluate five different detection results which are frequently used by previous works. The first detection results are Head Detection. We adopt the head detection sub-branch to locate the object head position and generate the object detection bounding box according to the ratio between the human head and body. The second detection results are Faster R-CNN[57], SDP[58] and Mask R-CNN [44] as the anchor based object detection results. The third detection results are Soft-Pose NMS[23], which introduce object pose information to optimize object detection result. The fourth detection results are the object bounding boxes generated by the JDE tracker [12]. The fifth detection results are based on the anchor-free detection branch in FairMOT [11]. The sixth detection result is center-point-pair detection strategy used in our tracker. In addition, to exclude the disturbance of other factors, we use PCB [59] to extract object feature and DeepSORT[10] tracking framework to generate object trajectory. Both Faster R-CNN and SDP directly use the detection results provided by the MOT Challenge. Other methods detection threshold have been turned to make it achieve the best tracking result.

The results are shown in Table 1. We can see that our center-point-pair detection strategy obtains higher MOTA, IDF1, and FN than other approaches. Our detection strategy improves 1.7 in MOTA, 1.5 in IDF1 with the second-best detection method and effectively reduces FN. These metrics can faithfully reflect the quality of detection results. In addition, it does not achieve the best tracking performance when only head or body detection features are used. By fusing the two detection feature and learning the hidden connection between body and head, our tracker improves almost all relevant tracking metrics, justifying our tracking framework. The results validate that center-point-pair detection branch is more suitable for tracking scenes than the strategies used in the previous works.

To evaluate the effectiveness of the auxiliary branches on the final tracking performance. We performed an ablation study of the Center-Point-Pair Detection branch, for which we designed the following three ingredients including (a) center-point-pair detection sub-branch (CPP), (b) body detection sub-branch (BDB) and (c) head detection sub-branch (HDB). We use the ResNet-50 to extract object feature and DeepSort to generate tracking result. Table 2 shows the result. We can see that Center-Point-Pair detection sub-branch improve the MOTA to 60.4, IDF1 to 61.5 and reduce FN to 41328. Although there is no body feature or head feature to help the network to refine the features of object, the guidance information from the head regions can still guided the network to learn the correlation between object head and object body. We also see that introducing body feature or head feature can improve the detection effect of the detection branch and improve the tracking performance. Finally, when both body detection sub-branch and head detection sub-branch are applied, the MOTA improve to 64.7, IDF1 improve to 63.9 and FN reduce to 32336. This result suggests that, in the proposed detection branch, the head feature and body feature information can guide the center-point-pair detection sub-branch to refine the full-body feature, thereby accurately estimating the bounding box of the object.

4.3.2. Re-identification in MOT

To demonstrate the contribution of the proposed Context-Aware Re-Identification branch in our tracker, we compare pedestrian feature representations learned by our tracker with PCB [59], Strong Baseline [60], JDE [12] and Fair MOT [11]. PCB and Strong Baseline are commonly used person Re-ID networks. JDE and FairMOT are end-to-end multiple object tracking methods. SpatialAttention and ChannelAttention indicate that only spatial and channel attention mechanisms are used in our proposed re-identification sub-branch, respectively. Note that the rest of the factors of these approaches are all controlled to be the same for a fair comparison. We use SDP [58] detection results, provided by MOTChallenge officially, to locate object bounding boxes and use DeepSORT [10] tracking framework to generate object trajectories.

The results are shown in Table 3. By comparing the results of Strong Baseline and FairMOT, we surprisingly find that specialized

Table 1
Evaluation results on MOT16 dataset with different detection method. The arrow each metric indicates that the higher (↑) or lower (↓) value is better. The best result for each indicator is bolded.

Detector	MOTA↑	IDF1↑	FP↓	FN↓
Head Detection + PCB	32.7	39.2	14893	50368
Faster R-CNN + PCB	40.2	52.6	11426	51234
SDP + PCB	60.7	62.4	**3417**	38041
Mask R-CNN + PCB	62.6	64.1	4867	39478
Soft-Pose-NMS + PCB	64.3	65.9	5115	35732
JDE + PCB	65.8	66.1	8649	48369
FairMOT + PCB	67.2	67.8	13158	31059
Ours + PCB	**68.9**	**69.3**	14683	**29548**

Table 2
The effect of different auxiliary detection sub-branches on tracking results.

CPP	BDB	HDB	MOTA↑	IDF1↑	FP↓	FN↓
	✓		58.9	59.2	**9328**	46859
✓			60.4	61.5	13148	41328
✓	✓		63.1	62.8	13487	39534
✓		✓	62.9	62.1	13257	38317
✓	✓	✓	**64.7**	**63.9**	14537	**32336**

Table 3
Evaluation results on MOT16 dataset with different feature representations.

Method	MOTA↑	IDF1↑	IDs↓
SDP + PCB	62.1	60.9	1061
SDP + StrongBaseline	63.3	63.5	837
SDP + JDE	63.9	64.8	785
SDP + FairMOT	64.6	65.7	768
SDP + SpatialAttention	65.9	66.4	729
SDP + ChannelAttention	65.2	65.9	743
SDP + Ours	**66.4**	**67.8**	**706**

person feature extraction networks do not perform well in multiple object tracking. For example, MOTA decreases by 1.3, IDF1 decreases by 2.2, and the number of IDs increases from 768 to 837, respectively. In addition, both spatial and channel attention mechanisms can improve the discriminative of extract appearance features compared to FairMOT. All these results suggest that using a specialized person feature extraction network cannot effectively improve the tracking performance. In contrast, the Context-Aware Re-Identification branch, which introduces the context information around objects, achieves a higher MOTA score and IDF1 score than other methods. More importantly, IDs decrease significantly from 768 to 706, suggesting that the context information around the person has clear advantages for multiple object tracking.

To validate the influence of the context information around the object on the final tracking results. We performed an analysis of the Context-Aware Re-Identification branch, for which we designed the following three variants. (a) Fair MOT is our baseline. (b) Previous-Frame based Context-Aware Re-Identification branch utilizes the previous frame image to guide the network to learn context information around the object. (c) Next-Frame based Context-Aware Re-Identification branch uses the next frame image to guide the network to extract object context information. We can see from Table 4 that the Next-Frame based Context-Aware Re-Identification branch has improved because it takes full advantage of all the context information around the object to extract more robust object features. However, adopting the next frame image for guidance does not conform to the typical pedestrian search pattern. Therefore, the Previous-Frame based Context-Aware Re-Identification branch achieves the best tracking performance. It means context information around the object helps improve the discriminative ability of the object feature.

4.3.3. Data association in MOT

This section evaluates the two trajectory management methods and the three ingredients in the data association step, including bounding box *IoU*, object appearance feature and *Kalman* Filter. We use the Similarity-Cluster Algorithm, and Hungarian Algorithm [61] to associate trajectories and detection, respectively. The above three ingredients are employed to calculate the similarity between detection and trajectories. Table 5 shows the results. We can see that our method achieves the best tracking results. This fully shows that the multi-dimensional affinity descriptors and the minority obeying the majority principle can effectively alleviate the interference of the noise in trajectory.

We also see that only applying bounding box *IoU* or object appearance feature causes decreases in tracking performance. This is because the occlusion between objects is frequent in crowd tracking scenes, and a single ingredient cannot accurately describe the similarity between trajectories and detections. Furthermore, the *Kalman* filter can obtain smooth trajectories, which effectively reduces the number of IDs. Therefore, it is essential to employ bounding box *IoU*, object appearance feature, *Kalman* filter and Similarity-Cluster algorithm to obtain good trajectory management performance.

4.3.4. Training data in MOT

We aim to study the impact of training data on the proposed end-to-end tracker. To this end, we compare the tracking performance of the tracker trained on different datasets, such as MOT17 [52], CUHK-SYS C[54], PRW [55] and CalTech [53]. We conduct experiments on the MOT17 test dataset. CUHK-SYSC and PRW are common person search datasets. CalTech is a commonly used dataset in pedestrian detection research. "MIX" represents the large-scale dataset generated by mixing the abovementioned datasets. These datasets provide object bounding boxes and identity. The experimental results are shown in Table 6. First, the tracker training on the "MIX" dataset outperforms other trackers. Second, due to the small number of persons and the simple scene, the tracking performance of tracker training on the PRW dataset decreases. On the contrary, the CalTech dataset enables the proposed model to fully train and improves the tracking performance with its vast data volume. The results validate that we can improve the tracking ability of our tracker by augmenting the training data and making it more competitive in real applications.

4.4. Experiment results On MOTChallenge

We compare our tracking framework to state-of-the-art (SOTA) methods, including both online and offline tracking methods.

Some published works of JDE [12], Track-RCNN [22] and FairMOT [11] jointly perform object detection and object feature embedding. We compare our tracker with both of them. Following the previous work, the testing dataset contains six videos from MOT15. We use the same data for training these trackers. In particular, since Track R-CNN requires segmentation labels to train the network, it only uses the four videos from MOT17 dataset, which

Table 4
The effect of different context information on tracking results.

Method	MOTA↑	IDF1↑	IDs↓
Baseline	65.2	66.3	782
Previous-Frame	**67.3**	**68.5**	**709**
Next-Frame	66.5	67.1	728

Table 6
The effects of different training dataset on the tracking performance.

Dataset	MOTA↑	IDF1↑	IDs↑
MOT17 [52]	67.9	70.1	593
CUHK-SYSC [54]	60.3	62.5	2692
CalTech [53]	64.6	66.8	1862
MOT17 + CUHK-SYSC	68.1	70.6	506
MOT17 + CalTech	69.5	72.3	489
MIX	**71.3**	**74.9**	**392**

Table 5
Comparison of different data association methods and different ingredients on the MOT16 dataset.

Method	Box IoU	Object Feature	Kalman Filter	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓
Hungarian Algorithm[61]	✓			60.3	61.8	28.60%	23.90%	2948
		✓		60.9	62.1	29.80%	22.10%	1982
		✓	✓	61.5	63.2	31.60%	20.50%	1359
	✓	✓	✓	62.6	64.9	33.90%	19.30%	1296
Similarity-Cluster Algorithm	✓			60.5	62.1	29.30%	23.20%	2864
		✓		61.1	62.5	31.50%	21.20%	1568
		✓	✓	62.3	64.4	33.50%	19.80%	1236
	✓	✓	✓	**64.1**	**65.3**	**35.30%**	**18.50%**	**1059**

has segmentation labels as training data. In this case, we also use the same data to train our tracker and other models. The results are shown in Table 7. We can see that our tracking framework remarkably outperforms other trackers. In particular, our tracker achieves a comparable MT, FP, FN and performs favourably against the other method in terms of MOTA, IDF1, ML and IDs. The results validate the effectiveness of the center-point-pair detection branch over the previous detection method.

To comprehensively benchmark our technique for multiple object tracking, we also evaluate the proposed tracker on MOT16 and MOT17. However, due to we do not use the official detection results, the private detector protocol is adopted. The final experimental results are evaluated by MOTChallenge.

Quantitative results and comparisons with the other tracking methods are shown in Table 8 and Table 9. As shown in Table 8, our tracking method achieves a comparable ML, FP, IDs, Frag score and performs favourably against the state-of-the-art methods in terms of MOTA, IDF1, MT and FN on the MOT16 dataset. Our tracker upgrades MOTA to 69.5, IDF1 to 72.3, MT to 40.3%, and reduces FN to 40631. Meanwhile, our tracker achieves the best performance in MT and IDs among online methods, demonstrating the merits of our tracker in object matching and the stability of multi-object tracking. MOTA and FN correspond to the object detection capability. Therefore, the improvement of MOTA and FN demonstrates the merits of our center-point-pair detection strategy in object locating for MOT. Similarly, Table 9 shows that our tracker outperforms existing online trackers on more than half of the metrics and achieves the best performance in terms of MOTA, IDF1, FN, IDs and Frag among online and offline methods on the MOT17 dataset.

In terms of tracking speed, our method outperforms previous state-of-the-art methods on several benchmark datasets. However, the tracking speed of our tracker is 4.3 FPS slower than FairMOT. That is because the proposed tracker needs to mine the object head position information and the context information around the object to improve the tracking performance. We believe that the speed penalty paid compared to the improvement in tracking performance is worthwhile, which makes our method more suitable for practical applications.

In addition, we believe that due to the designed detection branch can detect these small-scale pedestrians, occluded pedestrians, and pedestrians who are not recorded as tracking objects. Therefore, our detection branch will cause the phenomenon of high FP, as shown in Table 7 and Table 8, and the similar situation exists in [11,70,31] too. However, this phenomenon also reflects that the proposed detection branch can complement missing objects and reduce unreliable detection, which is more suitable for tracking scenes.

Table 7
Comparsion of the state-of-the-art end-to-end trackers on the MOT15 dataset.

Method	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
JDE[12]	67.9	67.1	35.4%	18.6%	1881	**2083**	218
Track R-CNN[22]	69.2	49.4	42.6%	16.5%	1354	2397	294
FairMOT[11]	70.3	65.8	**47.6%**	11.0%	**1263**	2598	108
Ours	**71.5**	**67.9**	46.3%	**10.3%**	1348	2084	**93**

Table 8
Comparing our tracking framework with state-of-the-art methods on MOT16 dataset. The arrow each metric indicates that the higher (↑) or lower (↓) value is better. The best result for each indicator is bolded.

Tracker		MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓	FPS↑
Offline	NLLMPa[62]	47.6	47.3	17.0%	40.4%	5844	89093	629	768	–
	LMP[19]	48.8	51.3	18.2%	40.1%	6654	86254	**481**	**595**	–
	NT[63]	47.5	43.6	19.4%	36.9%	13002	81762	1035	1408	–
	TNT[64]	56.1	49.2	17.3%	40.3%	8400	83702	606	882	
Online	FWT[24]	48.8	51.3	18.2 %	40.1 %	6654	86245	481	1534	0.6
	EAMTT[65]	52.5	53.3	19.9%	34.9%	4407	81233	910	1321	< 5.5
	Tracktor++[22]	56.2	54.9	20.7%	35.8%	**2394**	76844	617	1068	-
	DeepSORT_V2[10]	61.4	52.2	32.8%	18.2%	5119	63352	781	2008	< 6.4
	TMOH[66]	63.2	63.5	27.0%	31.0%	3122	63376	635	1486	0.7
	CNNMTT[67]	65.2	62.2	32.4%	21.3%	6578	55896	946	2283	< 5.3
	POI[68]	66.1	65.1	34.0%	20.8%	5061	55914	805	3093	< 5.0
	Tube_TK_POI[69]	66.9	62.2	39.0%	**16.1%**	11544	47502	1236	1444	–
	Soft_Pose_MOT[23]	67.7	66.4	37.9%	18.6%	11453	42494	579	902	< 5.8
	CTracker[70]	67.6	57.2	32.9%	19.0%	8934	48305	1897	3112	6.8
	FairMOT[11]	68.7	70.4	39.5%	17.5%	13501	41653	953	2399	**25.9**
	Ours	**69.5**	**72.3**	**40.3%**	17.9%	14538	**40631**	589	2406	21.6

Table 9
Comparing our tracking framework with state-of-the-art methods on MOT17 dataset. The arrow each metric indicates that the higher (↑) or lower (↓) value is better. The best result for each indicator is bolded.

Tracker		MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓	Frag↓	FPS↑
Offline	EDMT[71]	50.0	51.3	21.6%	36.6%	22875	252889	2314	2865	–
	NOTA[72]	51.3	54.7	17.1%	35.4%	20148	252531	2285	4080	–
	TNT[64]	58.0	51.9	23.5%	35.5%	37311	231658	2294	2917	–
Online	FWT[24]	51.3	47.6	21.4%	35.2%	24101	247921	2648	4279	0.6
	Tracktor++[22]	56.3	55.1	20.1%	35.3%	**8866**	235449	1987	3763	–
	TMOH[66]	62.1	62.8	26.9%	31.4%	10951	201195	1897	4622	0.7
	POI[68]	63.0	58.6	31.2%	19.9%	27060	177483	4137	5727	< 5.0
	Soft_Pose_MOT[23]	67.3	65.9	37.9%	20.7%	20574	195176	2031	3098	< 5.8
	CTracker[70]	66.6	57.4	32.2%	24.2%	22284	160491	5529	9114	6.8
	CTTrack17[73]	67.8	64.7	34.6%	24.6%	11498	160332	3039	6102	17.0
	FairMOT[11]	68.2	70.1	**38.1%**	**20.6%**	36541	141899	3303	4349	**25.9**
	Ours	**69.8**	**71.5**	38.0%	21.5%	37932	**134803**	**1835**	**2783**	21.6

Fig. 7. Tracking example of the proposed method from MOT17. The colour of bounding boxes identifies the estimated trajectories. The sampling frame results listed in each row follow the time sequence of the tracking video in MOT17.

In summary, comparing the results of two datasets, the tracking framework proposed in this paper can effectively improve the tracking performance.

4.5. Qualitative results

Fig. 7 shows the effect diagram of the tracker proposed in this paper on the MOT 17 test set. From the tracking results of MOT17-03 and MOT17-08, it can be seen that our method performs well in scenes with crowded pedestrians. In particular, our method can generate accurate bounding boxes when occlusions between pedestrians are frequent, which benefits from center-point-pair detection strategy. From the tracking results of MOT17-01, MOT17-06 and MOT17-07, it can be seen that our method can maintain the stability of the object trajectory well. With the help of contextual information around the object, our method can extract more discriminative object features. The tracking results of MOT17-12 and MOT17-14 show that our method can deal with large-scale variations of objects and cope with the issue of camera movement during tracking. MOT17-14 is similar to the automatic driving scene, which shows that our tracker is more appealing in real applications.

5. Conclusion

We present a novel end-to-end multi-object tracking method that optimizes three main components of most existing trackers, incluluding detection, feature extraction and data association. The tracker introduces the object head location information to locate the object. Then generating more accurate object bounding boxes by the proposed center-point-pair detection branch also helps alleviate typical difficulties in tracking, such as occlusion handling and trajectory offset. Here, with the guidance of the pre-

vious frame, the network can effectively learn the context information around the object and extract more discriminative object features. In addition, our proposed similarity-cluster trajectory management method extends the affinity descriptor that can accurately and comprehensively evaluate the similarity between detections and trajectories. Meanwhile, we adopt the principle of minority obeying majority for data association, which improves the generated trajectory's quality. Through extensive experiments, we have proved that the proposed tracking framework leads to competitive performance improvement.

CRediT authorship contribution statement

Xin Zhang: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Yunan Ling:** Formal analysis, Data curation, Writing - review & editing. **Yuanzhe Yang:** Writing - review & editing. **Chengxiang Chu:** Writing - review & editing. **Zhong Zhou:** Writing - review & editing, Supervision.

Data availability

The authors do not have permission to share data.

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.

Acknowledgment

This work was supported by the National Key R&D Program of China (Grant No.2018YFB2100603) and the National Natural Science Foundation of China (Grant No.61872024). The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions.

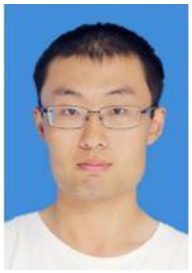
References

[1] Y. Rasekhipour, A. Khajepour, S. Chen, B. Litkouhi, A potential field-based model predictive path-planning controller for autonomous road vehicles, IEEE Trans. Intell. Transp. Syst. 18 (5) (2016) 1255–1267.
[2] Z. Li, X. Yu, P. Li, M. Hashem, Moving object tracking based on multi-independent features distribution fields with comprehensive spatial feature similarity, Visual Comput. 31 (12) (2015) 1633–1651.
[3] J. Janai, F. Güney, A. Behl, A. Geiger, et al., Computer vision for autonomous vehicles: Problems, datasets and state of the art, Foundations and Trends in Computer Graphics and Vision, vol. 12, no. 1–3, pp. 1–308, 2020.
[4] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.
[5] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, et al., A large-scale benchmark dataset for event recognition in surveillance video, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2011, pp. 3153–3160.
[6] N. Ran, L. Kong, Y. Wang, Q. Liu, A robust multi-athlete tracking algorithm by exploiting discriminant features and long-term dependencies, in: International Conference on Multimedia Modeling, Springer, 2019, pp. 411–423.
[7] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, Z. Zhang, Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2420–2440.
[8] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1428–1437.
[9] L. Zhang, M.D.L. Van, Preserving structure in model-free tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (4) (2013) 756–769.
[10] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, IEEE international conference on image processing (ICIP), IEEE 2017 (2017) 3645–3649.
[11] Y.F. Zhang, C.Y. Wang, X.G. Wang, W.J. Zeng, W.Y. Liu, Fairmot: On the fairness of detection and re-identification in multiple object tracking, Int. J. Comput. Vision 129 (11) (2021) 3069–3087.
[12] Z. Wang, L. Zheng, Y. Liu, et al., Towards real-time multi-object tracking, in: European Conference on Computer Vision, Springer, 2020, pp. 107–122.
[13] G. Ciaparrone, F.L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, F. Herrera, Deep learning in video multi-object tracking: A survey, Neurocomputing 381 (2020) 61–88.
[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, Centernet: Keypoint triplets for object detection, in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6569–6578.
[15] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Stat: spatial-temporal attention mechanism for video captioning, IEEE Trans. Multimedia 22 (1) (2019) 229–241.
[16] J. Zhu, H. Yang, Nian Liu, M. Kim, W. Zhang, and M. Yang, Online multi-object tracking with dual matching attention networks, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 366–382.
[17] W. Choi, Near-online multi-target tracking with aggregated local flow descriptor, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3029–3037.
[18] Z. Zhou, J. Xing, M. Zhang, W. Hu, Online multi-target tracking with tensor-based high-order graph matching, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 1809–1814.
[19] S. Tang, M. Andriluka, B. Andres, and B. Schiele, Multiple people tracking by lifted multicut and person re-identification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3539–3548.
[20] S. Tang, B. Andres, M. Andriluka, and B. Schiele, Subgraph decomposition for multi-target tracking, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5033–5041.
[21] K. Fragkiadaki, J. Shi, Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement, CVPR, IEEE 2011 (2011) 2073–2080.
[22] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B.B.G. Sekar, A. Geiger, and B. Leibe, Mots: Multi-object tracking and segmentation, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 7942–7951.
[23] X. Zhang, S. Wang, Y. Yang, C. Chu, Z. Zhou, Online multi-object tracking with pose-guided object location and dual self-attention network, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2021, pp. 223–235.
[24] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, Fusion of head and full-body detectors for multi-object tracking, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1428–1437.
[25] Z. Sun, J. Chen, M. Mukherjee, H. Wang, D. Zhang, An improved online multiple pedestrian tracking based on head and body detection, in: 2021 17th International Conference on Mobility, Sensing and Networking (MSN), IEEE, 2021, pp. 74–80.
[26] Y. Liu, J. Yin, D. Yu, S. Zhao, J. Shen, Multiple people tracking with articulation detection and stitching strategy, Neurocomputing 386 (2020) 18–29.
[27] X. Zhang, X. Wang, C. Gu, Online multi-object tracking with pedestrian re-identification and occlusion processing, Visual Comput. 37 (2021) 1089–1099.
[28] J. Yin, W. Wang, Q. Meng, R. Yang, and J. Shen, A unified object motion and affinity model for online multi-object tracking, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6768–6777.
[29] H. Liu, J. Feng, Jie. Z, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, Neural person search machines, in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 493–501.
[30] B. Munjal, S. Amin, F. Tombari, and F. Galasso, Query-guided end-to-end person search, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 811–820.
[31] L. Chen, H. Ai, Z. Zhuang, C. Shang, Real-time multiple people tracking with deeply learned candidate selection and person re-identification, IEEE International Conference on Multimedia and Expo (ICME) 2018 (2018) 1–6.
[32] C. Kim, F. Li, A. Ciptadi, and J. Rehg, Multiple hypothesis tracking revisited, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 4696–4704.
[33] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2013) 58–72.
[34] C. Huang, B. Wu, R. Nevatia, Robust object tracking by hierarchical association of detection responses, in: European Conference on Computer Vision, Springer, 2008, pp. 788–801.
[35] L. Zhang, Y. Li, R. Nevatia, Global data association for multi-object tracking using network flows, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
[36] H. Pirsiavash, D. Ramanan, C. Fowlkes, Globally-optimal greedy algorithms for tracking a variable number of objects, CVPR, IEEE 2011 (2011) 1201–1208.
[37] B. Yang, R. Nevatia, An online learned crf model for multi-target tracking, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2034–2041.
[38] X. Jiang, P. Li, Y. Li, and X. Zhen, Graph neural based end-to-end data association framework for online multiple-object tracking, arXiv preprint arXiv:1907.05315, 2019.

[39] G. Brasó and L. Leal-Taixé, Learning a neural solver for multiple object tracking, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6247–6257.
[40] C. Ma, Y. Li, F. Yang, Z. Zhang, Y. Zhuang, H. Jia, and X. Xie, Deep association: End-to-end graph-based learning for multiple object tracking with conv-graph neural network, in Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 253–261.
[41] S. Wang, H. Sheng, Y. Zhang, Y. Wu, and Z. Xiong, A general recurrent tracking framework without real data, in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13219–13228.
[42] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, Track to detect and segment: An online multi-object tracker, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12352–12361.
[43] S. Sun, N. Akhtar, H. Song, A. Mian, M. Shah, Deep affinity network for multiple object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2019) 104–119.
[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, Mask r-cnn, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[45] J. Redmon and A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767, 2018.
[46] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[47] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, Deep layer aggregation, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2403–2412.
[48] S. Torimitsu, Y. Makino, H. Saitoh, A. Sakuma, Namiko Ishii, Daisuke Yajima, Go Inokuchi, Ayumi Motomura, Fumiko Chiba, Rutsuko Yamaguchi, et al., Stature estimation from skull measurements using multidetector computed tomographic images: a japanese forensic sample, Legal Med. 18 (2016) 75–80.
[49] K.M. Kyllonen, T. Simmons-Ehrhardt, K.L. Monson, Stature estimation using measurements of the cranium for populations in the united states, Forensic Sci. Int. 281 (2017) 184–1e1.
[50] S. Kumar, R. Garg, K. Mogra, R. Choudhary, Prediction of stature by the measurement of head length in population of rajasthan, J. Evol. Med. Dental Sci. 2 (14) (2013) 1334–1339.
[51] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[52] M. Anton, L.T. Laura, R. Ian, R. Stefan, and S. Konrad, MOT16: A Benchmark for Multi-Object Tracking, arXiv e-prints, p. arXiv:1603.00831, Mar. 2016.
[53] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, IEEE conference on computer vision and pattern recognition, IEEE 2009 (2009) 304–311.
[54] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, Joint detection and identification feature learning for person search, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3415–3424.
[55] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, Person re-identification in the wild, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1367–1376.
[56] K. Bernardin, R. Stiefelhagen, Evaluating multiple object tracking performance: the clear mot metrics, EURASIP J. Image Video Process. 2008 (2008) 1–10.
[57] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.
[58] F. Yang, W. Choi, and Y. Lin, Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2129–2137.
[59] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 480–496.
[60] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[61] H.W. Kuhn, The hungarian method for the assignment problem, Naval Res. Logist. Quart. 2 (1–2) (1955) 83–97.
[62] M. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, et al., Joint graph decomposition & node labeling: Problem, algorithms, applications, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6012–6020.
[63] L. Wen, D. Du, S. Li, X. Bian, and S. Lyu, Learning non-uniform hypergraph for multi-object tracking, in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 8981–8988.
[64] G. Wang, Y. Wang, H. Zhang, R. Gu, and J. Hwang, Exploit the connectivity: Multi-object tracking with trackletnet, in Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 482–490.
[65] R. Sanchez-Matilla, F. Poiesi, A. Cavallaro, Online multi-target tracking with strong and weak detections, in: European Conference on Computer Vision, Springer, 2016, pp. 84–99.
[66] D. Stadler and J. Beyerer, Improving multiple pedestrian tracking by track management and occlusion handling, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10958–10967.
[67] N. Mahmoudi, S.M. Ahadi, M. Rahmati, Multi-target tracking using cnn-based features: Cnnmtt, Multimedia Tools Appl. 78 (6) (2019) 7077–7096.
[68] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, J. Yan, Poi: Multiple object tracking with high performance detection and appearance feature, in: European Conference on Computer Vision, Springer, 2016, pp. 36–42.
[69] B. Pang, Y. Li, Y. Zhang, M. Li, and C. Lu, Tubetk: Adopting tubes to track multi-object in a one-step training model, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6308–6318.
[70] P. Jinlong, W. Changan, W. Fangbin, W. Yang, W. Yabiao, T. Ying, W. Chengjie, L. Jilin, H. Feiyue, F. Yanwei, Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking, in: European Conference on Computer Vision, Springer, 2020, pp. 145–161.
[71] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, Enhancing detection model for multiple hypothesis tracking, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 18–27.
[72] C. Long, A. Haizhou, C. Rui, Z. Zijie, Aggregate tracklet appearance features for multi-object tracking, IEEE Signal Process. Lett. 26 (11) (2019) 1613–1617.
[73] X. Zhou, V. Koltun, and P. Krähenbühl, Tracking objects as points, arXiv, 2020.

Xin Zhang is a Ph.D. student at State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He received his B.S. and M.S. degree from North University of China, Taiyuan, China, in 2015 and 2018, respectively. His research interests include Multi-Object Tracking, Person Re-Identification and Computer Vision.

Yunan Ling received the B.E. degree in software engineering from Jilin University, Changchun, China, in 2021. He is currently pursuing the M.S. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. His research interests include Person Re-Identification and Deep Learning.



Yuanzhe Yang is a postgraduate, at State Key Lab of Virtual Reality Technology and Systems, Beijing University, Beijing, China. He received his B.S. degree from Jilin University, Changchun, China, in 2016. His main research interests include Computer Vision, Person Search, Person Re-Identification.

Chengxiang Chu received the B.S. degree in Software Engineering from Jilin University, Changchun, China, in 2020. He is currently pursuing the M.S. degree in electronic information with the State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. His major research interests include Vehicle Re-Identification, Object Tracking and Deep Learning.

Zhong Zhou Professor, Ph.D. adviser, State Key Lab of Virtual Reality Technology and Systems, Beihang University, Beijing, China. He got B.S. degree from Nanjing University and Ph.D. degree from Beihang University in 1999 and 2005 respectively. His main research interests include Virtual Reality/Augmented Reality/Mixed Reality, Computer Vision and Artificial Intelligence. He is member of IEEE, ACM and CCF.