Optics & Laser Technology 153 (2022) 108267

Contents lists available at ScienceDirect

Optics and Laser Technology

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

One-shot multi-object tracking using CNN-based networks with spatial-channel attention mechanism

Guofa Li [a], Xin Chen [a], Mingjun Li [b], Wenbo Li [c], Shen Li [d,*], Gang Guo [c], Huaizhi Wang [e], Hao Deng [f]

[a] Institute of Human Factors and Ergonomics, College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China
[b] State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University, Changsha 410082, Hunan, China
[c] College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China
[d] Department of Civil Engineering, Tsinghua University, Beijing 100084, China
[e] Guangdong Key Laboratory of Electromagnetic Control and Intelligent Robots, Shenzhen University, Shenzhen 518060, China
[f] Yi Xin Ke Ji (Shenzhen) Co., Ltd., Shenzhen 518060, China

ARTICLE INFO

Keywords:
Multi-object tracking
Deep learning
Joint detection and embedding
Attention mechanism
Autonomous vehicle

ABSTRACT

Deep learning algorithms for multi-object tracking have made great progress and have powered the emergence of state-of-the-art models to address multi-object tracking problems. Though a lot of efforts have been made, false detections (named "FP") and missed detections (named "FN") caused by inaccurate tracking still cannot be well addressed especially in extremely crowded driving situations. To address these problems, we develop a new online one-shot multi-object tracking system based on convolutional neural networks with spatial-channel attention mechanism. Firstly, we propose a feature combination module (FCM) that uses dilated convolution to obtain different receptive fields to adapt to the deformation of the targets, instead of introducing a large number of parameters to deal with the problem of target scale transformation like the recent feature pyramid network. Then, an attention mechanism network (AM-Net) is designed to allow the model to dynamically focus on certain parts of the input that help perform the task and ignore irrelevant information. Finally, we introduce a combination of triple loss and online instance matching loss (TOIM Loss) to distinguish similar instances within a class. Our proposed method is evaluated in three commonly used multi-object tracking datasets including 2DMOT15, MOT16 and MOT20. The results show that our proposed method is generally superior to the compared models.

1. Introduction

Autonomous driving systems can improve the safety of the entire transportation system since these systems can release drivers from the driving task by monitoring the traffic environment via sensors and making decisions by artificial intelligence algorithms with learning abilities [1,2]. The vision-based environment perception system in autonomous driving uses real-time information of the surrounding vehicles and pedestrians to provide environmental warnings and auxiliary driving instructions, thus avoiding the corresponding traffic accidents [3,4]. Moving target tracking are the prerequisite and key to environment perception in autonomous driving.

Many approaches have been developed in the literature for moving object tracking. For example, Henschel et al. [5] established an appearance representation model of pre-detected objects, which was then used to estimate similarities between objects across different frames in video sequences. Their results showed that the proposed model helped to effectively remove false positive head detections that are inconsistent with full-body detections. Lu et al. [6] combined detection and re-identification (re-ID) in a network and used greedy bipartite matching for online association. The results concluded the effectiveness of joint training over the method of training detection and re-ID models separately. In addition, the computational cost was also reduced. However, the features extracted by such models are not robust to illumination variations and intra-category occlusions [7]. Moreover, they cannot distinguish different objects with high similarities.

* Corresponding author.
E-mail addresses: hanshan198@gmail.com (G. Li), 15989492470@163.com (X. Chen), mingjunl@hnu.edu.cn (M. Li), liwenbo@cqu.edu.cn (W. Li), sli299@tsinghua.edu.cn (S. Li), guogang@cqu.edu.cn (G. Guo), wanghz@szu.edu.cn (H. Wang), hao.deng@yixinkeji-sz.com (H. Deng).

https://doi.org/10.1016/j.optlastec.2022.108267
Received 16 October 2021; Received in revised form 11 April 2022; Accepted 2 May 2022
Available online 8 May 2022

Therefore, multi-object tracking (MOT) is still commonly recognized as a challenging task.

Previous studies on MOT mainly focus on the tracking-by-detection paradigm, which often breaks the problem into two separate models: 1) a detection model to localize the targets of interest by rectangle bounding boxes in single video frames; and 2) an association model which extracts re-ID features for each detected target to determine whether the objects should be matched to the existing trajectories or not. For instance, Sheng et al. [8] presented a heterogeneous association graph, which integrated high-level detections and low-level image proofs for association. It was concluded that the heterogeneous association graph could effectively solve the detection failures due to occlusion or various poses. Wen et al. [9] proposed an algorithm based on non-uniform hypergraph, which could model different degrees of dependencies between trajectories in a unified target. The results showed that exploiting high-order dependencies between objects could improve the model efficiency when dealing with complex scenes. Since these approaches do not share features and need to apply the re-ID model for every detected target in single video frames, the corresponding studies cannot perform real-time inference and require at least two compute-intensive units. This property of the previous tracking-by-detection paradigm has brought severe challenges to building a real-time MOT system [10–12], in which real-time performance is a basic requirement.

The above-mentioned problems have been most successfully solved by joint detection and re-ID learning. Specifically, Voigtlaender et al. [13] added a re-ID branch on Mask R-CNN to obtain re-ID features, which enabled the model to simultaneously predict locations and extract appearance embedding features of interested objects. The inference time was reduced by using a weight-shared backbone network. Pang et al. [14] proposed a new algorithm which introduced the "bounding-tube" to indicate temporal-spatial locations of objects in video sequences. The results showed that the proposed method had the capability to solve occlusion problems to a certain extent. However, the tracking accuracies of the above methods were not as good as that of the two-step methods [13,14]. These results suggest that the combination of detection and re-ID should be further improved. Therefore, joint detection and embedding (JDE) [15] was proposed to explore and deliberately design the following fundamental aspects: training data, network architecture, learning objectives, optimization strategies (modeled as a multi-task learning problem [16]), and validation metrics. DarkNet-53 [17] and feature pyramid network (FPN) [18] are used as the backbone in JDE. The type of loss function with the best embedding features can be then determined through the learning process. However, since this method only forwards propagation through the DarkNet-53 network, the discriminative power of the feature map obtained through this network is not strong enough and may cause failure of tracking. Moreover, although the identified cross entropy loss can reflect the possibility of a positive prediction and avoid gradient dispersion, it does not give an explicit distance metric between the input features. In other words, it adopts an inter-class competition mechanism and is better at learning information within a class, but only cares about the accuracy of the prediction probability of the correct label without focusing on the other incorrect labels.

To alleviate these problems, we develop a method based on JDE to enhance its ability to capture detailed and multi-scale information to further improve its tracking performance. The main contributions of this study are summarized as follows:

(1) A novel framework based on JDE and hybrid dilated convolution is proposed to enhance the convolutional receptive field without generating any additional parameters for accuracy improvement in multi-object tracking.
(2) The attention mechanism together with JDE is newly developed to increase more powerful representation while not heavily damaging the efficiency. The comprehensively used attention mechanism not only identifies the focus of attention, but also improves the expression of interested targets.
(3) An effective new re-ID loss is developed to pull the feature vectors from the same person close and push the vectors from different people away from each other.

The experiment results on 2DMOT15, MOT16 and MOT20 demonstrate the advantage of our proposed method over the compared MOT methods. The remainder of this paper is organized as follows. The next section provides a review of the related work. Section 3 introduces our proposed method in detail. The experiment details and the obtained results are presented and discussed in Section 4 and Section 5, respectively. Finally, the conclusions are presented in Section 6.

2. Literature review

2.1. MOT based on non-deep learning methods

According to whether MOT is based on future frames, the non-deep learning methods can be divided into two categories, i.e., online methods and batch methods. The online methods only use the current and previous frames, while the batch methods use the entire sequence including the future frames. Most online methods focus on data association. For example, Bewley et al. [19] proposed to use Kalman filter [20] to predict future object locations, compute their overlap with the detected objects in future frames, and then adopt Hungarian algorithm [21] for tracking. The results concluded that involving the data association step could improve the tracking accuracy. Bochinski et al. proposed a high-speed MOT system [22] that directly correlated detections through spatial overlap of adjacent frames without using Kalman filter. The experimental results showed that due to its simplicity, this method could greatly reduce the computation cost to make the algorithm meet the real-time requirements. However, due to the lack of re-ID features, they may not be used in challenging scenes (e.g., crowded scenes and fast camera movement scenes). In [23], Bae et al. utilized linear discriminant analysis (LDA) to extract re-ID features for interested targets. The results showed that LDA could achieve more robust tracking than the compared methods. Xiang et al. [10] formulated the online MOT problem into a decision in Markov decision processes (MDP), and built an MDP model for each target. In MDP, learning a policy mainly involves learning the correlation of similar data, which is more conducive to the correlation between offline data and online data. At the same time, the framework can naturally handle the birth/death and appearance/disappearance of targets by treating them as state transitions in MDP while leveraging existing online single object tracking methods.

The batch methods usually perform better than the online methods because they can effectively realize global optimization in the entire sequence. Zhang et al. [24] established a graphical model in which nodes represented the detections in all frames for MOT. The results concluded that applying the specific structure of the graph could be able to reach the optimum faster than linear programming. In [25], Berclaz et al. took advantage of the k-shortest paths algorithm to solve data association task. The results showed that it significantly speeded up the computation speed and reduced the number of parameters that needed to be fine-tuned. Milan et al. [26] proposed an alternative formulation of MOT as the minimization of a continuous energy. This method focused on designing an energy function that corresponded to a more complete representation of the problem. It was concluded that the developed algorithm constructed a suitable optimization scheme that alternated between continuous conjugate gradient descent and discrete trans-dimensional jump moves. Shen et al. [27] established a novel multi-object tracking method based on submodular optimization. Firstly, the authors generated low-level tracklets from the initial detection results based on overlap criteria together with min-cost flow, following by an integration operation of the tracklets into a candidate set. Then, the authors transformed the tracking problem into an optimization problem

of the submodular function, and used the optimized function to pick the desired tracklets from the integrated candidate set to generate the final trajectory. In addition, a connecting process was also developed to handle the occlusion problem. Their results supported the effectiveness of the proposed method.

With the rapid development of deep learning technologies in the recent years, the advantages of deep learning methods (e.g., no need for manual feature design, good feature expression ability, excellent tracking accuracy) have significantly motivated the adoption of the related technologies in MOT to make deep learning based methods become the mainstream for MOT.

2.2. MOT based on deep learning methods

The rapid development of deep learning has prompted researchers to explore modern object detectors instead of using baseline detection results provided by benchmark datasets. Fang et al. [10] treated object detection and re-ID as two separate tasks. The object detection phase firstly applied CNN-based object detectors to locate all objects of interest in the input images. Then, in a separate step, the re-ID phase cropped the image according to the detected bounding box generated by the object detection phase and then fed it to the identity embedding network to extract re-ID features. The linking step usually follows the standard convention, which firstly calculates the cost matrix based on the intersection over union (IOU) of the re-ID features and the bounding box, and then uses the Kalman filter and the Hungarian algorithm to complete the link task. It was concluded that the two-step method can develop the most suitable model for each task separately. In [28], Wojke et al. integrated the appearance information to improve the performance of [19]. Due to this extension, the algorithm could be able to track objects through longer periods of occlusions and reduce the number of identity switches. Yu et al. [29] explored the high-performance detection and deep learning based appearance features, and showed that the proposed method could lead to significantly better MOT results in both online and offline tracking. In [30], MOT is formulated as a high-order graph matching problem, based on which a l_1-norm tensor power iteration solution is proposed. The experiment results showed that the proposed deep pair-wise appearance similarity metric could solve the weak discrimination problem in the situation where the bounding boxes heavily overlapped each other. In [31], Bergmann et al. exploited the bounding box regression of an object detector to predict the position of an object in the next frame, thereby converting a detector into a tracker. The results showed that none of the dedicated tracking methods were considerably better in dealing with complex tracking scenarios like small and occluded objects with missing detections. Sun et al. [32] employed deep learning for data association by jointly modeling the appearances of objects and their similarities between different frames in an end-to-end manner. The proposed deep affinity network learned the compact and comprehensive features of pre-detected objects at several abstract levels, and performed a detailed pairwise arrangement of these features in any two frames to infer the similarities of the objects. It was concluded that utilizing the resulting efficient affinity computations to associate objects in the current frame deep into the previous frames could be reliable on tracking. However, these methods are usually slow because the two tasks need to be completed separately. Therefore, it is difficult to achieve the video rate inference required in many applications.

To alleviate the above issue, Yin et al. [33] proposed a new triplet network UMA by unifying object motion and affinity model into a single network, which can also address the association-discriminative feature learning problem by incorporating an attention mechanism. Their results showed that the proposed method can help the tracker distinguish distractors, effectively improve the computational efficiency and simplify the training procedure by learning a compact feature that can distinguish object motion and affinity measure.

With the rapid development of multi-task learning in deep learning, one-shot MOT has begun to attract more research attention. The core idea is to simultaneously complete object detection and identity embeddings (re-ID features) in a single network to reduce the inference time. For example, Voigtlaender et al. [13] added a re-ID head based on Mask R-CNN [34] and regressed a bounding box and a re-ID feature for each proposal. The experimental results showed that the proposed method could achieve near real-time inference. Analogously, JDE was built on top of YOLOv3 [17], which could allow learning object detection and appearance features in the shared model. The results concluded that JDE significantly reduced the runtime for MOT, making it possible to be used in real-time systems. Meanwhile, the tracking accuracy of JDE is comparable with the part of two-shot MOT methods. This study demonstrated that one-shot MOT methods could also effectively predict the position of the interested targets in video frames. However, the accuracy of JDE was still lower than the state-of-the-art methods. Therefore, on the basis of JDE, we further strengthen the ability of feature representation to improve the tracking accuracy of the MOT system.

3. Methodology

3.1. Problem formulation

Assuming that the labeled dataset $D = \{x_i, y_i\}_{i=1}^{N}$ consists of N frames, where x_i and $y_i = (b_i, id_i)$ respectively represent the i-th image and its corresponding label information (i.e., the coordinates of the bounding box and its identity label). Here, $b_i \in R^{K\times 4}$ and $id_i \in Z^K$ respectively represent the bounding box annotations and corresponding identity labels for the K objects in each frame. The goal of our method is to output a more accurately predicted bounding box $\widehat{b_i}$ and more discriminate appearance embedding features $\widehat{f_i}$. The output bounding box should be as close to the labeled bounding box as possible. As for appearance embedding features, the observed Euclidean distance or Mahalanobis distance of the same target features in successive frames should be less than the distance between different target features.

3.2. Framework overview

The overall framework of our proposed method is shown in Fig. 1. The framework defines two steps of tracking, as shown in Fig. 1, Part A and Part B. The performance of a MOT system depends largely on how the backbone extracts and aggregates high-quality features. In our proposed method, we employ DarkNet-53 as the backbone in order to improve the performance of MOT by increasing the accuracy while minimizing the loss of efficiency in the first step shown in Part A. Furthermore, an FPN is applied behind the backbone to fuse the low-resolution feature maps with strong semantic information and high-resolution feature maps with rich spatial information. An input image frame is firstly forward-passed into the backbone network to obtain three different resolution feature maps, whose resolutions are 8, 16, and 32 times lower than the input size, respectively. Then, these three feature maps are fed into the FPN to become highly robust to the scale variations. Finally, the prediction heads, consisting of several stacked convolutional layers, are added upon fused feature maps at all the three scales and output three dense prediction maps with a size of $(6B + Emb\text{-}D) \times H \times W$, where B is the number of bounding boxes allocated to a certain ratio, and $Emb\text{-}D$ indicates the dimension of appearance embedding features. For the second step shown in Part B, the dense prediction map is further divided into two parts, one for target detection and the other for re-ID features. The detection branch outputs the classification score of each point in the three different feature maps, the coordinate offsets and the scale transformation corresponding to each anchor. The appearance embedding branch includes our proposed attention mechanism network (AM-Net) and a feature combination module (FCM). A detailed description of the proposed method is given in the following subsections.

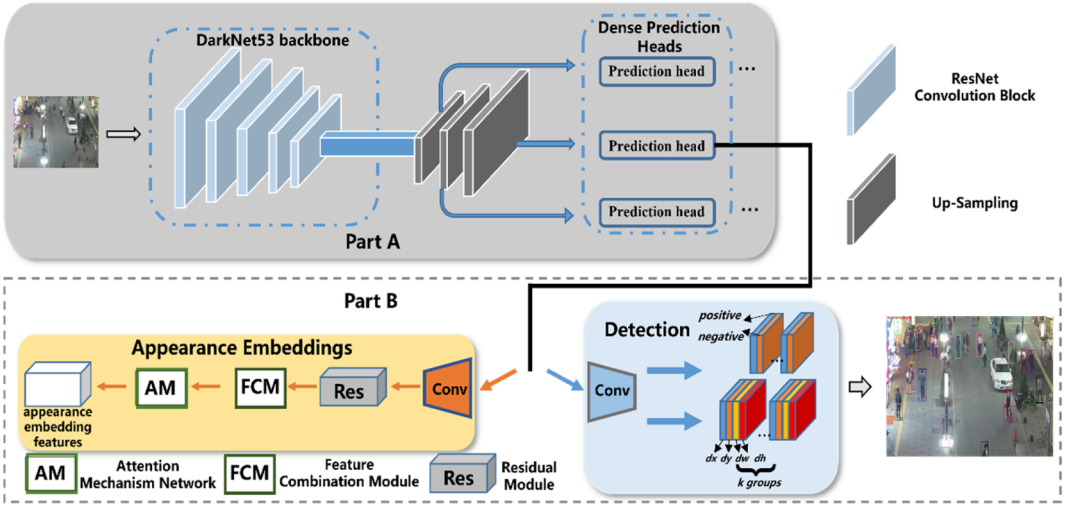


Fig. 1. The overall framework of the proposed method. The feature extractor is presented in Part A. The detection branch and the appearance embedding branch are shown in Part B.

3.3. The detection branch

Considering the prior knowledge [35], we set the aspect ratio of all anchors to 1:3 and set the number of anchor boxes on each pixel to 4. Moreover, consistent with the general settings in target detection tasks, the predicted bounding boxes *PBox* with an IOU (see Eq. (1)) score higher than 0.5 with the ground truth box *GTBox* are regarded as foreground, while those that have an IOU < 0.4 are regarded as background. The other bounding boxes are ignored.

$$IOU = \frac{PBox \cap GTBox}{PBox \cup GTBox}, \tag{1}$$

The target confidence score is set to 1 for the foreground and 0 for the background.

$$TConf = \begin{cases} 1, IOU > 0.5 \\ 0, IOU < 0.4 \\ -1, 0.4 < IOU < 0.5 \end{cases}, \tag{2}$$

where *Tconf* is the labeled target confidence score.

Our detection branch has two loss functions including the foreground/background classification loss L_{cls} and the bounding box regression loss L_{box}. Similar to the settings of JDE, L_{cls} is formulated as a cross-entropy loss and L_{box} is formulated as a smooth-L1 loss.

$$L_{cls} = -\frac{1}{N}\sum_{k=1}^{N} Tconf_k \left(\log\left(\frac{\exp(Pconf_k^i)}{\sum_i \exp(Pconf_k^i)} \right) \right), i = 0, 1, \tag{3}$$

$$L_{box}(\Delta PBox, \Delta GTBox) = \frac{1}{N}\sum_{i=1}^{N} \begin{cases} 0.5*(\Delta GTBox_i - \Delta PBox_i)^2, & |\Delta GTBox_i - \Delta PBox_i| < 1 \\ |\Delta GTBox_i - \Delta PBox_i| - 0.5, & otherwise \end{cases}, \tag{4}$$

where N is the number of targets, $\Delta PBox$ and $\Delta GTBox$ denote the corresponding offset values (i.e., Δx, Δy, Δw, Δh) of the predicted bounding box and the ground truth bounding box, respectively.

3.4. The appearance embedding branch

The goal of the appearance embedding branch is to extract features that can distinguish similar targets. To achieve this goal, we propose an AM-Net and an FCM approach, which can contribute to learn features that are sufficient to distinguish similar targets, so that the affinity between features of different objects is less than the affinity between features of the same object.

3.4.1. Attention Mechanism Network (AM-Net)

The attention mechanism enables the deep neural network to pay more attention to the important areas of the input image and ignore the interference of the irrelevant areas, so as to improve the performance of the algorithm in MOT. Inspired by [36,37], we propose an AM-Net approach which calculates the attention map of the features from the space domain and channel domain and performs feature adaptive learning by multiplying the input feature map and the attention map to highlight important features.

The included channel attention module and spatial attention module in AM-Net can weight the channel domain and spatial domain of the feature map to highlight the importance of different regions in the feature map. For a feature map $F \in R^{C\times H\times W}$ in the middle layer, AM-Net can sequentially infer a channel attention map $M_C \in R^{C\times 1\times 1}$ and a spatial attention map $M_S \in R^{1\times H\times W}$. The whole process is calculated as:

$$F' = M_C(F) \otimes F, \tag{5}$$

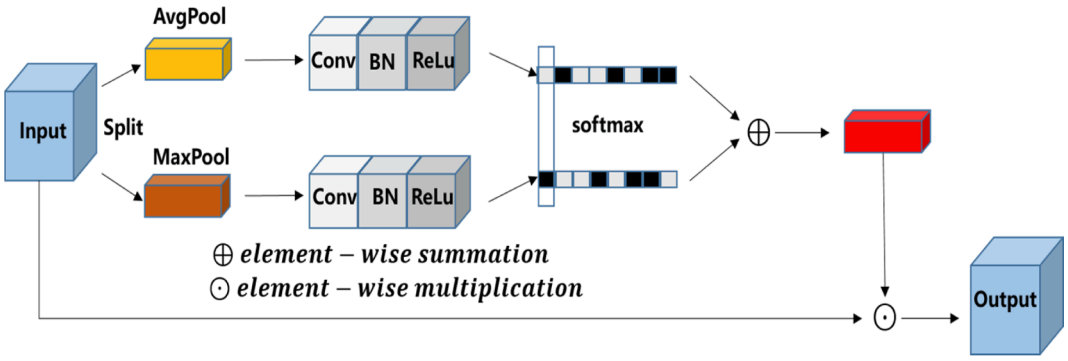
$$F'' = M_S(F') \otimes F', \tag{6}$$

where $\otimes$ represents the multiplication of the corresponding positions of the matrix elements, that is, the Hadamard product. F' and F'' represent the feature maps processed by the channel attention module and spatial attention module, respectively. The channel attention map is multiplied by the input feature map to get F', and the spatial attention map of F' is calculated and multiplied with F' to get the final output F''.

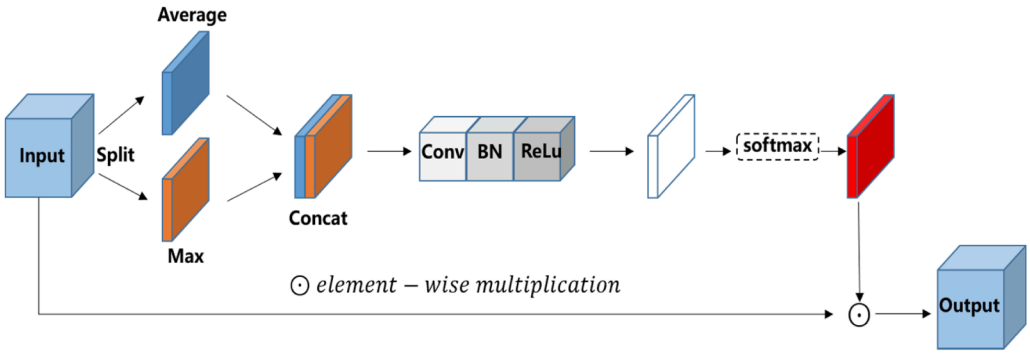
Fig. 2 shows the channel attention module and spatial attention module in the AM-Net. The channel attention module focuses on the channel domain of the feature map. First, it uses the global maximum pooling and global average pooling to compress the feature map in spatial dimensions to obtain two different spatial background descriptions F_{avg}^c and F_{max}^c, and then utilizes the convolutional neural network with shared weights to calculate and merge these two feature maps. The weight of the channel attention map $M_C \in R^{C\times 1\times 1}$ is the final output. The specific calculation process is defined as follows:

$$M_C(F) = \sigma(CNN(AvgPool(F)) + CNN(MaxPool(F))), \tag{7}$$

where σ represents the softmax function. We add a temperature of 5 into the softmax, such that the softmax output values are more discriminative. *CNN* denotes the convolutional neural network with shared



(a) spatial attention module



(b) channel attention module

Fig. 2. The structural details of our proposed attention mechanism network (AM-Net).

weights.

The spatial attention module mainly focuses on the location information. First, the global maximum pooling and global average pooling are used to compress the channel dimensions to obtain two different feature descriptions. Then, two feature descriptions are cascaded, and a convolutional layer, a batch normalization layer and an activation function are successively used to obtain a spatial attention map. The specific calculation process is given as follows:

$$M_S(F) = \sigma(CNN([AvgPool(F); MaxPool(F)])), \tag{8}$$

where σ is the same as of softmax function used in the channel attention module.

3.4.2. Feature Combination Module (FCM)

Inspired by [39], we propose an FCM method to alleviate the gridding effect caused by dilated convolution. The FCM consists of dilated convolution with different expansion rates. The convolutions with different expansion rates have different receptive fields so that FCM can capture the features of regions with different scales, improving the scale invariance of the final combined features. It is found in our experiment that increasing the diversity of dilations can improve the representability of features. Therefore, we select three different dilated convolutions, and their strides are set to $(x, y) \in \{(1,3),(2,2),(3,1)\}$, where x and y represent the expansion rates along the X-axis and Y-axis, respectively. The input features are convolved with three different expansion rates, thus three feature maps with different resolutions are generated. The feature maps are then interpolated to the same and appropriate resolution so as to facilitate feature fusion. The architecture of FCM is illustrated in Fig. 3.

3.4.3. Re-ID loss

Existing work typically employs the cross entropy loss or its variant-

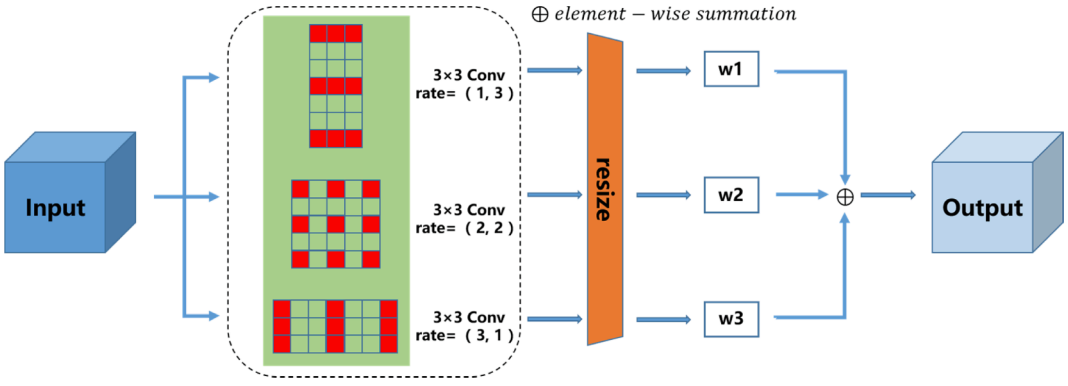


Fig. 3. The structure of feature combination module (FCM).

focal loss to optimize the re-ID branch. Although the cross entropy loss can reflect the possibility of a positive prediction, and the focal loss can solve the problem of a serious imbalance in the proportion of positive and negative samples in object detection by reducing the weight of a large number of simple negative samples in training, they still have the following limitations: First, these loss functions do not calculate the distances between different appearance embedding features from different targets. Second, they do not give an explicit distance metric between different feature pairs. Inspired by [38] to figure out these problems, we introduce a combination of triple loss and online instance matching (OIM) loss, namely TOIM Loss, to distinguish similar instances within a class. See Fig. 4.

If a bounding box is marked as the foreground, the corresponding embedding vector is extracted from the dense embedding graph. Then, we set up a look-up table (LUT) $V \in R^{D\times L} = \{v_1, v_2, \cdots, v_L\}$ to store all labeled appearance embedding features, where L denotes the number of labeled identities and D denotes the dimension of embedding features. In each iteration, given an input feature x with its corresponding ID label l, OIM computes the similarity between x and all the features in the LUT. The probability of x belong to the identity l is calculated as:

$$p_l = \frac{\frac{\exp\left(v_l^T x\right)}{\tau}}{\sum_{j=1}^{L}\frac{\exp\left(v_j^T x\right)}{\tau}}, \tag{9}$$

where$\tau = 0.1$ is a hyperparameter that controls the softness of the probability distribution. The objective of OIM is to minimize the expected negative log-likelihood.

$$L_{OIM} = -E_x[logp_t], t = 1, 2, \cdots, L, \tag{10}$$

where L is the number of identities in the entire dataset.

Furthermore, we propose a specially designed triplet loss, which could help to pull the feature vectors from the same person close and push the vectors from different people away. Specifically, the features in the LUT that have the same identity as the input features are found out and concatenate with the input features. If the label corresponding to the input features is consistent with the LUT, the input features will be used to update the features stored in LUT. The triplet loss is calculated as:

$$L_{tri} = \sum_{pos,neg}\left[M + D_{pos} - D_{neg}\right], \tag{11}$$

where M denotes the distance margin which controls the distinguishability between different targets, D_{pos} and D_{neg} denote the Euclidean distances between the positive pair and the negative pair, respectively.

Finally, since it is required to train multiple prediction heads for optimization of our overall one-shot MOT, we approach the work as a problem of multi-task learning. Specifically, the automatic balancing of multiple losses using uncertainty [40] is applied to the multi-task learning of object detection and re-identification:

$$L_{detection} = L_{cls} + L_{box}, \tag{12}$$

$$L_{identity} = L_{tri} + L_{OIM}, \tag{13}$$

$$L_{total} = \frac{1}{2}\left(\frac{1}{e^{w_1}}L_{detection} + \frac{1}{e^{w_2}}L_{identity} + w_1 + w_2\right), \tag{14}$$

where w_1 and w_2 are learnable parameters that automatically balance the two tasks, respectively.

3.5. Online association

Online association is a crucial process in MOT since the detected objects are interpreted into tracklets by the online association algorithm. The association algorithm aims to accurately track objects by maximizing the leverage of the appearance embedding features with the aid of the predicted object locations. We employ standard online tracking algorithms to associate the estimated bounding boxes. Firstly, the bounding boxes are generated by the detection branch and the corresponding appearance embedding features are generated by re-ID branch. Then, we remove the bounding boxes with confidence scores lower than predefined confidence threshold or high overlap rates according to the predefined Non-Maximum Suppression (NMS) [41]. The multiple trajectories are initialized based on the bounding boxes in the first frame, and the detected boxes are linked to the existing trajectories according to the Euclidean distance calculated based on the appearance embedding features in the successive frames. We also utilize the Kalman filter (KF) to further filter the bounding boxes in the current frame. If the distance between the tracking bounding boxes predicted by KF and the detected bounding boxes is too far, the trajectory information allocated to this target should be rejected. The appearance embedding features are then updated in each frame. If there is no observation allocated to the tracklet, the tracklet is marked as lost. If the lost time is greater than the track buffer, the lost tracklet will be deleted from the set of tracklets, otherwise it will be found again in the allocation step. The overall flow chart of the online association is shown in Fig. 5.

4. Dataset and experiment platform

4.1. Datasets

Six datasets is used to train our proposed network, including ETH [42], CityPersons [43], CalTech [44], MOT17 [45], CUHK-SYSU [46], and PRW [47]. ETH is a pedestrian detection dataset, which is captured from stereo rigs mounted on cars with a resolution of 640 × 480 and a framerate of 13–14 FPS. CityPersons is a subset of Cityscapes and consists only person annotations. The average number of pedestrians in each image of CityPersons is 7. The Caltech pedestrian dataset is

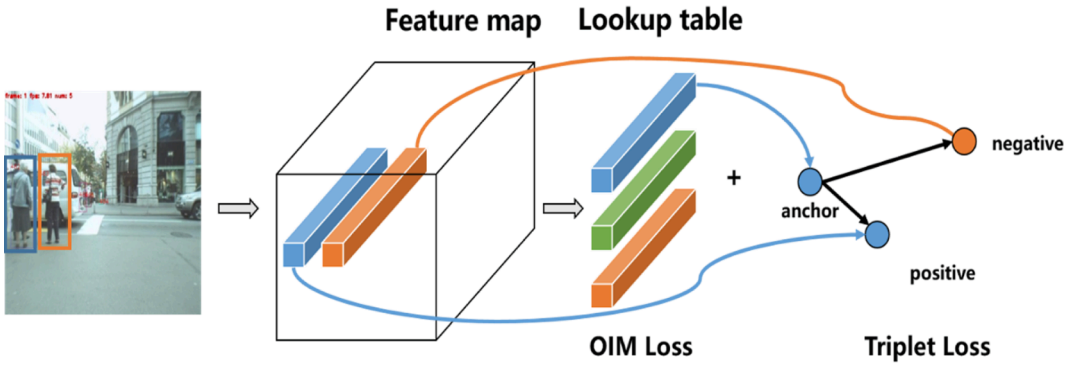


Fig. 4. Illustration of the TOIM loss.

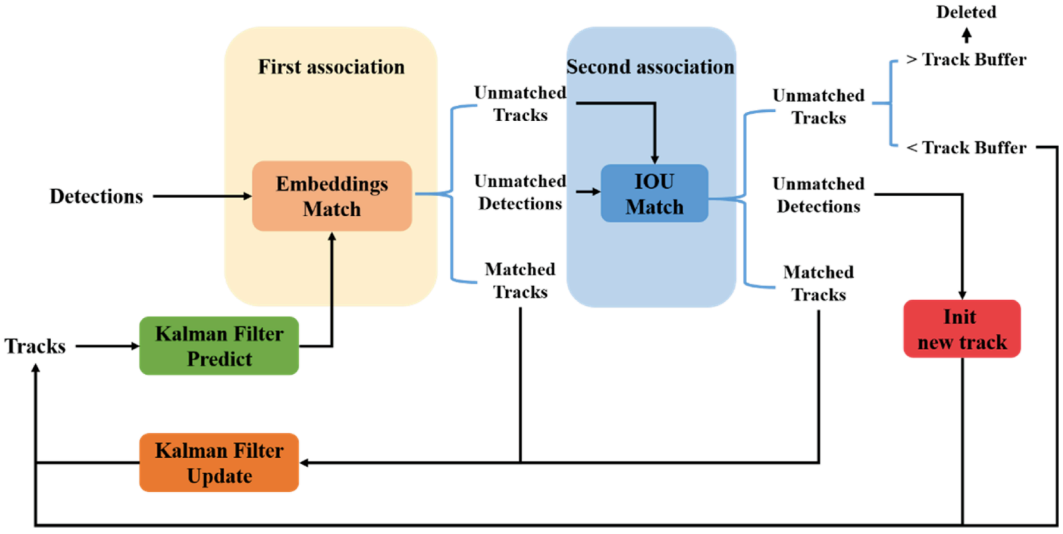


Fig. 5. The overall flow chart of online association.

constructed by approximately 10 h of 640 × 480 30 Hz video collected from a vehicle driving through regular traffic in an urban environment. MOT17 is a widely used dataset for MOT. The CUHK-SYSU dataset is a large-scale benchmark for person search, which is much closer to real application scenarios by searching person from whole images in the gallery. PRW is a person re-ID dataset in a wild environment. The ETH and CityPersons datasets only provide bounding box annotations, while the CalTech, MOT17, CUHK-SYSU and PRW datasets provide both bounding box and identity annotations.

4.2. Implementation details

In our experiment, DarkNet-53 is used as the backbone of our proposed network. The network is mainly composed of convolution and residual structures, but also contains skip and shortcut connections, upsampling, batch normalization [48] layer, and other structures. The activation function is leaky ReLU [49]. In order to reduce the training time, DarkNet-53 is used to be pre-trained on ImageNet [50] before applying the proposed method. Throughout the experiment, the sizes of all input images in the experiment are resized to 1088 × 608, and the resolutions of output are 136 × 76, 68 × 38 and 34 × 19 for our experiment. The network is trained with the standard stochastic gradient descent (SGD) [51] optimizer for 30 epochs (including six training datasets) and 60 epochs (including one training datasets) with an initial learning rate of 0.01. The learning rate is dropped by ten times at 0.5 × total epochs and 0.75 × total epochs. The batch size is set to be 8. Due to the insufficient number of images in the datasets for reliable training, several standard data augmentation techniques including rotation, scaling, and color jittering are applied to reduce the overfitting problem. The source code of our proposed method is available at: the link will be provided in the final accepted paper.

4.3. Experimental platform

The experiment is carried out on a platform with 2 pieces of NVIDIA GeForce RTX 3090 and 24 GB of memory. The machine is running Ubuntu 18.04.5 LTS with NVIDIA CUDA 11.0 and cuDNN 8.0 installed. The training and inference steps are done using the Pytorch framework. The proposed network performs tracking at 15 fps.

5. Results and discussions

The performance of our proposed method is examined on three commonly used MOT datasets including 2DMOT15 [52], MOT16 [45] and MOT20 [53]. To demonstrate the improvement of our proposed method, three multi-object tracking methods are used for comparison including the baseline JDE, the simple online and real-time tracking (SORT), and the deep affinity network for multi object tracking (SST). For fair comparison, all the examined methods are trained based on the same training set. The results are presented in the following subsections.

5.1. Quantitative results

Seven quantitative evaluation metrics are used to examined the effectiveness of our proposed method including multi-object tracking accuracy (MOTA) [54], identification F-Score (IDF1) [55], mostly tracked targets (MT), mostly lost targets (ML), false positive (FP), false negative (FN), and fragmentations (FM). Among all the metrics to evaluate the performance of an MOT method, MOTA is the most important metric that can generally reflect the tracking performance. A higher MOTA value indicates a better performance. Besides MOTA, IDF1 is also commonly used for evaluation, which is defined as the ratio of correctly identified detections over the average number of ground-truth objects and object tracks. Intuitively, a higher IDF1 score indicates that the images of an object are mostly mapped to the same identity without including that of other objects. MT and ML are defined as the ratios of ground-truth trajectories that are covered by a track hypothesis more than 80% and less than 20% of their life span, respectively. FP and FN are defined as the total number of false targets and missed targets, respectively. FM is the total number of times a trajectory is fragmented. To make fair comparisons with the three methods mentioned above, we firstly test our proposed method trained by MOT17.

The quantitative results on 2DMOT15 are presented in Table 1. The results show that our proposed method generally outperforms all the compared methods on 2DMOT15. By comparing with the baseline JDE method, almost all the examined metrics of our proposed method are the best except FP. Specifically, the MOTA, IDF1, and MT of our proposed method are 37.1%, 58.4%, and 353 higher than the numbers of JDE, respectively. The ML, FN, and FM of our proposed method are reduced

Table 1
Quantitative results of different methods on 2DMOT15-Train (only one dataset).

	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
SORT [19]	25.97	30.97	64	237	**6770**	21,990	1174
SST [32]	9.29	38.2	98	202	15,730	18,847	1984
JDE [15]	25	53.4	302	95	22,764	8805	1271
Ours	**37.1**	**58.4**	**353**	**80**	19,742	**6650**	**993**

by 15, 2155, and 278, respectively. When comparing with SST and SORT, our proposed method can still achieve the best performance in term of these evaluation metrics. The performance of our proposed method on FP is not the best probably because the extracted features of the falsely detected target and the correct target are very similar, and the classifier cannot correctly classify. In general, the results on 2DMOT15 validate the advancement of our proposed method over the compared methods.

The results in Tables 1 and 2 demonstrate that our method can effectively help reduce ML, FP, FN and increase MOTA, MT. The advancement of our proposed method is mainly because that our method utilizes the dilated encoder to capture different scale information instead of using the recent FPN, such as M2Det which contains a large number of parameters and has a slow inference speed. Therefore, our method is more likely to achieve a good balance between accuracy and efficiency.

5.2. Qualitative results

To visually illustrate the tracking performance for qualitative evaluation, Fig. 6 shows the tracking results of our proposed method on the training set of MOT16. From the results of MOT16-05 in Fig. 6, we can observe that our method can still assign the correct identity with the help of high-quality re-ID features when two pedestrians cross each other, while trackers without using the attention mechanism usually cause identity switches under these circumstances. The results of MOT16-04 show that our method performs well under crowded scenes. The results of MOT16-09 and MOT16-10 illustrate that our method can deal with large scale variations. This is mainly attributed to the use of dilated convolution to obtain different receptive fields and deal with the scale transformation of the targets. In particular, the occlusion of identity −1 and identity −2 in scene 09 is handled well by our proposed method. Although our tracker may temporarily misjudge the trajectory in the middle picture by assigning it a wrong identity, it is able to quickly recover from this situation by looking deeper into the previous frames. From the results of MOT16-02 and MOT16-11, we can observe that our method can keep both correct identities and correct bounding boxes when the pedestrians are heavily occluded. The results of MOT16-13 show that our method can also detect small objects accurately.

5.3. Ablation study results

The AM-Net and the FCM are the key components of our proposed method. We conducted an ablation analysis to examine their contributions to the task performance. As displayed in Tables 3 and 4, the MOTA of the baseline JDE is only 25% and 72.9% on 2DMOT15 and MOT16, respectively. The MOTA when using JDE + AM-Net is 34.1% and 75.4%, respectively, indicating that AM-Net contributes to 9.1% and 2.5% improvements in the corresponding testing sets. Comparing the numbers of JDE + AM-Net and JDE + AM-Net + FCM, the results show that the dilated convolution can contribute to 3.0% and 1.5% improvements in the MOTA, respectively. Besides the MOTA, our method can effectively help reduce ML, FP and FN. In summary, the presented results in Tables 3 and 4 demonstrate the necessity of each component in our proposed approach. Therefore, our proposed method could significantly improve both driving safety for drivers and traffic safety for vulnerable traffic participants by generating more accurate tracking results.

Table 2
Quantitative results of different methods on MOT16-Train (only one dataset).

	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
SORT [19]	24.9	29.7	41	271	12,916	68,904	**1493**
SST [32]	27.9	37.5	37	278	6914	71,552	2785
JDE [15]	72.9	**74.2**	330	25	8253	20,612	1889
Ours	**76.9**	72.3	**343**	**21**	**4652**	**19,969**	1979

Experiments on small datasets may lead to biased results, and when the same algorithm is applied to large datasets, the conclusion may not hold. Consequently, we also conduct experiments based on six publicly available datasets on pedestrian detection, MOT and person search (mentioned in section 4.1). The corresponding results are presented in Tables 5 and 6, respectively. The results show that our proposed method outperforms the baseline JDE. Specifically, the MOTA of our proposed method is 40.78% in 2DMOT15 and 62 % in MOT16, higher than the numbers of JDE. The above results verify the effectiveness of our method on both small datasets and large datasets.

To further verify the generalization ability of our proposed modules, we also evaluate them on a more recent dataset MOT20. The ablation results on MOT20 are presented in Table 7. The results show that our proposed method outperforms the baseline JDE. Specifically, the MOTA and IDF1 when using JDE + AM-Net are 21.3% and 22.6%, respectively, which proves that the automatic learning image attention mechanism can learn the weights adaptively and assign different weights to pixels in different dimensions. Comparing the performances of JDE + AM-Net and JDE + AM-Net + FCM, the results show that dilated convolution can contribute to 8.4% and 3.2% improvements on MOTA and IDF1, respectively. This indicates that the feature combination module adaptively assigns weights to feature maps produced from convolutions with different dilated rates according to the sizes of targets. The above results on three different datasets demonstrate the generalization ability of our proposed modules.

5.4. Performance when using different embedding loss functions

Embedding loss function is crucial for our proposed method to learn discriminative representation. In this paper, three embedding loss functions are implemented. The first is the cross-entropy loss (named "CE") which is used to determine how close the actual output is to the expected output. The second is the triplet loss (named "Tri") to train samples with small differences. The third is the triplet-guided online instance matching loss (named "TOIM") to pull the feature vectors from the same object close, and push the vectors from different object away. The corresponding results of these three embedding loss functions on 2DMOT15, MOT16 and MOT20 are shown in Tables 8, 9 and 10, respectively.

As expected, L_{TOIM} generally outperforms L_{CE} and L_{Tri}. Specifically, the MOTA of L_{TOIM} is 45.92% higher than the numbers of L_{CE} and L_{Tri} in Table 8, 66.92% higher than the numbers of L_{CE} and L_{Tri} in Table 9, and 30.8% higher than the numbers of L_{CE} and L_{Tri} in Table 10, respectively. A possible reason for the large performance gap is that TOIM not only combines the advantages of Tri and CE, but also makes full use of the historical information by creating a LUT, which can better perform hard mining and find the most representative inter-class distance and intra-class distance. As a result, TOIM can help to pull the feature vectors from the same person close, and push the vectors from different people away.

5.5. Limitations and future work

One challenge not addressed in this paper is the problem of inaccurate detection bounding box in crowded scenarios. High-quality bounding box may be filtered out by the bounding box with high classification score through non-maximum suppression due to high intersection over union between these two bounding boxes. Further, the definition of the positive and negative samples in the training phase is crucial for classification and regression, which will eventually affect the class score and regression accuracy of the bounding box output by the tracker. For future work, a more reasonable bounding box filtering mechanism could be used to select high-quality bounding boxes instead of only relying on IOU thresh. Further, how to define positive and negative training samples is important for current multi-object tracking methods to generate accurate bounding box and deserves further study.

Fig. 6. Example tracking results of our method on MOT16.

Table 3
Evaluation of the two ingredients on 2DMOT15-Train (only one dataset).

Method	AM-Net	FCM	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
JDE [15]	×	×	25	53.4	302	95	22,764	8805	1271
Ours	√	×	34.1	57.4	**381**	**57**	21,529	**6138**	1029
Ours	√	√	**37.1**	**58.4**	353	80	**19,742**	6650	**993**

Table 4
Evaluation of the two ingredients on MOT16-Train (only one dataset).

Method	AM-Net	FCM	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
JDE [15]	×	×	72.9	**74.2**	330	25	8253	20,612	**1889**
Ours	√	×	75.4	72.5	339	28	5703	20,415	2148
Ours	√	√	**76.9**	72.3	**343**	**21**	**4652**	**19,969**	1979

Table 5
Evaluation of the two ingredients on 2DMOT15-Train (six datasets).

Method	AM-Net	FCM	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
JDE [15]	×	×	38.44	59.05	288	61	16,843	7709	943
Ours	√	×	39.64	**59.64**	302	42	16,691	6849	**858**
Ours	√	√	**40.78**	59.28	**316**	**41**	**16,453**	**6648**	905

Table 6
Evaluation of the two ingredients on MOT16-Train (six datasets).

Method	AM-Net	FCM	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
JDE [15]	×	×	55.4	57	141	111	8571	39,284	2411
Ours	√	×	61.2	**64.9**	185	**70**	8433	**33,123**	2336
Ours	√	√	**62**	64.5	**191**	75	**6647**	34,200	**2223**

Table 7
Evaluation of the two ingredients on MOT20-Train (six datasets).

Method	AM-Net	FCM	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
JDE [15]	×	×	15.1	19.1	45	1524	33,743	1,093,306	29,040
Ours	√	×	21.3	22.6	78	1306	38,902	1,001,468	**26,244**
Ours	√	√	**29.7**	**25.8**	**135**	**992**	**13,994**	**908,974**	30,187

Table 8
Tracking performance of different embedding loss functions on 2DMOT15-Train (six datasets).

Embedding loss	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
L_{CE}	40.78	**59.28**	316	**41**	16,453	6648	**905**
L_{Tri}	41.9	53.9	365	55	17,186	7032	1014
$\boldsymbol{L_{TOIM}}$	**45.92**	57.71	**397**	47	**16,153**	**6344**	975

Table 9
Tracking performance of different embedding loss functions on MOT16-Train (six datasets).

Embedding loss	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
L_{CE}	62.03	**64.48**	191	75	**6647**	34,200	**2223**
L_{Tri}	62.02	53.84	194	57	9776	30,146	2775
$\boldsymbol{L_{TOIM}}$	**66.92**	60.57	**235**	**42**	7764	**27,149**	2658

Table 10
Tracking performance of different embedding loss functions on MOT20-Train (six datasets).

Embedding loss	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	FM↓
L_{CE}	29.7	25.8	135	**992**	**13,994**	908,974	30,187
L_{Tri}	18.9	18.4	55	1419	19,715	1,052,596	28,934
$\boldsymbol{L_{TOIM}}$	**30.8**	**29.1**	**147**	1012	16,220	**896,689**	**28,390**

In addition, sensor fusion technologies can also be considered for improvement in various environmental scenarios [56].

6. Conclusion

In this paper, we propose a convolutional neural network-based method with the spatial-channel attention mechanism for multi-object tracking, which is crucial for the design of intelligent systems to address the problem of inaccurate tracking caused by heavily occluded in pedestrians. Our method is a one-shot multi-object tracking system, which allows target detection and appearance features to be learned in a shared model and reduces inference time compared with two-shot system. Moreover, we introduce FCM, AM-Net, and TOIM Loss, which could provide more detailed cues for identity matching and have little effect on the running speed (reduced from 18 FPS to 16 FPS). With our approach, a discriminative re-identification system can be obtained,

which could pull the feature vectors from the same person close, and push the vectors from different people away. Our model can be trained end-to-end using the standard SGD optimization technique. The presented results demonstrate that our method outperforms the baseline JDE and is effective in various multi-object tracking scenarios. The sound results of our approach will promote the development efficiencies of intelligent systems, especially for autonomous vehicles to avoid collisions between vehicles and pedestrians in urban driving scenarios.

CRediT authorship contribution statement

Guofa Li: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Funding acquisition. **Xin Chen:** Methodology, Data curation, Validation, Visualization, Software, Formal analysis, Writing – original draft. **Mingjun Li:** Writing – review & editing. **Wenbo Li:** Writing – review & editing. **Shen Li:** Conceptualization, Writing – review & editing. **Gang Guo:** Supervision, Writing – review & editing. **Huaizhi Wang:** Writing – review & editing. **Hao Deng:** Resources.

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.

Acknowledgement

This study is supported by the Shenzhen Fundamental Research Fund (grant number: JCYJ20190808142613246, 20200803015912001).

References

[1] G. Li, Y. Yang, S. Li, X. Qu, N. Lyu, S.E. Li, Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness, Transport. Res. Part C: Emerg. Technol. 134 (2022) 103452.
[2] Y. Chen, G. Li, S. Li, W. Wang, S.E. Li, B. Cheng, Exploring Behavioral Patterns of Lane Change Maneuvers for Human-Like Autonomous Driving, IEEE Trans. Intell. Transp. Syst. (2021), https://doi.org/10.1109/TITS.2021.3127491.
[3] G. Li, Z. Ji, X. Qu, Stepwise Domain Adaptation (SDA) for Object Detection in Autonomous Vehicles Using an Adaptive CenterNet, IEEE Trans. Intell. Transp. Syst. (2022), https://doi.org/10.1109/TITS.2022.3164407.
[4] G. Li, Z. Ji, X. Qu, R. Zhou, D. Cao, Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach, IEEE Trans. Intell. Veh. (2022), https://doi.org/10.1109/TIV.2022.3165353.
[5] R. Henschel, L. Leal-Taixe, D. Cremers, 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 (CVPR) Workshops (2018), pp. 1428–1437.
[6] Z. Lu, V. Rathod, R. Votel, J. Huang, Retinatrack: Online single stage joint detection and tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 14668–14678.
[7] S.-H. Bae, K.-J. Yoon, Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 40 (3) (2018) 595–610.
[8] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, J. Zhang, Heterogeneous association graph fusion for target association in multiple object tracking, IEEE Trans. Circuits Syst. Video Technol. 29 (11) (2019) 3269–3280.
[9] L. Wen, D. Du, S. Li, X. Bian, S. Lyu, Learning non-uniform hypergraph for multi-object tracking, Proceedings of the AAAI Conference on, Artif. Intell. 33 (01) (2019) 8981–8988.
[10] K. Fang, Y. Xiang, X. Li, S. Savarese, Recurrent autoregressive networks for online multi-object tracking, in: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (2018), pp. 466–475.
[11] Y. Xiang, A. Alahi, S. Savarese, Learning to track: Online multiobject tracking by decision making, in: Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 4705–4713.
[12] L. Chen, H. Ai, C. Shang, Z. Zhuang, B. Bai, Online multi-object tracking with convolutional neural networks, in: Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP) (2017), pp. 645–649.
[13] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, B. Leibe, Mots: Multi-object tracking and segmentation, in: Proceedings of the 2019 European Conference on Computer Vision (2019), pp. 7942–7951.
[14] B. Pang, Y. Li, Y. Zhang, M. Li, 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.
[15] Z. Wang, L. Zheng, Y. Liu, Y. Li, S. Wang, Towards real-time multi-object tracking, arXiv preprint arXiv:1909.12605 (2019).
[16] I. Kokkinos, Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory, in: Proceedings of the IEEE/CVF Conference on Ccomputer Vision and Pattern Recognition (2017), pp. 6129–6138.
[17] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv: 1804.02767 (2018).
[18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[19] A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft, Simple online and realtime tracking, in: Proceedings of the 2016 IEEE International Conference on Image Processing, 2016, pp. 3464–3468.
[20] G. Welch, G. Bishop, An Introduction to the Kalman Filter (1995) 1–16.
[21] H.W. Kuhn, The hungarian method for the assignment problem, Naval Res. Logist. Quart. 2 (1-2) (1955) 83–97.
[22] E. Bochinski, V. Eiselein, T. Sikora, High-speed tracking-bydetection without using image information, in: Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2017, pp. 1–6.
[23] S.-H. Bae, K.-J. Yoon, Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1218–1225.
[24] L. Zhang, Y. Li, R. Nevatia, Global data association for multi-object tracking using network flows, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[25] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, Multiple object tracking using k-shortest paths optimization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (9) (2011) 1806–1819.
[26] A. Milan, S. Roth, K. Schindler, Continuous energy minimization for multitarget tracking, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 58–72.
[27] J. Shen, Z. Liang, J. Liu, H. Sun, L. Shao, D. Tao, Multiobject tracking by submodular optimization, IEEE Trans. Cybern. 49 (6) (2019) 1990–2001.
[28] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: Proceedings of the 2017 IEEE International Conference on Image Processing, 2017, pp. 3645–3649.
[29] F. Yu, W. Li, Q. Li, Y. Liu, X. Shi, J. Yan, Poi: Multiple object tracking with high performance detection and appearance feature, in: Proceedings of the 2016 European Conference on Computer Vision, 2016, pp. 36–42.
[30] Z. Zhou, J. Xing, M. Zhang, W. Hu, Online multi-target tracking with tensor-based high-order graph matching, in: Proceedings of the 2018 24th International Conference on Pattern Recognition, 2018, pp. 1809–1814.
[31] P. Bergmann, T. Meinhardt, L. Leal-Taixe, Tracking without bells and whistles, arXiv preprint arXiv:1903.05625 (2019).
[32] S. Sun, N. Akhtar, H. Song, A.S. Mian, M. Shah, Deep affinity network for multiple object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2021) 104–119.
[33] J. Yin, W. Wang, Q. Meng, R. Yang, 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.
[34] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision (2017), pp. 2961–2969.
[35] 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) (2017) 1137–1149.
[36] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[37] S. Woo, J. Park, J.Y. Lee, et al., Cbam: Convolutional block attention module, in: Proceedings of the 2018 European Conference on Computer Vision, 2018, pp. 3–19.
[38] Y. Yan, J. Li, J. Qin, S. Bai, S. Liao, L. Liu, F. Zhu, L. Shao, Anchor-free person search, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7690–7699.
[39] P. Wang, P. Chen, Y. Yuan, et al., Understanding convolution for semantic segmentation, in: Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 1451–1460.
[40] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491.
[41] A. Neubeck, L. van Gool, Efficient non-maximum suppression, in: Proceedings of the 18th International Conference on Pattern Recognition, 2006, pp. 850–855.
[42] A. Ess, B. Leibe, K. Schindler, L. van Gool, A mobile vision system for robust multiperson tracking, in: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[43] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 3213–3221.
[44] P. Doll r, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009), pp. 304–311.
[45] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831 (2016).
[46] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, X. Joint detection and identification feature learning for person search, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 3415–3424.

[47] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person reidentification in the wild, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1367–1376.
[48] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on Machine Learning (2015), pp. 448–456.
[49] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.
[50] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inform. Process. Syst. 25 (2012) 1097–1105.
[51] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. 22 (3) (1951) 400–407.
[52] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, K. Schindler, Motchallenge 2015: Towards a benchmark for multi-target tracking, arXiv preprint arXiv:1504.01942 (2015).
[53] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, L. Leal-Taixé, MOT20: A benchmark for multi object tracking in crowded scenes, arXiv preprint arXiv:2003.09003 (2020).
[54] K. Bernardin, R. Stiefelhagen, Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics, EURASIP J. Image nd Video Process. 2008 (2008) 1–10.
[55] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Proceedings of the 2016 European Conference on Computer Vision, 2016, pp. 17–35.
[56] G. Li, Y. Lin, X. Qu, An infrared and visible image fusion method based on multi-scale transformation and norm optimization, Inform. Fusion 71 (2021) 109–129.