IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 12 March 2023, accepted 23 March 2023, date of publication 27 March 2023, date of current version 30 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3262466*

RESEARCH ARTICLE

# Enhanced Detection Model and Joint Scoring Strategy for Multi-Vehicle Tracking

**ZIYI ZHAO[1], (Member, IEEE), ZHONGPING JI[1], YINGBIAO YAO[2], ZHIWEI HE[2], (Member, IEEE), AND CHENJIE DU[2,3], (Member, IEEE)**

[1]School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China
[2]Faculty of Electronics Information, Hangzhou Dianzi University, Hangzhou 310018, China
[3]Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China

Corresponding author: Chenjie Du (ducj@hdu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61571394 and Grant 62001149, in part by the Key Research and Development Program of Zhejiang Province under Grant 2020C03098, and in part by the Natural Science Foundation of Zhejiang Province under Grant LY22F020025.

**ABSTRACT** Multi-vehicle tracking is one of the most crucial components of an intelligent transportation system (ITS). However, when it comes to busy traffic flow, tracking targets robustly becomes more problematic due to occlusion, motion blur, high appearance similarity, etc. To achieve accurate and efficient tracking performance, we present a novel multi-vehicle tracking method based on the enhanced detection model and joint scoring strategy. Specifically, the former aims to (1) adopt lightweight yet efficient YOLOv5s to improve detection accuracy and running speed, and (2) incorporate the CBAM and transformer encoder modules into the detection model to generate the refined features for the target localization. The Latter preferentially provides high-confidence detections and tracklets for subsequent data association, significantly reducing the number of identity switches and redundant vehicle trajectories caused by mutual occlusion, similar object interference, etc. We evaluated the proposed multivehicle tracking approach on the UA-DETRAC vehicle tracking dataset and demonstrated its superior capabilities through intensive comparison and analysis. Moreover, our proposed method runs at 24.4 FPS on a single GPU and meets the real-time requirement.

**INDEX TERMS** Multi-vehicle tracking, YOLOv5s, transformer encoder, joint scoring strategy, high confidence.

## I. INTRODUCTION

Intelligent transportation system (ITS), which aims to consistently and accurately track numerous moving vehicles in realistic traffic scenes, has been one of the most important research areas [1], [2]. A robust and reliable ITS plays a vital role in numerous applications, such as visual surveillance, autonomous driving, and traffic flow estimation [3], [4], [5]. Multi-object tracking (MOT) based on vision, which generally follows the tracking-by-detection paradigm, has been a critical technique for ITS. Specifically, the detection stage is performed to localize target locations frame-by-frame. Then the tracking stage is carried out to associate these detection results for generating target trajectories across video frames [6]. Thus far, designing a robust MOT approach for ITS is still challenging. As illustrated in Fig. 1, the tracker undergoes the following challenging factors: occlusion, motion blur, viewpoint change, etc. This paper aims to alleviate the tracking deviation caused by unknown challenges in real-world traffic scenarios.

With the noticeable progress of deep learning, a huge variety of deep neural network-based trackers have been proposed for MOT tasks [7]. Among them, one-shot MOT methods, which simultaneously accomplish target detection and identity embedding re-identification [8], have begun to gain significant attention. Considering that one-short MOT approaches have the advantages of high reliability and low computation cost, they are very suitable for multi-vehicle tracking in practical traffic scenarios. As an outstanding representative of one-short MOT, the joint detection and

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

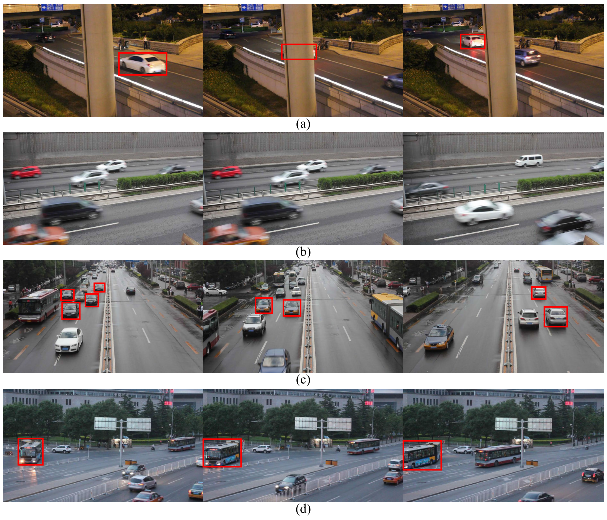



**FIGURE 1. Examples of challenging factors in real-world multivehicle tracking. (a) Occlusion. (b) Motion blur. (c) Similar object interference. (d) Varying viewpoints.**

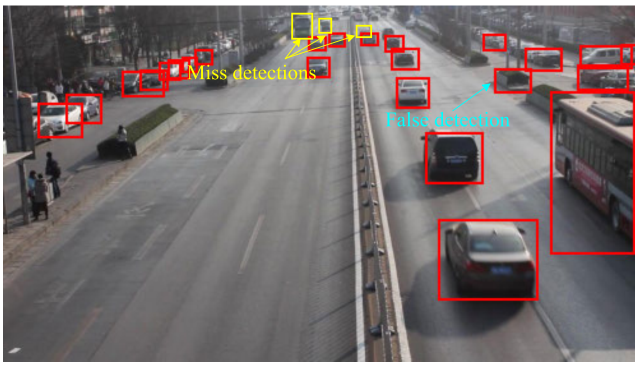



**FIGURE 2. The false and missed detection results of JDE.**

embedding (JDE) [9] develop a shared model that explicitly learns target detection and appearance embedding. Compared to the recent progress on MOT, the computation overhead of JDE is substantially decreased, which achieves real-time multiple target tracking.

Although the performance of JDE is superior, three main issues must be resolved. First, the JDE tracker may be confused by high appearance similarity among vehicles and can not easy to detect vehicles with small sizes, thereby inevitably increasing the number of false detections and missed detections over time, as illustrated in Fig. 2. Second, some extracted appearance and motion features, which may be redundant, cannot help to localize targets. Meanwhile, the JDE is prone to drift when encountering similar distractors (e.g., billboards, traffic lights). Third, frequent occlusion and interaction among vehicles in crowded traffic scenes may produce numerous identity switches, resulting in overall performance deterioration.

To improve the tracking accuracy of the JDE while maintaining an acceptable frame rate in real-world traffic scenarios, we present a novel multi-vehicle tracking approach and provide promising solutions, as shown in Fig. 3. Specifically, we investigate that the main problem in complicated scenes is the limited detection performance of the underlying detector, producing a series of false and missed detections. With this in mind, we adopt lightweight yet efficient YOLOv5s instead of JDE's detector to enhance the target detection capability. Nonetheless, we find that the detections yielded by directly utilizing the YOLOv5s are still unsatisfactory. To further boost the accuracy, we seamlessly integrate the CBAM and transformer encoder modules in our detection framework, thus enhancing the informative features and suppressing irrelevant yet confusing ones (e.g., the complicated background). As a result, the enhanced detection model regresses more precise target locations and is robust to interference information. Furthermore, we design an effective joint scoring strategy to evaluate the confidence of the detections and tracklets, preferentially pushing high-confidence detections and tracklets to the later data association stage. It is beneficial for decreasing the number of identity switches and improving identity preservation in complex interactions among vehicles. Meanwhile, the number of redundant vehicle trajectories can be effectively reduced.

Benefit from the proposal refinement, our proposed method achieves 22.7 PR-MOTA, 33.1 PR-MOTP, and 483.3 PR-IDs on the UA-DETRAC benchmark at 24.4 FPS, which outperforms state-of-the-art MOT methods in terms of both effectiveness and efficiency. In summary, the main contributions of this paper are three-fold:

- We introduce a novel enhanced detection model that integrates the plug-and-play CBAM, transformer encoder, and YOLOv5s into a unified network structure, significantly enhancing the network detection capability.
- We design a joint scoring strategy to preferentially provide high-confidence detections and tracklets for subsequent data association, effectively decreasing the number of identity switches and redundant vehicle trajectories.
- Our proposed method achieves superior tracking accuracy while maintaining high efficiency on a generic vehicle tracking benchmark.

The remainder of this paper is organized as follows. We first provide an overview of related work in Section II. Section III presents the specific design of the proposed approach in detail. The quantitative and qualitative experiments are presented in Section IV. Finally, Section V concludes this paper and future work.

## II. RELATED WORKS

In this section, we briefly review the existing MOT methods, which can be classified into three categories: classical MOT approaches, one-shot MOT approaches, and two-stage MOT approaches.

### A. CLASSICAL MOT APPROACHES

Classical MOT approaches mainly utilize traditional feature extractors to determine whether the tracked targets have appeared. Shu et al. [10] proposed the support vector

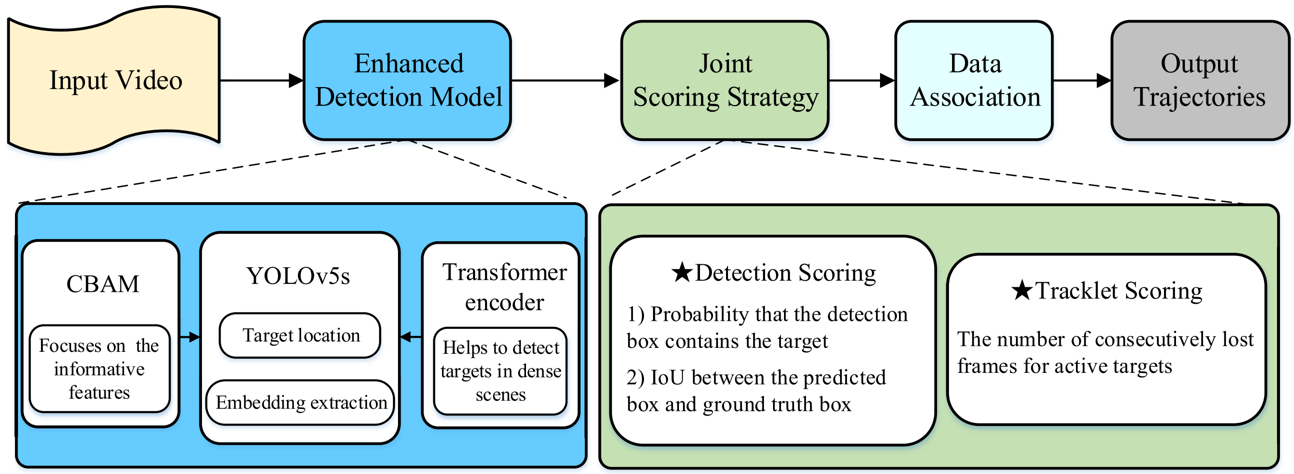



**FIGURE 3.** **The flowchart of the proposed approach.**

machine (SVM) classifier to handle occlusion between targets dynamically. Rezatofighi et al. [11] utilized joint probabilistic data association to tackle the uncertainty in association conditions. References [12] and [13] dealt with tracklet fragments based on the tracklet confidence and tackled similar object inference by discriminative appearance learning. Henschel et al. [14] tackled the graph labeling problem in the MOT system by fusing the head and full-body detectors. However, the appearance and motion features extracted by these classical approaches are not robust to occlusion and background clutter. Moreover, these trackers may fall short of distinguishing targets with high similarities. With the advent of deep learning, the MOT task is immediately dominated by the convolutional neural networks (CNNs)-based trackers, which can be classified into two-stage MOT approaches and one-shot MOT approaches.

### B. TWO-STAGE MOT APPROACHES

The two-stage MOT approaches primarily rely on two steps: 1) adopting the CNN-based detector to localize the objects of interest by a series of detect boxes, and then 2) cropping the image patches and feeding them to the identity embedding network for Re-ID feature extraction. For the detection part, Zakria et al. [15] achieved excellent results in processing remote sensing images through the introduction of a modified version of the YOLOv4. Additionally, inspired by Faster R-CNN [16], a novel evolving framework [17] was proposed to generate refined object boxes. In terms of feature extraction, a multi-level feature extraction approach [18] and a dataset augmentation methodology [19] were proposed to enhance the efficacy of the generated Re-ID. Simultaneously, numerous multi-target tracking methodologies have been proposed by integrating detection and feature extraction. The simple online and real-time tracking (SORT) proposed by Bewley et al. [20] performed favorably at a high frame rate. Wojke et al. [24] extended the work of [20], which exploited the Faster R-CNN to produce proposal detections and then associated them through a match strategy. Building on similar concepts, Tran et al. [21] improved the precision of target recognition and tracking by combining DeepSORT [22] and Yolov7 [23]. Meanwhile, a deep affinity network (DAN) [25] was employed to track the vehicles based on the generated object boxes. Following these works, Zhou et al. [26] proposed a novel dual-direction unit tensor power iteration to address the matching model issue. In the RAR16wVGG [27], the recurrent autoregressive network (RAN) coupled an external memory responsible for storing previous vehicle trajectories in the time window and an internal memory responsible for associating detections. Mahmoudi et al. [28] applied robust CNN-based features and a groups-based affinity measure to improve overall performance. However, However, due to the high computing cost, achieving real-time tracking with two-stage MOT approaches can be challenging in practical applications.

### C. ONE-SHOT MOT APPROACHES

With the flourishing development of multi-task learning, the recent research trend is heading towards applying one-short MOT approaches, which jointly treat detection and feature extraction to improve overall efficiency. This is achieved by extracting target features to depict the appearance and motion information in the current frame, thus inherently using it for tracking. JDE [9] was the first to integrate object detection and appearance embedding in a unified network, which obtained promising tracking accuracy and a high frame rate. Following [9], Voigtlaender et al. [29] proposed the Track-RCNN to jointly tackle the multi-object tracking and segmentation (MOTS) task with a single network, effectively incorporating temporal information and linking target identities as time passes on. Later, the research work in [30] designed a point-based model, namely CenterTrack, to calculate the offsets of detections and tracks according to the heatmap from the target center. Chained-Tracker [31] focused on the chained structure and attentive regression to output a pair of detection boxes in the adjacent frames. Lu et al. [32] presented a simple yet effective joint model referred to as RetinaTrack, which defined both detection and tracking as critical tasks. Additionally, the instance-level features were extracted to track the targets by modifying the one-shot RetinaNet. By integrating object detection and feature extraction in a unified framework, one-shot MOT approaches achieve competitive tracking results while being faster and involving less computation overhead. In this paper, we develop a novel one-shot MOT approach based on a more efficient architecture, achieving better tracking accuracy and higher efficiency than the one-shot approaches.

## III. THE PROPOSED METHOD

This section describes a detailed explanation of the proposed MOT method, which comprises two branches: enhanced detection model and joint scoring strategy.

### A. ENHANCED DETECTION MODEL

#### 1) YOLOv5s

Since the reliability of the tracking-by-detection paradigm heavily depends on the performance of a detector, we employ the YOLOv5s to distinguish multiple target vehicles in heavy traffic situations. The reasons why we choose the YOLOv5s are as follows. First, the YOLOv5s, which serves as a generic detector, can precisely localize moving vehicles in busy traffic flow. Second, because of the high efficiency of the YOLOv5s, our model can locate and track vehicles almost in real-time. Third, the lightweight YOLOv5s has vital portability, which can be deployed on unmanned air vehicle (UAV), vehicle-mounted cameras, home surveillance, etc.

As shown in Fig. 4, the network structure of the YOLOv5s is composed of input, backbone, neck, and prediction. During the input phase, the YOLOv5s employs k-means clustering to adaptively calculate the optimal anchor according to different classes of targets, making the network easier to choose better priors. The focus layer, which is embedded with the backbone network of the YOLOv5s, aims to reserve more complete downsampling target features by slice operation compared to the conventional convolution operation. Two cross-stage partial connections (CSP) structures are applied to the backbone and neck, which contribute to fusing the feature layers of different stages and then realizing the multi-scale feature maps. Based on the fact that the intersection over union (IoU) metric is incapable of dealing with non-overlapping bounding boxes, a novel generalized IoU (GIoU) metric [33] is employed to calculate the distance of two arbitrary convex shape boxes. Besides, the YOLOv5s utilizes the GIoU as bounding box regression loss, significantly improving the localization precision. The GIoU loss is defined as follows:

$$L_{GIOU} = 1 - \left(\frac{\left|R_p \cap R_g\right|}{\left|R_p \cup R_g\right|} - \frac{\left|R_m / (R_p \cup R_g)\right|}{|R_m|}\right) \tag{1}$$

where $R_p$ and $R_g$ are the predicted bounding box and ground truth box, respectively, and $R_m$ is the minimum enclosing rectangle surrounding $R_p$ and $R_g$. In addition, we adopt the SEBottleneck [34] to optimize the structure of CSP further. As shown in Fig. 5, this module enables the network to distinguish feature information more effectively by learning the weight coefficients of each channel.

#### 2) CBAM

As shown in Fig. 6, we seamlessly integrate the CBAM [35] into the network structure of the detector to highlight the informative features and suppress the redundant ones, thus generating refined features for localizing the target location.

Precisely, we separately execute average-pooling and max-pooling on the global features $U \in \mathbb{R}^{H \times W \times C}$, and then concatenate them in the channel dimension. After that, we execute dimensionality reduction through the first fully connected layer (Fc1) and activate it using the ReLU function. The second FC layer (Fc2) encodes the features to decrease the computational burden by compressing the $C$ channels into the $C/r$ channels, where r is the reduction ratio. Thereafter, the third FC layer (Fc3) restores the channel number of the features to $C$ channels, and utilize the sigmoid activation to obtain the required channel weight $M_C$ that represents the importance of different channels:

$$M_C = \sigma(g(z, W)) = \sigma(W_3\delta(W_2\delta(W_1 z))) \tag{2}$$

where $\delta$ and $\sigma$ refer to the ReLU and sigmoid functions respectively, and $W_1$, $W_2$, and $W_3$ represent the parameters of three FC layers respectively. Thereby, we can yield the recalibrated features $\tilde{U} \in \mathbb{R}^{H \times W \times C}$ as follows:

$$\tilde{U} = (1 + M_C)U \tag{3}$$

where $\tilde{U} = \left[\tilde{u}_1, \tilde{u}_2, \cdots, \tilde{u}_C\right]$and $M_C = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_C]$ refer to the channel-wise weight. Thus, we exploit the channel attention mechanism to increase the discriminability of features across different channels significantly.

After taking the features $\tilde{U}$ from the channel attention module, two pooled features are concatenated and then convolved by a general convolution operation to generate the spatial-wise weight $M_S$. Finally, the refined features can be formulated as follows:

$$\hat{U} = (1 + M_S)\tilde{U} \tag{4}$$

By accessing the spatial attention module, the features are further recalibrated in the spatial dimension, selectively highlighting the features of effective regions and suppressing the features of interference regions.

Therefore, we can effectively learn which feature information to highlight or suppress by using the complementary CBAM and transformer encoder modules. On the other hand, we achieve the detection accuracy boost via seamlessly integrating simple yet effective attention modules and YOLOv5s into a unified framework.

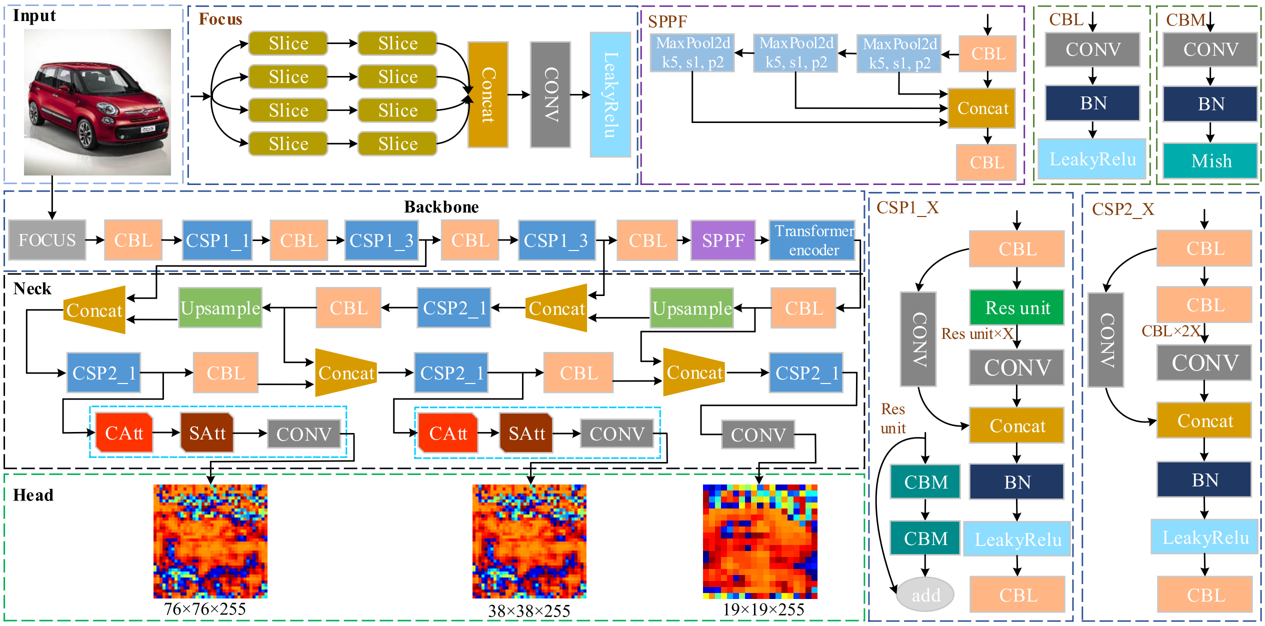



**FIGURE 4.** The network structure of enhanced detection model.

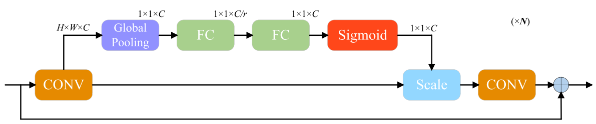



**FIGURE 5.** The structure of the SEBottleneck.

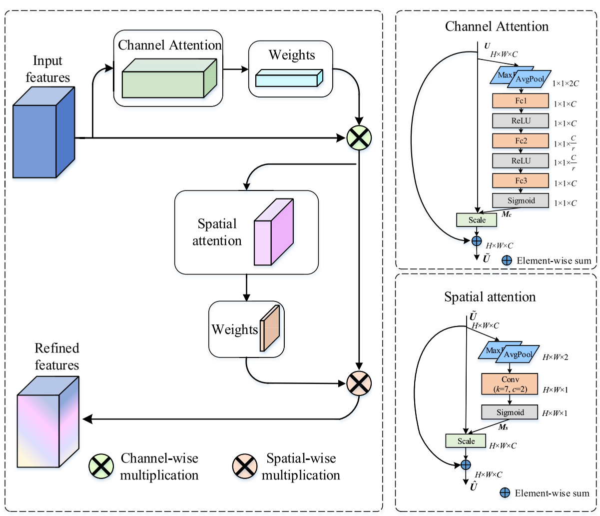



**FIGURE 6.** The introduced CBAM module makes the network focus on informative features.

### 3) TRANSFORMER ENCODER

We design a transformer encoder module and add it to the final layer of the Backbone, as shown in Fig. 7 The transformer encoder is mainly composed of multi-head attention and multi-layer perceptron (MLP). The multi-headed attention essentially executes multiple attention layers in parallel and then concatenates their outputs together. The MLP maps the features from a low-dimensional space to a high-dimensional space, and then compresses the sparse features to make them more stable. The transformer encoder plays an active role in detecting targets in dense scenes and reduces the expensive computation and memory costs.

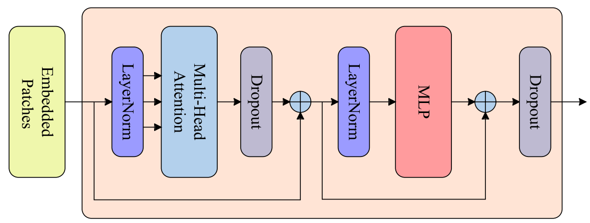



**FIGURE 7.** The structure of the transformer encoder.

## B. JOINT SCORING STRATEGY

In busy traffic scenarios, it is difficult to maintain a consistent identity due to numerous interactions among vehicles. Meanwhile, various scenes where other things (e.g., billboards, traffic lights) occlude the tracked vehicles often occur, which will produce many undesired identity switches. Additionally, plenty of redundant trajectories will be generated when the tracked vehicles leave or enter the view. To redress the above oversight, we propose a novel joint scoring strategy to filter out unreliable detections and tracklets and preferentially push high-confidence detections and tracklets to the later data association stage, which assists in maintaining target identities during the tracking duration.

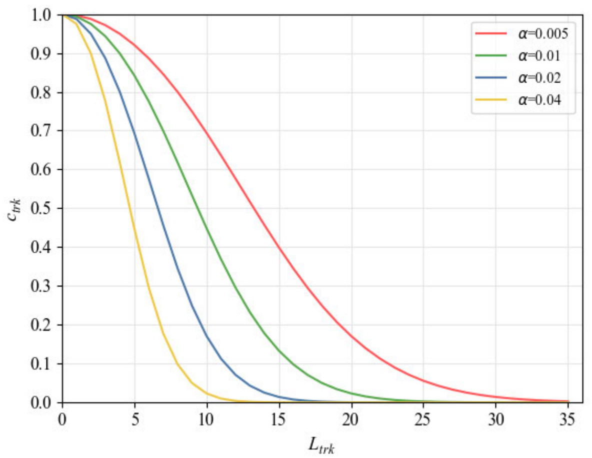



**FIGURE 8.** The tracklet confidence under the different sizes of $L_{trk}$.

#### 1) DETECTION SCORING

The detection confidence $c_{det}$ that evaluates the detections is defined as follows:

$$c_{det} = Pr(object) \times IoU(pred, gt) \tag{5}$$

where $Pr(object)$ represents the probability of whether the detection box contains the target. And $IoU(pred, gt)$ is the intersection-over-union between the region of predicted box $R^{pred}$ and the region of the ground-truth box $R^{gt}$, which can be computed as follows:

$$IoU(pred, gt) = \frac{\left|R^{pred} \cap R^{gt}\right|}{\left|R^{pred} \cup R^{gt}\right|} \tag{6}$$

The detection whose $c_{det}$ is less than the preset threshold $\delta_{det}$ is unsuitable for participating in follow-up data association and should be removed.

#### 2) TRACKLET SCORING

Here, we define $L_{trk}$ as the number of consecutively lost frames for a target. We define the temporal information-based tracklet confidence $c_{trk}$ as follows:

$$c_{trk} = \frac{4}{\pi} \arctan\left(e^{-\alpha L_{trk}^2}\right) \tag{7}$$

where $\alpha$ is a hyperparameter of the proposed tracklet scoring. An example of the tracklet confidence under different $L_{trk}$ is illustrated in Fig. 8. As we can see from Fig. 8, the smaller $L_{trk}$ indicates that the tracker has a higher probability of relocating the re-appear target after it suffers cover by the other distractors. In this case, we set such a tracklet to high confidence. When the $L_{trk}$ gradually increases, the probability of tracklet recovery in subsequent frames gradually decreases. Particularly, if the calculated confidence score $c_{trk}$ is less than the preset threshold $\delta_{trk}$, the active tracklet has been lost for a long time. We thus prevent the unreliable tracklet from participating in the subsequent data association.

For the remaining detections and tracklets, we will preferentially match the high-confidence detections and tracklets for optimizing the data association process.

### C. DATA ASSOCIATION

For the $i$-th detection and the $j$-th tracklet, the appearance features are extracted and denoted as $r_i$ and $v_j$. In addition to the visual features, we also consider the motion features by calculating the $i$-th detection box location $d_i$ and the $j$-th tracklet distribution $(y_j, S_j)$. Subsequently, we compute the assignment cost between the pair of $i$-th detection and $j$-th tracklet as follows:

$$s_{i,j} = 1 - r_i^T v_j + \lambda\left((d_i - y_i)^T S_j^{-1}(d_i - y_j)\right) \tag{8}$$

where $\lambda$ is a combination factor, and $T$ represents a transpose operation. According to the calculated assignment cost, we utilize the Hungarian algorithm [36] to associate detections to tracklets for generating reliable target trajectories. Then we assign a numerical ID to each specific target in the given video frame. For the matched tracklet, we update its motion state using the Kalman filter [37], and the appearance state in frame $t$ is updated as follows:

$$f^t = (1 - c_{det})f^{t-1} + c_{det} r^t \tag{9}$$

where $c_{det}$ and $r_t$ represent the detection confidence and the appearance features of the current matched detection, respectively. For the remaining detections that are not associated with any tracklet, we initialize new tracklets based on the location of detections. To sum up, we present the specific steps of the proposed MOT algorithm, as shown in Algorithm 1.

## IV. EXPERIMENTAL RESULTS AND ANALYSES

In this section, we first present the dataset, evaluation metrics, and implementation details in section IV-A. Then we compare our tracking method with several state-of-the-art methods in section IV-B. In section IV-C, we execute detailed ablation studies that emphasize the novelty of this work. In section IV-D, the discussion about our experiment is presented.

### A. EXPERIMENTAL SETUP

#### 1) BENCHMARK DATASET

We implement a detailed comparison experiment on a large-scale dedicated benchmark called UA-DETRAC [38], which is widely utilized to evaluate the tracking performance of MOT methods in traffic scenes. This dataset is composed of 100 video sequences with over 140K frames in total. Additionally, it contains 8,520 vehicles and 1.21 million densely labeled bounding boxes. Each video sequence comprises diverse challenges deriving from realistic traffic situations, such as occlusion, background clutter, viewpoint change, etc. In particular, several vehicles in various traffic scenarios (e.g., traffic junctions, urban highways) perhaps enter or leave the view at any time, hence increasing the difficulty of vehicle detection and tracking.

**Algorithm 1** The Proposed MOT Algorithm

**for** *frame* $= 1, \cdots, t$ **do**
  For $k$-th target of the input image $I_t$, estimate the target location $x_k^t$ and extract appearance features $r_k^t$;
  Calculate the detection confidence $c_{det}^t$ of each target (Eq. 5);
  **if** $c_{det}^t < 0.3$ **then**
    Remove this detection;
  **end if**
  **for** *each tracklet* **do**
    Calculate the tracklet confidence $c_{trk}^t$ (Eq. 7);
    **if** $c_{trk}^t < 0.1$ **then**
      Remove this tracklet;
    **end if**
    Predict new location $\hat{x}_k^t$ of tracklet using Kalman filter;
  **end for**
  Calculate the assignment cost $s_{i,j}$ between the $i$-th detection and the $j$-th tracklet;
  Associate each detection and tracklet using Hungarian algorithm and assign a numerical ID to each specific object;
  **for** *the tracklets associated with detections* **do**
    Update the estimated location using Kalman filter;
    Update the appearance state of tracklets;
  **end for**
  **for** *the remaining detections that are not associated with any tracklet* **do**
    Initialize new tracklets based on the location of detections;
  **end for**
**end for**

### 2) EVALUATION METRICS

Considering the effect of detection performance on the MOT system, we use the UA-DETRAC protocol for the overall performance evaluation, which is slightly different from the commonly used classification of events, activities and relationships (CLEAR) MOT metrics [39]. The UA-DETRAC metrics, which reflect the overall performance of trackers in the vehicle tracking task, are defined as follows:

- **PR-MOTA**: The PR-MOTA curve first characterizes the relationship between target detection performance and tracking performance. The PR-MOTA can be obtained by calculating the average MOTA score over the precision vs. recall (PR) curve. The PR-MOTA is generally selected as the primary evaluation metric.
- **PR-MOTP**: The misalignment between the predicted box and the ground-truth box over PR curve.
- **PR-MT**: The percentage of ground-truth trajectories that are correctly tracked in at least 80% of their life cycle over PR curve.
- **PR-ML**: The percentage of ground-truth trajectories that are correctly tracked in at most 20% of their life cycle over PR curve.
- **PR-IDs**: The number of the associated ID for the target is mistakenly changed over PR curve.
- **PR-FP**: The number of false positives over PR curve.
- **PR-FN**: The number of false negatives over PR curve.
- **FPS**: The overall tracking speed in the vehicle tracking scenes.

### 3) IMPLEMENTATION DETAILS

We train our model in an end-to-end manner within the smooth-L1 and cross-entropy loss. The CSPDarkNet-53 network is utilized as the backbone. Meanwhile, we train our model with standard stochastic gradient descent (SGD) for 50 epochs, and the batch size is set to 4. The learning rate is initialized as $10^{-3}$ and is reduced by a factor of 0.1 every 50,000 iterations. The training usually converses after 26 epochs. Additionally, we execute several data augmentation strategies (e.g., random rotating, random scaling, photometric distortion) to reduce overfitting. Additionally, 1) In section III-B, we set the hyperparameter $\alpha$ to 0.05. The thresholds $\delta_{det}$ and $\delta_{trk}$ are set to 0.3 and 0.1, respectively. 2) In section III-C, the combination factor $\lambda$ in Eq. (8) is set to 0.1. All experiments are implemented with PyCharm 2020.2 on a PC with i5-10600KF CPU and NVIDIA Geforce RTX 3070 GPU. The programming language is Python 3.10. The tracking speed of our method on the UA-DETRAC test sequence is 24.4 FPS on a single GPU.

## B. COMPARISON WITH THE STATE-OF-THE-ART METHODS

On the UA-DETRAC dataset, we compare the proposed method against the state-of-the-art MOT methods, including JDE [9], Chained-Tracker [31], EB [17]+DAN [25], EB [17]+SiamIOU [40], EB [17]+IOUT [41], R-CNN [42]+IOUT [41], Faster R-CNN [16]+DeepSORT [24], CompACT [43] +FAMNet [44], CompACT [43]+GOG [45], CompACT [43]+CMOT [12], CompACT [43]+H2T [46], R-CNN [42]+DCT [47], CompACT [43]+ IHTLS [48] and CompACT [43]+CEM [49]. All the compared methods are trained on the UA-DETRAC-train set and evaluated on the UA-DETRAC-test set. The comprehensive quantitative results of the compared approaches are summarized in Table 1. The best and second-best results are in bold and underlined, respectively.

As we can see from Table 1, the proposed method achieves 22.7 PR-MOTA, significantly outperforming existing methods. For instance, the proposed method obtains up to 5.6% relative improvements in PR-MOTA over the suboptimal approach. Moreover, the proposed method performs favorably over the state-of-the-art in terms of PR-MT, PR-ML, PR-FP, and PR-FN on the UA-DETRAC dataset. Additionally, we compare the computation cost of our proposed method with other methods, as shown in the FPS column of Table 1. Our tracker runs at 24.4 FPS, which is faster than most compared methods. In short, our proposed method has advantages over the compared methods in multiple performance indicators and is amenable to ITS demanding real-time tracking. Additionally, Fig. 9

**TABLE 1.** Quantitative results by our method and state-of-the-art methods on the UA-DETRAC dataset. ↑ Denotes that higher is better and ↓ represents the opposite. The best and second best results are in bold and underline, respectively.

| Backbone | PR-MOTA↑ | PR-MOTP ↑ | PR-MT ↑ | PR-ML ↓ | PR-IDs ↓ | PR-FP ↓ | PR-FN ↓ | FPS ↑ |
|---|---|---|---|---|---|---|---|---|
| JDE [9] | 18.9 | 31.2 | 13.9 | 21.3 | 602.4 | 13002.7 | 137428.3 | 19.1 |
| Chained-Tracker [31] | 20.1 | 30.3 | 12.8 | 23.9 | 616.8 | 10756.8 | 126235.7 | 28.6 |
| EB [17]+DAN [25] | 20.2 | 26.3 | 14.5 | 18.1 | 518.2 | 9747.8 | 135978.1 | 6.3 |
| EB [17]+SiamIOU [40] | 21.5 | 28.6 | 23.0 | - | 479.9 | 21137.8 | 169095.0 | 20.1 |
| EB [17]+IOUT [41] | 19.4 | 28.9 | 17.7 | 18.4 | 2311.3 | 14796.5 | 171806.8 | **6902.1** |
| R-CNN [42]+IOUT [41] | 16.0 | **38.3** | 13.8 | 20.7 | 5029.4 | 22535.1 | 193041.9 | - |
| Faster R-CNN [16]+DeepSORT [24] | 17.3 | 30.6 | 12.7 | 23.4 | 563.2 | 15201.2 | 142320.8 | 8.9 |
| CompACT [43]+FAMNet [44] | 19.8 | 36.7 | 18.2 | - | 617.0 | 14989.0 | 164433.0 | - |
| CompACT [43]+GOG [45] | 14.2 | 37.0 | 13.9 | 19.9 | 3334.6 | 32092.9 | 180183.8 | 389.5 |
| CompACT [43]+CMOT [12] | 12.6 | 36.1 | 16.1 | 18.6 | 285.3 | 57885.9 | 167110.8 | 3.8 |
| CompACT [43]+H2T [46] | 12.4 | 35.7 | 14.8 | 19.4 | 852.2 | 51765.7 | 173899.8 | 3.0 |
| R-CNN [42]+DCT [47] | 11.7 | 38.0 | 10.1 | 22.8 | 758.7 | 336561.2 | 210855.6 | 0.7 |
| CompACT [43]+IHTLS [48] | 11.1 | 36.8 | 13.8 | 19.9 | 953.6 | 53922.3 | 180422.3 | 19.8 |
| CompACT [43]+CEM [49] | 5.1 | 35.2 | 3.0 | 35.3 | **267.9** | 12341.2 | 260390.4 | 4.6 |
| **Ours** | **22.7** | 33.1 | 21.1 | **18.1** | 483.3 | **9445.1** | **110038.0** | 24.4 |

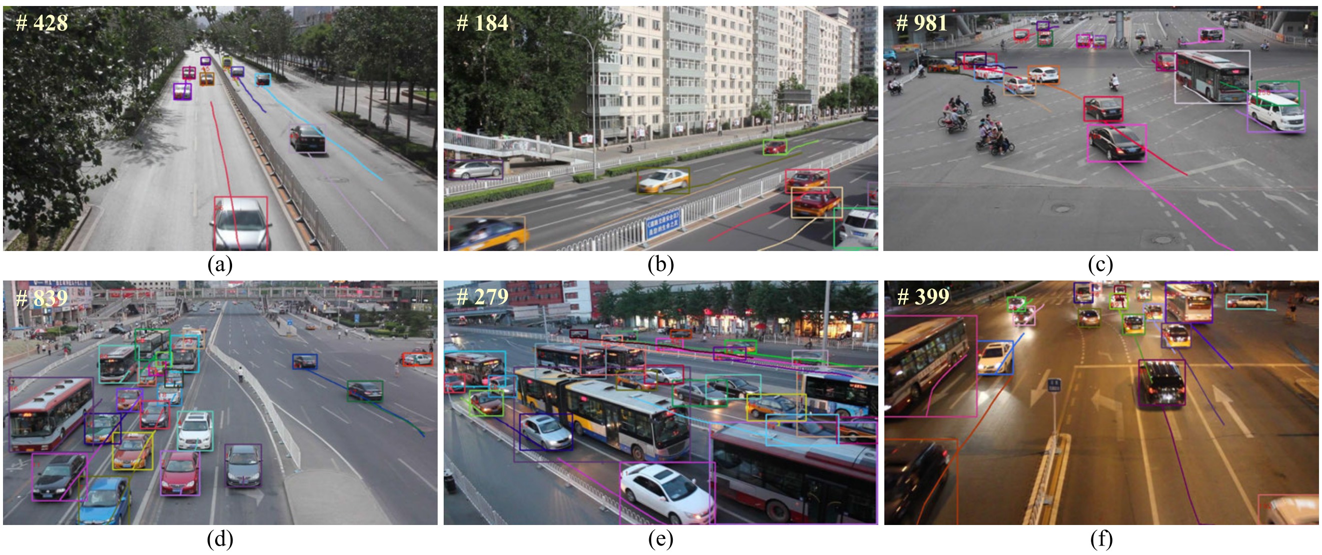



**FIGURE 9.** Qualitative results of our tracker on UA-DETRAC test dataset. (a) MVI 39031. (b) MVI 39371. (c) MVI 40701. (d) MVI 40714. (e) MVI 40742. (f) MVI 40771.

provides the exemplary output of the proposed approach on six challenging sequences, including MVI 39031, MVI 39371, MVI 40701, MVI 40714, MVI 40742, and MVI 40771. As we can see from Fig. 9, our proposed method achieves high tracking accuracy and augment the robustness of the tracker, making our system applicable to busy traffic scenarios.

### C. ABLATION STUDIES

To validate the effectiveness of each component in our model, we perform extensive ablation studies, as shown in Table 2. 1) *Baseline+enhanced detection model* performs better than *baseline*, which proves the effectiveness of the proposed detection model. There is an improvement in PR-MOTA, which increases from 18.9 to 21.4. Since the attention modules are integrated within YOLOv5s to enhance detection capacity, we can regress target locations more accurately compared to JDE. 2) *Baseline+enhanced detection model+joint scoring strategy* further outperforms *baseline+enhanced detection model*. Using the joint scoring strategy, we can ensure that the high-confidence detections and tracklets preferentially participate in the later data association, reducing the number of identity switches and redundant vehicle trajectories. Compared to the ablation study (case 1), we obtained the PR-MOTA improvement by about 6.1%. Particularly, the significant decline in PR-IDs demonstrates that our joint scoring strategy is beneficial for maintaining target identities. Therefore, the comparison with the basic JDE shows the utility of our enhanced detection model and joint scoring strategy to localize targets and improve identity preservation in complex traffic scenes.

Furthermore, to confirm the impact of the CBAM and transformer encoder modules in the enhanced detection model, we also perform an ablation experiment on the detection component, as presented in Table 3. 1) *YOLOv5s + CBAM* outperforms *YOLOv5s* in terms of accuracy and

**TABLE 2.** Ablation Studies on the UA-DETRAC Dataset. ↑ Denotes that higher is better and ↓ represents the opposite. The best result is in bold.

| Description | Enhanced detection model | Joint scoring strategy | PR-MOTA ↑ | PR-MOTP ↑ | PR-MT ↑ | PR-ML ↓ | PR-IDs ↓ | PR-FP ↓ | PR-FN ↓ | FPS ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| JDE(*baseline*) | × | × | 18.9 | 31.2 | 13.9 | 21.3 | 602.4 | 13002.7 | 137428.3 | 19.1 |
| **Ours**(*case 1*) | ✓ | × | 21.4 | 32.8 | 18.4 | 19.6 | 554.4 | 9674.4 | 112628.9 | 23.5 |
| **Ours**(*case 2*) | ✓ | ✓ | **22.7** | **33.1** | **21.1** | **18.1** | **483.3** | **9445.1** | **110038.0** | **24.4** |


(a) (b)

**FIGURE 10.** The detection results on MVI 20034 challenging sequence at frame 156. (a) JDE. (b) Ours. Our proposed method can react to some far-away vehicles and rectify the false detection.

**TABLE 3.** Ablation Studies on the detection components using UA-DETRAC Dataset. ↑ Denotes that higher is better and ↓ represents the opposite. The best result is in bold.

| Description | CBAM | transformer encoder | Precision ↑ | Recall ↑ | mAP@0.5 ↑ | mAP@0.5:0.95 ↑ |
|---|---|---|---|---|---|---|
| YOLOv5s | × | × | 0.776 | 0.714 | 0.672 | 0.519 |
| **Ours**(*case 1*) | ✓ | × | 0.778 | 0.721 | 0.682 | 0.526 |
| **Ours**(*case 2*) | ✓ | ✓ | **0.799** | **0.748** | **0.697** | **0.549** |

effectiveness. 2) *YOLOv5s + CBAM + transformer encoder* boosts the detection effect even more. The test results show that the detection performance of the model is improved by integrating YOLOv5s, CBAM and the transformer encoder into a unified network.

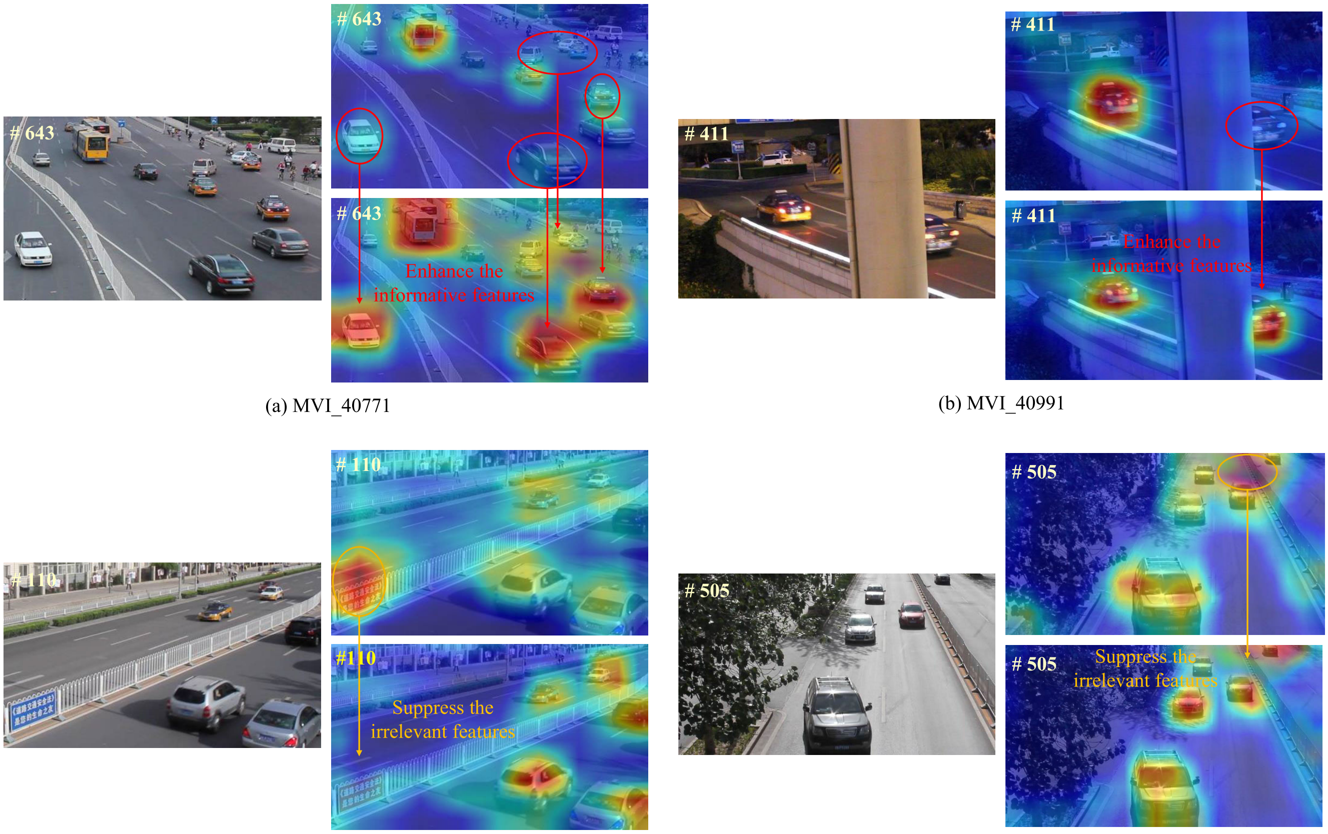
In addition to the above, we evaluate significant detection performance in busy traffic flow. The detection results of the proposed method and JDE on the MVI 20034 challenging sequence are shown in Fig. 10. As we can see from Fig. 10 (a), the false and missed detections occur due to the occlusions and similar object interference in dense clutter. As shown in Fig. 10 (b), by integrating the YOLOv5s and attention modules into a unified framework, we enhance the capability of detecting vehicles with small sizes and rectifying false detection, which is essential for ITS. Moreover, our model has less computing cost and is hence fast as compared to JDE. Additionally, to further verify the effectiveness of the CBAM and transformer encoder modules, we compare the heatmaps of our detection model and YOLOv5s, as shown in Fig. 11. By introducing the simple attention modules, we can enhance the informative features and suppress the irrelevant ones, effectively improving the discriminative capability of our method. To sum up, the ablation studies based on the UA-DETRAC benchmark indicate that our proposed method achieves a higher tracking accuracy while maintaining a higher speed than baseline JDE.

**TABLE 4.** The number of model parameters comparison.

| Description | Number of model parameters |
|---|---|
| JDE(*baseline*) | 7.30802e+07 |
| **Ours** | 1.18691e+07 |

### *D. DISCUSSION*

Lastly, according to the above quantitative and qualitative evaluations of vehicle tracking performance on the UA-DETRAC benchmark, we can conclude that: 1) As a pipeline depending on a lightweight yet efficient detection network, our method not only reduces false detections but also improves the detection ability of small targets; 2) The number of model parameters of our method and JDE is shown in Table 4. As we can see from Table 4, our method's model parameters are significantly reduced compared with JDE. This is because the model parameters of the lightweight YOLOv5s are much less than those of the JDE's detector. In addition, the number of model parameters brought by the attention modules is also tiny; 3) Considering the real-




(a) MVI_40771

(b) MVI_40991

(c) MVI_39371

(d) MVI_39031

**FIGURE 11.** **The heatmap comparison of our method and JDE.**

time implementation requirement of the intelligent traffic system, the proposed method achieves the tracking with a high frame rate owing to the low computing overhead of our model.

## V. CONCLUSION

Multi-vehicle tracking has been widely utilized in many fields. Nevertheless, when it comes to busy traffic flow, the performance of the basic JDE tracker remains needs to be improved. This paper proposes the enhanced detection model and joint scoring strategy. First, by integrating the lightweight YOLOv5s and attention modules, the enhanced detection model can effectively enhance the target localization capability and improve detection speed. Meanwhile, the false and missed detections caused by complicated challenges are decreased. Second, according to confidence scores of detection and tracking results, we preferentially push high-confidence detections and tracklets to the later data association stage, reducing the number of identity switches and redundant vehicle trajectories. The overall performance of our proposed method performs favorably against state-of-the-art on the UA-DETRAC benchmark, which helps advance the development of autonomous driving, traffic state estimation, collision avoidance, etc. In the future, we will try to design an end-to-end network to match the detections and tracklets.

## REFERENCES


[1] D. Song, R. Tharmarasa, G. Zhou, M. C. Florea, N. Duclos-Hindie, and T. Kirubarajan, "Multi-vehicle tracking using microscopic traffic models," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 149–161, Jan. 2019.

[2] Y. Zhang, B. Song, X. Du, and M. Guizani, "Vehicle tracking using surveillance with multimodal data fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 7, pp. 2353–2361, Jul. 2018.

[3] P. Barcellos and J. Scharcanski, "Part-based object tracking using multiple adaptive correlation filters," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[4] S. Sun, Y. Yin, X. Wang, and D. Xu, "Robust visual detection and tracking strategies for autonomous aerial refueling of UAVs," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 12, pp. 4640–4652, Dec. 2019.

[5] M. Jiang, R. Sogabe, K. Shimasaki, S. Hu, T. Senoo, and I. Ishii, "500-Fps omnidirectional visual tracking using three-axis active vision system," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[6] J. Xiang, G. Xu, C. Ma, and J. Hou, "End-to-end learning deep CRF models for multi-object tracking deep CRF models," *IEEE Trans. IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 275–288, Jan. 2021.

[7] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, Mar. 2020.

[8] Zakria, J. Deng, Y. Hao, M. S. Khokhar, R. Kumar, J. Cai, J. Kumar, and M. U. Aftab, "Trends in vehicle re-identification past, present, and future: A comprehensive review," *Mathematics*, vol. 9, no. 24, p. 3162, Dec. 2021.

[9] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020, pp. 107–122.

[10] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1815–1821.

[11] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. F. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3047–3055.

[12] S.-H. Bae and K.-J. Yoon, "Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1218–1225.

[13] S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 595–610, Mar. 2018.

[14] R. Henschel, L. Leal-Taixe, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1509–1518.

[15] Z. Zakria, J. Deng, R. Kumar, M. S. Khokhar, J. Cai, and J. Kumar, "Multiscale and direction target detecting in remote sensing images via modified YOLO-v4," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1039–1048, 2022.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[17] L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue, "Evolving boxes for fast vehicle detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1135–1140.

[18] J. Zakria, J. Cai, J. Deng, M. U. Aftab, M. S. Khokhar, and R. Kumar, "Efficient and deep vehicle re-identification using multi-level feature extraction," *Appl. Sci.*, vol. 9, no. 7, p. 1291, 2019.

[19] Zakria, J. Deng, J. Cai, M. U. Aftab, M. S. Khokhar, and R. Kumar, "Visual features with spatio-temporal-based fusion model for cross-dataset vehicle re-identification," *Electronics*, vol. 9, no. 7, p. 1083, Jul. 2020.

[20] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[21] D. N.-N. Tran, L. H. Pham, H.-H. Nguyen, and J. W. Jeon, "City-scale multi-camera vehicle tracking of vehicles based on YOLOv7," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–4.

[22] B. Veeramani, J. W. Raymond, and P. Chanda, "DeepSort: Deep convolutional networks for sorting haploid maize seeds," *BMC Bioinf.*, vol. 19, no. S9, pp. 1–9, Aug. 2018.

[23] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[24] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Mar. 2017, pp. 3645–3649.

[25] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, Jan. 2021.

[26] Z. Zhou, J. Xing, M. Zhang, and W. Hu, "Online multi-target tracking with tensor-based high-order graph matching," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1809–1814.

[27] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 466–475.

[28] N. Mahmoudi, S. M. Ahadi, and M. Rahmati, "Multi-target tracking using CNN-based features: CNNMTT," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 7077–7096, Mar. 2019.

[29] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: Multi-object tracking and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7934–7943.

[30] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 474–490.

[31] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu, "Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 145–161.

[32] Z. Lu, V. Rathod, R. Votel, and J. Huang, "RetinaTrack: Online single stage joint detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14656–14666.

[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 658–666.

[34] Z.-D. Zhang, M.-L. Tan, Z.-C. Lan, H.-C. Liu, L. Pei, and W.-X. Yu, "CDNet: A real-time and robust crosswalk detection network on Jetson nano based on YOLOv5," *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10719–10730, Feb. 2022.

[35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comupt. Vis. (ECCV)*, 2018, pp. 3–19.

[36] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[37] E. Neuburger and V. Krebs, "Einführung in die theorie des linearen optimalfilters (Kalman-filter) (introduction to linear optimal filtering theory (Kalman filter))," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC–6, no. 11, p. 796, Nov. 1976.

[38] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understand.*, vol. 193, Apr. 2020, Art. no. 102907.

[39] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "MOTChallenge 2015: Towards a benchmark for multi-target tracking," 2015, *arXiv:1504.01942*.

[40] A. Li, L. Luo, and S. Tang, "Real-time tracking of vehicles with Siamese network and backward prediction," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.

[41] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2017, pp. 1–6.

[42] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[43] Z. W. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2195–2211, Sep. 2020.

[44] P. Chu and H. Ling, "FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6171–6180.

[45] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1201–1208.

[46] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multiple target tracking based on undirected hierarchical relation hypergraph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1282–1289.

[47] A. Andriyenko, K. Schindler, and S. Roth, "Discrete-continuous optimization for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1926–1933.

[48] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2304–2311.

[49] A. Andriyenko and K. Schindler, "Multi-target tracking by continuous energy minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1265–1272.

**ZIYI ZHAO** (Member, IEEE) is currently pursuing the B.S. degree with the School of Computer Science, Hangzhou Dianzi University. His research interests include deep learning and computer vision, especially on object detection and object tracking.

**ZHONGPING JI** received the B.E. degree in mathematics from Northwestern Polytechnical University, in 2003, and the Ph.D. degree in mathematics from Zhejiang University, in 2008. He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include computer vision, machine learning, and image processing. He is a member of the Geometric Design and Computing Committee and the China Society for Industrial and Applied Mathematics.

**YINGBIAO YAO** was born in Songzi, Hubei, China, in 1976. He received the B.S. and M.S. degrees in electronic engineering from Xi'an Shiyou University, Shanxi, China, in 2003, and the Ph.D. degree in information and communication engineering from Zhejiang University, Zhejiang, China, in 2006. From 2008 to 2016, he was an Associate Professor with the School of Communication Engineering, Hangzhou Dianzi University, China. In 2011, he was a Visiting Scholar with RPI University, NY, USA. Since 2017, he has been a Professor with the School of Communication Engineering, Hangzhou Dianzi University. He is the author of three books and more than 50 articles, and the inventor of more than 30 inventions. His research interests include storage system design, wireless sensor networks, indoor localization, and media signal processing. He was a recipient of the Third Level of Zhejiang Provincial ''151'' Talent for Excellence, in 2015, the first class Prize of Zhejiang Provincial Scientific and Technological Progress Award, in 2016, and the Zhejiang Provincial Young and Middle-Aged Academic Leaders, in 2017.

**ZHIWEI HE** (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in communication and information systems from Zhejiang University, in 2001 and 2006, respectively. He was a Research Assistant with Hong Kong Polytechnic University, in 2004, and a Senior Visiting Scholar with Wayne State University, in 2012. Currently, he is a Professor with Hangzhou Dianzi University, Hangzhou, China. He is the author of more than 80 articles and the inventor of more than 50 inventions. His research interests include deep learning, computer vision, and media signal processing.

**CHENJIE DU** (Member, IEEE) received the B.E. and M.E. degrees in electronics and information engineering from Hangzhou Dianzi University, Hangzhou, China, in 2013 and 2016, respectively, and the Ph.D. degree from the School of Electronic and Information, Hangzhou Dianzi University, in 2022. Currently, he is a Lecturer with Ningbo University, Ningbo, China. He has published more than ten papers in refereed journals and conferences. His research interests include media signal processing, object detection, and visual target tracking.