sustainability

MDPI

Article

Vehicle Tracking Algorithm Based on Deep Learning in Roadside Perspective

Guangsheng Han [1], Qiukun Jin [2], Hui Rong [1,*], Lisheng Jin [2] and Libin Zhang [1]

[1] CATARC (Tianjin) Automotive Engineering Research Institute Co., Ltd., No.68, Xianfeng East Road, Dongli District, Tianjin 300300, China
[2] School of Vehicle and Energy, Yanshan University, Qinhuangdao 066104, China
* Correspondence: ronghui@catarc.ac.cn

Abstract: Traffic intelligence has become an important part of the development of various countries and the automobile industry. Roadside perception is an important part of the intelligent transportation system, which mainly realizes the effective perception of road environment information by using sensors installed on the roadside. Vehicles are the main road targets in most traffic scenes, so tracking a large number of vehicles is an important subject in the field of roadside perception. Considering the characteristics of vehicle-like rigid targets from the roadside view, a vehicle tracking algorithm based on deep learning was proposed. Firstly, we optimized a DLA-34 network and designed a block-N module, then the channel attention and spatial attention modules were added in the front of the network to improve the overall feature extraction ability and computing efficiency of the network. Next, the joint loss function was designed to improve the intra-class and inter-class discrimination ability of the tracking algorithm, which can better discriminate objects of similar appearance and the color of vehicles, alleviate the IDs problem and improve algorithm robustness and the real-time performance of the tracking algorithm. Finally, the experimental results showed that the method had a good tracking effect for the vehicle tracking task from the roadside perspective and could meet the practical application demands of complex traffic scenes.

Keywords: multi-vehicle tracking; roadside perception; roadside view; one-shot; real-time

check for updates

Citation: Han, G.; Jin, Q.; Rong, H.; Jin, L.; Zhang, L. Vehicle Tracking Algorithm Based on Deep Learning in Roadside Perspective. *Sustainability* **2023**, *15*, 1950. https://doi.org/10.3390/su15031950

Academic Editor: Shuai Su

Received: 23 November 2022
Revised: 30 December 2022
Accepted: 10 January 2023
Published: 19 January 2023

1. Introduction

In recent years, Multiple Object Tracking (MOT) technology has become the focus of research in the field of computer vision [1,2]. By processing the video image information obtained by the vision sensor, the target appearance characteristics, motion and other information are obtained and the relationship between different targets at the same time is analyzed, thus forming the continuous motion trajectory of each target. According to its task characteristics, multi-target tracking technology has great practical significance in the field of automatic driving, intelligent transportation, intelligent monitoring and other fields, among which, vision-based vehicle tracking technology has been widely used in bayonet monitoring, road monitoring and other equipment [3]. The MOT algorithm can be divided into deep learning-based multi-target tracking and traditional multi-target tracking according to whether deep learning technology is adopted or not. As intelligent equipment upgrades, and with vehicles' rapid growth, vehicle tracking technology based on traditional image processing has been unable to meet the demands of complex traffic scenes; by means of the development of deep learning technology, breakthrough vehicle tracking technology has become more suitable for the complex reality of road traffic, and has obtained a good vehicle tracking effect.

In 2020, Wang et al. embedded the apparent model into a single-order target detection network to share the network weight, so that the model could output detection results and corresponding apparent features at the same time. This algorithm was the first real-time

multi-target tracking system [4]. In 2020, Zhou et al., the authors of CenterNet [5], inspired by the JDE model, designed the CenterTrack multi-target tracking algorithm based on the CenterNet target detection network [6] and transformed the problem into tracking based on the target center point. However, its limitation lies in that CenterTrack does not extract recognition features, and IDs (ID Switch) is more frequent when the target is lost for a long time. Therefore, FairMOT [7], proposed by Zhang et al., aiming at the problem of low accuracy in dense scenes caused by the presence of anchor frames, proposed a frame-free design similar to CenterNet [5], which greatly improved the algorithm effect and running speed. However, the method [4–7] is designed for the pedestrian tracking scene and has a poor tracking effect for objects moving at speed and with strong appearance similarity. Meanwhile, it does not consider the mutual occlusion among objects and ignores the mutual relationship between objects in complex scenes, thus leaving some space for the improvement of the vehicle tracking effect.

Different from the work in the above literature, this paper proposes an end-to-end vehicle tracking algorithm for complex traffic scenes from the roadside perspective. By optimizing the feature extraction network, it has a better feature extraction ability and detection effect for large rigid objects such as vehicles. In the optimized feature extraction network, smooth loss function is used, an anti-bottleneck structure is designed, an attention mechanism is introduced [8] and a Ghost module [9] is used to improve the effect of network feature extraction, balance the speed and accuracy of the network calculation and meet the real-time and accuracy demands of vehicle tracking in complex traffic scenes.

Our contributions are as follows:

(1) An optimized DLA-34 feature extraction network is proposed, and the Basic block-N module is designed to improve the detection ability of the network for large- and medium-sized objects, better to adapt to the vehicle tracking algorithm and meet the algorithm accuracy and processing speed.
(2) In view of the complex block traffic scene, the introduction of the joint loss function in complex traffic roadside monitoring perspective frames inside the car with a discrimination frame between similar vehicles and the target vehicle leads to the algorithm learning to target more discriminant features and expressing the ability to improve the matching success rate, reduce the IDs and improve the effect of vehicle tracking.
(3) Our method is verified on the UA-DETRAC [10] vehicle tracking dataset. Compared with current mainstream multi-target tracking methods, good tracking results are achieved on most evaluation indices, which proves the effectiveness and robustness of the proposed method in the field of vehicle tracking.

2. Related Work

2.1. Multiple Object Tracking

In the booming field of deep learning technology, MOT based on deep learning has better precision and speed balance than traditional MOT. At present, most of the mainstream MOT algorithm framework is based on detection and focuses on data association. The MOT algorithms based on deep learning can be roughly divided into detection-based tracking (DBT) and joint detection tracking (JDT), according to whether the algorithm framework is end-to-end or not [11].

Detection-based tracking is the main method of visual MOT based on deep learning. Deep network detectors such as SSD [12] and YOLO [13] are introduced into visual MOT. However, from the perspective of the structure of the deep neural network, the submodules of DBT, such as feature extraction, can be integrated into the target detection network. Based on the fusion of the neutron modules of DBT, joint detection and tracking, namely the JDT mode, using a deep network framework to achieve visual MOT, is a new trend of the last two years. The JDT class algorithm framework not only reduces the complexity of the DBT class framework, but also improves the accuracy of MOT.

In 2019, Bergmann et al. proposed a new joint detection Tracktor++ framework, which uses a simple and lightweight data association algorithm to match the tracking frame and observation frame, uses a deep detection network to generate the whole tracking sequence result and integrates the two modules of target detection and data association for the first time [14]. Sun et al. also proposed the DAN model [15] in 2019, which learns the compact and comprehensive features of the pre-detected object at multiple abstraction levels and makes a detailed pairing arrangement of these features in any two frames to infer the affinity of the object. In 2020, Gao et al. proposed a real-time target tracking algorithm for manifold twin networks [16], which combines correlation filtering, twin networks and popular information into a three-branch convolutional network to achieve end-to-end real-time tracking. In 2021, Yan et al. proposed an optimized DeepSort vehicle tracking algorithm with Gaussian YOLO v3 as the detector, which significantly improved the vehicle detection accuracy [17]. In 2021, Zhu proposed multi-sensor vehicle tracking based on the federal Kalman filter for vehicle tracking tasks in roadside occlusion scenarios [18] to improve the robustness of the algorithm. In 2022, Song et al. considered probe sampling under multiple roadside equipment units while minimizing sample exchange volume to maximize tracking performance [19]. In the field of vehicle tracking, the appearance and size of the vehicle will change greatly in a short time due to fast speed, the change of the shooting angle, the change of the lens from far to near, from near to far and other objective factors that produce the phenomenon of target deviation. At the same time, the shape and color of the car remain similar and shielded. Bad weather conditions, chaotic backgrounds and changes in lighting conditions between day and night will also affect the vehicle tracking effect.

2.2. Data Association

The data association operation is an important operation to connect different frames with the same target track, which directly affects whether the common target and tracking target can be distinguished correctly. At present, MOT data association methods based on deep learning mostly adopt the cost matrix combining the apparent feature and motion feature. Reasonable data association technology can effectively increase the robustness and accuracy of the algorithm and deal with complex scenes.

In 2018, Kim et al. proposed a new bilinear LSTM to improve the long-term apparent model, aiming at the difficulty of the previous long-term apparent model to effectively solve the dilemma of serious object occlusion and multiple missed detection. The tracking performance was partially improved, but the spatial-temporal characteristics of the object were not fully learned [20]. Xu directly based the TrctrD15 algorithm, proposed in 2020, on the visual evaluation index characteristics of the multi-target tracking set loss function, designed a deep Hungarian network, entered the distance between the adjacent frame target matrix, directed output to improve the cost of the evaluation index matrix and the training method to improve the multi-target tracking performance. However, the replacement of the Hungarian algorithm with the deep Hungarian network brings more calculations and affects the running speed of the algorithm [21].

In this paper, considering the real-time and occluding problems of vehicle tracking in complex traffic scenes, we chose the multi-target tracking algorithm based on the Hungarian algorithm of joint detection and tracking to balance the real-time and tracking accuracy.

3. Proposed Method

3.1. Overall Network Architecture

Most existing multi-object tracking algorithms adopt a two-step method: (1) the target is detected through a target detection algorithm; (2) it is matched through the Re-ID model and linked to an existing trajectory according to specific measures defined on the feature. With the improvement of target detection and Re-ID, the two-step method is also significantly improved in target tracking. However, it is difficult to calculate the video rate because the feature map of the detection algorithm and Re-ID cannot be shared. Therefore,

the one-shot method of simultaneously detecting targets and learning re-ID features has gradually entered researchers' field of vision, and can greatly reduce the calculation time. In this paper, an integrated tracking network framework is adopted. Video sequence images are input to two parallel branches: the detection branch and the Re-ID branch, through the feature network. The detection branch uses an anchor-free detector to process the heatmap of the output feature image, the size of bounding boxes and the offset of the center point. At the same time, the re-ID branch improves the ability to distinguish objects by using high-dimensional features. The overall tracking network architecture is shown in Figure 1; the integrated framework uses multiple branches to share network features, reduce parameter redundancy, improve the algorithm reasoning rate and meet the real-time demand of traffic scenes.

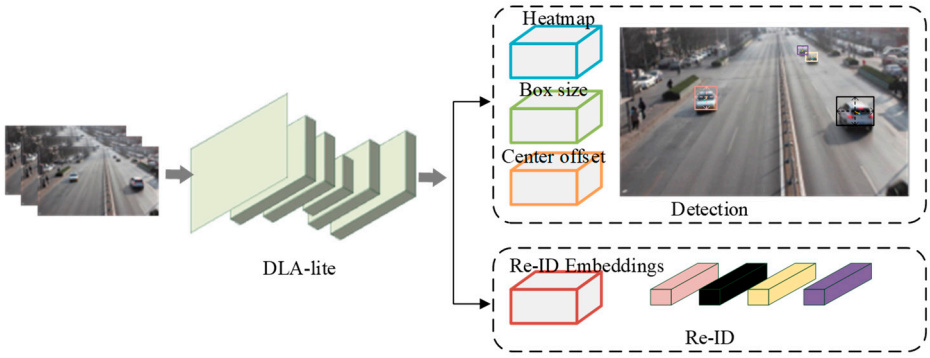

Figure 1. Overall framework of the algorithm.

3.2. DLA-Lite

DLA [22] is a general architecture that can be easily integrated into the existing CNN structure to complete a variety of computer vision tasks. CenterNet [5] uses DLA-34 (a convolutional network structure with a depth of 34 layers) as the feature extraction network of the object detection network. However, DLA-34 has a complex network structure and many parameters that consume a lot of computing resources. In order to reduce the number of network calculation parameters, improve computing efficiency and enhance network generalization, this paper uses MobileNetV2 [23] and GhostNet [9] ideas to improve DLA-34 and design a high-performance DLA-lite network structure. The improvement idea is as follows: (1) Refer to the inverted convolution layer similar to that proposed in MobileNetV2, which is constructed into a new basic block module in the form of small dimension–large dimension–small dimension. (2) After the base layer of the DLA-34 network, the Convolution Block Attention Module (CBAM) [8] is introduced to perform in the channel and space dimensions to make it pay more attention to the target area. (3) Replace the ReLU activation function [24] with a smoother Mish activation function [25], so that the network has better accuracy and generalization. The formula is as follows:

$$Mish = x \times \tan h(ln(1 + e^{x})), \tag{1}$$

(4) Referring to the GhostNet network, replace 2D standard convolution with the Ghost module, and use the Ghost module to generate the same number of feature maps as the common convolution layer. The resulting Basic block-N module is shown in Figure 2.

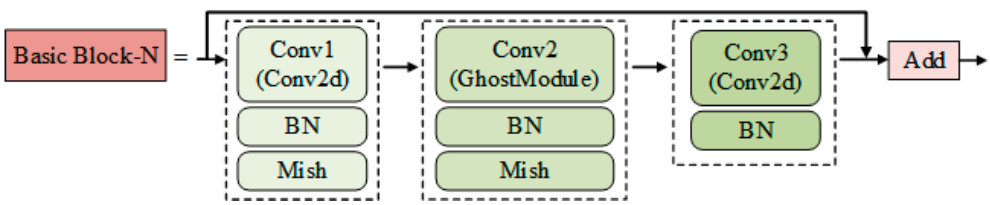

Figure 2. Basic block-N module.

The basic Block-N is simple, lightweight and consists of only three layers of convolution, which are used to gradually extract deeper features. Firstly, the image is downsampled twice by Conv1 to reduce the number of network learning parameters, increase the receptive field and increase the number of channels to obtain more feature maps. Then, BN standardizes the data of each layer, improves the convergence speed of the model and makes the model more robust. Next, the number of Conv2 output channels is 4 times the number of input channels, and the Ghost module is used instead of Conv2d to generate more feature maps with fewer parameters. The Ghost module divides Conv2 into two parts. The first part is general convolution. According to the inherent feature diagram of the first part, a series of simple linear operations are applied to generate more feature diagrams, as shown in Figure 3.

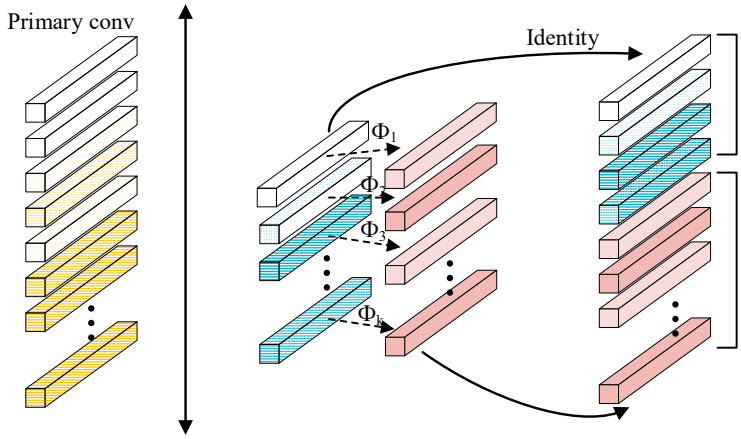

Figure 3. Ghost module.

The total number of Conv2 parameters and the computational complexity are reduced without changing the size of the output feature graph. The Mish activation function increases nonlinearity and improves network generalization. Finally, the number of input channels in Conv3 is 4 times that of output channels, and further double downsampling is performed to increase the receptive field. At the same time, according to the number of channels, it forms a small-dimensional–large-dimensional–small-dimension form with Conv1 and Conv2 to avoid the information loss caused by the compression dimension when the information is converted between the feature spaces of different dimensions and adopts jump connection to alleviate the degradation of the model performance caused by the deepening of the network depth. Further, the CBAM [8] consists of two independent sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM).

After the base layer of the DLA-34 network, the CBAM module is introduced to pay attention to the channel and space, respectively. The target can increase the expressive force through the CBAM module and, at the same time, it can make the network pay attention to important features and restrain unnecessary features, emphasizing the meaningful features in the two dimensions of space and channel. This facilitates information flow on the network, as shown in Figure 4.

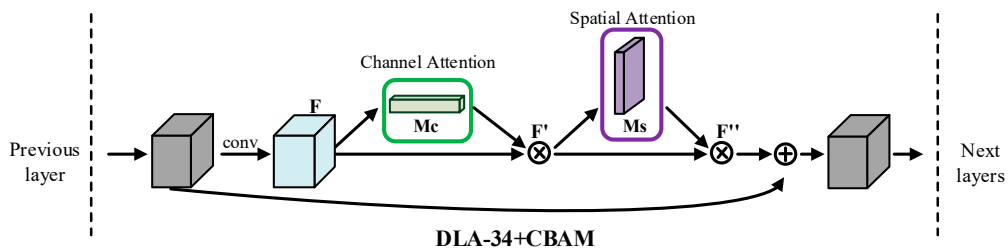

Figure 4. The CBAM module is embedded in the DLA-34 network.

A DLA-lite network is formed. Figure 5 shows the network structure; the network selectively extracts and fuses multi-layer scale features through convolutional blocks to better describe the features of rigid objects with similar appearance, such as color, shape, size and other high-level semantic features and spatial features. The parameter calculation caused by the increase in the number of convolution layers is slowed down by a cheap calculation operation, which improves the network generalization and maintains the computational efficiency.

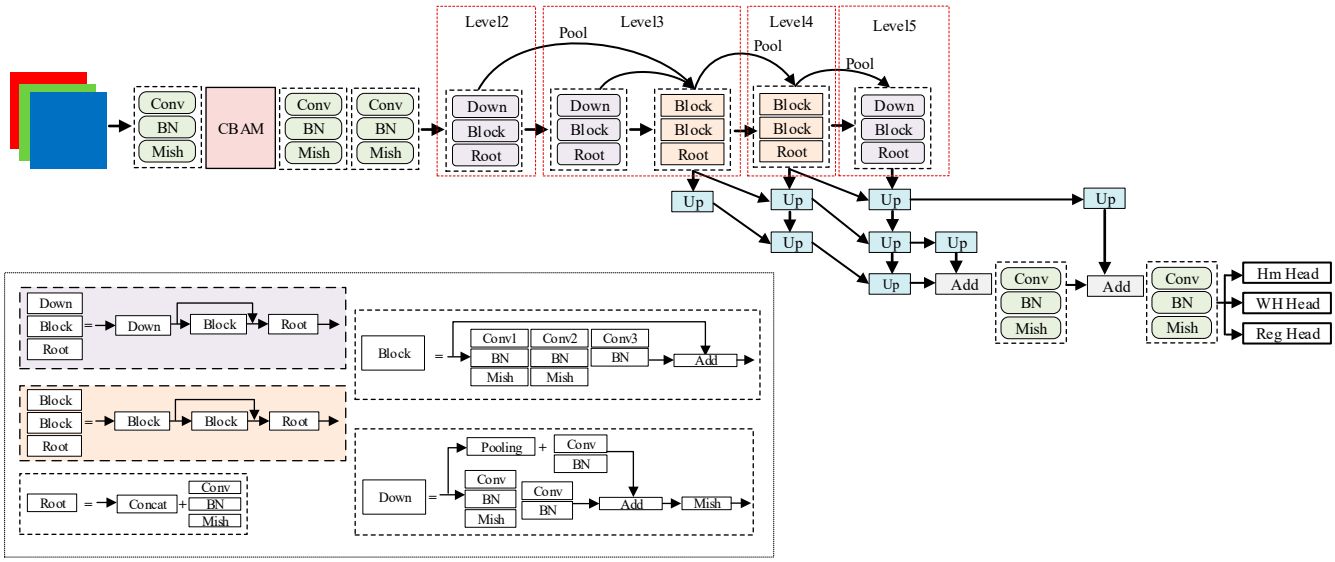

Figure 5. The DLA-lite Network Architecture.

3.3. Joint Loss Function

Depth cosine metric learning is used for similarity between two feature vectors, which is an important step in the pedestrian re-recognition algorithm. However, it is not available for multi-vehicle tracking. First, the vehicle in the process of driving gives the appearance of the same shape, but the angle difference is obvious; secondly, the colors of modern vehicles are extremely similar, and the similarity of in-class features is large. In view of the above problems, the Center Loss function [26] applied in the field of face recognition is introduced, which not only includes the distance between classes, but also considers reducing the difference within classes.

The Re-ID branch in FairMOT adopts cross-entropy loss function, namely Softmax loss function, which is often used in binary classification tasks. However, for the classification of objects with similar features, such as vehicles, features extracted from the network should have strong differentiation, namely, intra-class compact and inter-class separability. However, the cross-entropy loss function has separability between classes, but cannot constrain the feature of the in-class compact, resulting in a poor effect.

The center loss function can directly constrain the distance between sample features: the distance between individual features and average features of the same class is small enough, requiring similar features to be close to the center point of the class, as shown in Formula (2).

$$L_C = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{yi}\right\|_2^2 \tag{2}$$

m is the size of the mini-batch; x_i represents the characteristics before the full connection layer; c_{y_i} is the mean value of all depth features in the y_i category; the L_C gradient and the c_{y_i} update formula are shown in (3) and (4).

$$\frac{\partial L_c}{\partial x_i} = x_i - c_{yi} \tag{3}$$

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(y_i = j)(c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \tag{4}$$

The center loss function randomly generates a center in the network for each class, then calculates the distance between the sample center and the class center in each mini-batch of sample training and adds the value to the class center for parameter correction. If a sample feature in the class is far from the center, it needs to be punished.

By the weighted summation of the center loss function and the cross-entropy loss function, the joint loss function is formed to realize intra-class aggregation and inter-class separation. The expressions are shown in (5).

$$L = L_s + \lambda L_c = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{h_{it}}}{\sum_{j=1}^{n} e^{h_t}} + \lambda \frac{1}{2}\sum_{i=1}^{m} \left\| x_i - c_{yi} \right\|_2^2 , \tag{5}$$

λ represents the proportion of the difference within the class to the whole objective function. The training details are as follows: (1) center loss function and cross-entropy loss function are calculated; (2) reverse propagation is carried out to obtain the gradient; (3) update weight; (4) update the feature center; (5) update the initialization parameters of the convolution layer. The results show that λ = 1 has a better effect. Compared with the single cross-entropy loss function, the joint loss function can better reflect the intra-class and inter-class relationships, which is helpful for the network to discriminate and train the vehicle targets in traffic scenes, narrowing the distance between similar vehicle targets and increasing the similarity.

4. Experiment Results and Discussion

4.1. Datasets and Metrics

We first use the MS COCO (Microsoft Common Objects in Context) dataset [27] to train, verify and test the DLA-lite feature extraction network to prove that the optimized DLA-34 network gives the detector a better detection effect. Secondly, the UA-DETRAC dataset [9] is used to train and test the optimized vehicle tracking model based on FairMOT.

The MS COCO dataset contains 200,000 images, 80 different categories of objects and more than 500,000 object annotations, making it the most widely available target detection dataset. UA-DETRAC is a challenging real-world multi-target detection and multi-target tracking benchmark. Recorded at 960 × 540 pixels, the UA-DETRAC dataset contains more than 140,000 frames, 8250 hand-annotated vehicles and a total of 1.21 million marked object boundary boxes. The vehicles in the dataset are divided into four categories, namely cars, buses, vans and other vehicles; weather conditions include cloudy, nighttime, rainy, sunny; traffic scenes are from different shooting angles.

The MOT evaluation index is an effective measure to reflect the performance of multi -target tracking algorithms. To measure the performance of a MOT algorithm, the following should be considered: (1) the target appearing in every frame of the image should be found as far as possible; (2) the predicted target position should be as close as possible to the real target position; (3) different targets are assigned different IDs, and the ID of the same target remains the same in different frames of images.

Multi-object Tracking Accuracy (MOTA) intuitively represents the performance of the tracking algorithm when detecting objects and maintaining trajectories, which is the most important evaluation index. The closer MOTA is to 1, the better the performance of the tracking algorithm is. ID Switch (IDs) is the sum of ID switching numbers in the whole video track; MT (Mostly Tracked) is the percentage of tracks tracked that satisfy GT that are matched at least 80% of the time. Mostly Lost (ML) represents most missing tracking and represents the proportion of tracks satisfying GT that are successfully matched less than 20% of the time in all tracking targets; Identification F-Score (IDF1) is the ratio of correctly identified tests to the real number and the average of calculated tests.

4.2. Implementation Details

We use Ubuntu18.04 system implementation and three Titan-XP GPU cards. Use Anaconda3 to create a virtual environment and configure Python3.7 and Pytorch1.2 deep learning frameworks.

We first use the COCO dataset to train, verify and test the DLA-lite feature extraction network to prove that the DLA-lite network gives the detector a better detection effect. Among them, the training set is 118,287 pictures of train in COCO; the verification set is 5000 pictures of val in COCO; the test set consists of 5000 images in the test section of COCO. In the process of DLA-lite network training, the input image to be checked is zoomed. After zooming, the input resolution is 512 $\times$ 512 and the output resolution is 128 $\times$ 128. The training parameters are set as follows: the epoch is 160, the batch size is 24 and the initial learning rate is 1.25×10^{-4}, reduced to 1/10 in the 90th and 130th rounds, respectively. During training, the Adam optimizer is selected and the momentum parameter is set to 0.9.

Secondly, the UA-DETRAC dataset is used to train and test the optimized vehicle tracking model based on FairMOT. The training set consists of 60 video sequences in the train part of the UA-DETRAC dataset. The test set consists of four representative video sequences (MVI_39031/MVI_39311/MVI_40714/MVI_40771) from the TEST part of the UA-DETRAC dataset, which are: daytime positive view, daytime busy intersection, daytime intensive road and night positive view. In the training process, the training parameters are set as follows: the epoch is 30, the batch size is 4 and the initial learning rate is 1×10^{-4}, which is reduced to 1/10 of the original in the 20th round. During training, the Adam optimizer is selected and the momentum parameter is set to 0.9. At the same time, common data enhancement strategies are adopted, including rotation, scaling, color dithering and other operations.

4.3. Ablative Studies

Firstly, we conducted the DLA-lite network ablation experiment to show the contribution of different improvements to the model by studying the changes in the model reasoning speed and the accuracy of 5000 pictures in the COCO test set after each improvement. The change of training loss is shown in Figure 6. As can be seen from the total_loss change curve, the training process of the optimized feature extraction network is smoother, convergence is faster and the training effect is better. By comparing the loss curves of each branch, it can be seen that DLA-lite is better for center point migration, height, width and thermal map training and has good fitting performance. Four ablation experiments were conducted to compare the evaluation indices such as FPS and AP. The experimental results are shown in Table 1, indicating that the DLA-lite network shows great improvement compared with the original network and is more suitable for vehicle tracking application scenarios.

First, the activation function was replaced. When the borderless Mish function was selected to ensure the flow of information, the FPS decreased 7.4%, while other indicators increased significantly: AP increased 0.9%, AP50 increased 1.4%, AP75 increased 1.1%, APs increased 0.8%, APM increased 1.4% and APL increased 1.3%. Then, the CBAM module was inserted in the front end of the network to further improve some indicators: AP increased by 0.1% and APM increased by 0.5%, indicating that the module has a good effect on medium-size targets: vehicle targets. Then, Basic block-n was used, which reduced the computation time and improved the FPS from 21.6 to 22.8. The other six metrics improved. Compared with the original feature extraction network, DLA-34, the improved network DLA-lite has a lower FPS, but still meets the real-time requirements. All indicators are improved, especially for large- and medium-sized objects such as vehicles, and accuracy is significantly improved. In conclusion, the DLA-lite network proposed in this paper achieves a good balance between inference speed and accuracy and can be used as the feature extraction network of vehicle tracking algorithms for subsequent applications.

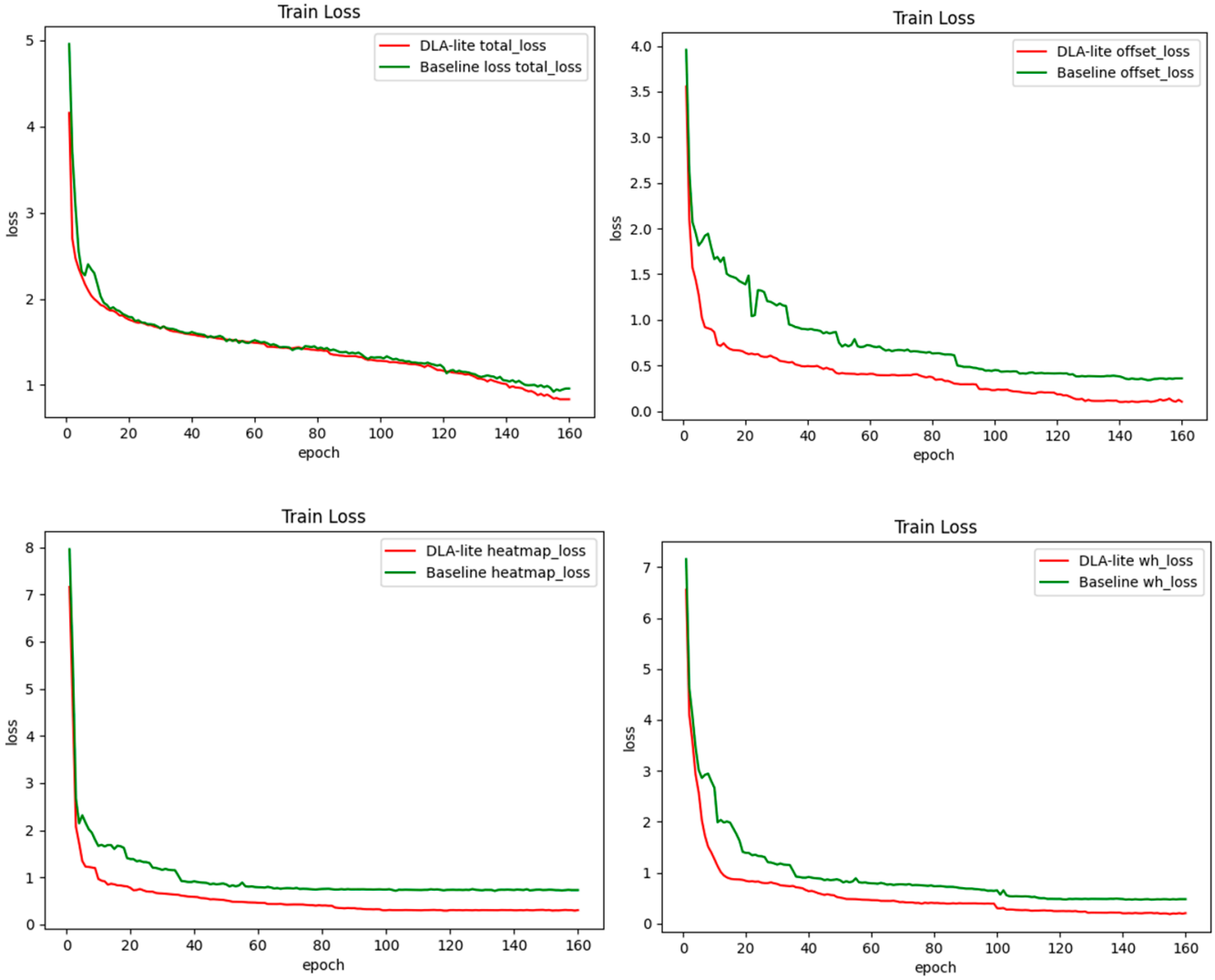

Figure 6. Training Loss.

Table 1. The DLA-lite feature extraction network ablation experiment.

	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	32.8	36.6	54.3	39.2	16.6	39.9	53.7
+Mish	25.4	37.5	55.7	40.3	17.4	41.3	54.4
+Mish + CBAM	21.6	37.6	55.5	40.2	17.2	41.8	54.1
+Mish + CBAM + Basic Block-N	22.8	38.2	56.2	41.0	17.8	42.0	56.1

In order to prove the detection effect of the DLA-lite network on vehicles, part of the UA-DETRAC datasets were visualized by the heat maps of the detection results, as shown in Figure 7. It can be seen that the DLA-lite network has good feature extraction capability in different traffic scenarios. Under different shooting angles, occlusion degrees and illumination conditions it can reduce vehicle-missing detection and reduce center point deviation.

As shown in Figure 8, the recall (maxDet = 100) of DLA-lite increases by 2.59% compared with DLA-34, and DLA-lite brings about a small increase. At the same time, it can be seen from the comparison diagram of the model volume and parameters that the model volume is reduced by about 24% and the model parameters by 16% without loss of detection accuracy.

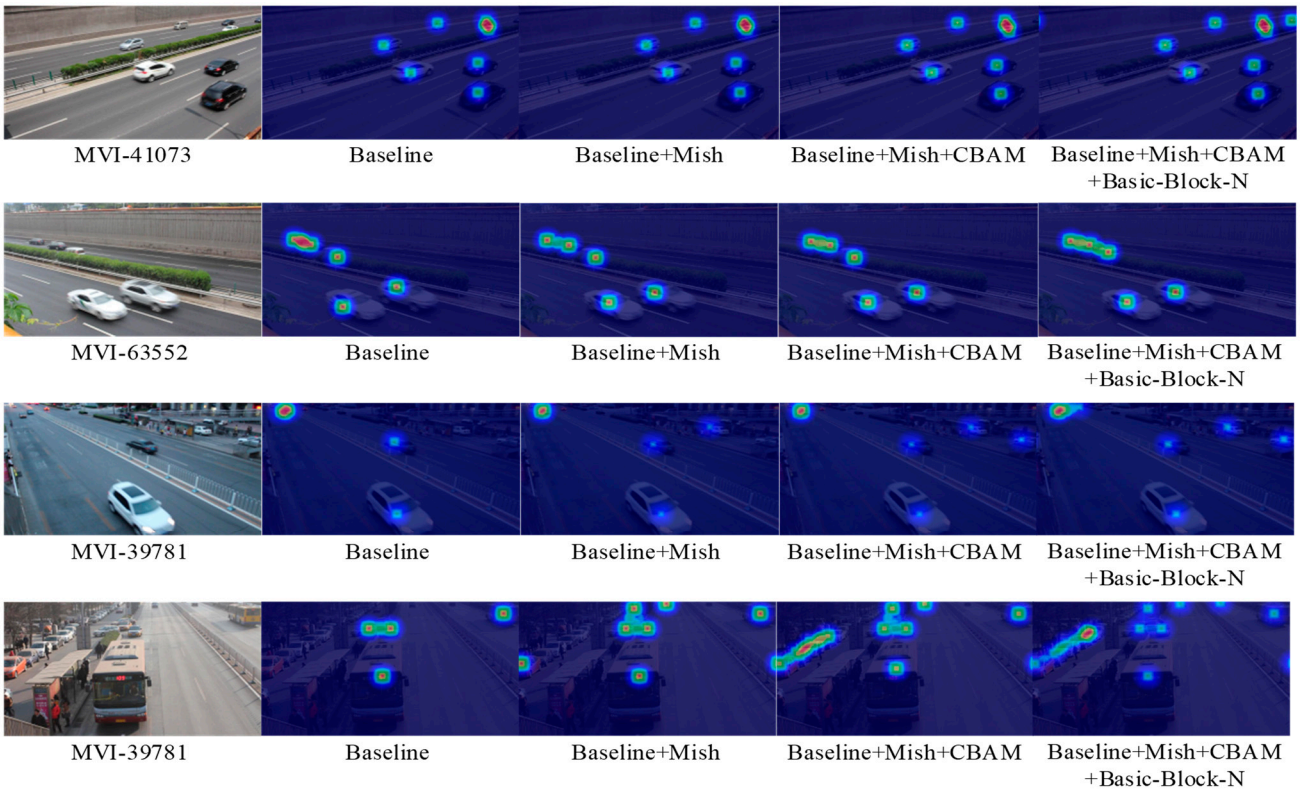

Figure 7. Heatmap Visualization.

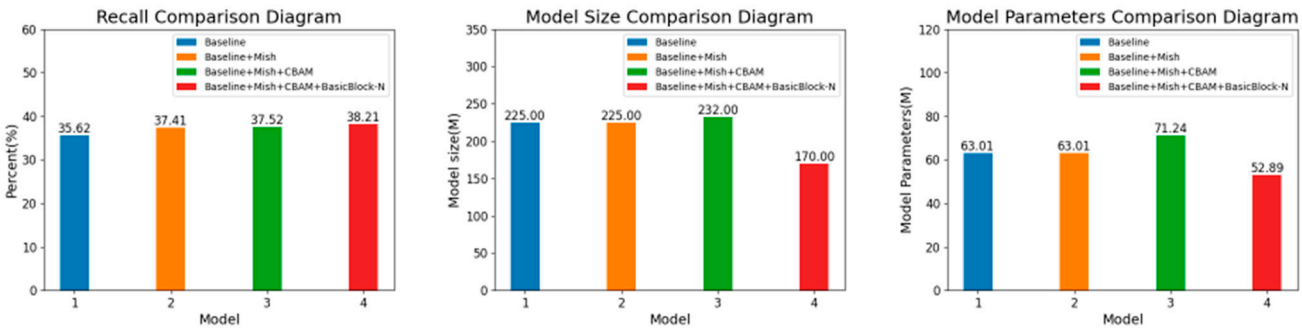

Figure 8. Chart Column.

Secondly, we conducted an ablation test on the optimized FairMOT-based vehicle tracking algorithm and compared the inference speed and five multi-objective tracking evaluation indices (FPS, MOTA, IDF1, IDS, MT and ML) after each improvement through four ablation tests to show the contribution of different improvements to the model. Table 2 shows the experimental results (Note: ↑ indicates that the larger the value of this evaluation index is, the better the effect is; ↓ indicates that the smaller the value of this evaluation index is, the better the effect is).

Table 2. Ablation experiments of vehicle tracking algorithm.

	MOTA↑	IDF1↑	MT↑	ML↓	IDS↓	FPS↑
Baseline	77.7	84.1	160	4	47	20.37
Baseline + DLA-lite	79.3	84.8	164	3	57	18.95
Baseline + Center Loss	78.6	84.2	160	4	45	20.53
Baseline + DLA-lite + Center Loss	78.9	84.3	158	4	48	19.23

Firstly, the optimized feature extraction network DLA-lite was introduced to perform the ablation experiment based on the baseline. Compared with the original network, DLA-

lite increased the number of network layers, resulting in a slight decrease in FPS, a 1.5% increase in MOTA, a 0.7% increase in IDF1 and an increase in MT and ML indices. In other words, the tracking effect is improved on the premise that the real-time performance is also met, but the number of IDs increases, which may be because the optimized feature extraction network is not strong enough to distinguish distant small target vehicles, and then the number of IDs increases due to mutual occlusion factors. Secondly, by introducing the central loss function to the baseline, MOTA increased by 0.9%, IDF1 increased by 0.1%, MT and ML did not change, the number of IDs decreased and FPS slightly increased, indicating that introducing the central loss function can improve the resolution between vehicles, reduce IDs and further improve the overall vehicle tracking effect. Finally, based on the baseline, DLA-lite and the center loss function were introduced to perform ablation experiments, and some indices were improved, indicating that the discriminant and tracking accuracy of vehicle objects were improved by adopting optimized feature extraction networks and introducing center loss functions. There was an improved roadside view of the vehicle tracking scene.

In order to better demonstrate the effectiveness of our proposed vehicle tracking algorithm based on the roadside perspective, the proposed method was compared with four current mainstream methods on the same UA-DETRAC dataset test video sequence, as shown in Table 3. In terms of MOTA, the most important index in multi-target tracking tasks, the method proposed by us achieves the best results in MOTA, up 78.9%. However, the real-time performance of the algorithm is poor. The reasons are as follows: since the data association step of the SORT algorithm only uses IoU technology, the real-time performance of the algorithm is greatly improved. However, due to its simple data association method, a large number of ID jumps are caused and tracking accuracy is reduced. DeepSORT designed cascading matching to increase the accuracy of data association and reduce ID hopping. However, due to the mediocre effect of its detector, the evaluation index of the algorithm was low. FairMOT is second only to the method in this paper in terms of the MOTA index. This method is for a pedestrian tracking algorithm, and it has a good effect on slow-moving objects and pedestrians with large differences in appearance. However, it ignores the related problems of fast-moving objects such as vehicles, so the index of this algorithm is worse than the method in this paper. CenterNet has excellent detection results, so it indirectly improves the accuracy of the tracking algorithm. However, because its data association steps are simple and similar to SORT, the algorithm has poor robustness. The method in this paper starts with the vehicle traffic scene angle optimization algorithm from the roadside perspective and the feature extraction network to optimize the network structure for large objects such as vehicles. At the same time, the joint loss function is introduced to further improve the discriminant ability of vehicles of similar appearance, alleviate the ID jump caused by vehicles with similar distances and optimize the vehicle tracking algorithm from multiple angles. Therefore, compared with the four mainstream tracking algorithms, the proposed algorithm achieves the optimal effect.

Table 3. Tracking results of mainstream tracking algorithms on UA-DETRAC test.

	MOTA↑	**IDF1↑**	**MT↑**	**ML↓**	**IDS↓**	**FPS↑**
SORT	70.5	79.8	146	4	65	35.67
DeepSORT	74.3	82.6	153	3	55	26.85
FairMOT	77.7	84.1	160	4	47	20.37
CenterTrack	76.9	84.7	155	4	50	25.56
Ours	78.9	84.3	158	4	48	19.23

4.4. Visualization Results

In order to show the performance of the algorithm more directly, we give some visualization results of the algorithm in this paper on four video sequences in the UA test.

It can be observed from Figure 9 below that the video sequence of MVI-39031 is in the daytime under strong lighting conditions. In frame 404 and frame 419, when there

are fewer vehicles, the distant small target vehicles can also be tracked well and maintain the ID number stably, and the tracking effect is good. In the case of a large number of vehicles, in frame 1309 and frame 1383, the vehicle with the ID number 49 in frame 1309 can be successfully tracked, but it cannot be detected and tracked in frame 1383 due to the occlusion of trees. In frame 1309, the two vehicles with the ID numbers 61 and 66 cannot be detected due to serious occlusion. Even the human eye cannot recognize them, but in frame 1383 the middle of the blocked vehicle slightly out of frame is immediately given a new tracking ID. In the case of partial occlusion, the trajectory can be successfully tracked when most of the vehicles in the pictures of the two frames have a similar appearance. This shows that the multi-vehicle tracking algorithm optimized by depth cosine measurement learning can achieve better tracking results in dense scenes. It proves that the weighted summation of the center loss function and cross-entropy loss function can effectively achieve intra-class aggregation and inter-class separation. In the MVI-39311 video sequence at a complex intersection, as shown in frames 217 and 316, a special type of vehicle—a bus—can also be tracked. In frame 217, vehicle No. 3 is successfully detected despite being heavily obscured, and in frame 316 the same ID number is successfully maintained after the occlusion disappears. In frames 1100 and 1151, the bus is not successfully tracked because it is out of range and heavily obscured. It can be observed from MVI-40714 that the video sequence is in the traffic scene of the main road with serious occlusion. In frames 99 and 136, vehicle No. 38 is well tracked, but the size of the target frame of vehicle No. 41 is incorrect because the vehicle has just entered the field of view of the camera. However, the vehicle is well tracked under the condition of dense vehicles in the left lane of the two pictures. In frames 828 and 849, each vehicle correctly maintains its ID number. MVI-40771 can observe that the video is a traffic scene at a busy intersection at night. The visualization results show that the optimized vehicle tracking algorithm can also be applied well to night traffic scenes. Meanwhile, based on scene measurement in the evening, the optimized algorithm has a good tracking effect for distant vehicles and dense traffic flow, and the ID remains stable.

Figure 9. Visual output result of video sequence.

5. Conclusions

Taking intelligent transportation systems from the roadside perspective as the entry point, and taking deep learning as the basic technology, focusing on multi-vehicle tracking tasks in the field of computer vision, a vehicle tracking model suitable for complex traffic scenes from the roadside perspective was constructed. We proposed an optimized DLA-lite feature extraction network and designed a Basic Block-N module for vehicle rigid objects to improve the feature extraction and detection ability of large- and medium-sized objects. At the same time, the joint loss function was proposed to improve the discrimination ability of vehicles of similar appearance and color by the weighted center loss function and cross-entropy loss function. In the UA-DETRAC open vehicle tracking dataset, compared with the current popular tracking algorithms, the algorithm we designed achieved the optimal tracking effect of comprehensive indicators.

Author Contributions: G.H. wrote the main manuscript text. Q.J. established the experimental platform and provided financial support. L.J. and H.R. conducted the experiments. L.Z. improved the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China: 52072333; S&T Program of Hebei: 21340801D.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used and analyzed during this study are available from the corresponding author by request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
2. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
3. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Zhao, X.; Kim, T.-K. Multiple Object Tracking: A Literature Review. arXiv: Computer Vision and Pattern Recognition. *arXiv* **2014**, arXiv:1409.7618.
4. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. *arXiv* **2019**, arXiv:1909.12605.
5. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
6. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 474–490.
7. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
8. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the ECCV 2018, Munich, Germany, 8–14 September 2018; Lecture Notes in Computer Science. Springer: Cham, Switzerland, 2018; Volume 11211.
9. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
10. Wen, L.; Du, D.; Cai, Z.; Lei, Z.; Chang, M.C.; Qi, H.; Lim, J.; Yang, M.H.; Lyu, S. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Comput. Vis. Image Underst.* **2020**, *193*, 102907. [CrossRef]
11. Xu, Y.; Zhou, X.; Chen, S.; Li, F. Deep learning for multiple object tracking: A survey. *IET Comput. Vis.* **2019**, *13*, 355–368. [CrossRef]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2016**, arXiv:1506.02640.
14. Bermann, P.; Meizhardt, T. Tracking without bells and whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 941–951.
15. Sun, S.; Akhtar, N.; Song, H.; Shah, M. Deep Affinity Network for Multiple Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 104–119. [CrossRef] [PubMed]

16. Gao, M.; Jin, L.; Jiang, Y.; Guo, B. Manifold siamese network: A novel visual tracking ConvNet for autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 1612–1623. [CrossRef]
17. jin, L.; Hua, Q.; GUO, B.; Xie, X.-Y.; YAN, F.-G.; WU, B.-T. Multi-object tracking of forward vehicle based on optimized DeepSort. *J. Zhejiang Univ. (Eng. Sci.)* **2021**, *55*, 1056–1064.
18. Zhu, J. *Research on Vehicle Recognition and Tracking Technology in Roadside Occlusion Scene*; Southeast University: Nanjing, China, 2021.
19. Song, J.; Hyun, S.H.; Lee, J.H.; Choi, J.; Kim, S.C. Joint Vehicle Tracking and RSU Selection for V2I Communications with Extended Kalman Filter. *IEEE Trans. Veh. Technol.* **2022**, *71*, 5609–5614. [CrossRef]
20. Kim, C.; Li, F.; Rehg, J.M. Multi-object tracking with neural gating using bilinear lstm. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 200–215.
21. Xu, Y.; Ban, Y.; Alameda-Pineda, X.; Horaud, R. DeepMOT: A Differentiable Framework for Training Multiple Object Trackers. *arXiv* **2019**, arXiv:1906.06618.
22. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412.
23. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
24. Agarap, A.F. Deep learning using rectified linear units (relu). *arXiv* **2018**, arXiv:1803.08375.
25. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* **2019**, arXiv:1908.08681.
26. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A comprehensive study on center loss for deep face recognition. *Int. J. Comput. Vis.* **2019**, *127*, 668–683. [CrossRef]
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.