remote sensing

MDPI

Article

Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure

Yeneng Lin [1], Mengmeng Wang [1], Wenzhou Chen [1], Wang Gao [2], Lei Li [2] and Yong Liu [1,*]

[1] Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China
[2] Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing 100191, China
* Correspondence: yongliu@iipc.zju.edu.cn

Abstract: The task of multi-object tracking via deep learning methods for UAV videos has become an important research direction. However, with some current multiple object tracking methods, the relationship between object detection and tracking is not well handled, and decisions on how to make good use of temporal information can affect tracking performance as well. To improve the performance of multi-object tracking, this paper proposes an improved multiple object tracking model based on FairMOT. The proposed model contains a structure to separate the detection and ReID heads to decrease the influence between every function head. Additionally, we develop a temporal embedding structure to strengthen the representational ability of the model. By combing the temporal-association structure and separating different function heads, the model's performance in object detection and tracking tasks is improved, which has been verified on the VisDrone2019 dataset. Compared with the original method, the proposed model improves MOTA by 4.9% and MOTP by 1.2% and has better tracking performance than the models such as SORT and HDHNet on the UAV video dataset.

Keywords: deep learning; multi-object tracking; data association; object detection; temporal information; remote sensing data; video sequence

Citation: Lin, Y.; Wang, M.; Chen, W.; Gao, W.; Li, L.; Liu, Y. Multiple Object Tracking of Drone Videos by a Temporal-Association Network with Separated-Tasks Structure. *Remote Sens.* **2022**, *14*, 3862. https://doi.org/10.3390/rs14163862

Academic Editors: Yingying Dong, Chenghai Yang, Giovanni Laneve and Wenjiang Huang

Received: 13 June 2022
Accepted: 6 August 2022
Published: 9 August 2022

1. Introduction

In recent years, with the rapid development of artificial intelligence technology, computer vision [1–3] has also penetrated into various fields in society, and the application of remote sensing data has become more and more popular [4–6]. Making full use of remote sensing videos and computer vision technology can greatly improve the efficiency of environmental monitoring and security monitoring. The use of computer vision technology to accomplish multi-object detection and tracking tasks in UAV images has gradually become one of the research hotspots. Multiple object tracking (MOT) refers to the object detection of all targets in each frame in the continuous frame sequence of the videos and obtaining the positions of targets in the images, the sizes of bounding boxes, and the speed of each object, as well as assigning individual ID identifications for every object in each frame. In the current mainstream research, the MOT model system can often be divided into two different paradigms, tracking by detection (TBD) [7–10] and joint detection and tracking (JDT) [11–14].

In the task of tracking video sequences, the accuracy of target detection [15–28] and data association can affect the performance of the model. Our research mainly pays attention to these two aspects. In the current research, target detection based on deep learning has been an important part of MOT. The anchor-free target detection algorithm has also become one of the more popular frameworks recently. The main idea of the algorithm is to perform target detection based on the center point of every object, such as CenterNet [29], FCOS [30], and Centerpoint [31]. The task of multi-object tracking can

also be achieved based on the anchor-free framework. In the framework of JDT, such as RetinaTrack [32] and CenterTrack [33], these two methods combine detection and tracking, simplifying the model structure and improving the real-time performance of calculation. Figure 1 is the process of two paradigms of MOT. Compared with our research, it is mainly the use of a single frame without using temporal information. Although the combination of detection and tracking tasks can improve training efficiency, the two tasks have different concerns, affecting the accuracy of detection or tracking performance. The model in our paper separates the detection and tracking tasks to a certain extent in the phases of training and inference for this problem and achieves end-to-end multi-object tracking. In addition, the use of temporal information will also affect the performance of the model on the MOT. Wu [34] and Liang [35] proposed models to improve the detection ability by using temporal information with multiple frames. Except for the detection performance, our research mainly focuses on the use of temporal information to improve the feature representation of the objects so as to improve the accuracy of the data association.

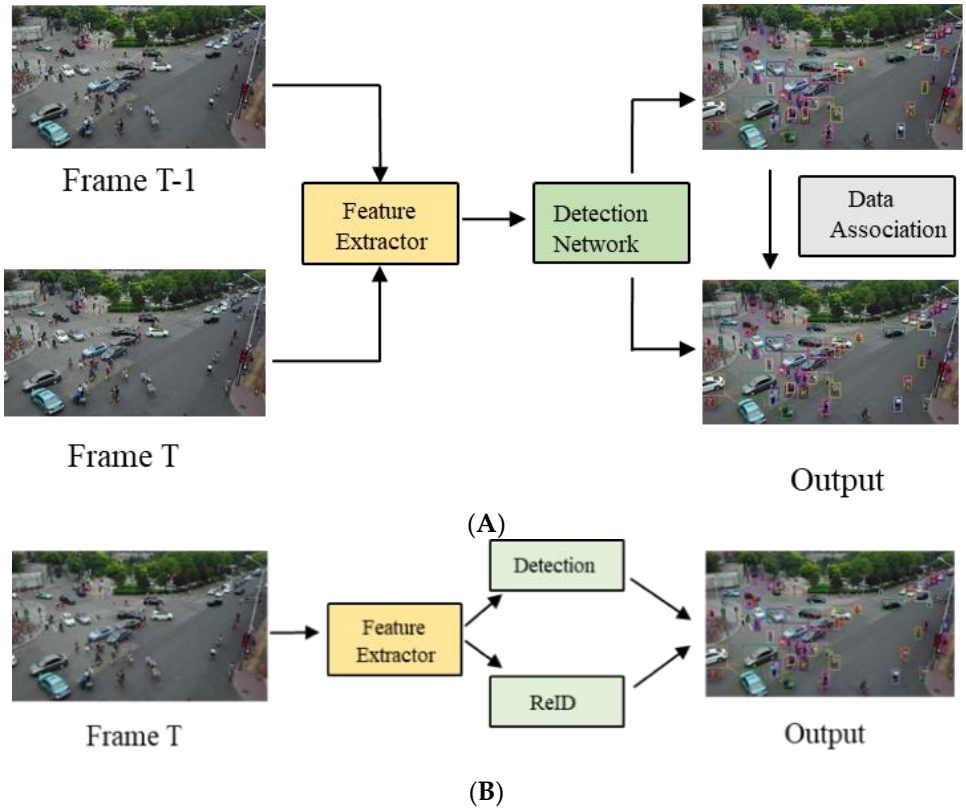


Figure 1. The process of two paradigms of MOT. (**A**) is the process of tracking by detection, and (**B**) is the process of joined detection and tracking.

MOT based on deep learning has also been used in the field of remote sensing, and the video interpretation of drones has been used as the research direction. This research direction is the main topic of this article, which is also a current research hotspot. Some research [36–42] combines trackers and detection models to realize the MOT tasks. Although separating the detection and tracking tasks can train models independently, it is more difficult to adjust one of the structures of a part to adapt the another. Our research aims to set up an improved end-to-end MOT model for remote sensing data. Furthermore, to improve the problems of small size and diverse backgrounds in remote sensing data, Jin [43] and Kraus [44] made use of features in different aspects of the tracked objects to improve the problems. Compared with the above research, temporal information can also be used in the MOT for remote sensing videos except for the feature of every object in our research.

MOT is a cross-frame video interpretation task, and most models in the current research do not make good use of the information in time series. There are certain limitations in relying only on the information of a single frame. The object lacks the links between the frames. For example, if an object is occluded in a certain frame, and if the model only relies on a single frame of information for data association, there will often be situations where the same object has different characterization information, which may lead to the ID switch (IDS) problem, thereby reducing the accuracy of the model. Therefore, using temporal information can significantly improve the model's performance for this task. In addition, although the JDT paradigm model combines object detection and data association for joint training to achieve end-to-end MOT, object detection and tracking are often two different vision tasks. Object detection needs to distinguish multiple categories, which needs to maximize the distance between different types and minimize the distance between the same type, to improve object detection accuracy. However, object tracking needs to maximize the distance among all objects in the same category. Therefore, if the two subtasks share many parameters during training, the training efficiency of the model may be reduced, and the performance of the trained model, in some cases, may get worse.

Given using temporal information and the conflict of two subtasks during the training phase, we propose an improved MOT model, using FairMOT [45] as the baseline. For the overall model structure, the detection part and the ReID part are disassembled. Compared with the detection part, the generation part of embedding is on an additional branch. A feature enhancement structure based on temporal information is added to the branch to improve the model's ability to discriminate ReID information. In the calculation process of model loss, compared with single-frame loss calculations, we perform double-frame output in the output part of the model detection and perform loss calculations on the output of two adjacent frames simultaneously.

Our contributions can be summarized as follows:

1. We change the single frame output of the original model to an output of two adjacent frames to improve the training efficiency.
2. We improve the conflict problem of two subtasks, including object detection and tracking during the training phase.
3. We constructed a feature enhancement structure based on temporal information to improve the representation of ReID information, enhancing the training efficiency of the ReID head of the model and ensuring data association performance.

2. Materials and Methods

This chapter mainly explains the detailed structure of the model proposed in this paper and the overall process of using the model to complete the MOT task. The model in this study uses FairMOT [45] as the baseline, and we improve it to achieve the functions mentioned in this article. Compared with the original structure, we have improved the feature extraction part of the model for the UAV videos, separated its detection and embedding parts, and added a feature enhancement structure in the ReID head. The detection head is changed to generate outputs of two adjacent frames for loss calculation. The following parts will describe each block in detail as a subsection.

2.1. The Structure of FairMOT

In the research field of MOT, many tracking algorithms can achieve MOT tasks based on detection results. The authors believe that object detection and ReID should be parallel visual tasks, so an anchor-free multi-target tracking algorithm called FairMOT is constructed. FairMOT is an end-to-end anchor-free MOT framework built on CenterNet. The model adopts a simple network framework, which is mainly composed of detection and ReID modules. FairMOT contains an anchor-free target detection framework, which can output heatmaps, sizes of bounding boxes information and offset information. In the Reid branch, the appearance feature of each pixel can be obtained, and it is used as the feature of the object with the pixel as the center point. The two functions can be per-

formed during stages of training and inference at the same time, which achieves a balance between detection and ReID functions and has a better MOT performance. The feature extraction part of the model is shown in Figure 2. DLA-34 [46] is used as the backbone network to perform feature extraction on two-dimensional video images. The structure of the encoder-decoder network is shown in Figure 2B. Then, multiple heads will be used according to different vision tasks, namely the heatmap head, offset head, the size head for bounding boxes, and ReID head. These branches share the same feature map after the feature extraction structure.

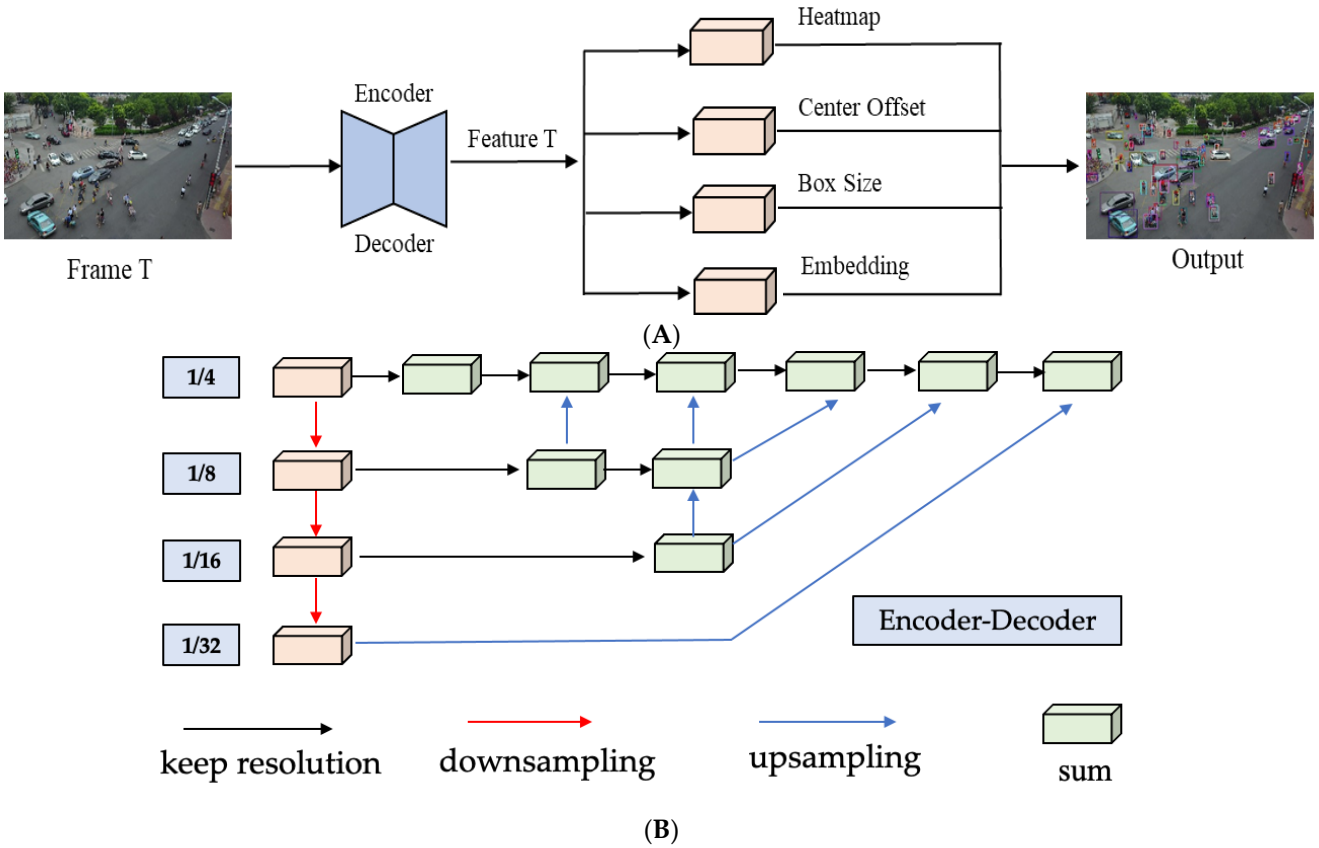


Figure 2. (**A**) is the process and structure of FairMOT. The blue part is the backbone of FairMOT, and the orange part includes different heads to form detection and ReID functions. (**B**) is the structure of the encoder-decoder network, which is used as the encoder and decoder of the FairMOT (the figure is revised from [45]).

The heatmap branch is used to generate a response map of the center point of every object, which represents the probability of objects. The size of the heatmap is HxWxC, where C represents the number of object categories. The response value of each pixel in the heatmap can reflect the probability of the object appearing, and the value of the pixel response value decreases as the distance from the center point increases. The offset branch represents the possible offset of each point compared to the original point after being encoded from the original image to the feature maps. This branch is responsible for accurately locating objects. Since the size of the feature map in the calculation graph varies greatly, this will produce some quantization errors, which will affect the locations of predicted objects and the extraction accuracy of ReID features of different objects. The size branch represents the width and height of the bounding box corresponding to each object. The three branches consist of the part responsible for object detection in the MOT framework.

The ReID branch generates an embedding of a specific length vector for each point, which is used to characterize the unique information of each point. Each pixel on the feature map contains a vector with a depth of 128 to represent the appearance features of

the object for that point, and finally, an embedding map with the size of 128 × W × H will be obtained. In the training phase of the model, each head performs loss calculations separately. The three branches of the detection part, including the heatmap, offset, and size head, are mainly calculated by focal loss and L1 regression loss. In contrast, the loss of the ReID branch is calculated through classifiers.

2.2. The Structure of the Proposed Model

Compared with the feature extraction part of the FairMOT framework, we consider that there is a certain degree of conflict between target detection and ReID tasks during training; that is, the target detection task is to maximize the distance of different categories and minimize the distance of the same category, and the ReID task is to maximize the distance of objects with the same category. Therefore, it is necessary to adjust the model structure for this problem.

Given the conflict between ReID and target detection, the model in this paper is mainly to separate the two branches. The functions of the two structures on the FairMOT are realized through four branches, and these branches share the same encoder and decoder. The adjustment of the model in this article is mainly to improve the decoder part of the backbone network. Because the original model with this decoder can realize detection and ReID functions well, we change it to two decoders with the same structure as the original model, which is used for target detection and ReID, respectively. Furthermore, the same structure can meet conditions of finetuning. Still, there is no parameter sharing in these two decoders, which reduces the mutual influence between parameters of two tasks during model training. The encoder and decoder structures have been shown in Figure 2B, which is used to extract image features. The overall structure of the model is shown in Figure 3. In general, the model in this paper mainly uses two adjacent frames as input and can achieve multiple associations and utilization of temporal information in object detection and tracking tasks. From the overall structure of the model, the separation of the object detection and the ReID structure can make the two parts of the structure better realize their respective functions. In the ReID part, the temporal feature association structure is added to the ReID structure. This structure mainly integrates the historical frame information with the current frame to improve the robustness of the model while processing temporal series information. Compared with the single-frame input, the proposed model changes the input to two adjacent frames, which can make good use of temporal information. Furthermore, the output of two adjacent frames can improve the training efficiency of the model compared with the single-frame input. In the test phase, the model has two adjacent frames as inputs, and if it is the first frame of the video sequence, the model's input will be two images of the first frame. In the input part, the processing method of two frames is parameter sharing.

After the feature extraction of the encoder, the obtained features are simultaneously input into two decoders, and the processing of target detection and ReID information are performed, respectively. In the target detection part, the functions of the heatmap branch, offset branch, and size branch are similar to the original model, which is indicated in Section 2.1. Compared with the original structure, the structure of the heatmap branch is adjusted. Firstly, the feature of the previous frame obtained by the decoder is followed by a multi-layer convolution, which generates a centered map, and we combine this map with the features of the current frame obtained by decoder B. Then, the output of the branch can be obtained through the heatmap branch. The only difference is the output results of these two branches. Compared with the single-frame output, the model's output in this paper includes two predicted results of adjacent frames in two branches, which perform loss calculations simultaneously during training.

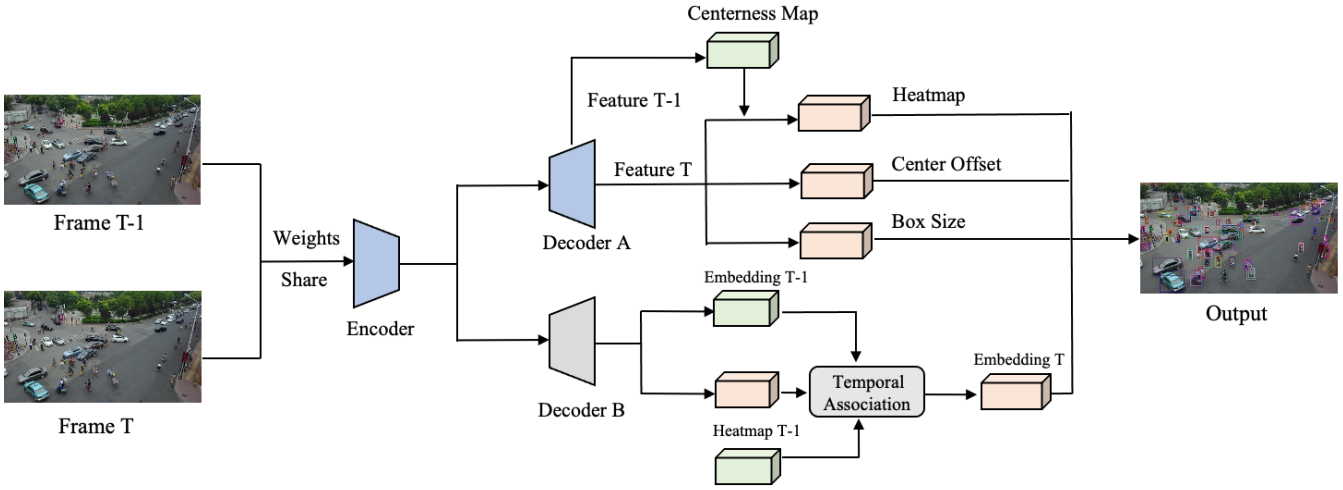


Figure 3. The structure of the proposed model. The blue part is the backbone of the model to extract features of the inputs, and there are two decoders with blue and grey colors. The orange part includes different heads to form detection and ReID function. To distinguish the information from different frames, the green part mainly makes use of the information from the T-1 frame. The detailed structure and process of the temporal association are explained in Figure 4, which can integrate the temporal information of two adjacent frames.

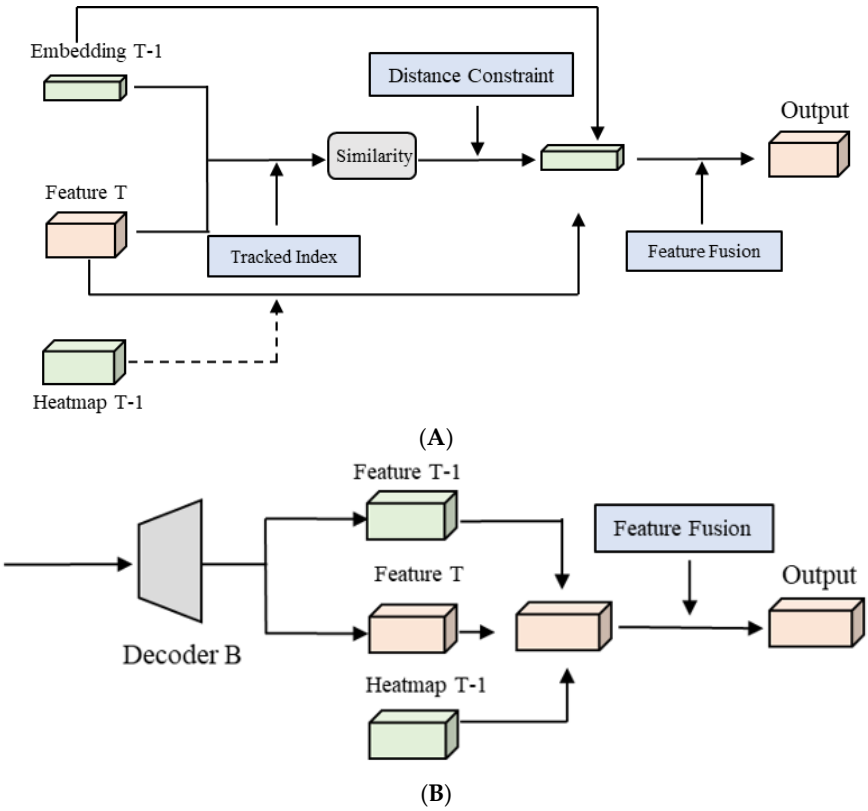


Figure 4. The two detailed structures of the module for the temporal association are in Figure 3. (**A**) is the indication of structure A, and (**B**) is the indication of structure B. Structure A can add to the model directly, and feature T and embedding T-1 are obtained by decoder B. The dotted arrow means that heatmap T-1 is not regarded as one of the inputs which can provide tracked indexes during the testing phase. Structure B combines the information of feature T-1, feature T obtained by decoder B, and heatmap T-1.

On the ReID branch, the branch adds a feature enhancement structure for the temporal association, which uses the features of two adjacent frames obtained by decoder B and the heatmap of the previous frame as the input information of the feature module. The structure and process of the temporal association are in the following sections. The temporal association can match and integrate the information of two adjacent frames to enhance the robustness of the ReID branch to get the final output of the ReID branch. Making use of temporal information on the ReID branch can improve the robustness of the appearance features generated by the model. Objects in a single frame may be deformed or occluded by other objects and background information, and the appearance features will change greatly, which will affect subsequent data associations. The use of temporal information can make some objects fused with historical feature information, which makes it possible to reduce the IDS problems caused by the influence of appearance features when problems such as occlusion and deformation of predicted objects occur in the current frame.

2.2.1. The ReID Branch with Structure A

As shown in Figure 4A, firstly, we associate the obtained feature T-1 with feature T. In the training phase, the calculation process is performed by inputting label information. The input includes the number of objects that exist simultaneously and the corresponding location index in two adjacent frames. Tracked indexes in Figure 4A mean that the heatmap T-1 can provide the detailed location and number of detected objects in the last frame, which can be used to extract the embedding information. We use this information to obtain the feature at the corresponding position of feature T-1 and calculate the similarity between it and feature T to obtain the feature similarity between each object in the previous frame and each point in the current frame; after that, we keep the point with the smallest distance. Feature similarity can be used as a measure of the degree of matching among embeddings. Additionally, it is regarded as the possible position of the object where the object of the previous frame may exist in the current frame. In this study, we mainly associate objects of two adjacent frames by calculating the similarity of the ReID feature, and the corresponding calculation process is shown in Equation (1). By calculating the similarity between the objects in the two frames, the possible position of each object in the previous frame in the current frame is obtained, and the feature fusion can be performed with the corresponding objects in the current frame. In this study, we also tried to achieve a similar effect through the point multiplication calculation of embeddings, which can improve the training and inference speed of the model. After the position information is obtained, feature fusion is performed, and the fusion method selected in this step is to add the corresponding feature matrix to the average. In the inference stage of the model, since there is no label information, the heatmap of the previous frame obtained by the model is used as auxiliary information. The number of possible targets in the previous frame is obtained from the heatmap.

$$similarity(x, y) = \frac{x \cdot y}{|x| \cdot |y|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (1)$$

where n represents the length of embeddings in the branches, and x and y represent the features of different objects between two frames.

The ReID embedding of the corresponding positions of these targets is input into the feature module as the feature of frame T-1. Since there is a situation in the inference phase where the object that appeared in the previous frame may disappear in the current frame, for this situation, a distance constraint needs to be added at this time. A threshold value needs to be set; namely, if the center point of the previous frame is far away from the center point matched in the current frame and exceeds the threshold, the matched point is considered unreliable. It should be ignored, and only a matched point with high reliability is retained. Distance constraints can be applied to the inference stage of the model to filter the predicted object positions. For the scaled size of the dataset images in this study, we set that if the distance of two predicted points is more than 50 pixels, the association can be

considered unreliable matching, and we set it as the filtering threshold. The feature fusion method consistent with the training phase is performed to generate the final output of the ReID branch.

2.2.2. The ReID Branch with Structure B

The way that the structure makes use of temporal information is shown in Figure 4A. The information used in the structure includes three blocks, namely the heatmap of the previous frame and the feature t-1 and feature t of the two adjacent frames obtained by the decoder B. After the channel is spliced, the multi-layer convolution of the ReID branch is used to generate the final output on the branch. In the model's training phase, the previous frame's heatmap is also provided with label information. In the inference phase, the heatmap obtained from the model detection part is used as one of the inputs of the ReID branch.

2.3. The Post-Processing Part

The post-processing part functions mainly through adjusting the process of SORT [41] and DeepSORT [42] to complete the data association. Compared with single-category MOT, this research changes the post-processing part on multiple categories. Unlike the training stage, the post-processing stage does not assign an ID to each category; objects of various categories are assigned IDs together in sequence. The inference part uses DeepSORT as the main process framework, and there is a round for every three frames. The first frame normalizes the heatmap and standardizes ReID features obtained by the model, and performs non-maximum suppression processing on the heatmap according to the threshold to filter out possible objects. We assign an ID to objects in the first frame.

The second frame repeats the operation of the first frame; after getting the possible objects, it matches the object by the IoU value of the bounding boxes in the first frame, retains the expected detection, assigns the same ID, and includes those that are not matched. In the third frame, the ReID feature is added to the second frame, the cosine distance of the ReID feature is calculated on the detection target of the two adjacent frames, and the Kalman filter [47] is used for motion prediction. The appearance and motion characteristics are combined for data association. After that, the unmatched objects in the third frame and the objects in the second frame are subjected to IoU calculation. If it is smaller than a fixed threshold, it is regarded as a new target, and a unique ID is assigned. Finally, repeat the above steps for each subsequent frame to complete the post-processing steps of video MOT.

2.4. Training Strategy

In the model's training, a combination of multiple loss functions is used for training, and different training strategies are used primarily according to other branches of the model. The output of the heatmap of the target detection part is mainly used to train the model through focal loss. The corresponding label of the heatmap uses the label information to provide the center point position of every object. The corresponding response of the heatmap is obtained through Gaussian distribution processing, which is used as the training label of the heatmap, and the loss function is shown as follows. The information of offset and width and height branches are trained using the loss of L1 regression [48]. On the branch of ReID, the primary training method is classification type. After normalizing the target features obtained by the ReID head, classification training is carried out through classifiers. Since it is a multi-category MOT, multiple classifiers are also set at this time to train the model. The number of categories is the tota

$$L_H = -\frac{1}{n}\sum_{m}\begin{cases} (1-\hat{x})^a \log(\hat{x}), & 1 \\ (1-x)^b(\hat{x})^a \log(1-\hat{x}) & \text{otherwise} \end{cases} \quad (2)$$

where m represents the number of points in the heatmap, $\hat{x}$ represents the model's predicted values, x represents the label's value, and a and b represent weights of focal loss.

$$L_B = \sum_{i=1}^{n} \|off^i - \hat{off^i}\|_1 + \|wh^i - \hat{wh^i}\|_1 \quad (3)$$

$$L_{ID} = -\sum_{i=1}^{n}\sum_{j=1}^{J} A^i(j)\log(p(j)), \quad (4)$$

where off^i represents the values of the label, $\hat{off^i}$ and $\hat{wh^i}$ represent the predicted outputs from the heads of the model, $A^i(j)$ represents the class label, and $p(j)$ is a class distribution obtained by the model.

$$L_T = \frac{1}{2}\left(\frac{1}{e^{w1}}\left(L_H + \frac{1}{2}(L_{B1} + L_{B2})\right) + \frac{1}{e^{w2}}L_{ID}\right) \quad (5)$$

where $w1$ and $w2$ represent the weights for different losses, L_{B1} represents the loss of Frame T, and L_{B2} represents the loss of Frame T-1.

The experiment of this study used pre-trained parameters, that is, the pre-trained parameters of CenterNet on the coco dataset [49]. Since CenterNet does not contain ReID parameters, only the parameters of the target detection part need to be used. After the parameter migration, training is performed on the training set of visdrone2019, and the learning rate is performed in an attenuated method, which will be reduced after a particular epoch, and we use Adam [50] as the optimizing process. In the testing phase, models are tested on the validation set with a single RTX 3090ti to compare the performance of the models.

3. Experimental Results

3.1. Data Introduction and Processing

This research dataset uses the UAV video sequences of visdrone2019 [51]. The dataset includes a training set and validation set. The training set contains 56 UAV video sequences. The training set has a total of 24,201 frames, and the validation set contains 7 video sequences, which have a total of 2819 frames. Each video is an optical image containing different scenes and targets. Each frame in the same video has the same size and format, and other video sequences have different image sizes and shooting methods. There are 10 categories in total. The main target categories of this dataset are pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. The corresponding label format of the dataset used this time is the coco format. Table 1 is the distribution of the VisDrone2019 dataset in this study. Compared with the target detection dataset, the label of this dataset has ID information. Therefore, the information provided by the original data of the label mainly includes the serial number of the frame number, the target ID, the coordinates of the top-left vertex of the bounding box, the width and height of the bounding box, and whether the target is occluded and whether it needs to be ignored.

Table 1. The distribution of the VisDrone2019 dataset in this study.

	Training Set	Validation Set
Video Sequences	56	7
Number of Frames	24,201	2819
Category	10	10

The number of frames is the total frames in the sequences, and the category is the number of an object class in the video dataset.

Compared with the single-category MOT tasks, multi-category MOT requires additional processing on the dataset. This research is set to a total of 10 categories for the original label, so the training data construction needs to deal with the ID of multi-category. The

primary way is to count the ID according to the category; that is, the ID of each category starts counting from 0. Compared with single-category tasks, it is necessary to count the number of target IDs of each category that appears in the entire video sequence and use them as the number of classes for the ReID branch during the model training. The model was trained with data augmentation, and each image imported into the model was processed by rotation and scaling, which improved the training effect of the model.

Evaluation Index

The evaluation indicators involved in this research mainly contain CLEAR metrics [52], including MOTA, MOTP, MT, ML, and IDF1 [53]. The corresponding calculation equations are as follows. MOTA represents the comprehensive performance of the model in MOT, and MOTP represents the average degree of overlaps of all tracked targets; MT and ML, respectively, represent the number of trajectories that are successfully tracked and the number of courses that have failed to follow. FN represents the number of detected objects missing in the target detection, FP represents the number of erroneously detected objects in the detection, and IDS represents the number of ID switches for the same objects.

$$\text{MOTA} = 1 - \frac{\sum_t (N_t + P_t + ids_t)}{\sum_t gt} \tag{6}$$

$$\text{MOTP} = \frac{\sum\limits_{i,t} T_t^i}{\sum\limits_t d_t} \tag{7}$$

where N_t represents the number of missed objects in frame T, P_t represents the number of the wrong predictions in frame T, idst represents the number of the ID switch in frame T, T_t represents the number of matches in frame T between objects and the hypothesis, and dt represents the distance between targets and the hypothesis.

MOTP has two different calculation and evaluation criteria in this paper. The criteria are that if the tracked match is perfect, MOTP is 100%, and if it deviates completely, it is 0. The larger the MOTA result, the better the overall performance of the model on multi-target tracking tasks, and the maximum value of MOTA is 100%. FN and FP represent the error of target detection, and IDS represents the number of ID switches in the tracking task. The larger the value of these indicators, the worse the MOT effect. The indicators used in this study are mainly from clear mot metrics. By measuring the differences between the indicators, the detection and tracking performance of the models can be evaluated well.

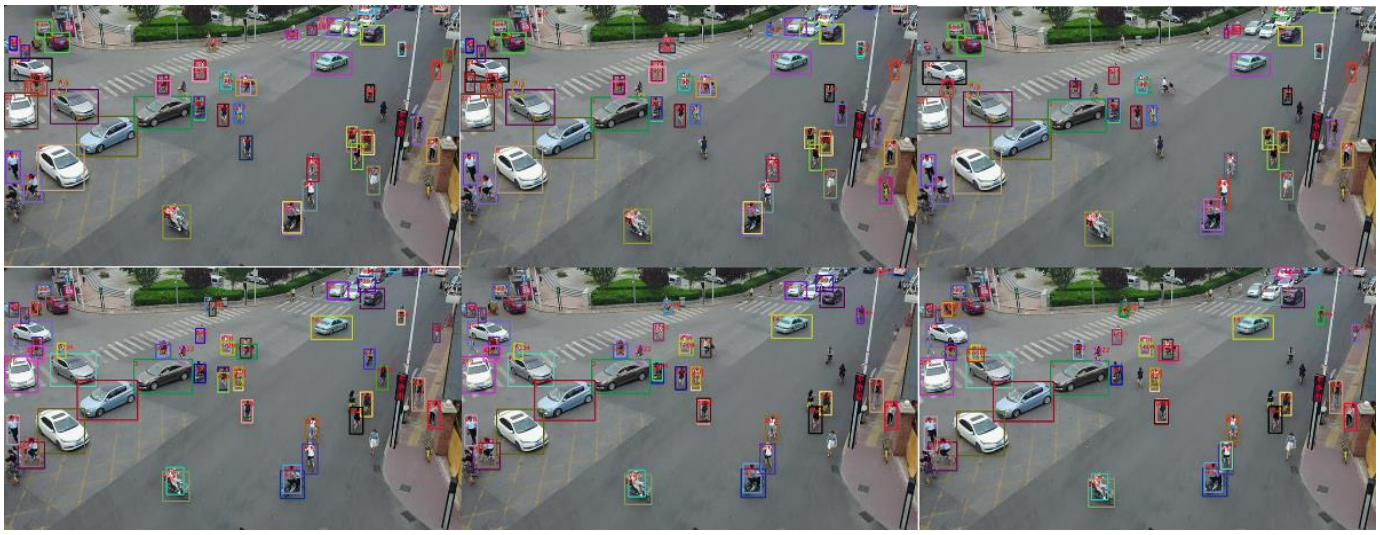
3.2. Experimental Results

This section mainly compares the performance of two feature enhancement structures constructed with temporal information. We add the two structures to the model for experiments, guarantee the same training and testing environment, and compare the performance of the two structures. The comparison method of the experiment in this section is mainly carried out by evaluating the difference of each indicator and comparing the performance shown by the visualization effect of each model.

The detailed information and structures of the modules are shown in Figure 4. Structure A combines the temporal information of two frames, including the embedding and heatmap obtained from the last frame. Structure B uses the features of two frames from the decoder part. In this part, the heatmap T-1 can be provided by the label information during the training phases. During the testing phase, the heatmap information can be obtained from the model in the last frame.

Figures 5 and 6 show the visualization effects of two different structures in the video sequences of the testing set. Figure 6 selects a part of the results of Figure 5 to enlarge the display, which can make the visualization effect of the model in the MOT task clearer. Table 2 shows the statistics of each indicator value on the testing set after adding two structures to the proposed model. The results in Table 2 show that the feature enhance-

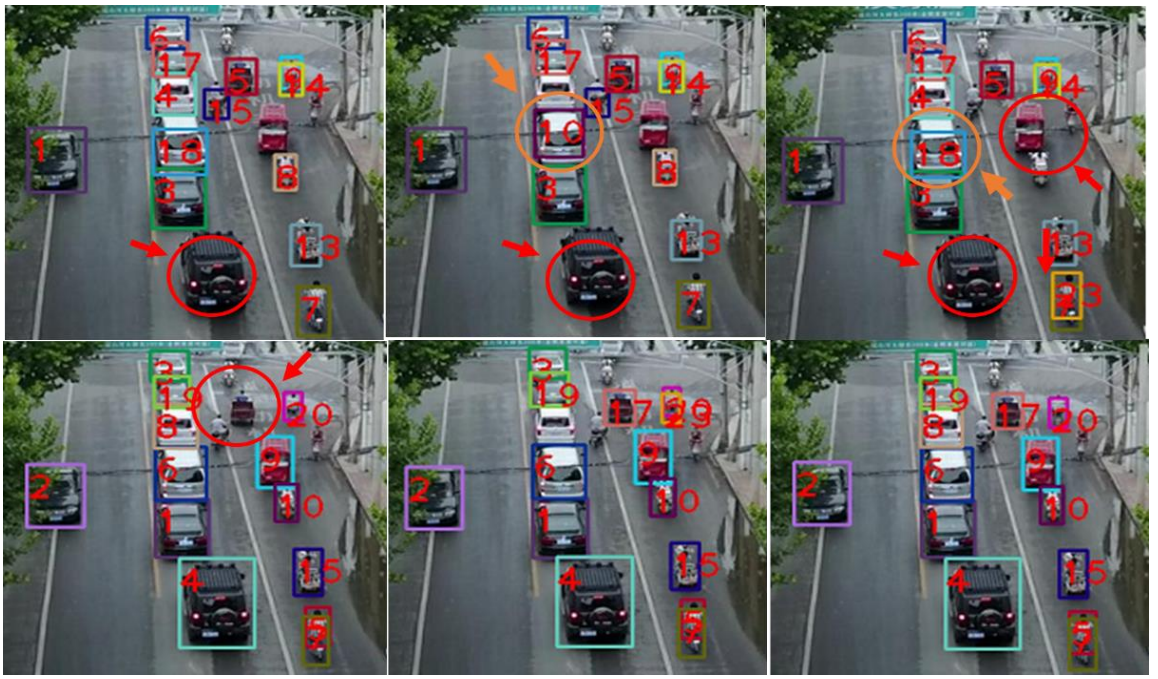
ment of structures A and B do not have much difference in the performance of the model's detection function. Structure A has a smaller total number of FN and FP than structure B. There is a difference in the function of ReID. It can be seen from the statistical table that the two models with different structures in the experiment have a specific difference in the IDS problem. When the FN is not notably different, structure A has a smaller IDS value than B, and structure A improves MOTA by 1.9% and MOTP by 0.3%, so it has a better improvement in ReID. The smaller number of IDS means fewer ReID errors in the testing phase, which provides better visual performance.



(**A**)

(**B**)

Figure 5. The visual results of MOT obtained by the two structures added to the ReID head, which displays two video sequences, including (**A**,**B**), in the validation set. The first row shows the visual results of structure B, and the second row shows the visual results of structure A. Every detected object is positioned by a bounding box of different sizes and is assigned an ID. In the three frames above, the same object is assigned the same ID in every frame. There are ten categories of objects detected in the video dataset at this time. The visualization of this multi-target tracking task only shows specific ID information, location and bounding boxes of every detected object.



(**A**)

(**B**)

Figure 6. (**A**,**B**) are partially enlarged results of (A,B) in Figure 5. The arrangement of the images is the same as in Figure 5, which displays two video sequences in the validation set. The first row shows the visual results of structure B, and the second row shows the visual results of structure A. In every image, the red circle and red arrow mean that the object to be detected is missed in this image, and the orange circle and arrows mean that the object has an ID switch between two adjacent frames.

Table 2. Quantitative comparison of two structures used in the ReID head of the proposed model.

	MOTA	MOTP	IDF1	MT	ML	FN	FP	IDS
Structure A	**34.7**	**74.5**	45.2	164	265	**57,848**	**14,385**	**2349**
Structure B	32.8	74.2	**45.9**	**175**	**257**	57,939	16,106	2706

Structures A and B are used for feature enhancement added to the ReID branch, and the details are explained in Figure 4. The bold values in the table mean the best results.

As shown in Figures 5 and 6, the Figures illustrate the visual results of two structures in the two video sequences. The visualization results show that the two structures have similar effects on target detection. Most targets with different classes can be detected in the following samples, which means that the models in this experiment can achieve the object detection task. Only a part of the targets is not well detected. Compared with the larger and closer targets, some distant targets and occluded objects are more challenging to be accurately identified.

As for another aspect, there also are some IDS issues in the visualization results. From Figure 6, we can see that the IDS problem always appears when some objects are occluded in the last frame, and these objects appear in the next frame. If some objects are occluded, the embeddings of the object may be changed or not precious for the same object, which will influence the process of ReID and the MOT performance of the model. In the visualization results, compared with structure B in the task of ReID, structure A has a better ReID effect, and there are fewer IDS issues.

In order to verify whether different model structures have an impact on the performance of MOT, in this section, we have added corresponding ablation experiments. Table 3 counts the numerical results of the ablation experiment. Methods of this ablation experiment include: (1) FairMOT, which is the baseline for this experiment; (2) adjusted FairMOT, which only splits the detection and ReID branches; (3) the model changes the outputs from a single frame to the two adjacent frames during the training phase; (4) based on (3), the model is improved by adding the heatmap information of the previous frame and adding feature enhancement structure A.

Table 3. Quantitative comparison of the ablation experiment.

	MOTA	MOTP	IDF1	MT	ML	FN	FP
baseline	29.8	73.3	**46.1**	**183**	279	58,657	17,683
+split structure	32.6	73.3	44.9	164	278	59,865	14,855
+two-frame output	32.5	74.2	45.3	167	**260**	58,695	15,376
+centermap, attention	**34.7**	**74.5**	45.2	164	265	**57,848**	**14,385**

The structures of different models in the table are explained in the following paragraph. The bold values in the table mean the best results.

The result shows that compared with the original baseline, the final model significantly improved target detection and tracking performance. In the target detection part, the number of missed and wrong detections are reduced, the total number of FN and FP has dropped, and the effect has increased by 4.9% and 1.2% on MOTA and MOTP, respectively, compared with the baseline. Model (2) splits the detection and ReID heads, which improves the efficiency of multi-object detection and tracking tasks during training and can obtain more representative features for the MOT. From the result, we can see that the detection performance has been improved compared with the baseline. As for Model (3), the output part has been changed from a single frame to two adjacent frames. Compared with Model (2), Model (3) improved the MOTP by 0.9%, though there is no noticeable improvement effect in object detection from the numbers of FN and FP. Figures 7 and 8 show the visual comparison of the performance of the proposed model compared with the FairMOT under this validation dataset. The visualization effect shows that the proposed model of this paper has a certain degree of improved performance in multi-category MOT compared with the baseline.

The results show that in the MOT task based on UAV video, the baseline model did not detect the most targets. As for detecting some small objects and objects that may be occluded, the model may miss these objects. On the other hand, it can be seen from the visualization results that the ID switch has appeared in the samples in the function of ReID of the baseline model. The problem of ID switch mainly occurs when two objects meet,

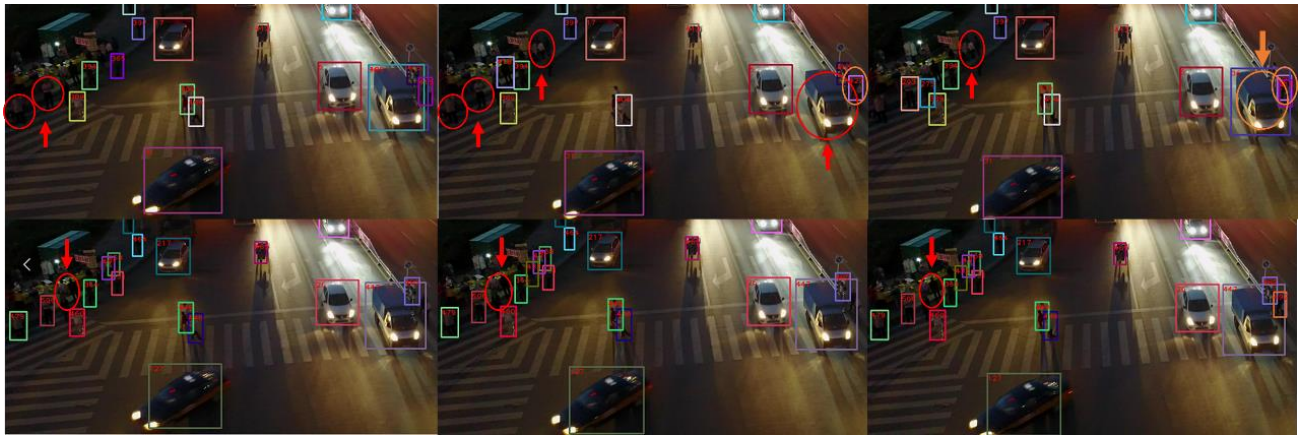
which may change the feature representation of the object between two frames. Therefore, the use of temporal information can improve this problem.

The numerical results of the quantitative comparison experiment are presented in Table 4, which includes an MOT evaluation matrix. Compared with the models that complete the MOT task on this dataset, the proposed model increases MOTA by 1.8% and IDF1 by 2.9% compared with HDHNet [54] and increases MOTA by 4.9% and MOTP by 1.2% compared with FairMOT. The proposed model has a smaller number of the sum of FN and FP than the other. The proposed model has a better object detection and ReID effect in the visualization results, which shows that the model can complete the MOT task. Thus, from the numerical and visualization results, the proposed model has improved the performance of the multi-category MOT task in this dataset.

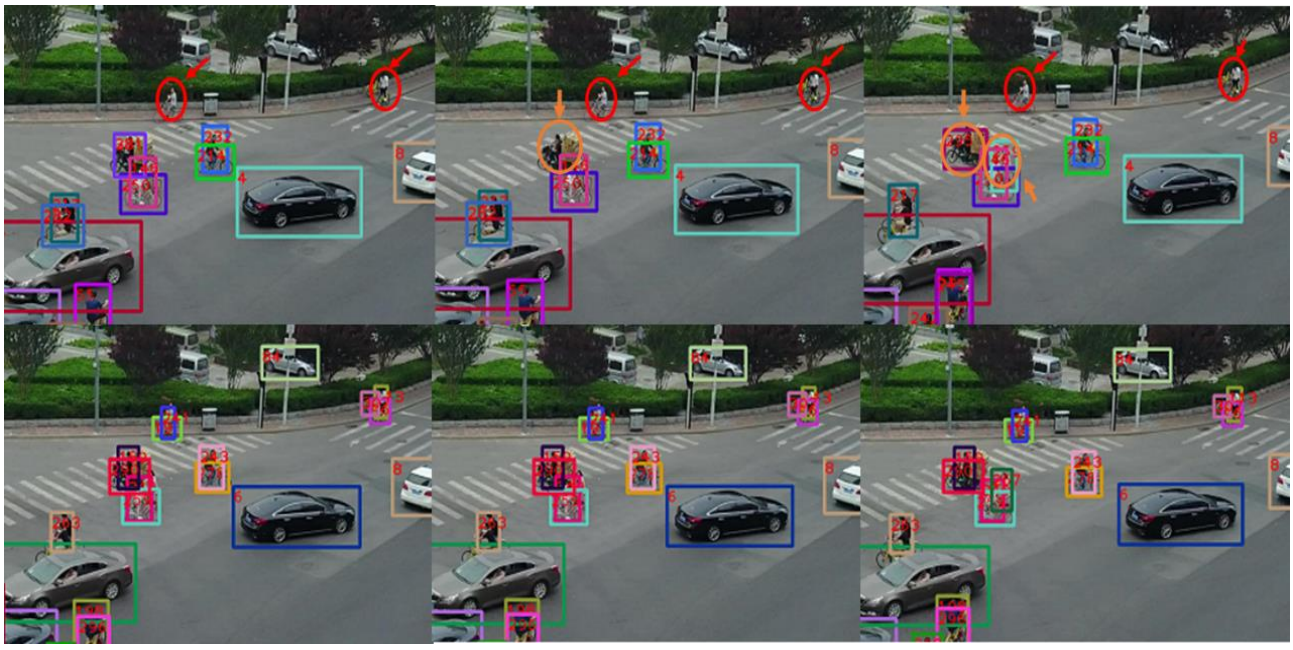
(**A**)

(**B**)

Figure 7. The visual results of MOT obtained by the two models, including the baseline and the proposed model, which displays two video sequences in the validation set. (**A**,**B**) are two examples in the two video sequences. The first row shows the visual results of FairMOT, and the second row shows the visual results of the proposed model. Every detected object is positioned by a bounding box of different sizes and is assigned an ID. In the three frames above, the same object is assigned the same ID in every frame.



(**A**)



(**B**)

Figure 8. (**A**,**B**) are partially enlarged results of (A,B) in Figure 7. The arrangement of the images is the same as Figure 7, which displays two video sequences in the validation set. The first row shows the visual results of structure B, and the second row shows the visual results of structure A. In every image, the red circle and red arrow mean that the object to be detected is missed in this image, and the orange circle and arrows mean that the object has an ID switch between two adjacent frames.

Table 4. Quantitative comparison of the different models in the VisDrone2019 dataset.

	MOTA	MOTP	IDF1	MT	ML	FN	FP
GGDTRACK [55]	23.4	/	**48.1**	/	/	42,917	12,630
SORT	18.1	65.1	32.2	/	/	78,467	104,453
HDHNet	32.9	**76.9**	42.3	/	/	35,686	80,454
FairMOT	29.8	73.3	46.1	**183**	279	58,657	17,683
Proposed Model	**34.7**	74.5	45.2	164	**265**	**57,848**	**14,385**

Some indicators of the models mentioned in this table are not provided in the corresponding references. The bold values in the table mean the best results.

4. Discussion

In this research, the multi-category multi-object tracking task based on UAV video sequences is realized using the proposed model in this paper. It can be seen from the

results of the first part of the ablation experiment that making good use of the temporal information and adjusting the corresponding network structure can improve the model's performance in detecting and tracking MOT tasks. Compared with the baseline structure, the structure that separates the detection and ReID branches improves the detection ability and reduces missed and false detection problems.

Different function parts can be separated to improve the structure and train the model on detection and ReID tasks. The original design integrates detection and ReID modules, and there will be some training conflicts during the training phase; that is, detection minimizes the distance between objects of the same category, and ReID maximizes the distance of objects within the same categories. In this study, separating the detection and ReID branches of the model in the encoder part can make the training of the two tasks more independent. From the overall structure of the model, the model uses the information of two adjacent frames in the input, and in the detection part, it realizes more independent training and inference based on the use of historical features, which weakens the influence of the ReID structure. In the output part of the model's detection branches, it changes the output format from single-frame output to a double-frame format, which can make the output of adjacent frames perform loss calculation at the same time, improves the training efficiency of the model, and uses the temporal information in the detection part during training. The model with the output structure of adjacent frames has an improvement of about 2.7% in MOTA performance compared with the baseline.

The proper way to use temporal information can help improve the ReID performance of the model. According to the results of the ablation experiment, by adding feature enhancement of structure A, the MOT performance of the model has been improved to a certain extent. In contrast, the feature enhancement structure B did not improve the model's performance in the ReID branch. After analyzing the structure, structure B needs to use the heatmap of the previous frame as auxiliary information for consideration, which is directly input into the network for calculation. At this time, the label information is used. However, the training efficiency of the model can be maintained during training; during the testing phase, the model will use the heatmap provided in the previous frame as the input of the branch, and the accuracy of the heatmap offered by the model is not as accurate as the label information, which calculates it directly as features will cause a specific deviation in the result. In subsequent experiments, other solutions were also implemented, such as now using the heatmap generated by the model as input during training. Firstly, using label information as training input after a particular round of iteration, then replacing it with the model's output. After comparing the two schemes, the effect has not been improved. As for structure A, the temporal information of the videos is also used. The object's center position tracked in the previous frame is input into the prediction as auxiliary information for the current frame. The ReID feature between the two frames is matched by the similarity of the embedding quality. From the result, it can be seen that structure A is better than structure B in the data association performance of MOT.

Adding temporal data to the model can also improve detection ability on the MOT task. From the data association process of DeepSORT, the matched target first needs to be detected. If an object in the video sequence is occluded or the size has changed, the detection result will change, and the target may be classified as the background. There is a case of missing detection or a change in the ReID embedding, which makes the model match the wrong object during the inference phase. This module expands the range of feature matching, merges embedding with the feature with the most significant similarity on the feature map, and combines the apparent elements of the previous frame, making the ReID features obtained during the training and testing phases of the model more sequential, which can improve the performance of the ReID task to a certain extent. Compared with the structure of the heatmap branch of the baseline, the model in this paper can improve the response to the position of the object by generating the centered map of the previous frame and adding it to the feature map of the current frame, which is used as auxiliary information in the target detection part of the model. It can be seen from the structure of the

model that, compared with the original structure, the model in this paper has been adjusted in the input and calculation process. The performance of the multi-target tracking of the model is improved, and the speed of training and inference of the model itself is slightly reduced compared with the original model. For example, in the post-processing process, with a single 3090ti, the processing speed of the video streams can be maintained at about 15 FPS, which is lower than the original structure. However, the real-time performance of the model for MOT tasks of video sequences can still be guaranteed, and the performance of multi-target tracking can be improved better at the expense of a little calculation speed compared with the original model. We will also try to improve the model algorithm and post-processing flow in the future to explore methods to enhance speed and accuracy.

5. Conclusions

Our study proposes an improved MOT model based on FairMOT, which can realize end-to-end detection and tracking of multi-category objects in UAV video sequences. As for the original structure, the target detection and ReID tasks may have some conflicts during training. We separate the detection and ReID branches to make the two parts more independent and improve the detection accuracy. Additionally, the model in this paper uses temporal information in target detection and the ReID head, combines the central point features of historical frames, and includes a feature enhancement structure to improve the tracking performance of the model on UAV video sequences. Finally, compared with other MOT models on this study's drone video dataset, the use of the proposed model can achieve better multi-category and multi-object tracking performance. Although making good use of temporal information can improve the tracking performance of the model, there are still some scenes where objects have similar appearance characteristics, which will affect the results of data association during the process of history frame association. Therefore, it is necessary to focus on the association of similar objects in historical frames in subsequent research. Furthermore, the temporal information contained in adjacent frames is still limited. In the follow-up research, we will also try to explore the use of multi-frame and long-term information and apply it to tracking tasks to improve long-span tracking tasks while ensuring the accuracy and real-time performance of the model.

Author Contributions: Conceptualization, Y.L. (Yong Liu) and Y.L. (Yeneng Lin); methodology, Y.L. (Yeneng Lin); software, Y.L. (Yeneng Lin); validation, Y.L. (Yeneng Lin); formal analysis, Y.L. (Yeneng Lin); writing—original draft preparation, Y.L. (Yeneng Lin); writing—review and editing, M.W., W.C., L.L., Y.L. (Yong Liu) and W.G.; visualization, Y.L. (Yeneng Lin); supervision, M.W. and W.C.; project administration, Y.L. (Yong Liu); All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Key Research and Development Project of Zhejiang Province under Grant 2021C01035.

Data Availability Statement: Publicly available dataset was analyzed in this study. VisDrone 2019 dataset can be found here: http://aiskyeye.com/. The detailed indication of the dataset presented in this study is available in reference [51].

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

References

1. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
2. Patrício, D.I.; Rieder, R. Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Comput. Electron. Agric.* **2018**, *153*, 69–81. [CrossRef]
3. Janai, J.; Güney, F.; Behl, A.; Geiger, A. Computer Vision for Autonomous Vehicles: Problems, Datasets and State of the Art. *Found. Trends Comput. Graph. Vis.* **2020**, *12*, 1–308. [CrossRef]
4. Ballesteros, R.; Intrigliolo, D.S.; Ortega, J.F.; Ramírez-Cuesta, J.M.; Buesa, I.; Moreno, M.A. Vineyard yield estimation by combining remote sensing, computer vision and artificial neural network techniques. *Precis. Agric.* **2020**, *21*, 1242–1262. [CrossRef]
5. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]

6. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable ConvNet with Aspect Ratio Constrained NMS for Object Detection in Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
7. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
8. Ahmed, I.; Ahmad, M.; Ahmad, A.; Jeon, G. Top view multiple people tracking by detection using deep SORT and YOLOv3 with transfer learning: Within 5G infrastructure. *Int. J. Mach. Learn. Cybern.* **2020**, *12*, 3053–3067. [CrossRef]
9. Wang, Z.; Miao, D.; Zhao, C.; Luo, S.; Wei, Z. A Robust Long-Term Pedestrian Tracking-by-Detection Algorithm Based on Three-Way Decision. In Proceedings of the International Joint Conference on Rough Sets (IJCRS), Debrecen, Hungary, 17–21 June 2019; pp. 522–533.
10. Xie, Y.; Huang, Y.; Song, T.L. Iterative joint integrated probabilistic data association filter for multiple-detection multiple-target tracking. *Digit. Signal Process.* **2018**, *72*, 232–243. [CrossRef]
11. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards Real-Time Multi-Object Tracking. In *Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioin-Formatics)*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12356. [CrossRef]
12. Munjal, B.; Aftab, A.R.; Amin, S.; Brandlmaier, M.D.; Tombari, F.; Galasso, F. Joint detection and tracking in videos with identification features. *Image Vis. Comput.* **2020**, *100*, 103932. [CrossRef]
13. Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; Luo, P. TransTrack: Multiple-Object Tracking with Transformer. *arXiv* **2012**, arXiv:2012.15460.
14. Lin, X.; Guo, Y.; Wang, J. Global Correlation Network: End-to-End Joint Multi-Object Detection and Tracking. *arXiv* **2021**, arXiv:2103.12511.
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27–30 June 2016; pp. 779–788.
16. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
17. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
19. Wang, X.; Shrivastava, A.; Gupta, A. A-Fast-RCNN: Hard positive generation via adversary for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3039–3048.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
22. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9905. [CrossRef]
23. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12346. [CrossRef]
24. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Honolulu, HI, USA, 21–26 July 2017.
26. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016.
27. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar BB, G.; Geiger, A.; Leibe, B. Mots: Multi-object tracking and segmentation. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7934–7943.
28. Hu, Y.T.; Huang, J.b.; Schwing, A.G. MaskRNN: Instance level video object segmentation. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017.
29. Zhou, X.; Austin, U.T.; Wang, D.; Berkeley, U.C.; Austin, U.T. Object as Point. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
30. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, ICCV, Seoul, Korea, 27 October–2 November 2019; pp. 9626–9635.
31. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3D Object Detection and Tracking. In Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 11779–11788.

32. Lu, Z.; Rathod, V.; Votel, R.; Huang, J. RetinaTrack: Online single stage joint detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 14656–14666.
33. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking Objects as Points. *arXiv* **2020**, arXiv:2004.01177.
34. Wu, J.; Cao, J.; Song, L.; Wang, Y.; Yang, M.; Yuan, J. Track to Detect and Segment: An Online Multi-Object Tracker. *arXiv* **2021**, arXiv:2103.08808.
35. Liang, C.; Zhang, Z.; Zhou, X.; Li, B.; Lu, Y.; Hu, W. One More Check: Making "Fake Background" Be Tracked Again. *arXiv* **2021**, arXiv:2104.0944. [CrossRef]
36. Wu, J.; Su, X.; Yuan, Q.; Shen, H.; Zhang, L. Multi-Vehicle Object Tracking in Satellite Video Enhanced by Slow Features and Motion Features. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–26. [CrossRef]
37. Xuan, S.; Li, S.; Han, M.; Wan, X.; Xia, G.-S. Object Tracking in Satellite Videos by Improved Correlation Filters With Motion Estimations. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 1074–1086. [CrossRef]
38. Yang, X.; Wang, Y.; Wang, N.; Gao, X. An Enhanced SiamMask Network for Coastal Ship Tracking. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
39. Shao, J.; Du, B.; Wu, C.; Zhang, L. Tracking Objects From Satellite Videos: A Velocity Feature Based Correlation Filter. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7860–7871. [CrossRef]
40. Lei, L.; Guo, D. Multitarget Detection and Tracking Method in Remote Sensing Satellite Video. *Comput. Intell. Neurosci.* **2021**, *2021*, 7381909. [CrossRef]
41. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the International Conference on Image Processing, ICIP, Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
42. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the International Conference on Image Processing, ICIP, Beijing, China, 17–20 September 2017; pp. 3645–3649.
43. Wu, J.; Cao, C.; Zhou, Y.; Zeng, X.; Feng, Z.; Wu, Q.; Huang, Z. Multiple Ship Tracking in Remote Sensing Images Using Deep Learning. *Remote Sens.* **2021**, *13*, 3601. [CrossRef]
44. Kraus, M.; Azimi, S.M.; Ercelik, E.; Bahmanyar, R.; Reinartz, P.; Knoll, A. AerialMPTNet: Multi-Pedestrian Tracking in Aerial Imagery Using Temporal and Graphical Features. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 2454–2461.
45. 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]
46. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the 31st IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412.
47. Welch, G.; Bishop, G. An introduction to the Kalman filter. *Course Notes ACM SIGGRAPH* **1995**, *8*, 127–132.
48. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Image Restoration With Neural Networks. *IEEE Trans. Comput. Imaging* **2017**, *3*, 47–57. [CrossRef]
49. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755.
50. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Wen, L.; Zhu, P.; Du, D.; Bian, X.; Ling, H.; Hu, Q.; Zheng, J.; Peng, T.; Wang, X.; Zhang, Y.; et al. VisDrone-MOT2019: The vision meets drone multiple object tracking challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 189–198.
52. Bernardin, K.; Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
53. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Volume 9914, pp. 17–35.
54. Huang, W.; Zhou, X.; Dong, M.; Xu, H. Multiple objects tracking in the UAV system based on hierarchical deep high-resolution network. *Multimed. Tools Appl.* **2021**, *80*, 13911–13929. [CrossRef]
55. Ardo, H.; Nilsson, M. Multi-target tracking from drones by learning from generalized graph differences. In Proceedings of the IEEE International Conference on Computer Vision, ICCV, Seoul, Korea, 27 October–2 November 2019; pp. 46–54.