symmetry

MDPI

Article

Attention Modulated Multiple Object Tracking with Motion Enhancement and Dual Correlation

Yifeng Wang [1], Zhijiang Zhang [1], Ning Zhang [2] and Dan Zeng [1,*]

[1] Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China; wyf1129@shu.edu.cn (Y.W.); zjzhang@staff.shu.edu.cn (Z.Z.)

[2] JD AI Research, Mountain View, CA 94040, USA; ning.zhang@jd.com

* Correspondence: dzeng@shu.edu.cn; Tel.: +86-131-2230-0551

Abstract: The one-shot multiple object tracking (MOT) framework has drawn more and more attention in the MOT research community due to its advantage in inference speed. However, the tracking accuracy of current one-shot approaches could lead to an inferior performance compared with their two-stage counterparts. The reasons are two-fold: one is that motion information is often neglected due to the single-image input. The other is that detection and re-identification (ReID) are two different tasks with different focuses. Joining detection and re-identification at the training stage could lead to a suboptimal performance. To alleviate the above limitations, we propose a one-shot network named Motion and Correlation-Multiple Object Tracking (MAC-MOT). MAC-MOT introduces a motion enhance attention module (MEA) and a dual correlation attention module (DCA). MEA performs differences on adjacent feature maps which enhances the motion-related features while suppressing irrelevant information. The DCA module focuses on decoupling the detection task and re-identification task to strike a balance and reduce the competition between these two tasks. Moreover, symmetry is a core design idea in our proposed framework which is reflected in Siamese-based deep learning backbone networks, the input of dual stream images, as well as a dual correlation attention module. Our proposed approach is evaluated on the popular multiple object tracking benchmarks MOT16 and MOT17. We demonstrate that the proposed MAC-MOT can achieve a better performance than the baseline state of the arts (SOTAs).

Keywords: multiple object tracking; deep learning; attention mechanism

Citation: Wang, Y.; Zhang, Z.; Zhang, N.; Zeng, D. Attention Modulated Multiple Object Tracking with Motion Enhancement and Dual Correlation. *Symmetry* **2021**, *13*, 266. https://doi.org/10.3390/sym13020266

Academic Editor: Stuart Rubin
Received: 12 January 2021
Accepted: 1 February 2021
Published: 4 February 2021

1. Introduction

Multiple object tracking (MOT) aims at estimating the locations of multiple objects and providing identifications. The MOT plays a fundamental role in many applications, such as intelligent driving [1], human–computer interaction, and pedestrian behavior analysis [2,3].

Currently, MOT can be summarized into two paradigms: two-stage and one-shot frameworks. The two-stage framework follows the tracking-by-detection paradigm [4–7], which divides MOT into two separate tasks: (i) object detection in terms of getting the objects bounding boxes, and (ii) object feature extraction and association in terms of finding the objects' tracking traces across the time course. At the detection stage, convolution neural network (CNN) detectors such as Yolov3 [8] and Faster-RCNN [9] are applied to localize all objects of interest in the images by a number of bounding boxes. Then, these images are cropped and resized into the same resolution. At the feature extraction and association stage, data association across frames is conducted based on the re-identification.

In this two-stage approach, earlier literature works use color histogram [10–12] and pixel-based template representation [13,14] features to describe the objects. Lately, feature extraction based on CNNs has been popular. There have been many works with different structures, such as the Siamese network [15], which are capable of learning discriminative

features cropped from the detected object bounding boxes. Some works use the appearance-based cost function while others pursue more complicated association mechanisms, such as motion and topology cues. This group of the methods [16–18] has achieved promising results on the public benchmark datasets. Finally, the Hungarian algorithm [19] is used to optimize the cost function to accomplish the optimal association results. Although effective in performance, the two-stage paradigm is time-consuming [20]. This is because the representation model needs to extract the ReID features from each individual bounding box. Furthermore, the two-stage paradigm depends on the performance of the detector.

On the other hand, the one-shot paradigm has drawn more attention in the MOT research community, with its maturity of multi-task learning in deep learning. The one-shot model integrates the detection and representation tasks into a unified system. This would reduce the inference time, which makes the system more real-time. However, the tracking accuracy of the one-shot paradigm is not as high as in the two-stage paradigm. The one-shot approach combines the detection and re-identification, which overlooks the specificity and commonality of different tasks in learning and the assistance role of motion information. The one-shot approach is capable of providing the target location and its associated embedding feature simultaneously, derived from a shared feature map. Nevertheless, there is an issue regarding the design. This is that two tasks of detection and re-identification are trained at the same time, using the same image input. The inherent differences between the detection and the re-identification tasks are overlooked. This creates a competition due to the different optimization processes. Specifically, the loss of the detection task is to narrow the feature distance between the different objects, while that of the re-identification is the opposite. This would result in learning a suboptimal solution, which degrades the performance. Additionally, the other problem is that current one-shot paradigms have a lack of motion information. Motion information plays an important role in understanding the objects' movement. The object motion is often different from the background movement, due to the camera motion. Hence, a lack of a motion modulated model would miss the critical motion information for the objects.

To alleviate the above issues of the one-shot paradigm, we propose a novel Motion and Correlation-Multiple Object Tracking (MAC-MOT) model to improve the performance of the one-shot tracking. First, symmetry plays an essential role in designing our proposed framework. The details are explained in the following: (i) The backbone is a Siamese-based deep learning network, where two branches of the Siamese are symmetric in design. (ii) The input of dual stream images has two input branches, where images are fed in consecutive order into the system in a symmetric manner. (iii) A dual correlation attention module, which is proposed to balance the two output streams of the proposed framework. All of these reflect the idea of symmetry.

Second, the MAC-MOT framework has the following components. Initially, it takes adjacent frames as input and uses the multi-scale convolution network to extract features, respectively. We propose a motion enhance attention module (MEA), which utilizes motion information as a guide to focus on important features. MEA exploits global average pooling (GAP) to squeeze the feature maps to only focus on the channel-level importance. Subsequently, it uses the temporal difference as an approximate motion map. To solve the competition between detection and ReID tasks, we introduce a dual correlation attention module (DCA) to decouple the input. Inspired by the Dual Attention Network (DANet) [21], we use the self-attention mechanism to obtain the task-related features. Meanwhile, we use the correlation mechanism to protect the commonality of features.

In summary, the main contribution of our work is two-fold:

(1) We propose a motion enhance attention module (MEA) to model the motion-relate feature, which obtains the weight for the feature channel.
(2) We introduce a dual correlation attention module (DCA) in order to reduce ambiguity in learning different tasks. DCA makes the one-shot method more adaptive to multi-task-based representation learning.

This paper is presented in the following structure. First, related work is introduced in Section 2. The proposed model is explained in Section 3. Experimentation and results are discussed in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

The two-stage method focuses on individual tasks at each stage. The individual stage can focus on a single task without risking other tasks' performance. However, the two-stage method is usually slow because both object detection and re-identification feature embedding need an inference effort. For this reason, the two-stage method cannot be used in many practical applications.

In order to overcome the shortcomings of the two-stage method, the one-shot method integrates the detection and re-identification into a unified framework, which simultaneously outputs the bounding box and the embedding feature. They can share the weight of the backbone in different tasks to reduce the time of inferencing. In detail, the one-shot framework consists of three parts, the feature extractor, the detection branch, and the re-identification branch. For instance, the Joint Detection Embedding (JDE) [20] approach models the training process as a multi-task learning problem with anchor classification, box regression, and embedding learning. The JDE architecture chooses the Feature Pyramid Network (FPN) [22] as its base architecture in order to extract multi-scale features. In the detection branch, it exploits the anchor-based detection paradigm, which sets a certain number of anchors of different sizes in each position in the feature map. Meanwhile, in the re-identification branch, consistent with general re-identification, it regards the re-identification as a classification task for network design.

However, the object may appear in the middle of two anchors, resulting in an ambiguity between the detection task and re-identification task. Therefore, the anchor-based method will cause the bounding boxes and embedding features to fail to align. Some studies [23,24] have shown that the anchor-free method is more suitable for the detection branch of the one-shot framework. They directly classify the object and background from the feature map and regress the bounding box. Although the one-shot methods have the advantage in regard to time, their accuracy is usually lower than that of the two-stage methods [24]. They tend to overlook the conflicts caused by a joint learning of the different classification and re-identification tasks. Additionally, they do not take into account the importance of the motion features on which the tracking performance often hinges. Inspired by the other computer vision tasks, such as action recognition [25] and semantic segmentation [21], we have improved the performance of one-shot MOT tracking with the proposed MAC-MOT model.

3. Materials and Methods

3.1. Overview

In this section, we introduce the MAC-MOT in detail. The overall framework of the network is described in Figure 1. The MAC-MOT consists of five components, a backbone, a motion-enhanced attention module, a dual correlation attention module, a detection branch, and a ReID branch. MAC-MOT takes adjacent frames as input, denoted by I_t and I_{t-1}. The backbone is exploited to extract multiple scale features from I_t and I_{t-1}. Then, the MEA utilizes the temporal difference between the features from the two frames to enhance the motion information. We are able to obtain the features that contain more motion information, denoted by φ_t. Next, the DCA module is proposed to decouple φ_t into a detection feature F_{det} and a re-identification F_{reid}. Finally, the detection branch tasks the F_{det} as input to get the bounding boxes of the objects, while the re-identification branch takes the F_{reid} as input to get the embedding feature of each object.

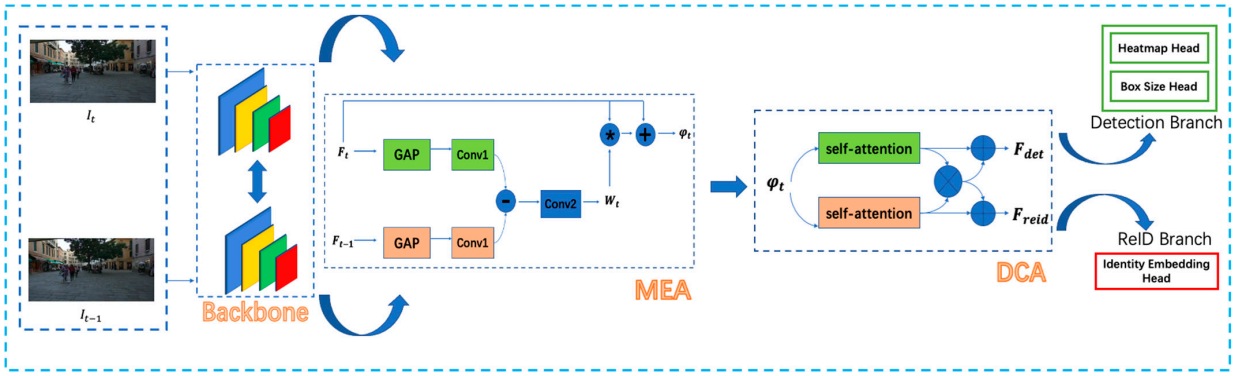

Figure 1. Pipeline of the Motion and Correlation-Multiple Object Tracking (MAC-MOT) framework. The adjacent frames, I_t and I_{t-1}, are fed to the backbone to obtain the multi-scale features. Then, the motion-enhanced attention module (MEA) makes the objects' motion feature more prominent. The dual correlation attention module's (DCA's) disassembly feature makes it more suitable for different tasks. * represents the element-multiplication. Ultimately, the detection branch outputs the locations and embedding features of the objects.

3.2. Motion Enhance Attention

We designed a MEA module in the one-shot method to focus on the motion salient features, while suppressing the irrelevant information in the background. MEA uses the temporal difference operation of the adjacent frame level features to enhance the motion related features in a channel-wise manner. Figure 2 shows two feature maps, F_t and F_{t-1}, with a tensor dimension of $C * H * W$, in which C is the channel, H is the height, and W is the width. F_t and F_{t-1} are the output of the backbone features extracted from the previous section. We first use the global adaptive pooling (GAP) to aggregate the feature maps across the spatial dimensions. This gives us a result of $C * 1 * 1$, which generates the channel-wise importance weights. This average pooling is along the H and W dimension. Then, these intermediated vectors go through two weight-sharing 2D convolutions with a kernel size of $\frac{C}{r} * 1 * 1$, which are exploited to convert the channel dimensions to $\frac{C}{r}$. This is effectively a process of feature dimension conversion. This conversion process reduces the channel size to refine the import features. We use an aggregation parameter r to control the output sizes.

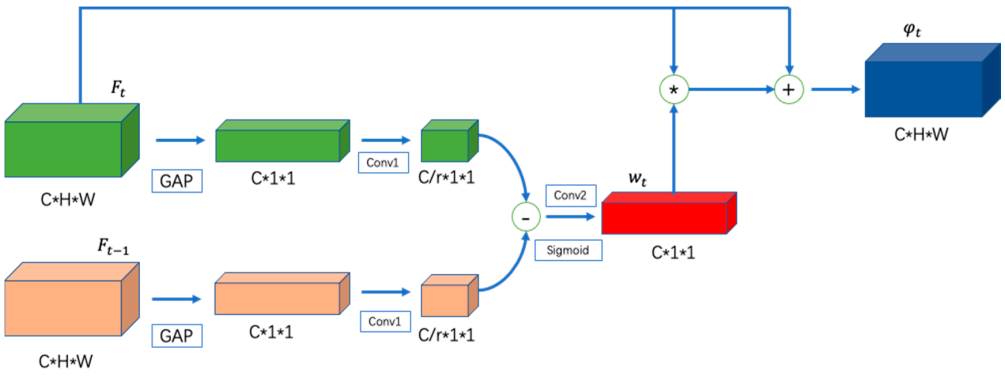

Figure 2. Structure of the motion enhance attention module. The features of the adjacent frames are aggregated by the global adaptive pooling (GAP). We use convolution operations to change the dimension of the features and reduce the redundancy. The difference operation is used to extract the temporal changes of objects from the adjacent frames. The channel importance weights act on the feature of the current frame in order to obtain more distinguishable features.

The pipeline of the difference calculation can be formulated as:

$$diff = Conv1(GAP(F_t)) - Conv1(GAP(F_{t-1})), \quad (1)$$

Subsequently, another 2D convolution with a kernel size of $C * 1 * 1$ is applied on the diff operator, in order to recover the dimension of the diff as for the input F_t. The attention weight is obtained by:

$$w_t = \sigma(Conv2(diff)), \tag{2}$$

The operator σ represents the sigmoid function, which is used to normalize the values of the weights to 0–1. We utilize channel-wise multiplication to enhance the motion-salient features φ_t.

$$\varphi_t = (1 + w_t) \cdot F_t, \tag{3}$$

In summary, the MEA uses the channel differences as the weight, in order to obtain better temporal information and motion features.

3.3. Dual Correlation Attention

Multi-task learning often suffers from task incompatibility, which leads to a decline in performance and even task failure. We propose a dual correlation attention to learn the commonalities and specificities of features for the detection and ReID tasks. For specificities learning, we use the self-attention mechanism to obtain the feature representation for each task. For commonalities learning, two tasks can be learned by an elaborately designed correlation mechanism.

The structure of our dual correlation attention module is presented in Figure 3. The feature φ_t, which is the output of the MEA module depicted in Figure 2, is split into two branches through two convolutions of the same structure with a kernel size of $C * 3 * 3$. The outputs are denoted by φ_1 and φ_2. Following the convolution, a reshape operator is implemented. Specifically, φ_1 and φ_2 are then reshaped to M_1 and M_2. The reason for this reshaping operation is due to a dimension matching in doing the matrix multiplication computation. As a result, the spatial two-dimensional distribution of the feature map is transferred to a one-dimensional distribution, where $N = H' * W'$. Subsequently, we perform the matrix multiplication on M_1 or M_2 and its corresponding transpose. A row softmax function is exploited to obtain the self-attention weight maps M_{T_1} and M_{T_2} for each task. Again, we perform the matrix multiplication between M_{T_1} and M_{T_2} to learn the commonalities between the different tasks, and then a row softmax is followed to generate the correlation attention weight maps M_c. M_c is added to M_{T_1} and M_{T_2}, aiming to get the enhanced representation, denoted by M_{S_1} and M_{S_2}. The enhanced representations are fused with φ_t by the attention to prevent information loss.

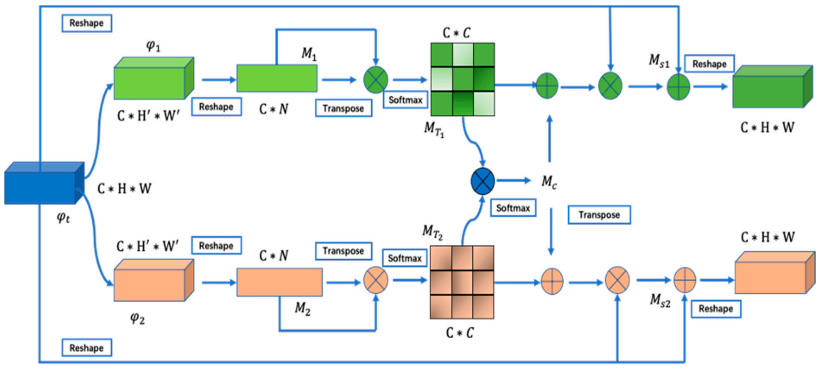

Figure 3. Structure of the DCA module. The shared features are divided into types that are applied to a detection task and a re-identification task. We use two convolutions with the same structure on the input feature map to aggregate local features. Then, we reshape the feature map to flatten it in the spatial dimension, thereby facilitating the calculation of matrix multiplication to obtain the self-attention weight and correlation weight. The self-attention mechanism is used to eliminate the ambiguity caused by different task requirements, while the correlation mechanism is used to obtain common parts from two tasks. Finally, we operate and fuse the original feature map and the obtained weights to output features that are suitable for different tasks.

3.4. Loss Function

The one-shot methods are composed of the detection task and re-identification task. Thus, we need to design corresponding losses for the two tasks at the same time.

Detection Branch. For a bounding box $b^i = \{x_1^i,\ y_1^i,\ x_2^i,\ y_2^i\}$, we need to know its center point coordinates, width and height. The detection branch contains two heads, a heatmap head and a box size head. The heatmap head is responsible for estimating the locations of the object's center. The response at a location where the object is in the heatmap is expected to be one. We define $\{C_x,\ C_y\}$ to be the center point of the object. Then, its location on the feature map is obtained by dividing the stride as follows:

$$\left(\widetilde{C_x},\ \widetilde{C_y}\right) = \left(\left\lfloor \frac{C_x}{stride} \right\rfloor \left\lfloor \frac{C_y}{stride} \right\rfloor\right), \tag{4}$$

The stride represents the down sampling rate of the network. The value of our backbone network is 4. The heatmap response at the location $\{x, y\}$ is computed as follows:

$$M_{xy} = \sum_{i=1}^{N} exp^{\frac{-\left(x-\widetilde{c_x}^i\right)^2+\left(y-\widetilde{c_y}^i\right)^2}{2\sigma^2}}, \tag{5}$$

where N represents the number of objects in the image and σ represents the standard deviation. The heatmap loss is defined as follows:

$$L_{heatmap} = -\frac{1}{N}\sum_{xy}\begin{cases} \left(1-\hat{M_{xy}}\right)^{\alpha} log\left(\hat{M_{xy}}\right),\ if\ M_{xy}=1 \\ \left(1-M_{xy}\right)^{\beta}\left(\hat{M_{xy}}\right)^{\alpha} log\left(1-\hat{M_{xy}}\right)\ otherwise \end{cases}, \tag{6}$$

$\hat{M_{xy}}$ is the estimated heatmap, M_{xy} is the ground truth, and α and β are the parameters. For each bounding box $b^i = \{x_1^i,\ y_1^i,\ x_2^i,\ y_2^i\}$, we define the size of the bounding box as $s^i = \{x_2^i - x_1^i,\ y_2^i - y_1^i\}$. The box size head output is denoted as $\hat{s}^i$. The box loss is defined as follows:

$$L_{box} = \sum_{i=1} \left\| s^i - \hat{s}^i \right\|, \tag{7}$$

Re-identification Branch. The goal of the identity embedding is to generate features that can distinguish different objects. We hope that the difference in the same object is much smaller than that of the different object. We apply a convolution layer to exact the identity embedding features for each location. We usually train a re-identification representation model as a classification task. All object instances of the same identity in the training set are treated as one class. The identity embedding loss is defined as follows:

$$L_{identity} = -\sum_{i=1}^{N}\sum_{k=1}^{K} L(K)^i log(p(K)), \tag{8}$$

where $p(K)$ represents the identity embedding features, $L(K)^i$ represents the distribution of the identity embedding features, and K is the number of classes.

Total Loss. We train the entire network to optimize the parameters of the two branches together. We add two trainable parameters to balance the two tasks, denoted by P_{det} and P_{reid}. The total loss is formulated as:

$$L_{total} = e^{-P_{det}} * \left(L_{heatmap} + L_{box}\right) + e^{-P_{reid}} * L_{identity} + \left(P_{det} + P_{reid}\right), \tag{9}$$

4. Experiments

In this section, we firstly introduce the relevant experimental settings. The proposed MAC-MOT has shown robust results on the MOT17 [26] and MOT16 [26] compared with the SOTA. Ablation study is also conducted on the MOT17 [26] testing dataset to verify the

effectiveness of the motion enhance attention (MEA) and dual correlation attention (DCA). We also demonstrate the impact of the parameters on network performance. Visualization of the tracking results is also presented to illustrate the superior performance of the proposed approach.

4.1. Experimental Settings

The MOT16 [26] and MOT17 [26] datasets are used in our experiments. We first adopted the pre-trained Deep Layer Aggregation 34 (DLA-34) [23] model on the Common objects in Context (COCO) dataset [27] to initialize our backbone. The MOT17 training set is used to train this proposed end-to-end network. A total of 30 epochs are trained, with the Adam optimizer. The batch size is set to 12. The initial learning rate is 1×10^{-4}, and we decay the learning rate to 1×10^{-5} at the 20th epoch. The hardware configuration is a 4-GPU machine, with RTX 2080Ti GPUs.

Hyperparameter r, used in the MEA model as depicted in Figure 2, controls the degree of the aggregation of features. We use two sizes of convolution kernels $64 * 1 * 1$ and $\frac{64}{r} * 1 * 1$ in our experiments. We have determined the value of r through experiments. The hyperparameter in the DCA module controls the kernel size of the convolution operator; we chose the convolution with a kernel size of $64 * 3 * 3$ for this experiment.

For the evaluation metrics, we follow the MOT challenge [28] to evaluate the MOT performance, including multiple object tracking accuracy (MOTA), multiple object tracking precision (MOTP), Identity F1 score (IDF1), the total number of ID switches (IDs), the MT (number of mostly tracked), and the ML (number of mostly lost). We evaluate our network on the MOT16 and MOT17 test sets and do not use additional training data.

4.2. Comparison Experiments

We compare our MAC-MOT with the SOTA approaches on the MOT16 and MOT17. Tables 1 and 2 show the results on the MOT16 and MOT17 benchmarks, respectively. The best performance is highlighted in a red color, while second place is highlighted in a blue color. Table 1 shows the comparisons on the MOT16 benchmark. MAC-MOT improves at least 2.4 points on the MOTA. Table 2 shows the performance of the MAC-MOT on the MOT17; our approach has a minimum of a 0.3 points improvement on the MOTA and 0.8 gains on the IDF1. Regarding the MT and ML in the two tables, we still report these performances to provide the entire picture of the proposed method, as well as other SOTAs.

There are two reasons that the proposed model has a better performance. On the one hand, the MEA enhances the moving feature of the object, which makes it more distinguishable. On the other hand, the DCA module reduces the conflict between the detection and re-identification during training. At the same time, we also notice that our approach leads to an increase in IDs in both the MOT16 and MOT17. The reason for this degraded performance is that adding two modules causes the stability of the one-shot framework to decrease.

Table 1. Tracking result on MOT16. Red color represents best performance. Blue color represents second place.

Method	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓
DeepSort-2 [29] ICIP 2017	61.4	62.2	32.8	18.2	781
RAW16wVGG [30] CACV 2018	63.0	63.8	39.9	22.1	482
TubeTK [31], CVPR 2020	64.0	59.4	33.5	19.4	1117
JDE [20] ECCV 2020	64.4	55.8	35.4	20.0	1544
HOGM [32] ICPR 2018	64.8	73.5	40.6	22.0	1544
CNNMTT [33] CMTA 2019	65.2	62.2	32.4	21.3	946
POI [34] ECCV 2016	66.1	65.1	34.0	21.3	805
CTrackerV [35] ECCV 2020	67.6	57.2	32.9	23.1	1897
FairMOT [24]	69.3	72.3	40.3	16.7	815
MAC-MOT(Ours)	71.7	70.7	39.3	18.3	1393

Table 2. Tracking result on MOT17. Red color represents best performance. Blue color represents second place.

Method	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓
TubeTK [31] CVPR 2020	63.0	58.6	31.2	19.9	4137
CTracker [35] ECCV 2020	66.6	57.4	32.2	24.2	5529
CenterTrack [23] ECCV 2020	67.8	64.7	34.6	24.6	3039
DeepSort [29] ICIP 2017	60.3	61.2	31.5	20.3	2442
FairMOT [24]	69.8	69	39.4	21.8	3960
MAC-MOT(Ours)	70.1	69.8	38.2	20.0	4392

4.3. Ablation Studies

In order to verify the role of each module, we studied each component in our tracking framework. All experiments are evaluated on the MOT17 testing set. The results of the experiments are shown in Table 3. We regard the network without the MEA and DCA as the baseline. The impact of the MEA and DCA is evaluated and discussed both in their individual and combined form. As can be seen from the Table 3, the baseline result has a 69.8 MOTA and 69 IDF1. By adding only the MEA module, it achieves a 0.1 point improvement on the MOTA and 0.6 gains on the IDF1. This demonstrates an improvement contributed by adding the attention on using the motion information. Adding the DCA module achieves a 0.6 point gain on the MOTA and a 1.1 gain on the IDF1. This result showcases the performance gain using the self-attention and correlation attention. When combining the MEA and DCA simultaneously, the result achieves a 0.3 point improvement on the MOTA and a 0.8 gain on IDF1. This combination approach does not combine the individual gain. This could be due to a combined attention effect which could cancel the other's attention mechanism.

Table 3. Experiment results of the ablation studies.

Method	MOTA↑	MOTP↑	IDF1↑	MT↑	ML↓	IDs↓
baseline	69.8	80.3	69	39.4	21.8	3960
MEA	69.9	80.3	69.6	40.3	19.5	4665
DCA	70.4	80.1	70.1	40.3	20.4	4416
MEA + DCA	70.1	80.3	69.8	38.2	20.0	4392

4.4. Parameter Analysis

There is a feature dimension conversion coefficient r in the MEA module, which reflects the degree of the feature aggregation. In order to find the suitable parameters, we conducted experiments on the MOT17 verification set. As shown in Figure 4, we chose four values of r: 4, 8, 16, and 32. We find that the MOTA is the highest when r is 8, and the IDF1 continues to decrease as r increases. Here, we conclude that with the current experiment settings, the hyperparameter r equaling to 8 provides the best performance.

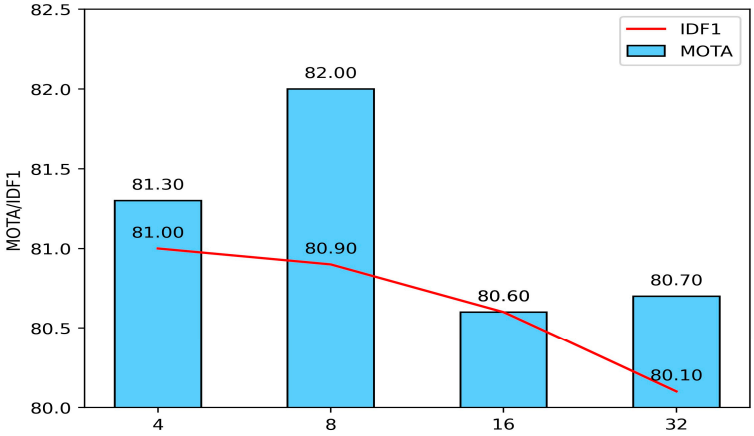

Figure 4. Tracking performance with different conversion rate r in the MOT17 verification set.

4.5. Track Visualization Results

The MAC-MOT has a better tracking ability than the SOTA one-shot method [24] in a real scene sequence. As shown in Figure 5, when the object is far away from the camera and the object is blurry, the simple one-shot method will fail to track the object; the object is missed. Our approach has improved this situation due to the enhanced motion-related feature.

Figure 5. (**a**) demonstrates the tracking performance of FairMOT [24]; (**b**) represents the tracking performance of MAC-MOT.

5. Conclusions

In this paper, we propose a novel network named MAC-MOT, which has a moduled attention in two parts: a motion enhance attention module (MEA) and a dual correlation attention module (DCA). The proposed MAC-MOT improves on previous one-shot approaches. The MEA plays a role in capturing the motion information which is ignored by the previous approaches. When an object is far away from the camera or the object is blurred due to motion, our approach can still keep track of the target objects while the previous counterpart fails. Meanwhile, the previous approach has a separate detection and re-identification task that produces ambiguity when the network parameters are optimized in a unified framework. The MAC-MOT uses the dual correlation module to obtain features that are more suitable for different tasks simultaneously. This strategy reduces the competition conflict caused by multi-task joint learning during training. In MOT challenges, our approach has achieved the competitive performance on the MOT16 and MOT17 datasets. Additionally, we also find that the simple combination of the two modules does not achieve the superposition of benefits. In the future, we will further investigate and improve the structure of the network to achieve a better tracking performance.

Author Contributions: Conceptualization, Y.W. and D.Z.; methodology, Y.W.; software, Y.W.; validation Y.W., Z.Z., N.Z. and D.Z.; formal analysis, Y.W. and N.Z.; investigation, Y.W.; resources, D.Z.; data curation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, N.Z. and D.Z.; visualization, Y.W.; supervision Z.Z. and D.Z.; project administration D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Chen, C.; Seff, A.; Kornhauser, A.; Xiao, J. Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
2. Girdhar, R.; Ramanan, D. Attentional pooling for action recognition. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 34–45.
3. Ross, P.; English, A.; Ball, D.; Upcroft, B.; Corke, P. Online novelty-based visual obstacle detection for field robotics. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015.
4. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. Mots: Multi-object tracking and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7942–7951.
5. Zhang, Z.; Cheng, D.; Zhu, X.; Lin, S.; Dai, J. Integrated object detection and tracking with tracklet conditioned detection. *arXiv* **2018**, arXiv:1811.11167.
6. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Detect to track and track to detect. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
7. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. *arXiv* **2019**, arXiv:1903.05625.
8. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
9. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
10. Choi, W.; Savarese, S. Multiple target tracking in world coordinate with single, minimally calibrated camera. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010.
11. Le, N.; Heili, A.; Odobez, J.M. Long-term time-sensitive costs for crf based tracking by detection. In Proceedings of the 11th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
12. Leibe, B.; Schindler, K.; Cornelis, N.; Van Gool, L. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1683–1698. [CrossRef] [PubMed]
13. Wu, Z.; Thangali, A.; Sclaroff, S.; Betke, M. Coupling detection and data association for multiple object tracking. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
14. Pellegrini, S.; Ess, A.; Schinedler, K.; van Gool, L. You'll never walk alone: Modeling social behavior for multi-target tracking. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
15. Leal-Taixé, L.; Canton-Ferrer, C.; Schindler, K. Learning by tracking: Siamese cnn for robust target association. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Boston, MA, USA, 8–12 June 2016.
16. Fernando, T.; Denman, S.; Sridharan, S.; Fookes, C. Tracking by prediction: A deep generative model for multi-person localization and tracking. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018.
17. Wan, X.; Wang, J.; Kong, Z.; Zhao, Q.; Deng, S. Multi-object tracking using online metric learning with long shot-term memory. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018.
18. Sun, S.; Akhtar, N.; Song, H.; Mian, A.; Shah, M. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 104–119. [CrossRef] [PubMed]
19. Kuhn, H.W. The hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
20. Wang, Z.; Zheng, L.; Liu, Y.; Wang, S. Towards real-time multi-object tracking. *arXiv* **2019**, arXiv:1909.12605.
21. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
23. 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.
24. Zhan, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. A Simple Baseline for Multi-Object Tracking. *arXiv* **2020**, arXiv:2004.01888.
25. Liu, Z.; Luo, D.; Wang, Y.; Wang, L.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Lu, T. TEINet: Towards an Efficient Aechitecture for video Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
26. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. Mot16: A benchmark for multi-object tracking. *arXiv* **2016**, preprint. arXiv:1603.00831.
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; pp. 740–755.
28. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 1–10. [CrossRef]
29. Wojke, N.; Bewley, A.; Paulus, D. Simple online and real time tracking with a deep association metric. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
30. Fang, K.; Xiang, Y.; Li, X.; Savarese, S. Recurrent autoregressive networks for online multi-object tracking. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 466–476.

31. Pang, B.; Li, Y.; Zhang, Y.; Li, M.; Lu, C. Tubetk: Adopting tubes to track multi- object in a one-step training model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hilton Head, SC, USA, 13–15 June 2020; pp. 6308–6318.
32. Zhou, Z.; Xing, J.; Zhang, M.; Hu, W. Online multi-target tracking with tensor based high order graph matching. In Proceedings of the 2018 24th International Conference on Pat- tern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1809–1814.
33. Mahmoudi, N.; Ahadi, S.M.; Rahmati, M. Multi-target tracking using cnn based features: Cnnmtt. *Multimed. Tools Appl.* **2019**, *78*, 7077–7096. [CrossRef]
34. Yu, F.; Li, W.; Li, Q.; Liu, Y.; Shi, X.; Yan, J. Multiple object tracking with high performance detection and appearance feature. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2016; pp. 36–42.
35. Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.