drones

MDPI

Article

Lightweight Spatial-Temporal Contextual Aggregation Siamese Network for Unmanned Aerial Vehicle Tracking

Qiqi Chen [1,2], Jinghong Liu [1,*], Faxue Liu [1,2], Fang Xu [1] and Chenglong Liu [1]

[1] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: liujinghong@ciomp.ac.cn

Abstract: Benefiting from the powerful feature extraction capability of deep learning, the Siamese tracker stands out due to its advanced tracking performance. However, constrained by the complex backgrounds of aerial tracking, such as low resolution, occlusion, similar objects, small objects, scale variation, aspect ratio change, deformation and limited computational resources, efficient and accurate aerial tracking is still difficult to realize. In this work, we design a lightweight and efficient adaptive temporal contextual aggregation Siamese network for aerial tracking, which is designed with a parallel atrous module (PAM) and adaptive temporal context aggregation model (ATCAM) to mitigate the above problems. Firstly, by using a series of atrous convolutions with different dilation rates in parallel, the PAM can simultaneously extract and aggregate multi-scale features with spatial contextual information at the same feature map, which effectively improves the ability to cope with changes in target appearance caused by challenges such as aspect ratio change, occlusion, scale variation, etc. Secondly, the ATCAM adaptively introduces temporal contextual information to the target frame through the encoder-decoder structure, which helps the tracker resist interference and recognize the target when it is difficult to extract high-resolution features such as low-resolution, similar objects. Finally, experiments on the UAV20L, UAV123@10fps and DTB70 benchmarks demonstrate the impressive performance of the proposed network running at a high speed of over 75.5 fps on the NVIDIA 3060Ti.

Keywords: aerial tracking; atrous convolution; Siamese tracker; temporal context

Citation: Chen, Q.; Liu, J.; Liu, F.; Xu, F.; Liu, C. Lightweight Spatial-Temporal Contextual Aggregation Siamese Network for Unmanned Aerial Vehicle Tracking. *Drones* **2024**, *8*, 24. https://doi.org/10.3390/drones8010024

Academic Editor: Diego González-Aguilera

Received: 24 November 2023
Revised: 15 January 2024
Accepted: 16 January 2024
Published: 19 January 2024

1. Introduction

Object tracking is an important computer vision task in which a tracker receives the target information in the first frame of a video and computationally recognizes the target in the following video frames. As a branch of tracking task, aerial tracking is one of the important tasks in remote sensing detection, which is applied in many fields, such as intelligent surveillance [1,2], aerial detection [3], motion object analysis [4,5], etc. Therefore, many researches focus on aerial tracking [6–10].

There are two mainstream methods used for aerial tracking, correlation filter-based tracker and Siamese-based tracker. Since MOSSE (Minimum Output Sum Square Error) [11], the correlation filter-based tracker has been widely used in the field of object tracking due to its high-speed computation. However, the correlation filter-based tracker is limited by its poor feature expression ability, which makes it difficult to cope with complex tracking scenarios in aerial tracking such as low-resolution, small targets, occlusion, scale variation, similar object interference, and so on [12,13]. Meanwhile, due to the powerful feature expression ability of the convolutional neural network (CNN), deep learning has been widely referenced in many fields [14]. The Siamese tracker [15] benefits from the powerful feature expression capability of deep learning to realize target tracking by template matching, which improves tracking performance by utilizing deep features.

However, compared with common target tracking, aerial tracking still has many problems to be solved due to the complexity of the scene and platform limitations. Compared with general tracking, aerial tracking often suffers from challenges due to the unique perspective of aviation [16], such as low resolution, occlusion, aspect ratio change, small targets, similar object interference and scale variation. Firstly, dealing with challenges such as occlusion, aspect ratio variation and scale variation requires the tracker to have high feature extraction capabilities. However, most of the existing lightweight backbone feature extraction networks adopt CNN networks with fixed convolution kernels, which makes it difficult to meet the above requirements of aerial tracking. Besides, the heavyweight feature extraction networks are too computationally intensive to balance tracking speeds. Secondly, due to the challenges of low resolution, small object and similar interference, accurate and effective feature extraction becomes very difficult. The temporal context information of consecutive frames contains rich motion information, which we believe can help the tracker mitigate these challenges. However, most existing aerial tracking trackers only use the template frame and the current frame information but ignore the temporal context information of consecutive frames. Finally, since aerial tracking needs to be deployed on platforms with limited computational resources, how to balance computation and accuracy is also a problem for aerial tracking.

Avoiding high computation to extract effective features is important for aerial tracking. We take advantage of the fact that atrous convolution [17] can change the receptive field of convolution without adding extra parameters and computation, and design a parallel atrous module (PAM) based on the traditional backbone. The proposed PAM uses a series of 3×3 atrous convolution branches with different dilation rates in parallel, to aggregate multi-scale features with spatial context information while ensuring the tracking speed. The spatial context information refers to the interaction between neighboring and long-range pixels.

Secondly, to add the temporal context information to the network, we propose an adaptive temporal context aggregation module (ATCAM). The proposed ATCAM adopts the temporal context information of consecutive frames to adaptively adjust the response map of the current frame through the encoding and decoding structure. Furthermore, the proposed ATCAM cleverly utilizes the results of the network computation of the previous frames, which introduces the temporal context information while increasing a small amount of computation. Our main contributions are as follows.

(1) A Siamese aerial tracking network is proposed, which jointly extracts feature maps with multi-scale information and continuous temporal context information to improve aerial tracking performance, at the same time balances the model parameters and computational effort to achieve efficient and lightweight aerial tracking.
(2) The PAM is designed, which effectively aggregates the multi-scale features and increases the spatial context information while increasing the computation amount in a small amount to increase the expression ability of the features.
(3) The ATCAM is designed for storing the response maps of consecutive frames and adaptively adjusting the response map of the current frame, which can introduce the temporal context information of consecutive frames while hardly increasing the computational amount.
(4) Experimental results on several aerial tracking benchmarks, DTB70, UAV20L and UAV123@10fps demonstrate that the algorithm has superior aerial ground tracking performance compared with other cutting-edge methods and can reach 75.5 fps on NVIDIA 3060Ti.

The rest of the article is organized as follows. Section 2 briefly introduces the related work on aerial tracking. In Section 3, we introduce the proposed aerial tracking network and detail the proposed PAM and ATCAM. Section 4 presents the experiments of this work. Section 5 gives the conclusion.

2. Related Work

2.1. Siamese Trackers for Aerial Tracking

In the field of target tracking, the Siamese tracker benefits from the powerful feature expression ability of deep learning, which has aroused a lot of attention. SINT [18] first proposes a Siamese deep neural network for target tracking, choosing template matching to solve the target tracking problem and providing advanced tracking performance.

The Siamese network contains three parts: the feature extraction network, cross-correlation fusion network, and prediction head. The feature extraction network is composed of two branches with the same structure and weights, which are used to extract the target template and search area features, respectively. The cross-correlation fusion network takes the target template features as the convolution kernel and performs the cross-correlation on search area features to generate the response map. The prediction head analyzes the response map to obtain the tracking results.

SiamFC [15] proposes a fully convolutional Siamese network with end-to-end training to achieve real-time, accurate tracking performance. SiamRPN [19] adopts a region proposal network and adds regression branches to the Siamese structure to obtain more accurate bounding boxes. SiamCAR [20] chooses the anchor-free mechanism to avoid the hyper-parameter tuning brought about by the anchor-based mechanism and reduces the amount of computation. Therefore, a Siamese-based network is widely used in aerial tracking. For example, SiamAPN [21] proposes an adaptive anchor frame generation and refinement of the Siamese tracker, which reduces the computational effort and improves the robustness of aerial ground tracking in complex scenarios. SiamAPN++ [6] designs an attention enhancement module consisting of a self-attention mechanism and mutual attention and proposes an attention twin network to improve tracking performance. SiamSA [22] adds a paired-scale attention module to the Siamese structure to mitigate the serious scale variation problem in aerial ground tracking.

The above trackers design various schemes on Siamese network structure to improve tracking performance, which demonstrates the potential of Siamese network structure in aerial tracking. In this work, based on the Siamese network structure, we propose a lightweight spatial-temporal contextual aggregation Siamese network for unmanned aerial vehicle tracking.

2.2. Atrous Convolution

The receptive field of a convolution represents the area of the input feature map that the convolution can "see", the larger the receptive field, the more global features can be obtained after convolution. For traditional convolution, the way to increase the perceptive field is to increase the size of the convolution kernel, but with the increase in the size of the convolution kernel comes an increase in the number of parameters and the amount of computation. Therefore, we turned our attention to atrous convolution. Compared with the traditional convolution with a fixed receptive field, atrous convolution can increase the receptive field by adding voids between neighboring elements of the convolution kernel, which does not introduce additional parameters and computation. The distance between two neighboring valid elements of the convolution kernel (excluding voids) is denoted as the dilation rate. The schematic of the atrous convolution is shown in Figure 1.

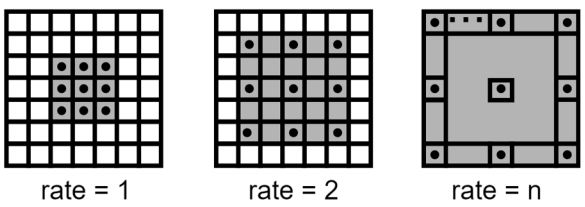


Figure 1. The schematic of the atrous convolution, 'rate' means the dilation rate.

As shown in Figure 1, the receptive field of the atrous convolution changes with the dilation rate, and the equation of the receptive field of the atrous convolution versus the expansion rate is as follows,

$$RF = [r(n-1)+1] \times [r(n-1)+1] \quad (1)$$

where r refers to the dilation rate, n is the size of the convolution kernel, RF means the receptive field. Since the insertion of voids does not increase the number of parameters and computation of the convolution, there are many studies utilizing atrous convolution for feature extraction. Bolme et al. [11] introduce the atrous convolution to build the feature pyramid for extracting multi-scale features. CFPNet [23] uses asymmetric atrous convolution to build a channel-based feature pyramid for image segmentation. Cao et al. [24] use dilation convolution to form a multi-dilation component to extract multi-scale information. Wang et al. [25] improve the residual learning model by dilation convolution, which not only propagates high-frequency information but also eliminates color differences.

Atrous convolution can increase the convolutional receptive field to a certain extent, while a fixed dilation rate still brings a fixed receptive field. Therefore, we design a parallel atrous module (PAM), which uses a series of 3×3 atrous convolutional branches with different dilation rates in parallel to aggregate multiple scales of features and increase the global information.

2.3. Temporal Context Information for Object Tracking

Object tracking is a vision task that deals with consecutive frames, and the rich temporal information in its video is crucial. During aerial tracking, trackers often face challenges such as low resolution and small object, when the temporal information between consecutive frames is particularly critical to achieve robust aerial tracking.

Currently, studies have been conducted to add temporal context to the network for object tracking. For example, AutoTrack [26] introduces temporal regularization for automatic hyperparameter optimization and uses temporal information in the tracking process to improve tracking performance. Most approaches in the Siamese network architecture choose to use transformers to integrate temporal information into the network. TrDiMP [27] designed a transformer-based encoder and decoder structure to aggregate the temporal context information of consecutive frames, and PTSEFormer [28] adopted the Transformer structure, which utilizes the attention layer to aggregate the current frame and temporal context information. TCTrack uses a transformer architecture to enhance feature extraction and improve response maps using temporal information. However, the Transformer structure is computationally very expensive and is not suitable for aviation tracking contexts with limited computational resources.

In this work, we design a lightweight temporal context aggregation module, ATCAM, which saves the response maps of consecutive frames and adaptively adjusts the response map of the current frame, which introduces the temporal context information of consecutive frames with almost no increase in computation and improves the tracking robustness in the context of aerial tracking.

3. Method

This section describes in detail the network proposed in this paper, as shown in Figure 2 the proposed network contains four main parts: feature extraction network, cross-correlation fusion network, the ATCAM, and prediction head. At the end of this section, the training and testing process of the proposed network is described.

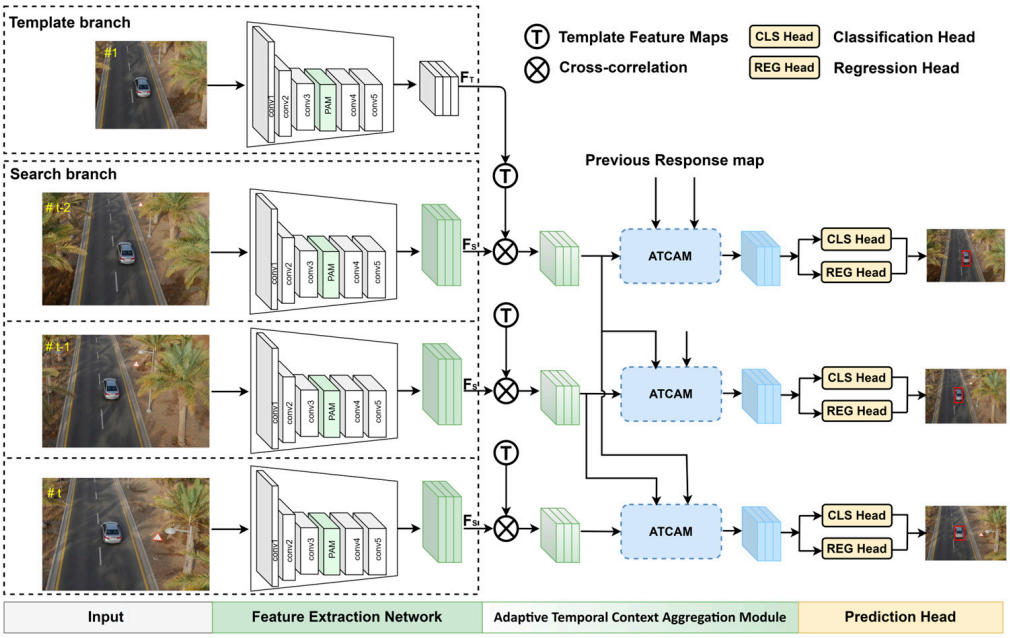


Figure 2. The schematic structure of the proposed tracking network.

3.1. Feature Extraction Network

Feature extraction, as a fundamental task in computer vision tasks, is generally regarded as a critical aspect. For aerial tracking, the extraction of discriminative and effective features can effectively improve the robustness of the tracker in complex backgrounds and is crucial for the tracker. As shown in Figure 2, the feature extraction network of the Siamese tracker contains two branches, the template branch and the search branch, the two branches are used to perform feature extraction on the target template and the search region, respectively. The inputs of the template branch and search branch are the template patch $Z \in \mathbb{R}^{127\times127\times3}$ and the search patch $S \in \mathbb{R}^{255\times255\times3}$, where 127 and 256 are the size of the input patches.

To ensure the tracking speed of aerial tracking, we do not choose the heavyweight backbone like ResNet [29], but choose the lightweight AlexNet [30] with a depth of only five layers. However, the lightweight feature extraction network implies a limited feature extraction capability, so to improve the feature extraction capability of the AlexNet network and to cope with the challenges of aerial tracking, such as occlusion, deformation, scale variation, etc., we propose a parallel atrous module (PAM). By parallelizing multiple atrous convolutions with different dilation rates, the PAM takes into account the features of different scales and increases the global sensing ability of the convolution, which is used to extract discriminative features in the complex context of aerial tracking. The structure of the proposed PAM is shown in Figure 3. For the input feature map $x \in \mathbb{R}^{w\times h\times c}$, where w and h are the size of input and x includes c channels. 1×1 convolution is firstly used to reduce its dimension and simultaneously transported to k parallel atrous convolution branches with different dilation rates for feature extraction. Each atrous branch contains two 3×3 atrous convolutions with the same dilation rate, and we design a small feature pyramid structure [31] in each branch, the output feature maps of the two convolutions are jump-spliced to obtain the output results of the branch $x_i \in \mathbb{R}^{w\times h\times(c/k)}$, $i \in \{1, 2, \ldots, k\}$. The dilation rates in different atrous branches are different. We use the same method for each branch to obtain the multi-scale calculation results. The computational results of all the branches are combined to get $x' \in \mathbb{R}^{w\times h\times c}$. The final output $y \in \mathbb{R}^{w\times h\times c}$ is obtained by summing the inputs x and x'.

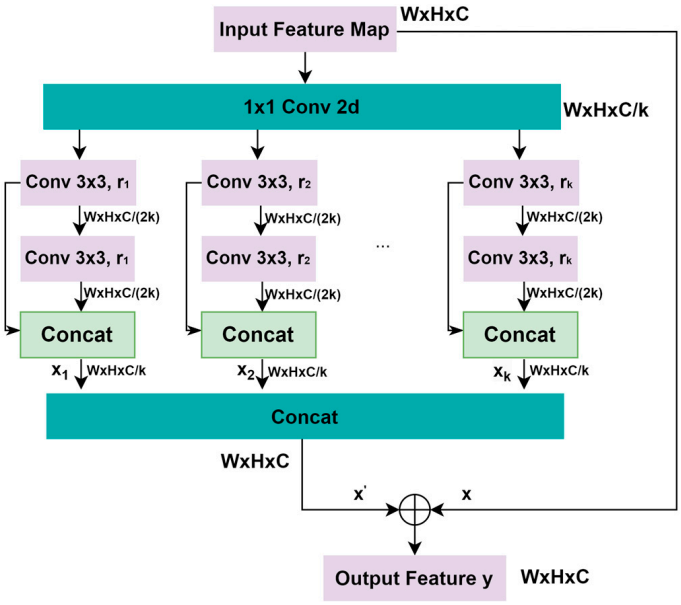


Figure 3. The schematic structure of the proposed parallel atrous module.

Each atrous branch has a different dilation rate, which means that for the same feature map, each branch focuses on a different-sized region. It is easy to understand that branches with small dilation rates are more concerned with detailed features, while branches with large dilation rates can easily obtain semantic features with long-range contextual information. Thus, with the proposed PAM, we can aggregate multi-scale features with global information, which improves the feature extraction capability of the model and helps the tracker to cope with challenges such as occlusion, deformation, scale variation, etc. In our method, the dilation rates are 1, 2, 3, and 5, and the receptive fields of each branch are 3, 5, 7, and 11, respectively.

The proposed PAM does not change the shape of the feature map, so we can regard the proposed PAM as a layer of the feature extraction network. We take the outputs of PAM and the last two convolutional layers together as the output of the feature extraction network. The output feature maps of the two branches are noted as $F_{Ti} \in \mathbb{R}^{7\times7\times256}, i \in 1,2,3$ and $F_{Si} \in \mathbb{R}^{31\times31\times256}, i \in 1,2,3$. Where 7 and 31 are the size of the template and search feature map and each feature map contains 256 channels.

3.2. Cross-Correlation Fusion Network

After the feature extraction network, we get the target template features F_T and search region features F_S. The role of the cross-correlation fusion network is to get the response maps containing the similarity information of the target template and the search region by fusing and matching F_T and F_S. Before the cross-correlation fusion network, the template branch and the search branch are independent of each other, so the cross-correlation fusion network realizes the communication of information between the two branches, which is a key step in the tracking network.

We know that convolutional neural networks compute feature maps serially, layer by layer. The shallow feature layer contains high-resolution but low-semantic features, while the deeper feature layer is more concerned with low-resolution but high-semantic information features. If we only use the last layer of the convolutional neural network feature map for fusion, the shallow features will be ignored. Therefore, to take into account both deep and shallow features and realize more comprehensive information fusion, we use multi-layer feature layers to generate response maps. The structure is shown in Figure 4.

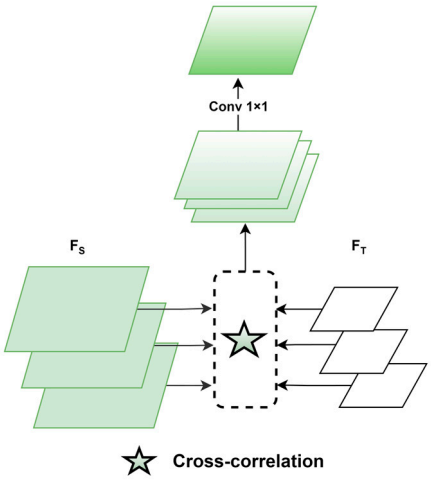


Figure 4. The schematic structure of the cross-correlation fusion network.

Specifically, the inputs of the cross-correlation fusion network are $F_{Ti} \in \mathbb{R}^{7\times7\times256}$, $i \in 1,2,3$ and $F_{Si} \in \mathbb{R}^{31\times31\times256}, i \in 1,2,3$. We use the template feature maps F_{Ti} as the kernel and perform the cross-correlation on F_{Si} to generate the response maps. The formula is shown below,

$$R_i = F_{Ti} \bigotimes F_{Si}, i = 1,2,3 \tag{2}$$

where $\otimes$ replaces the cross-correlation operation and $R_i \in \mathbb{R}^{25\times25\times256}, i \in 1,2,3$ are the response maps. Then, we fuse the three feature maps R_i with 1×1 convolution, the process can be formulated as

$$R = \text{Conv}1 \times 1(\text{Concat}(R_i)) \tag{3}$$

where $R \in \mathbb{R}^{25\times25\times256}$ is the response map.

3.3. Adaptive Temporal Context Aggregation Module

To effectively utilize the temporal context information of videos, we design an adaptive temporal context aggregation module (ATCAM). Its structure is shown in Figure 5. The ATCAM can be divided into a temporal context encoder and a temporal context decoder. Firstly, the encoder obtains the temporal priori information R_e from combining the context information of the response map of the target frame and the response maps of the neighboring frames, and the decoder realizes the temporal context aggregation of the response map of the target frame by concat-fuse the priori information. In the encoding process, we choose the response map for temporal information extraction; this is because the response map as a result of the fusion of the template branch and the search branch contains the information of the two branches of the neighboring frames instead of a single template information or search region information. In addition, choosing the response map for temporal information encoding is also efficient compared to the feature extraction session, requiring less storage and computational resources.

Specifically, the previous response maps $R_{t-i} \in \mathbb{R}^{25\times25\times256}, i \in 1,\ldots,n$ and the current response map $R_t \in \mathbb{R}^{25\times25\times256}$ are first fed into the ATCAM encoder, which aggregates the input response maps to obtain the a priori information $R_e \in \mathbb{R}^{25\times25\times256}$ through the linear-Concat-fusion structure. The process can be formulated as

$$Re = \text{Conv}(\text{Concat}(\text{Linear}(Rt,\ldots,Rt-n))) \tag{4}$$

The decoder uses Concat-fusion to perform a priori information aggregation on the current response map and outputs $R_d \in \mathbb{R}^{25\times25\times256}$. The formula is as follows,

$$Rd = \text{Conv}(\text{Concat}(Re, \text{Linear}(Rt))) \tag{5}$$

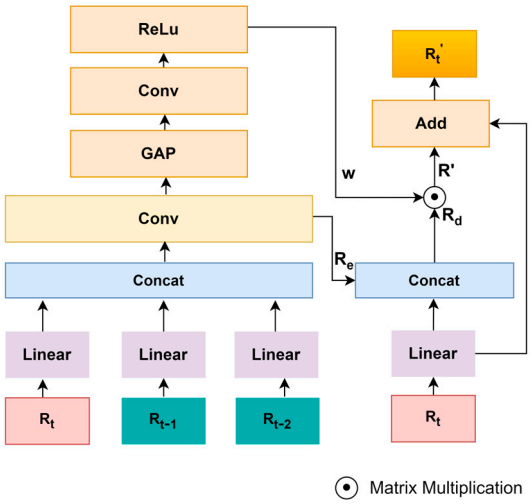


Figure 5. The schematic structure of the proposed adaptive temporal context aggregation module.

Moreover, not all neighboring frames in the tracking process are informative to the current frame tracking, and some contextual information, such as blurring and occlusion, are not only useless but even harmful to the tracking of the current frame. Therefore, to have screening for adaptive enhancement of the response map of the current frame, in this work, we design an adaptive temporal context regulator. The adaptive temporal regulator is a structure that performs a global average pooling (GAP) operation on the encoded information R_e and feeds it into a feed-forward network to obtain a regulation parameter $w \in \mathbb{R}^{1\times1\times256}$, which is used in the decoder for feature-selective aggregation.

$$w = \text{ReLu}(\text{GAP}(\text{Conv}(Re))) \tag{6}$$

By multiplying regulation parameter w and R_d, we obtain R'. The process can be formulated as

$$R' = w \cdot Rd \tag{7}$$

Finally, to prevent network degradation, residual connections are used in the decoder to ensure that the upper and lower frame information only enhances the current frame and does not interfere excessively. The output of the proposed ATCAM is $R_t' \in \mathbb{R}^{25\times25\times256}$

$$Rt' = R' + \text{Linear}(Rt) \tag{8}$$

3.4. Prediction Head

After the feature extraction network, cross-correlation fusion network, and ATCAM, we obtain the response map R_t' with temporal context information. The prediction head contains a classification branch and a regression branch, the classification branch is used to predict each location in the response map and determine whether the point is the target foreground or background, while the regression branch is used to calculate the bounding box of each location. The specific process is as follows.

Firstly, the response map R_t' is fed into the classification branch to obtain the classification feature map $A_{cls} \in \mathbb{R}^{H\times W\times2}$, where H, W represent the height and width of the classification feature map, and 2 means each point (i, j) of the feature map contains two scores, which are the foreground and background scores. Similarly, the regression branch outputs a regression map $A_{reg} \in \mathbb{R}^{H\times W\times4}$, 4 means each point (i, j) of the feature map contains a 4D vector. Each point (i, j) of the regression map corresponds to a part of the search patch, and we will mark the center of the patch as (x, y). The 4D vector represents the distance of (x, y) to the four edges of the prediction box centered. We denote the 4D vector as $t(i, j) = (l, t, r, b)$. (l, t, r, b) represent the distances from (x, y) to the left, top, right and bottom four edges of the prediction box, respectively.

4. Experiments

In this section, we first introduce the implementation details of the proposed network. Then, we perform comparison experiments with 19 state-of-the-art (SOTA) [6–8,19–22,26,32–42] trackers on three authoritative tracking benchmarks, such as the UAV20L [43], UAV123@10fps and the DTB70 [44]. Thirdly, we set up the ablation experiments to validate the effectiveness of the proposed PAM and ATCAM in this work. Besides, we do the tracking speed comparison experiment with four trackers for aerial tracking. Finally, visual comparisons and case discussions were conducted to visualize the tracking results.

4.1. Implementation Detail

In the training session, we used COCO [45], GOT10K [46], LaSOT [47], and VID [48] four datasets for model training. The training was performed using RTX3080 under Python 3.7 as well as pytorch version 1.7.1. We use Stochastic Gradient Descent (SGD) with momentum 0.9 as the optimizer, and the total number of training rounds was 20 epochs. We set the batch size to 64. Since this paper uses contextual information from consecutive frames for tracking, a multi-frame training approach is required in the training session. In the same training video, video frames are selected as target template frames and search frames, while neighboring frames are randomly selected as reference frames to train the proposed ATCAM.

The training loss function includes classification loss and regression loss. For the classification, each point (i, j) in the classification feature map $A_{cls} \in \mathbb{R}^{H \times W \times 2}$ contains the foreground and background scores δ_{pos} and δ_{neg}, and we use the cross-entropy loss [49] function to calculate the classification loss

$$L_{cls} = 0.5 \times (L_{\mathrm{BCE}}(\delta_{pos}, I) + L_{\mathrm{BCE}}(\delta_{neg}, I)) \tag{9}$$

where I represents the ground truth of the position (i, j) and L_{BCE} is the binary cross-entropy loss function. Let (x_0, y_0) and (x_1, y_1) represent the coordinates of the upper-left and lower-right corners of the ground truth box, and (x, y) is the coordinates of the search region corresponding to the regression mapping point (i, j).

$$\begin{array}{l} l' = x - x_0, t' = y - y_0 \\ r' = x_1 - x, b' = y_1 - y \end{array} \tag{10}$$

In (10), l', t', r', b' represent the distance between the coordinates (x, y) and the left, top, right, and bottom four boundaries of the ground truth box, respectively. To balance the number of positive and negative samples, we do not use all the points of the regression feature map when calculating the regression loss, but choose the points (i, j) whose coordinates (x, y) of the corresponding search area fall in the ground truth box as regression samples, which can be expressed as $l', t', r', b' > 0$. The process can be formulated as

$$\mathbb{I}(i, j) = \begin{cases} 1 & l', t', r', b' > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\mathbb{I}_{(i,j)}$ is a binary function that represents whether point (i, j) participates in the regression loss or not. Let $T_{(x,y)}$ be the ground truth box and $t(i, j)$ be the calculated predicted box, then the regression loss is calculated as follows.

$$L_{reg} = \frac{1}{\sum\limits_{(i,j)} \mathbb{I}_{(i,j)}} \sum_{(i,j)} \mathbb{I}_{(i,j)} \times L_{\mathrm{IoU}}\left[T_{(x,y)}, t_{(i,j)}\right] \tag{12}$$

$$L_{\mathrm{IoU}}\left[T_{(x,y)}, t_{(i,j)}\right] = -\ln\left(\frac{T_{(x,y)} \cap t_{(i,j)}}{T_{(x,y)} \cup t_{(i,j)}}\right) \tag{13}$$

where L_{reg} is the regression loss and L_{IoU} is the IoU (Intersection over Union) loss [50] of $T_{(x,y)}$ and $t(i,j)$. The regression loss is used to calculate the distance between the predicted bounding box and the ground truth box. We choose the IoU loss to calculate the distance.

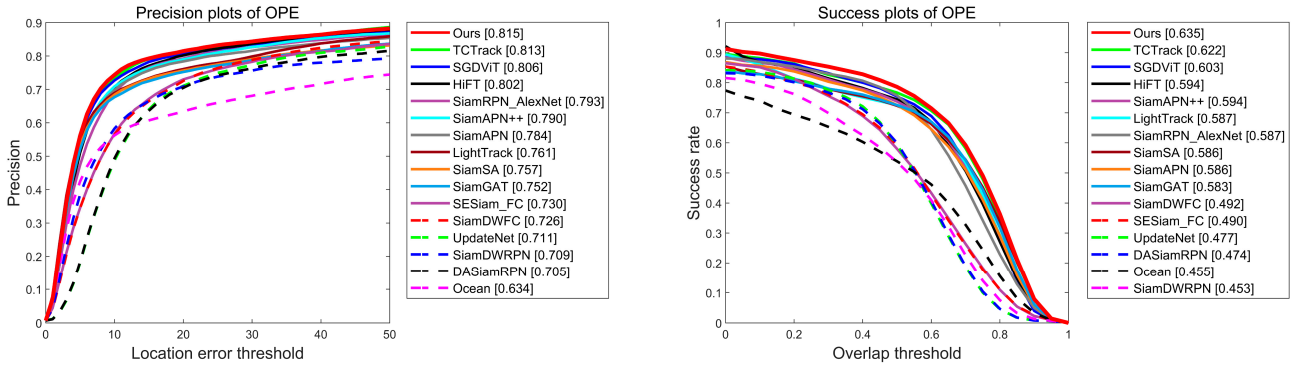
In the test session, we use the ground truth box of the first frame of a video as the target template and the rest of the frames as the search region. Both the target template and the search region are tracked in the order of Figure 2. The only thing that needs to be emphasized is the use of the proposed ATCAM. In chronological order, the first frame has no reference temporal context information available, and the ATCAM is not used in the first frame, but the response map is retained for use in subsequent search frames. In this work, we choose the first two frames of the target frame as the temporal context reference information.

4.2. Comparison with the State-of-the-Arts

Experiments on UAV20L, UAV123@10fps and DTB70 benchmarks are used to evaluate the tracking performance of the proposed method. The evaluation metrics contain a success rate and a precision rate. The success rate is calculated by the IoU score (Intersection over Union) of the prediction box and ground truth box. The success plot displays the proportion of frames whose success rate is higher than the predetermined threshold. The precision plot shows the percentage of frames whose precision rate is less than the predetermined thresholds, and the precision rate is determined by calculating the center location error between the predicted bounding box and the ground truth box. The precision rate threshold in this work is 20 pixels, while the success rate threshold is 0.5.

4.2.1. DTB70 Benchmark

DTB70 contains 70 tracking videos obtained in aerial scenarios, the videos contain many challenges under aerial tracking, so we chose DTB70 to verify the robustness of the tracker, as can be seen from Figure 6, our tracker is ranked first in terms of both precision (0.815) and success rate (0.635) compared with the SOTA trackers.



(**a**) (**b**)

Figure 6. Precision plots (**a**) and success plots (**b**) on DTB70 benchmark.

Additionally, from the attributed-based plot of the success rate shown in Figure 7, we can see that the proposed tracker performs superior in the DTB70 dataset for multiple challenges. This demonstrates that the PAM and ATCAM proposed in this paper can improve the feature extraction capability of the tracker as well as introduce adaptive contextual information to improve aerial tracking performance. In particular, the proposed tracker outperforms TCTrack by 0.2% and 1.3% in precision and success rate, respectively. Through transformer architecture, TCTrack aggregated temporal information into tracking processes.

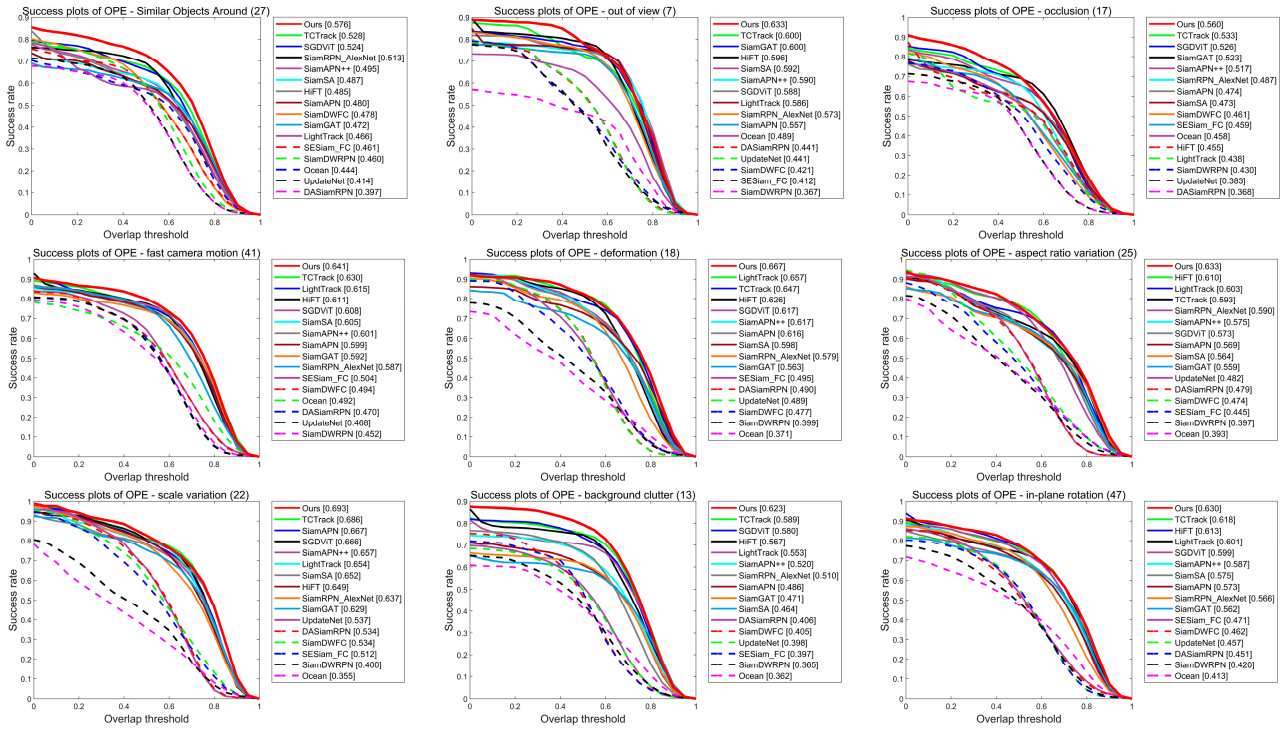


Figure 7. The success rate comparison with 15 SOTA trackers of 9 attributes on the DTB70 benchmark.

Figure 7 shows the attributed-based evaluation performance of the proposed tracker and the SOTA trackers. We can see from Figure 7 that the proposed tracker and TCTrack are doing better than the rest of the compared trackers when challenged with similar object, out of view, occlusion, and fast camera motion. It is easy to understand that the tracker has difficulty extracting accurate and effective features when faced with similar targets, out of view, occlusion and fast motion. We believe that the proposed tracker and TCTrack work well because of the addition of temporal context information, which helps the tracker track the target based on the valid information of consecutive frames. TCTrack is not as effective as the rest of the comparison algorithms when, e.g., deformation, aspect ratio variation and scale variation. We conclude that it is because the feature extraction ability of TCTrack needs to be improved. Deformation and scale variation require the tracker to extract multi-scale features. The proposed tracker adds PAM to the feature extraction network and extracts multi-scale features with global contextual information, so the proposed tracker outperforms TCTrack.

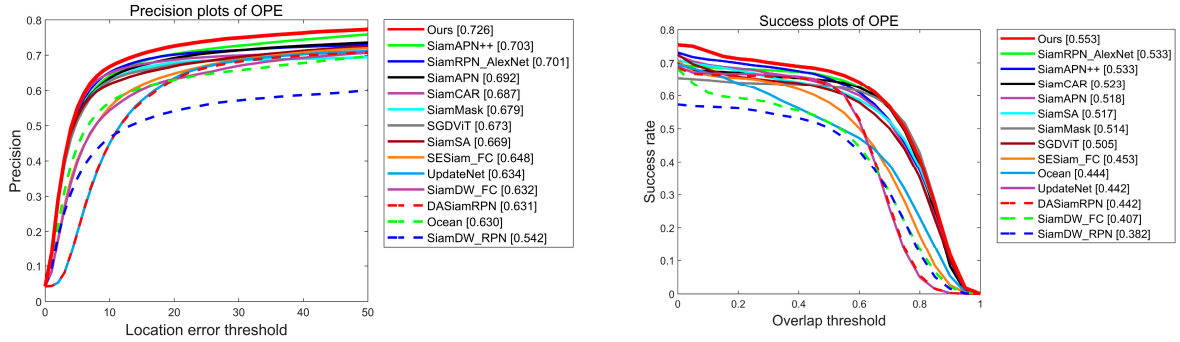
To summarize, the proposed tracker improves the feature extraction capability and aggregates temporal context information, and thus outperforms the compared SOTA tracker in DTB benchmark.

4.2.2. UAV20L Benchmark

The UAV20L benchmark contains 20 long-term object tracking videos from UAV viewpoints. Compared with short-time tracking, long-term tracking increases the probability of multiple interferences, such as aspect ratio change, camera motion, target occlusion, scale variation, etc. UAV20L benchmark is very suitable for testing the tracker performance for long-term tracking, and it can also better evaluate the robustness of the tracker. Therefore, in this paper, UAV20L is selected for tracker performance evaluation. The evaluation metrics for UAV20L are success rate and precision rate.

As shown in Figure 8, we compare several SOTA trackers for aerial tracking. We can see from Figure 8 the tracker proposed in this paper is doing best than the comparison trackers. Specifically, for both the overall precision rate and success rate, the proposed

tracker is ranked first and outperforms SiamRPN and SiamAPN++ in success rate (2%) and precision rate (2.3%).



(**a**) (**b**)

Figure 8. Precision plots (**a**) and success plots (**b**) on UAV20L benchmark.

Secondly, we performed an attribute-based evaluation, as can be seen from Figure 9, the proposed tracker outperforms all the comparison trackers in the cases of similar object, partial occlusion, low resolution, camera motion, aspect ratio change, and scale variation, which are often seen in aerial tracking. It demonstrates that the proposed tracker can cope with these challenges that arise in long-term aerial tracking and improve tracking performance.

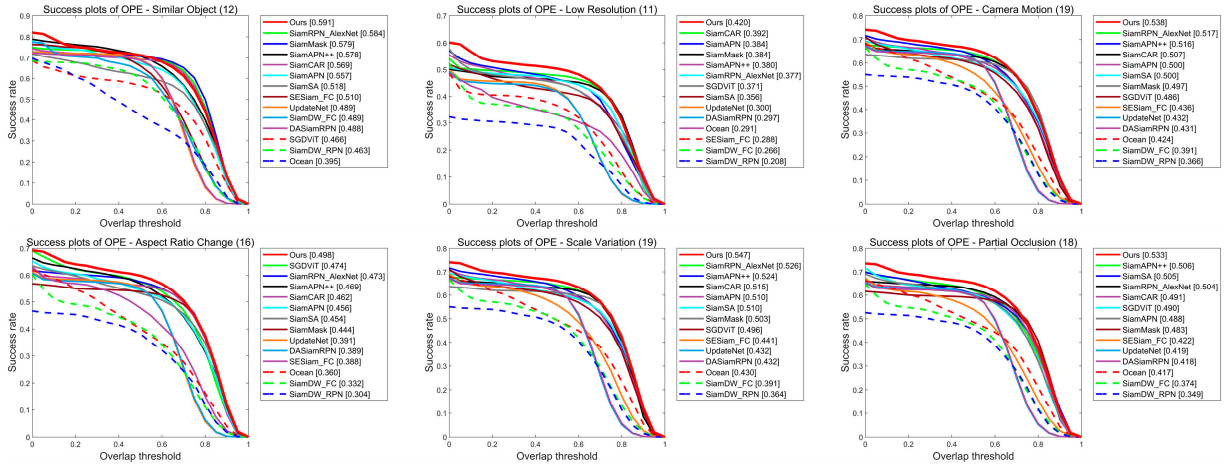


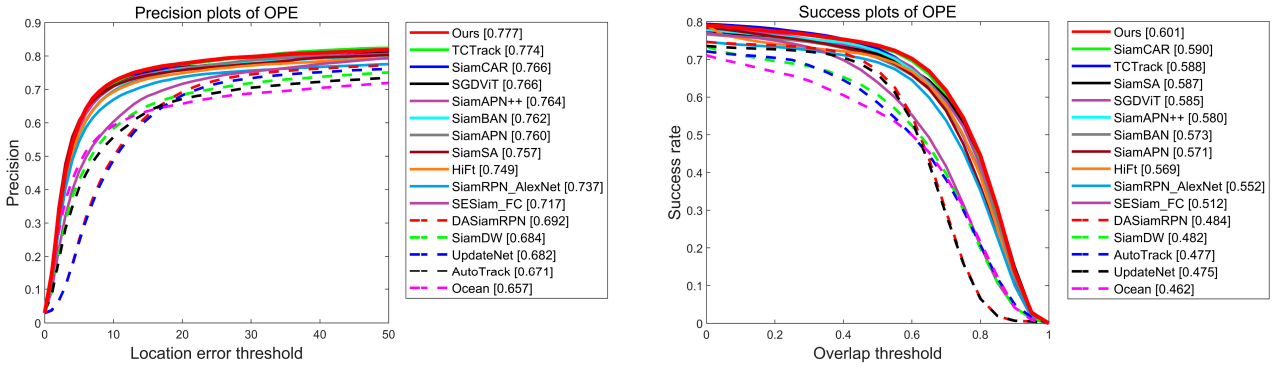
Figure 9. The success rate comparison with 14 SOTA trackers of 6 attributes on the UAV20L benchmark.

4.2.3. UAV123@10fps Benchmark

In contrast to the UAV20L benchmark, UAV123@10fps is a benchmark for short video tracking, which contains more common tracking scenarios than UAV20L, with a total of 123 tracking videos. UAV123@10fps uses an image rate of 10fps, where the target's movement is more dramatic and faster. So, UAV123@10fps is suitable to be used in measuring the tracker's performance evaluation for drastic target changes. UAV123@10fps also uses precision and success rate to evaluate the tracker.

As shown in Figure 10, in comparison with the SOTA trackers the proposed tracker still outperforms the others in terms of both precision (0.777) and success rate (0.601). In particular, the proposed tracker outperforms the comparison tracker SiamCAR by 1.1% in terms of precision and success rate. We should emphasize that the feature extraction network used in SiamCAR is ResNet50, whose computational and parametric quantities are much higher than AlexNet used in our work, which proves that the proposed tracker achieves a certain degree of balance between accuracy and speed. Moreover, Figure 11

shows the attribute-based evaluation on UAV123@10fps benchmark also demonstrates the robustness of the proposed tracker.



(**a**) (**b**)

Figure 10. Precision plots (**a**) and success plots (**b**) on UAV123@10fps benchmark.

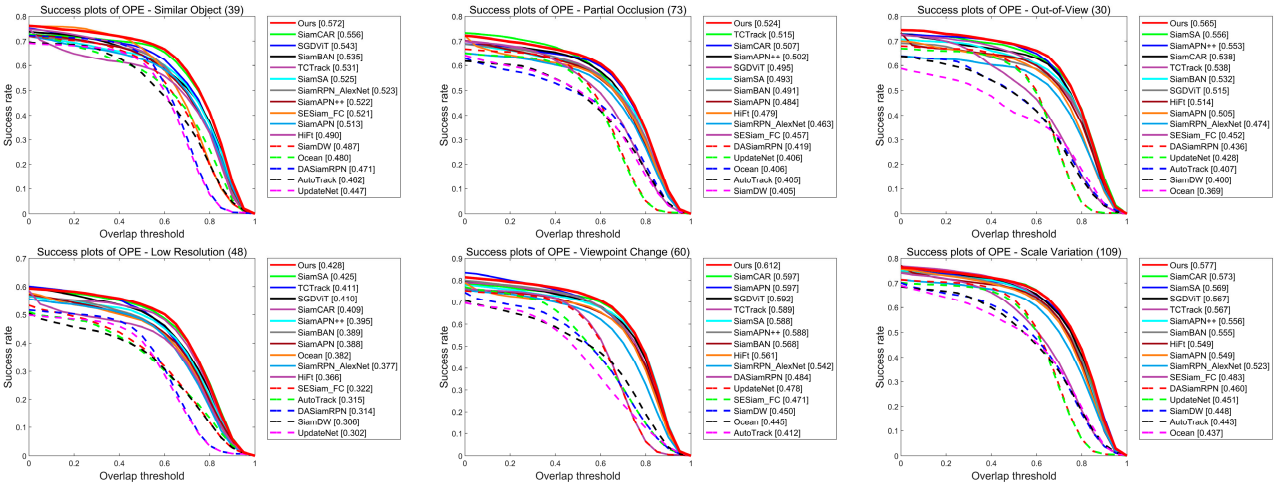


Figure 11. The success rate comparison with 15 SOTA trackers of 6 attributes on the UAV123@10fps benchmark.

4.3. Ablation Experiments

To validate the effectiveness of the proposed PAM and ACTAM in this paper for aerial tracking, ablation experiments are conducted on the tracking benchmark UAV123@10fps where the baseline tracker is a Siamese tracker consisting of an AlexNet backbone network, a cross-correlation fusion network, and a prediction head. We compare the changes in attributed-based success rates and overall tracking accuracy of the UAV123@10fps benchmark after adding the proposed PAM and ATCAM to the baseline tracker.

As shown in Table 1, common challenges in aerial tracking are: partial occlusion, scale variation, aspect ratio change, low resolution, similar object, and fast motion. Trackers with PAM or ATCAM added all have varying degrees of performance improvement. Partial occlusion, scale variation, and aspect ratio change all will change the appearance of the target. Compared with the baseline tracker, the tracker with the proposed PAM improves the success rate under the above challenges by 1.9%, 2.1%, and 2.2%. The improvement suggests that the PAM enables the feature extraction capability of the model by aggregating multi-scale features and global information, which can effectively deal with the multiple challenges that lead to the target deformation under aerial tracking and improve the tracking performance.

Table 1. Ablation study of the proposed PAM and ATCAM. ↑ means the improvement in success rate.

Model	Baseline	Baseline + PAM	Baseline + ATCAM	Baseline + PAM + ATCAM
Partial Occlusion	48.4%	50.3% 1.9%↑	50.8% 2.4%↑	52.4% 4%↑
Scale Variation	54.3%	56.4% 2.1%↑	56.3% 2.0%↑	57.7% 3.4%↑
Aspect Ratio Change	52.8%	55.0% 2.2%↑	54.2% 1.4%↑	55.4% 2.6%↑
Low Resolution	40.6%	42.0% 1.4%↑	42.7% 2.1%↑	42.8% 2.2%↑
Similar Object	50.6%	53.8% 3.2%↑	55.0% 4.4%↑	57.2% 6.6%↑
Fast Motion	45.9%	48.0% 2.1%↑	49.8% 3.9%↑	51.3% 5.4%↑
Precision	74.8%	76.9% 2.1%↑	76.7% 1.9%↑	77.7% 2.9%↑
Success	56.9%	58.7% 1.8%↑	58.1% 1.2%↑	60.1% 3.2%↑

Meanwhile, when facing the challenges of low resolution and similar object, it is difficult for the tracker to extract effective features with strong discriminative effectiveness. The proposed ATCAM can use temporal context information to help the tracker cope with the above challenges. We can see that under the above challenges, the tracking success rate also increases by 2.1% and 4.4%, respectively, with the addition of ATCAM compared to the baseline, which verifies that the proposed ATCAM can improve the tracking performance by incorporating temporal context information when facing challenges from which the tracker is unable to extract effective features. In addition, the tracking network, with the addition of ATCAM, also achieves a large improvement under the fast motion challenge. We hold the opinion that fast target motion often brings about target blurring, making feature extraction difficult. At this point, the temporal context from consecutive frames gives guidance on the target position in the search frame, so the tracking accuracy is improved after adding the proposed ATCAM.

Finally, the tracking performance was further improved with the addition of both the proposed PAM and ATCAM in the baseline. As can be seen from Table 1, the overall precision and success rate increased by 2.9% and 3.2% after the simultaneous addition of PAM and ATCAM. Compared to the addition of PAM and ATCAM alone, the precision increased by 0.8% and 1%, and the success rate increased by 1.4% and 2%, respectively. In addition, in terms of attribute-based performance, the tracking performance with the addition of both PAM and ATCAM is also better than that with the addition of PAM and ATCAM alone when facing the challenges mentioned above, which indicates that the proposed PAM and ATCAM complement each other in improving the tracking performance.

4.4. Tracking Speed Comparison

Since aerial tracking needs to run on platforms with limited computational resources such as UAVs, it is not enough to focus on tracking accuracy improvement, but tracking speed improvement is also needed. Therefore, the comparison of tracking speed is conducted in this section. We used the DTB70 benchmark to measure the tracking speed, as shown in Table 2, and we tested four trackers SiamSA, HiFT, TCTrack, and SGDViT used for aerial tracking for the comparison experiment. By testing experiments on the same platform RTX3060Ti, the results show that the algorithm proposed in this paper outperforms the comparison algorithm in both tracking accuracy and speed.

Table 2. Tracking speed comparison on DTB70 benchmark.

	DTB70		
Model	**Precision (%)**	**Success (%)**	**Fps**
SiamSA	75.7%	58.6%	64.3
HiFT	80.2%	59.4%	65.8
TCTrack	81.3%	62.2%	72.1
SGDViT	80.6%	60.3%	72.6
Ours	81.5%	63.5%	75.5

4.5. Visual Comparisons and Case Discussion

To intuitively show the tracking performance of the proposed tracker, we visualize some tracking results, as shown in Figure 12. In the Sheep2 tracking sequence, the resolution of the tracking image is very low, and the target is small, with similar target interference and viewpoint changes. During the tracking process, SiamAPN, SGDViT and SiamAPN++ successively experience tracking drift. Only the proposed tracker and TCTrack remain correct in recognizing the target in frames 101–139, which validates our idea. We believe that small targets accompanied by similar target interference and low resolution are difficult to track consistently if only the feature extraction capability is improved. The proposed tracker and TCTrack achieve better tracking results in such cases because of the addition of temporal context information. However, as the target and the camera move, TCTrack also loses the target, and only our tracker still tracks the target, which proves that the proposed tracker has higher robustness.

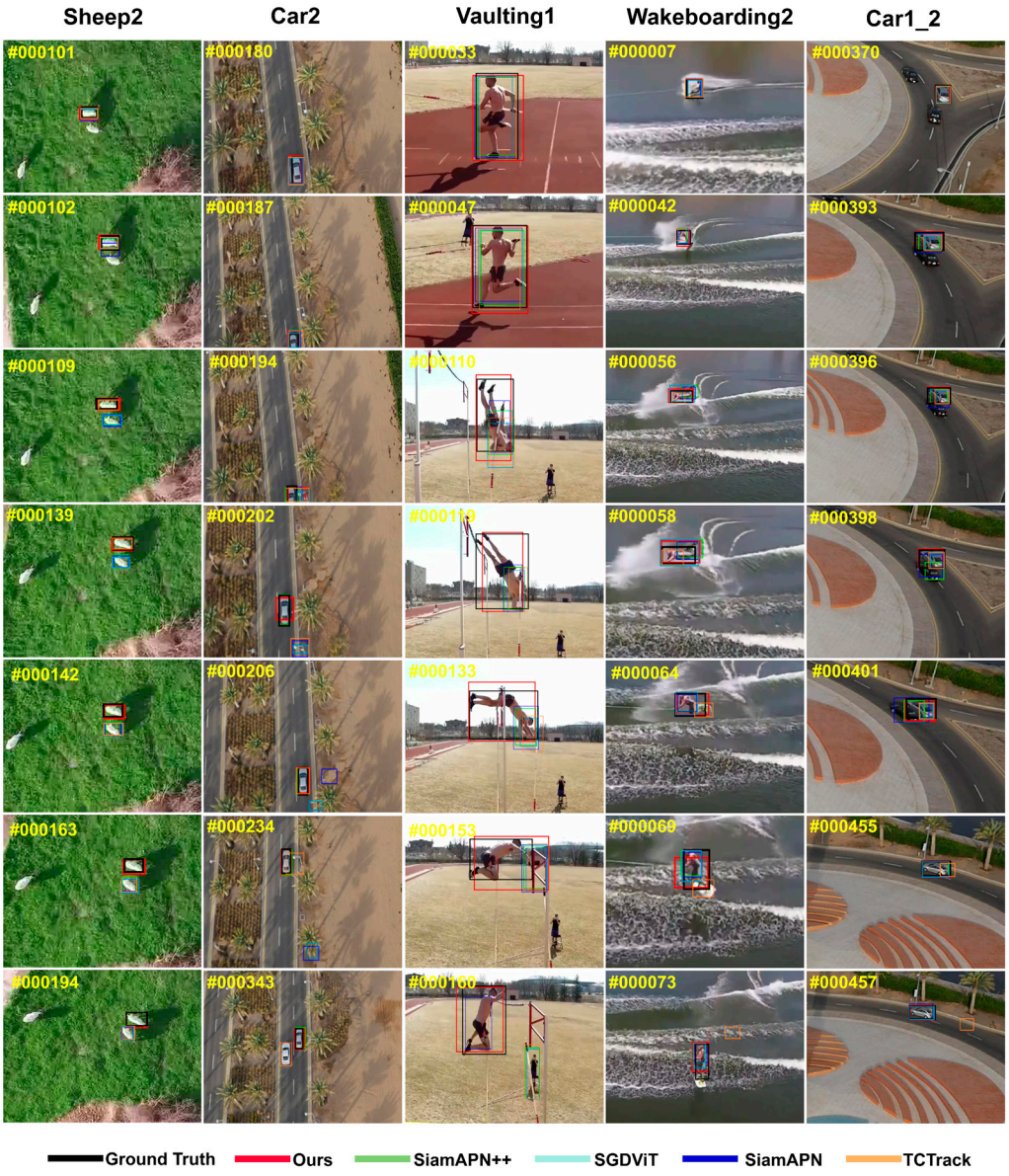


Figure 12. Visual Comparisons of our network and four SOTA trackers on five tracking videos.

In the Car2 tracking sequence, the target experiences common challenges in aerial tracking such as out of view, background interference and similar target interference. In frames 194–206, the target experiences a large change in appearance due to an occlusion,

while SGDViT, SiamAPN, and TCTrack experience tracking drift due to street light interference. In addition, only our tracker and SiamAPN++ succeed in tracking the target when similar target interference occurs in frame 343, which we attribute to the proposed attention Mechanism module in SiamAPN++ that can extract global features. Similarly, the proposed tracker consistently tracks the target, which we believe is because the proposed PAM extracts multi-scale features with global contextual information.

In the Vaulting1 tracking sequence, the target undergoes a large deformation during pole vaulting, accompanied by occlusion and similar interference, SGDViT, SiamAPN, SiamAPN++, and TCTrack are unable to recognize the target accurately from frame 110 to frame 153. Moreover, SiamAPN++ and TCTrack experience tracking drift due to the interference of a similar target in frame 160, and the proposed tracker always tracks the target accurately.

In the Wakeboarding2 tracking sequence, the target also faces the challenges of low-resolution, background interference and scale variation. From the tracking results, it can be seen that SiamAPN++, SGDViT, SiamAPN, and TCTrack all receive different degrees of interference, while our tracker performs optimally in the comparison trackers and can track the target more accurately. We believe that this is due to the multi-scale features extracted by the proposed PAM, which allows the tracker to adapt to the more drastic scale variation and get an accurate prediction box.

Our tracker performs optimally among the comparative trackers and can track the target more accurately. In the Car1_2 tracking sequence, the target undergoes occlusion and similar target interference while turning to produce deformation, and from the tracking results, the tracker proposed in this paper can effectively cope with the challenge and track the target consistently and accurately. Through the visual comparisons and case discussion, we can conclude that the tracker proposed in this paper can effectively cope with the challenges of low resolution, occlusion, similar targets, deformation, etc. in aerial tracking, which confirms that the PAM proposed in this paper can extract multiscale features, increase feature discrimination, and effectively identify the target, and at the same time, in low-resolution images accompanied by similar target interference such as Sheep2 the ATCAM plays a key role for the tracker to accurately track the target.

5. Conclusions

In this paper, a Siamese tracker for aerial tracking is designed to improve the tracking accuracy in two ways in response to the challenges that often occur during aerial tracking, such as low resolution, aspect ratio change, similar object, fast motion, occlusion, and scale variation. Firstly, in order to improve the feature extraction ability of the model in the context of aerial tracking, we design a parallel atrous module (PAM), which uses a series of atrous convolutional branches with different dilation rates in parallel to perceive different sizes of image regions in the same feature layer and aggregates multi-scale features with global information while balancing the efficiency of the model. Meanwhile, in order to cope with the difficulty of feature extraction in low resolution, occlusion, and small object under aerial tracking, we design an adaptive temporal context aggregation module (ATCAM), which adaptively introduces temporal context information to the target frame to help the tracker resist interference and recognize the target when it is difficult to extract high-resolution features. The ablation studies, comparison with SOTA trackers on DTB70, UAV20L and UAV123@10fps benchmarks, and tracking speed comparison demonstrate that the proposed network in this paper can efficiently and accurately realize aerial tracking.

Author Contributions: Conceptualization, Q.C.; methodology, Q.C.; validation, Q.C. and J.L.; investigation, Q.C., F.L. and J.L.; resources, F.X. and C.L.; writing—original draft preparation, Q.C.; writing—review and editing, J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Jilin Province: YDZJ202101 ZYTS048.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors are grateful for the anonymous reviewers' critical comments and constructive suggestions.

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

References

1. Panahi, F.H.; Panahi, F.H.; Ohtsuki, T. A Reinforcement Learning-Based Fire Warning and Suppression System Using Unmanned Aerial Vehicles. *IEEE Trans. Instrum. Meas.* **2022**, *72*, 1–16.
2. Bai, Y.; Song, Y.; Zhao, Y.; Zhou, Y.; Wu, X.; He, Y.; Zhang, Z.; Yang, X.; Hao, Q. Occlusion and Deformation Handling Visual Tracking for UAV via Attention-Based Mask Generative Network. *Remote Sens.* **2022**, *14*, 4756.
3. Li, J.; Jiang, S.; Song, L.; Peng, P.; Mu, F.; Li, H.; Jiang, P.; Xu, T. Automated optical inspection of FAST's reflector surface using drones and computer vision. *Light Adv. Manuf.* **2023**, *4*, 1–11.
4. 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.
5. Su, Y.; Liu, J.; Xu, F.; Zhang, X.; Zuo, Y. A Novel Anti-Drift Visual Object Tracking Algorithm Based on Sparse Response and Adaptive Spatial-Temporal Context-Aware. *Remote Sens.* **2021**, *13*, 4672. [CrossRef]
6. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. SiamAPN++: Siamese attentional aggregation network for real-time UAV tracking. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 3086–3092.
7. Cao, Z.; Huang, Z.; Pan, L.; Zhang, S.; Liu, Z.; Fu, C. TCTrack: Temporal contexts for aerial tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14798–14808.
8. Yao, L.; Fu, C.; Li, S.; Zheng, G.; Ye, J. SGDViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. *arXiv* **2023**, arXiv:2303.04378.
9. Javed, S.; Danelljan, M.; Khan, F.S.; Khan, M.H.; Felsberg, M.; Matas, J. Visual object tracking with discriminative filters and siamese networks: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6552–6574.
10. Fu, C.; Cao, Z.; Li, Y.; Ye, J.; Feng, C. Onboard real-time aerial tracking with efficient Siamese anchor proposal network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13.
11. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
12. Kiani Galoogahi, H.; Fagg, A.; Lucey, S. Learning background-aware correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1135–1143.
13. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef] [PubMed]
14. Zuo, C.; Qian, J.; Feng, S.; Yin, W.; Li, Y.; Fan, P.; Han, J.; Qian, K.; Chen, Q. Deep learning in optical metrology: A review. *Light Sci. Appl.* **2022**, *11*, 39. [PubMed]
15. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Proceedings, Part II 14. pp. 850–865.
16. Hua, X.; Wang, X.; Rui, T.; Shao, F.; Wang, D. Light-weight UAV object tracking network based on strategy gradient and attention mechanism. *Knowl.-Based Syst.* **2021**, *224*, 107071.
17. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
18. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
19. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
20. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6269–6277.
21. Fu, C.; Cao, Z.; Li, Y.; Ye, J.; Feng, C. Siamese anchor proposal network for high-speed aerial tracking. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 510–516.
22. Zheng, G.; Fu, C.; Ye, J.; Li, B.; Lu, G.; Pan, J. Siamese Object Tracking for Vision-Based UAM Approaching with Pairwise Scale-Channel Attention. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 10486–10492.

23. Lou, A.; Loew, M. Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1894–1898.
24. Cao, J.; Song, C.; Song, S.; Xiao, F.; Zhang, X.; Liu, Z.; Ang, M.H. Robust Object Tracking Algorithm for Autonomous Vehicles in Complex Scenes. *Remote Sens.* **2021**, *13*, 3234.
25. Wang, Q.; Fan, H.; Sun, G.; Cong, Y.; Tang, Y. Laplacian pyramid adversarial network for face completion. *Pattern Recognit.* **2019**, *88*, 493–505. [CrossRef]
26. Li, Y.; Fu, C.; Ding, F.; Huang, Z.; Lu, G. AutoTrack: Towards high-performance visual tracking for UAV with automatic spatio-temporal regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11923–11932.
27. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1571–1580.
28. Wang, H.; Tang, J.; Liu, X.; Guan, S.; Xie, R.; Song, L. Ptseformer: Progressive temporal-spatial enhanced transformer towards video object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 732–747.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 84–90. [CrossRef]
31. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
32. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
33. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.
34. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
35. Zhang, L.; Gonzalez-Garcia, A.; Weijer, J.v.d.; Danelljan, M.; Khan, F.S. Learning the model update for siamese trackers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4010–4019.
36. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
37. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-aware anchor-free tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 771–787.
38. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. Hift: Hierarchical feature transformer for aerial tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15457–15466.
39. Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; Shen, C. Graph attention tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9543–9552.
40. Sosnovik, I.; Moskalev, A.; Smeulders, A.W. Scale equivariance improves siamese tracking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 2765–2774.
41. Yan, B.; Peng, H.; Wu, K.; Wang, D.; Fu, J.; Lu, H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15180–15189.
42. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R.; Tang, Z.; Li, X. SiamBAN: Target-aware tracking with siamese box adaptive network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5158–5173. [CrossRef] [PubMed]
43. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. pp. 445–461.
44. Li, S.; Yeung, D.-Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
45. 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13, pp. 740–755.
46. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef] [PubMed]

47. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.
48. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
49. Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **2004**, *32*, 56–85. [CrossRef]
50. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.