drones

MDPI

Article

SiamMAN: Siamese Multi-Phase Aware Network for Real-Time Unmanned Aerial Vehicle Tracking

Faxue Liu [1,2], Xuan Wang [1,*], Qiqi Chen [1,2], Jinghong Liu [1] and Chenglong Liu [1]

[1] Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun 130033, China; liufaxue21@mails.ucas.ac.cn (F.L.); chenqiqi20@mails.ucas.ac.cn (Q.C.); liujinghong@ciomp.ac.cn (J.L.); liuchenglong@ciomp.ac.cn (C.L.)
[2] College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: wangxuan@ciomp.ac.cn

Abstract: In this paper, we address aerial tracking tasks by designing multi-phase aware networks to obtain rich long-range dependencies. For aerial tracking tasks, the existing methods are prone to tracking drift in scenarios with high demand for multi-layer long-range feature dependencies such as viewpoint change caused by the characteristics of the UAV shooting perspective, low resolution, etc. In contrast to the previous works that only used multi-scale feature fusion to obtain contextual information, we designed a new architecture to adapt the characteristics of different levels of features in challenging scenarios to adaptively integrate regional features and the corresponding global dependencies information. Specifically, for the proposed tracker (SiamMAN), we first propose a two-stage aware neck (TAN), where first a cascaded splitting encoder (CSE) is used to obtain the distributed long-range relevance among the sub-branches by the splitting of feature channels, and then a multi-level contextual decoder (MCD) is used to achieve further global dependency fusion. Finally, we design the response map context encoder (RCE) utilizing long-range contextual information in backpropagation to accomplish pixel-level updating for the deeper features and better balance the semantic and spatial information. Several experiments on well-known tracking benchmarks illustrate that the proposed method outperforms SOTA trackers, which results from the effective utilization of the proposed multi-phase aware network for different levels of features.

Keywords: aerial tracking; Siamese tracker; multi-phase aware network; feature fusion

Citation: Liu, F.; Wang, X.; Chen, Q.; Liu, J.; Liu, C. SiamMAN: Siamese Multi-Phase Aware Network for Real-Time Unmanned Aerial Vehicle Tracking. *Drones* **2023**, *7*, 707. https://doi.org/10.3390/drones7120707

Academic Editors: Wen Yang, Huai Yu, Jinyong Chen and Gang Wang

Received: 13 October 2023
Revised: 6 December 2023
Accepted: 8 December 2023
Published: 13 December 2023

1. Introduction

The task of aerial tracking is a challenging task, aiming at determining the target's position in subsequent frames and generating predicted boxes with the information of the initial position of the target in the first frame. Originally, it was a task that simulated human cognitive mechanisms; recently benefiting from the rapid development of cross-disciplines, it is widely used in video surveillance [1,2], UAV applications [3–5], and intelligent transportation [6,7], etc. To achieve efficient and accurate tracking, we need to distinguish between the two properties of the target foreground and background. Distinguished from general tracking, aerial tracking faces many challenges introduced by the UAV's shooting perspective which are not present in general tracking tasks, such as occlusions that require high long-range dependence on shallow features containing more spatial information, scenarios with scale and viewpoint changes that require a high global generalization of mid-level features, and small-target or low-resolution tracking scenarios which are more sensitive to pixel-level context optimization updating of deeper semantic features. Based on the aforementioned analysis of the properties of aerial tracking, one question is raised naturally: Can we design a new multi-phase aware framework to adapt the characteristics of different levels of the features to adaptively integrate regional features and the corre-

sponding long-range relevance information to improve the feature representation capability for pixel-level tracking?

In recent years, the Siamese tracker-based methods [8–11] have become highly efficient approaches to addressing aerial tracking tasks, with huge performance improvements and a balance between accuracy and real-time performance, becoming a hot research area in deep learning-based methods [12–15]. The core idea of the Siamese tracker-based method is to use two branches of the same feature extraction network for the target template and the search region, respectively and transform the tracking problem into a similarity matching problem between the features of the two branches through the process of the correlation operation. Finally, the best matching search area is obtained by the subsequent classification regression network. The development trend of the Siamese tracker reflects that how to effectively utilize different levels of features is the key to improving performance. One way is through linear multiscale context fusion. For example, some works [16,17] achieve feature fusion by direct summation or channel cascading of feature blocks extracted by the backbone network. Other works [18,19] enable the network to obtain richer dependency information by designing efficient local modeling encoders or expanding the receptive field by decomposing the feature information. While the existing approach of feature utilization enables the tracker to utilize the dependency information, linear fusion or local modeling does not take full advantage of the global view of feature information, and the pixel-level intercorrelation between features at different levels is often neglected, which is necessary for accurate tracking.

To address the above problems, designing the adaptive aware network for different levels of features is an effective and feasible approach. We propose a new Siamese multi-phase aware network called SiamMAN for aerial tracking tasks, as shown in Figure 1. It contains a multi-phase aware network adapted to the features at different depth levels to better capture the dependencies between features at different levels and improve the utilization of information from shallow spatial location features and deep semantic features. In the two-stage aware sub-network, the three feature blocks 3,4,5 extracted by the backbone network are first sent to the proposed cascade splitting encoder (CSE) to break the receptive field limitations and obtain the distributed long-range relevance among the sub-branches by the splitting of feature channels. Then, the multi-level contextual decoder (MCD) using a pooling strategy is used to achieve further global dependency fusion. Finally, in the similarity matching sub-network, we designed the response map context encoder (RCE) network utilizing long-range contextual information in backpropagation to accomplish pixel-level updating for the deeper features and better balance the semantic and spatial information. Our main contributions can be summarized as follows:

(1) We propose a novel multi-phase Siamese tracking method, SiamMAN, to enhance the network's ability to distinguish feature representations for the task of aerial tracking to improve accuracy in scenarios with high requirements at different feature levels. Specifically, the response map context encoder (RCE) module achieves optimization of deep semantic features by means of non-local perceptual modeling, and the multi-level contextual decoder (MCD) module achieves global relevance aggregation of features using an improved transformer structure. The cascaded splitting encoder (CSE) module can obtain long-range relevance information through channel splitting.
(2) A muti-phase aware framework adapted to different depth features is proposed to learn the dependency information between the channels in a global view, and we propose solutions to achieve better feature representation and utilization for different depth-level features, relying on the rich dependency information obtained from different levels to significantly improve the tracking results.
(3) We achieve the best performance compared with SOTA trackers on several well-known tracking benchmarks containing challenging scenes, including UVA123, UVA20L, DTB70, and LaSOT. Experiments show that the proposed SiamMAN can effectively improve the tracking performance in challenging scenes, such as those with low resolution and scale variation.

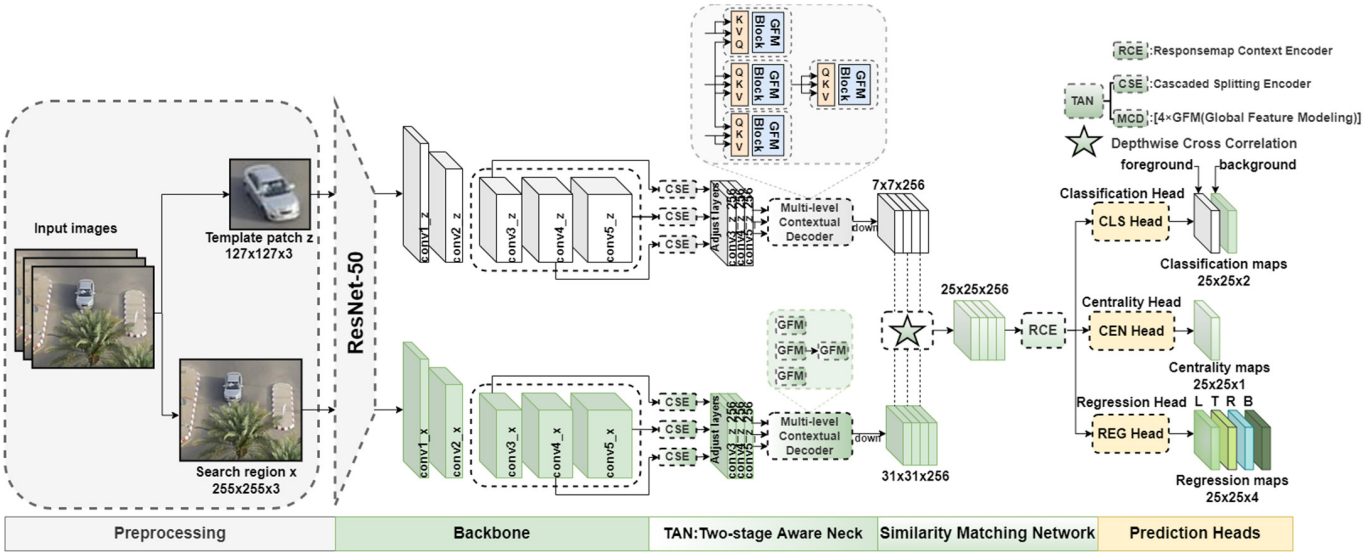

Figure 1. The overall framework of the proposed tracker.

2. Related Work

In this part, we briefly review the research related to our work in recent years, including a summary of the Siamese tracker and fusion networks.

2.1. Siamese Trackers

In recent years, Siamese network-based trackers have stood out from the crowd of trackers with excellent tracking performance; before that, correlation filtering-based approaches [20–22] received widespread attention for their efficient processing and easy deployment with low computation, driving the development of the aerial tracking field. However, the lower performance caused by artificially designed features makes it difficult to cope with challenging scenarios. In contrast, Siamese trackers emerged with many model variants for enhancing contextual information aggregation trying to achieve more efficient feature utilization and better performance. The early Siamese tracker was not specifically designed to solve the aerial tracking tasks but rather was designed to solve the challenges of the general target tracking tasks in pursuit of model generalization. The first algorithm to apply Siamese networks to address the tracking task was SINT [23], and the subsequent SiamFC [24] first introduced a correlation layer to unite feature maps, pioneering an end-to-end deep learning-based tracking approach, but the operation of the correlation operation required the network to satisfy strict translational invariance. Inspired by Faster R-CNN, Li et al. proposed SiamRPN [25] to avoid the process of multi-scale extraction of feature maps by introducing RPN networks [26] commonly used in target detection tasks, and the subsequent DaSiamRPN [27] achieved further performance improvements, but they extracted features at shallow depths. SiamRPN++ [28] applies a simple and efficient spatial perception strategy to achieve a deeper feature extraction network application, but it is sensitive to parameters such as pre-defined anchors. To address these problems, trackers such as SiamCAR [16], SiamBAN [17], and SiamFC++ [29] that redesign regression networks using an anchor-free strategy have been proposed, but the interference of unbalanced samples on features at different levels still exists. Later, the Siamese tracker designed for aerial tracking tasks began to appear in the field of target tracking such as the SOTA tracker SiamAPN [30] and SiamAPN++ [31] in the field of aerial tracking; they have enhanced the ability to cope with unbalanced samples through the study of adaptive anchors, but the strategy of adaptive anchors still cannot cope well with the need for multi-level feature utilization in challenging aerial scenarios.

2.2. Transformer and Fusion Networks

Transformer was first proposed in the literature [32], and the transformer structure has been widely used in the field of NLP in recent years, driving breakthroughs in the research of many tasks in the field of artificial intelligence [33]. The core goal of Transformer is to select the information that is more critical to the current task goal from a large amount of information, and the essential idea is to selectively filter a small amount of important information from a large amount of information and focus on the important information, ignoring most of the unimportant information. For example, NonLocal [34] proposed a non-local information statistical mechanism based on capturing dependencies between long-range features, which directly integrates global information and provides richer semantic information while obtaining global information through multiple convolutional layers. DaNet [35] proposed a pixel-level optimization module based on a self-attentive mechanism to capture global contextual dependency for image segmentation tasks, which achieved good results. Later, ViT [36] and MobileViT [37] were the first to introduce a more effective transformer architecture into computer vision tasks, breaking the limitation that CNNs can only acquire local information and ignore global information, thus enabling the modeling of dependencies between distant pixels. SiamHAS [38] proposed a tracking method with a hierarchical attention strategy that makes better use of the global relevance of features through the introduction of a multi-layer attention mechanism to achieve more accurate tracking. SE-SiamFC [39] used a scale model to break the limits of translational invariance and enhance the accuracy of the output prediction frame results of the classification regression network. SiamTPN [18] and HiFT [19] use the transformer structure directly in feature fusion networks but do not take into account the effect of adapting features at different depth levels, and the transformer structure is still limited by the receptive field of local modeling and could not achieve global contextual modeling for feature optimization at multilevel scales. SGDViT [40] applies a large-scale transformer attention structure designed specifically for aerial tracking tasks and is the current SOTA tracker in the field of aerial tracking. Unlike the above Trackers that employ various attention networks, we design the CSE module in the shallower feature level to acquire the distributed long-range dependencies of each branch through the process of channel splitting, to better cope with the requirements for long-range dependencies in scenarios such as occlusions. In addition, we designed the MCD module to further learn the global dependencies of the middle-level features to cope with the demand for global generalizability of features in common scale-view change scenarios in aerial tracking and to further solve the problem that the CSE module is unable to fully explore the global information due to the splitting of feature channels. Finally, we design the RCE module to complete the pixel-level updating of deep features by utilizing the contextual information and the characteristics of receptive field mapping in backpropagation, so that the network achieves a better balance between deep semantic information and spatial information, and better copes with scenarios such as small-target tracking and low-resolution scenes, which are particularly sensitive to semantic information. To summarize, SiamMAN proposes a multiphase awareness network strategy, where each special network designed to solve the aerial tracking challenges at different depth levels is well integrated into the framework, which has a greater advantage over the tracker using an attention network in challenging scenarios of aerial tracking. Comprehensive empirical experimental results validate the effectiveness of our proposed method.

3. Proposed Approach

In this section, we specify the general framework of the proposed network and then describe the designed two-stage aware network and response map context encoder for obtaining rich pixel-level global contextual information, respectively. Finally, we present the efforts made to adapt different levels of features for further optimization and the loss function during training.

3.1. Overall Architecture

The overall framework of the tracking algorithm proposed in this paper is shown in Figure 1.

The Siamese multi-phase aware network (SiamMAN) tracker consists of the following four main sub-networks: feature extraction backbone network, two-stage aware neck, similarity matching network, and prediction heads. The feature extraction backbone ResNet50 network takes a pair of images consisting of two branches of the target template and the search region as the inputs and uses a model that has been trained on ImageNet as its initial pre-trained model. The backbone network extracts the feature maps of the target template branch image patch Z and the search region branch image patch X, respectively, and uses the extracted feature blocks in the 3rd, 4th, and 5th as the input of the subsequent CSE block in the TAN module. In the backpropagation of the training process, the parameters are shared between the two branches of the search region and template in the Siamese network. In the model, the two-stage aware neck part achieves global contextual information aggregation of features using transformer architecture designed to adapt to different scale features. The adjustment layers use multilayer convolutional layers to dimensionally adjust the output data of the CSE block of each branch, and the number of channels of the three-layer feature blocks is uniformly adjusted from the original [512, 1024, 2048] channels to [256, 256, 256] channels to reduce the subsequent parameters and computation. In the similarity matching sub-network, the depth-separable correlation operation is used to achieve the fusion of the deep and shallow features in the output response maps by convolving the target template and the corresponding layers of 3, 4, and 5 of the search region layers. The process of deep intercorrelation operation can be described as

$$R = \varphi(X) \star \varphi(Z) \tag{1}$$

where $\star$ denotes the depth-separable correlation operation. The feature fusion module achieves the optimization of response map features based on the modeling of dependencies between long-range pixel features, which provides richer semantic information and better balances the utilization of deep feature information. Finally, the classification regression network with an anchor-free strategy is used to obtain the binary attribute classification results and the prediction box size information for each pixel point.

3.2. Two-Stage Aware Neck

Some challenging scenarios such as viewpoint change and target occlusion that may exist in different frames in aerial tracking tasks require high demand for multi-layer feature utilization and algorithmic robustness. Existing trackers such as SiamCAR and SiamBAN that utilize linear summation or cascade fusion strategies can neither fully utilize contextual information nor cope well with scale changes of small targets. Therefore, we propose a two-stage aware neck feature fusion network that contains two functional components before and after the adjustment layers: a cascaded splitting encoder and a multi-level contextual decoder. For the cascade splitting encoder, the computation process of the third feature block extracted by backbone in the target template branch Z, for example, is a branch computation process with 512 channels of input, decomposed into 4 sub-branches with 128 channels of input and two 512 channels of input sub-branches after four convolution operations, one pooling layer, and gamma function, to obtain distributed long-range information under each sub-branch by channel decomposition and cascading additive information exchange between sub-branches. The detailed calculation process is shown in Figure 2.

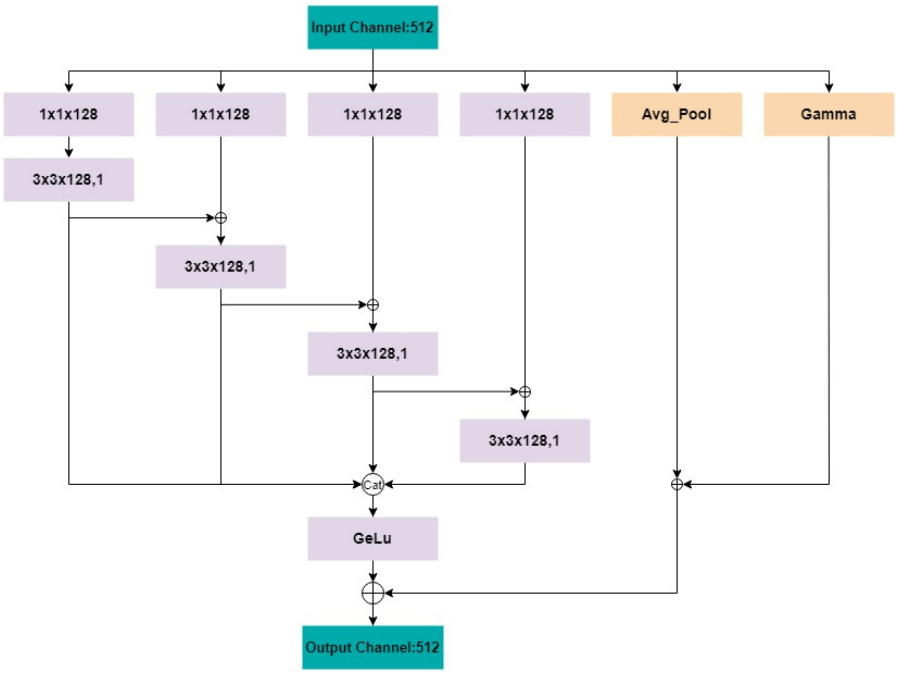

Figure 2. The structure of the proposed cascade splitting encoder.

For the first four branches, it is equivalent to dividing the input features $x(Z)$ into 4 subsets. Each subset of channels has the same size and is denoted as $x_i(Z) \in \mathbb{R}^{C\times H\times W}$, where i takes values in the set {1, 2, 3, 4}, and H, W, and C denote the shape of the input operational tensor data, the number of channels, and the height and width of each feature map, respectively. The first subset is sent into the 3×3 deep convolution, and the output is added to the next subset and used as the input of the next branch. The output of each sub-branch is represented separately as F_i, After that, we concatenate and sum it with the output of the fifth sub-branch as the final output of the module. In the fifth branch operation, the input can eliminate part of the noise interference after the average pooling layer and finally find the optimal fusion of the network through the continuous adjustment of the gamma parameter function in the training process to achieve better feature utilization. The specific formulas are as follows.

$$\begin{aligned} F_i &= conv_{3\times3}(x_i); i = 1 \\ F_i &= conv_{3\times3}(x_i + F_{i-1}); 2 \le i \le 4 \end{aligned} \tag{2}$$

$$F_i' = \text{Concat}(F_i); i \in \{1,2,3,4\} \tag{3}$$

$$F_5 = \text{AvgPool}[x(Z)] + \text{Gama}[x(Z)] \tag{4}$$

Finally, after cross-channel information optimization, we obtain the output feature map x':

$$x' = F_i' + F_5 \tag{5}$$

Compared with the traditional convolutional operation, the cascade splitting encoder can obtain the distributed long-range relevance among the sub-branches by the splitting of feature channels, break the limitation of the receptive field in the traditional CNN structure, and make full use of the multi-scale features between different levels of features to enhance the recognition ability of the network at relatively shallow feature layers that contain more spatial information.

For the Multi-level Contextual Decoder (MCD), after the CSE modules and subsequent dimensional adjustment layers, the feature blocks of each layer are flattened into sequence information using convolutional operations and used as the input of subsequent MCD

modules. Inspired by the global dependency modeling capability of Transformer, we design the global feature modeling (*GFM*) network to obtain global dependency relationships between channels over long distances using a muti-head awareness component to achieve further global dependencies fusion and address the problem that the cascade splitting encoder method does not fully explore global information due to the splitting of feature channels. The MCD blocks in the target template and search region each contain four of the proposed *GFM* modules. Specifically, the adjusted feature block L4 corresponding to the fourth layer feature block of the feature extraction backbone network is used as the query variable Q input of the three-way *GFM* modules, respectively, to realize the mutual aware mechanism and information exchange between different branches for a better global dependency modeling of long-range location and semantic information. The key–value pair inputs of each path correspond to the dimensionally adjusted output features L3, L4, and L5 at each level, respectively. The output tensor T whose key–value input is L4 is sent to another *GFM* block to achieve a better balance of deep and shallow features through a two-layer calculation to obtain the final output tensor L_4', which can be expressed as

$$\begin{aligned} T &= GFM(L_4, L_4, L_4) \\ L_4' &= GFM(T, T, T) \end{aligned} \tag{6}$$

For the proposed *GFM* module, specifically, in contrast to the traditional Transformer encoder structure, we use an averaging pooling strategy on the ternary input side of the aware computation as a preprocessing mechanism to optimize the input data for the K and V parameters. To further optimize for a more lightweight structure for aerial tracking tasks, we replace the position-encoding step in the traditional Transformer with the image itself, encoding sequence information using a zero-padding strategy to ensure the integrity of the sequence information. The *GFM* module consists of a multi-head aware module, a feed-forward network, and a normalization layer, whose core process is the processing of the input ternary data. The multi-head aware strategy enables the model to pay joint attention to the information from different representation subspaces at different locations. The calculation process of QKV ternary inputs illustrates the broadly theoretical implementation process of muti-head awareness, where Q, K, and V represent the query variables, the values of keys, and the values in the initial key–value pairs. This module calculates the similarity of Q and K, and multiplies V by the normalized distribution weights to achieve the feature enhancement of V. The final output is obtained with the same dimensionality as the original input. Softmax is used to obtain probability values about specific value parameters to norm the layers. To prevent the network from degradation, we add the input of the residual term to the output of the computation and perform hierarchical normalization after the residual connection. The overall calculation process of the *GFM* module can be summarized as

$$\begin{aligned} &GFM(Q, K, V) = \mathrm{Norm}(\mathrm{I} + \mathrm{MLP}(\mathrm{I})) \\ &\mathrm{I} = \mathrm{Norm}(Q + \mathrm{MutiHead}(Q, \mathrm{Norm\&AvgPool}(K, V))) \end{aligned} \tag{7}$$

3.3. Responsemap Context Encoder

After performing deep correlation to obtain the response maps, we design the response map contextual information encoder utilizing long-range context information in backpropagation to accomplish pixel-level updating for the deeper features and better balance the semantic and spatial information. This could make the model break through the local modeling limitation, and its structure is shown in the following Figure 3.

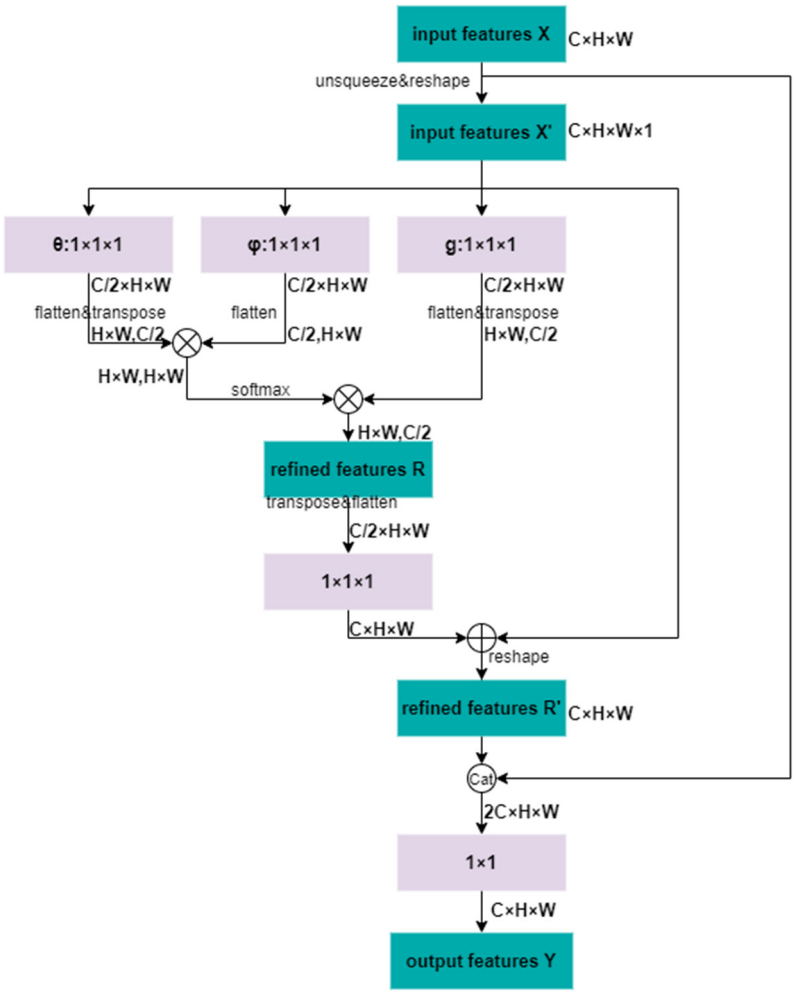

Figure 3. The structure of the proposed response map context encoder.

Specifically, the response map feature information sent into the module is first converted into four dimensions by a linear mapping process of unsqueezing and reshaping to fit the subsequent high-dimensional convolutional optimization. The three-branch input feature maps are then adjusted to half the original number of channels by three convolutional processes. The feature block of the θ-branch is multiplied with the feature block of the φ-branch after flattening and transpose operations, and the result is normalized by a softmax layer to obtain the distribution score, which is applied to the feature block of the g-branch after flattening and transpose, and multiplied with it to obtain the optimized feature R. The above process can be summarized in theoretical modeling as

$$R_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{8}$$

where x denotes the input feature maps, i represents the spatial and temporal index of the corresponding features, the f function calculates the similarity of i and j, the g function computes the representation of the feature map at position j, and the response factor $C(x)$ is used to normalize the output to obtain the final output. The temporal information obtained in the training phase through the temporal index could break the limitations of the local receptive field to obtain long-range relevance information, which is important for scenes with occlusion and a low resolution.

After that, R is transposed and flattened by a convolution layer, and then dimensionally expanded X' features are added, and the result is reshaped into the dimension of the features sent into the branches as feature R'. The above process can be summarized as

$$R' = \text{Reshape}(X' + conv(\text{transpose\&flatten}(R))) \tag{9}$$

Finally, R' is cascaded with the initial input feature X. Then, the channel dimension is adjusted by a 1×1 convolution layer to be consistent with the input X as the final output Y. The above process can be summarized as

$$Y = conv_{1\times1}(\text{Concat}\,(R', X)) \tag{10}$$

Compared with the constantly stacked convolution and RNN operator, the above operation can quickly capture the long-range dependence by directly computing the relationship between two spatial–temporal locations, and the high-dimensional global modeling of long-range dependence can effectively improve the feature expression of deep response maps, achieve the effect of pixel-level deep and shallow information balance and semantic information optimization, and have higher computational efficiency.

3.4. Training Loss

For the prediction heads, after similarity matching and feature optimization of the response map context encoder, the output tensor of dimension $25 \times 25 \times 256$ is used as the input data for each head. For the regression head, it outputs the regression maps $F_{reg} \in \mathbb{R}^{H\times W\times 4}$, where W denotes the width of the output feature map and H denotes the height, both of which are 25; each pixel position of the 4-channel feature maps records the distance from each corresponding position point to the 4 edges of the bounding box, noted as the four-dimensional vector $t(i, j) = (l, t, r, b)$, which can be calculated as follows:

$$\begin{array}{l} l = x - x_0; t = y - y_0 \\ r = x_1 - x; b = y_1 - y \end{array} \tag{11}$$

where (x, y) denotes the location coordinates (i, j) of the search area corresponding to that point, and (x_0, y_0) and (x_1, y_1) denote the coordinates of the ground truth.

For the classification head, we use the cross-entropy loss BCE to calculate the classification loss:

$$L_{cls} = 0.5 \times BCE(\delta_{pos}, I) + 0.5 \times BCE(\delta_{neg}, I) \tag{12}$$

where I is the ground truth, when calculating the BCE loss, and the fit with I denotes the foreground and background scores corresponding to the specific location of the search area branch, respectively.

We use the regression target boundary box $T_{(i,j)}$ and the prediction boundary box $t_{(i,j)}$ to calculate the regression loss, which can be calculated by the following equation:

$$L_{reg} = \frac{1}{\sum\limits_{(i,j)} I(i,j)} \sum_{(i,j)} I(i,j) \times L_{IOU}[T_{(i,j)}, t_{(i,j)}] \tag{13}$$

where $L_{IOU}\left[T_{(i,j)}, t_{(i,j)}\right]$ means that the IOU loss of $T_{(i,j)}$ and $t_{(i,j)}$, $I(i,j)$ is an indicator function defined by:

$$I(i,j) = \begin{cases} 1, (l,t,r,b > 0) \\ 0, (otherwise) \end{cases} \tag{14}$$

For the centrality head, it outputs a single channel of size 25 × 25 centrality feature map $F_{cen} \in \mathbb{R}^{H \times W \times 1}$ recording the centrality score $C(i, j)$ at the corresponding position. The centrality score is calculated as

$$L_{cen} = \frac{1}{\sum I(i,j)} \sum_{I(i,j)=1} C(i,j) \times \log R(i,j) + (1 - C(i,j)) \times \log(1 - R(i,j)) \quad (15)$$

where $C(i, j)$ denotes the predicted centrality score for a specific location and $R(i, j)$ represents the actual centrality score for this location.

The overall loss of the algorithm is as follows:

$$L = L_{cls} + \alpha_1 L_{cen} + \alpha_2 L_{reg} \quad (16)$$

where L_{cls}, L_{cen}, and L_{reg} represent the classification loss, centrality loss, and regression loss, respectively. α_1 and α_2 are used as weight hyperparameters to adjust the network and are set to 1 and 3, respectively, during the training process.

4. Experiments and Discussion

4.1. Experiment Setup

4.1.1. Implementation Details

The experimental environment for the algorithms in this paper is set up as follows: the operating system of the platform used is Windows 10, CUDA version 11.8, and the Python 3.7 + pytorch 1.13 programming framework is used to train and verify the algorithm performance. The hardware platform used is AMD Ryzen5 5600 for CPU and Nvidia GeForce RTX3080 for GPU. The parameters in the training process were set as follows: We trained the proposed network using COCO [41], GOT-10K [42], VID, and LaSOT [43] datasets. To evaluate the generality and robustness of the proposed algorithm from multiple perspectives, the model is trained by applying a stochastic gradient descent SGD optimizer with a momentum of 0.9, the batch size is set to 12, and a warm-up [44] training strategy is used to freeze the ResNet50 backbone network in the first ten rounds of training and unfreeze the backbone network in the second ten rounds for training, for a total of 20 iterations of the process. In our testing experiments, the traditional one-pass evaluation (OPE) setup was used. That is, we run the tracker from the first frame to the end frame. The tracker is initialized with the position of the first frame of the target in ground truth, and then the tracker is run to obtain the average precision and success rate, during which it is not initialized again.

4.1.2. UAV123 Benchmark

UAV123 [45] contains 123 image sequences collected by low-altitude UAVs, including image sequences with various challenging features, including scale variation, low resolution, occlusion, etc., and is one of the authoritative datasets in the field of aerial tracking. The UAV123 dataset involves the following attributes in general and aerial tracking scenes: aspect rotation changes (ARC), background clutter (BC), fast motion (FM), full occlusion (FOC), illumination variation (IV), out of view (OV), partial occlusion (POC), similar object (SOB), and scale variation (SV), especially for aerial difficulties there are camera motion (CM), low resolution (LR), and viewpoint change (VC). Success rate and Precision are used as evaluation metrics. The center position error between the prediction box and the ground truth within 20 pixels or a region overlap ratio within 50% are used as the criteria to discriminate successful tracking in terms of Precision and success rate, respectively. The ratio

of the number of frames judged to be successful to the total number of frames is defined as the Precision and success rate, respectively. The Precision is calculated as follows:

$$\begin{aligned} CLE &= \sqrt{(x_{pr} - x_{gt})^2 + (y_{pr} - y_{gt})^2} \\ f &= \begin{cases} 1, CLE < 20 \\ 0, CLE \geq 20 \end{cases} \\ Precision &= \frac{\sum_{i=1}^{N} f}{N} \end{aligned} \tag{17}$$

In this formula, (x_{pr}, y_{pr}) and (x_{gt}, y_{gt}) refer to the coordinates of the centers of the prediction box and ground truth, and CLE refers to the Euclidean distance between their centers. Accordingly, the success rate is calculated as follows:

$$\begin{aligned} S &= IOU \frac{\left|R_{pr} \cap R_{gt}\right|}{\left|R_{pr} \cup R_{gt}\right|} \\ f &= \begin{cases} 1, S \geq 0.5 \\ 0, S < 0.5 \end{cases} \\ Success &= \frac{\sum_{i=1}^{N} f}{N} \end{aligned} \tag{18}$$

4.1.3. UAV20L Benchmark

UAV20L is the definitive benchmark for evaluating and analyzing long-duration aerial tracking with 20 different long-duration video sequences of urban neighborhood scenes. These 20 long-duration sequences include complex scenes in various types of urban neighborhoods and challenging frame intervals, such as target occlusion, scale changes, and disappearance of targets.

4.1.4. DTB70 Benchmark

DTB70 [46] contains 70 UAV video sequences and is one of the most commonly used authoritative benchmarks for testing the comprehensive generalization performance of algorithms in the field of aerial tracking. The video sequences contain numerous comprehensive and challenging scenarios such as occlusion, scale variation, and low resolution. DTB70 also uses Precision and success rate as evaluation parameter metrics.

4.1.5. LaSOT Benchmark

LaSOT is a large-scale, high-quality, comprehensive benchmark for evaluating long-term tracking performance, and is a commonly used authoritative dataset in the field of target tracking, containing 280 long-term test video sequences of 70 object classes in a variety of scenarios. The LaSOT dataset still uses Precision and success rates for tracking effectiveness evaluation.

4.2. Ablation Studies

To verify the effectiveness of the proposed multi-phase aware strategy, CSE, MCD, and RCE modules in this paper, we conduct a comprehensive analysis and discussion of the effectiveness of the proposed method in aerial tracking scenes under the UAV123 benchmark and UVA20L benchmark, respectively, and conduct comprehensive and detailed ablation experiments. First, we add a two-stage aware network (TAN) including CSE and MCD to the framework for two benchmark evaluation experiments and compare the tracking results before and after adding the TAN network; then, we verify the effectiveness of adding the response map contextual encoder (RCE) proposed in this paper to the framework and compare the tracking results on the two benchmark; finally, we add the proposed TAN and RCE networks to the framework together and compare the tracking results to determine the better performance improvement that would be achieved by adding both together. As shown in Table 1, on the UAV123 benchmark, the RCE network improves the success rate and Precision of the tracker by 0.6% and 1.3% compared to no addition, reaching 62.1% and 80.5%, respectively. The MCD network improves the success rate and

Precision of the tracker by 1.4% and 2.5% compared to no addition, reaching 62.9% and 81.7%, respectively. Adding all the networks ultimately improves the success rate and Precision of the tracker by 2.4% and 3.8% compared to no addition, reaching 63.9% and 83.0%, respectively. On the UAV20L benchmark, the RCE network improved the success rate and Precision of the tracker by 0.3% and 1.1% compared to non-addition, reaching 55.5% and 71.5%, respectively. The CSE network improves the success rate and Precision of the tracker by 1.0% and 1.5%, reaching 56.2% and 71.9%, respectively, while adding all the networks eventually improves the success rate and Precision of the tracker by 2.6% and 4.7%, reaching 57.8% and 75.1%, respectively, compared to non-addition.

Table 1. Ablation study on RCE, CSE, and MCD modules. The symbol ✘ means that we add the corresponding module to the baseline model.

NO	RCE	CSE	MCD	UAV123		UAV20L	
				Pre (%)	Succ (%)	Pre (%)	Succ (%)
1				79.2	61.5	70.4	55.2
2	✘			80.5	62.1	71.5	55.5
3		✘		80.4	62.3	71.9	56.2
4			✘	81.7	62.9	73.5	56.7
5	✘	✘	✘	83.0	63.9	75.1	57.8

In summary, through the ablation experiments, we can conclude that the RCE, CSE, and MCD modules contribute to the Precision and success rate improvement of the framework to different degrees. The best performance can be obtained by using RCE, CSE, and MCD simultaneously.

4.3. Comparison with State-of-the-Art Methods

We compare our proposed model with plenty of state-of-the-art methods in the field of aerial tracking, including SiamCAR [16], SiamBAN [17], SiamTPN [18], HiFT [19], SiamRPN [25], DaSiamRPN [27], SiamRPN++ [28], SiamFC++ [29], SiamAPN [30], SiamAPN++ [31], SiamHAS [38], SE-SiamFC [39], SGD-ViT [40], SiamMask [47], Ocean [48], ATOM [49], SiamDW [50], MDNet [51], SiamAttn [52], ECO [53], Neighbor Track [54], TCTrack [55], ARTrack [56], MixFormer [57] and several recent traditional SOTA trackers TMCS [58] and CFIT [59]. For a fair comparison, all the tracking results are provided by the authors or achieved using available codes.

4.3.1. UAV123 Benchmark

(a) Overall performance:

Table 2 shows the success rate (Succ.) and Precision (Pre.) of the comparison trackers. Compared with the tracker with a similar architecture design such as SiamTPN, the proposed SiamMAN has a great improvement, the success rate of which increased from 59.3% to 63.9%, and the Precision increased from 79.0% to 83.0%. It is observed that our proposed SiamMAN ranks first in success rate and precision, outperforming all the selected state-of-the-art trackers. This is mainly because the adaptive awareness networks used in SiamMAN have advantages in measuring the edges of objects and in scenarios such as scale changes, resulting in an advantage in terms of Precision and success rate. Furthermore, we design a new architecture to adapt the characteristics of different levels of features in challenging scenarios to adaptively integrate regional features and the corresponding global dependency information; the multiphase awareness network adopted by our SiamMAN can realize long-distance contextual information aggregation, which can complete pixel-level measurements more accurately and determine the centre position of the target more precisely, and thus has an advantage in terms of success rate. The above description illustrates the effectiveness of the multiphase awareness network used in SiamMAN. Also,

SiamMAN ensures a high success rate and real-time requirements at faster speeds of 43 FPS with the hardware RTX3080. Also, we perform a fair speed comparison experiment on our RTX3080 platform based on the accessible codes of SiamBAN, SiamCAR, SiamHAS, Ocean, SiamTPN, and HiFT with the same environmental parameter settings. The experiment shows that SiamMAN achieves the same level of tracking speed and real-time performance as the mainstream Siamese trackers. Additionally, SiamMAN obtains a score of 64.6% in AUC and outperforms Ocean by 7.2 percentage points, which is a huge improvement relative to the mainstream, non-large model SOTA tracker.

Table 2. UAV123 benchmark comparison results. The bold font is the best score.

Metrics	SiamMAN	SiamBAN	SiamHAS	Ocean	SiamCAR	SiamTPN	HiFT	SiamFC ++	Neighbor Track	ARTrack	Mix-Former
Succ (%)	**63.9**	63.1	62.7	62.1	61.4	59.3	58.9	54.9	-	-	-
Pre (%)	**83.0**	82.8	82.0	82.3	80.4	79.0	78.7	76.5	-	-	-
AUC (%)	64.6	62.5	63.1	57.4	60.8	-	-	-	**72.5**	71.2	70.4
Hardware for FPS Test	RTX3080	RTX3080	RTX3080	RTX3080	RTX3080	RTX3080	RTX3080	GTX2080Ti	-	RTX 3090	GTX1080Ti
FPS	43	45	46	56	49	108	**135**	70	-	45	25

(b) Performance under different challenges:

Tables 3 and 4 show the comparison of the success rate and Precision of the trackers in ten groups of video sequences, including LR, POC, OV, VC, CM, and SOB challenges. Compared with the contrast trackers, it can be observed that, in the majority of cases, our proposed SiamMAN tracker achieves the best or second evaluation results compared to the state-of-the-art tracker, such as viewpoint change, fast movement, scale change, and low resolution, demonstrating the effectiveness of the proposed method in improving the performance in challenging scenarios. Specifically, for scenes including LR, POC, and VC attributes, such as Bike3, Car15, Person21, and Car1_s, SiamMAN obtains the best scores in success rate. For scenes including SOB, POC, and CM attributes, such as Bike3, Person21, and Car1_s, SiamMAN obtains the best scores in precision.

Table 3. The success rate achieved by the SimMAN tracker and other eight trackers on ten videos in the UAV123 benchmark. The best and the second-best results are highlighted in red and green, respectively.

Video	Attribute	SiamMAN	SiamBAN	SiamHAS	Ocean	SiamCAR	SiamTPN	HiFT	TMCS	SiamFC++
Bike3	LR POC	66.5	14.1	44.2	55.8	16.3	50.7	18.0	17.8	17.6
Boat5	VC	88.6	88.5	85.8	88.5	87.4	87.7	81.6	38.7	89.0
Building5	CM	78.8	42.3	54.7	59.6	71.4	81.6	81.9	99.8	89.1
Car15	LR POC SOB	69.9	68.0	65.7	5.0	63.9	3.5	5.1	49.1	39.9
Person21	LR POC VC SOB	49.1	41.4	26.3	37.4	33.6	22.8	26.2	28.7	19.2
Truck2	LR POC	32.7	31.5	78.2	79.3	34.2	19.9	31.8	88.5	65.7
Uav4	LR SOB	23.6	6.4	8.4	49.7	7.7	2.5	8.6	8.9	8.1
Wakeboard2	VC CM	73.2	75.3	74.7	74.5	73.5	73.7	69.8	26.1	18.8
Car1_s	POC OV VC CM	74.1	30.3	33.0	35.5	39.5	37.9	30.8	23.2	26.7
Person3_s	POC OV CM	79.4	79.3	77.7	78.3	73.2	80.4	72.1	48.3	39.9

Why could SiamMAN effectively cope with the challenging attributes? Thanks to the designed multi-phase awareness network adapted to different levels of feature characteristics, our SiamMAN has a powerful multi-level global long-range dependency modelling capability, which meets the demands of long-range contextual relationships well in scenes such as target occlusion and scale viewpoint change. In addition, the designed RCE module utilizes contextual information to accomplish pixel-level updating of deeper semantic features, which is critical for small targets and low-resolution scenes that are extremely sensitive to deep semantic information. Especially in the uniquely challenging attributes of aerial tracking, SiamMAN's performance shows great advantages. Compared to the traditional method such as TMCS, SiamMAN demonstrates absolute superiority under the

vast majority of attribute challenges. SiamMAN offers a novel and efficient approach in the field of aerial tracking. In addition, we observe that the proposed method is inferior to Ocean in terms of success rate and precision in the attributes of POC and CM, which may be because the template update strategy adopted by Ocean can better adapt to the real-time changes of the target aspect ratio features, and the incorporation of the online update feature extraction strategy will be an important further research direction for further study and improvement of our research.

Table 4. The precision achieved by the SimMAN tracker and other eight trackers on ten videos in the UAV123 benchmark. The best and the second-best results are highlighted in red and green, respectively.

Video	Attribute	SiamMAN	SiamBAN	SiamHAS	Ocean	SiamCAR	SiamTPN	HiFT	TMCS	SiamFC++
Bike3	LR POC	94.6	34.7	91.7	92.2	34.5	74.4	49.1	65.5	17.6
Boat5	VC	92.6	93.8	90.6	92.6	92.1	93.2	90.1	37.6	89.0
Building5	CM	92.4	87.0	92.3	91.2	93.4	92.7	93.8	99.8	89.1
Car15	LR POC SOB	96.7	96.3	96.7	11.2	96.5	8.0	11.1	99.7	39.9
Person21	LR POC VC SOB	85.9	73.2	50.1	66.6	61.3	39.3	63.3	73.9	19.2
Truck2	LR POC	41.4	42.2	94.4	93.7	41.4	32.3	41.1	99.7	65.7
Uav4	LR SOB	42.7	19.9	20.2	76.8	20.2	5.3	21.0	19.8	8.1
Wakeboard2	VC CM	87.7	88.7	89.0	89.5	89.2	90.4	88.3	64.6	18.8
Car1_s	POC OV VC CM	91.7	36.5	40.8	42.3	48.2	42.6	41.5	21.0	26.7
Person3_s	POC OV CM	80.7	79.7	78.2	82.0	71.6	80.3	75.9	55.8	39.9

(c) Qualitative evaluation:

To visualize the actual tracking performance of the proposed SiamMAN in various challenging scenarios compared with other advanced trackers and further discuss its performance, we visualized and compared the tracking results of seven video sequences containing various challenging scenarios in the UAV123 benchmark test, as shown in Figure 4. In the video sequences containing target occlusion scenes such as bike2_1, car7_1, person21, etc., only SiamMAN completes tracking in the face of the occlusion scenes, while all other trackers show tracking drift or failure, verifying that the TAN network can enhance tracking performance in scenes lacking target information using long-range global dependency modelling. In the video sequences containing low-resolution, fast-moving scenes such as Uav1_1 and Uav3_1, SiamMAN overcomes the effects of low-resolution and background clutter to complete tracking, while the rest of the trackers all fail to track. This further verifies the robustness of the SiamMAN in this paper in the face of complex scenes containing multiple challenging scenes and also demonstrates the effectiveness of the RCE network to extract global contextual features to accomplish pixel-level updating of deeper semantic features for tracking performance improvement in small-target tracking and low-resolution scenes that are extremely sensitive to deep semantic information.

The above seven tracking sequences demonstrate that the SiamMAN proposed in this paper has excellent robustness and tracking performance in scenarios such as scale change, occlusion, background clutter, and fast motion. It can be seen that the tracker's tracking accuracy is greatly improved when measuring small targets and in low-resolution scenarios, which is mainly due to the RCE module that can utilize the contextual information for pixel-level updating of deep semantic features. As for the occlusion and scale viewpoint change scenarios, the anchor-free strategy and the powerful long-range dependency modeling capability of the TAN module could measure the objects more accurately, which not only overcomes occlusion and other interferences but also further improves the tracking accuracy of the prediction boxes. SiamMAN provides a new and efficient tracking method for the aerial tracking field.

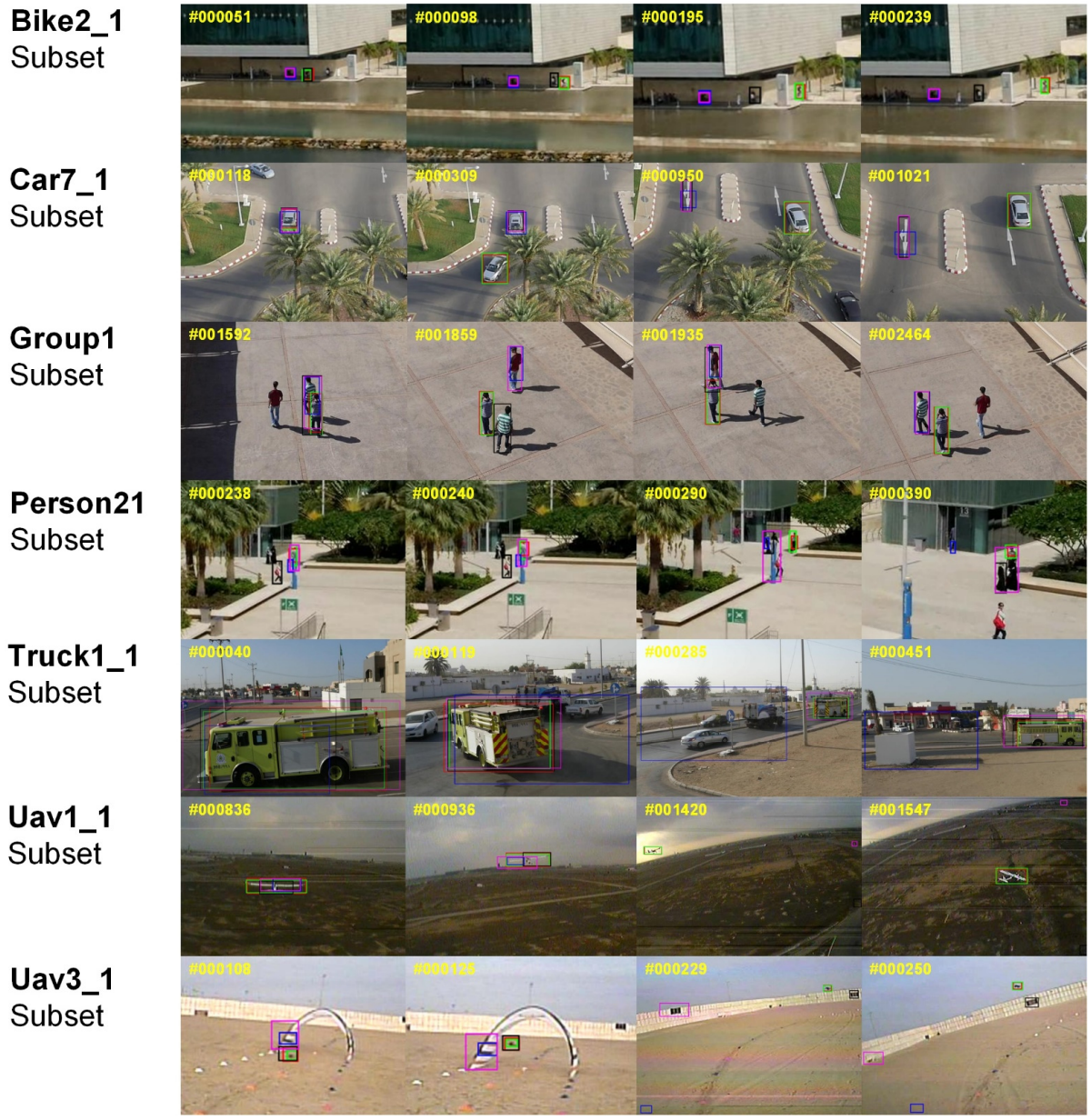

Figure 4. Visual case comparisons. Green: Ground truth; Red: SiamMAN; Black: SiamTPN; Pink: SGDViT; Blue: SiamCAR.

4.3.2. UAV20L Benchmark

To evaluate the performance of the proposed SiamMAN in a long-time aerial target tracking scenario, we compared it with eight other state-of-the-art trackers in the UAV20L benchmark, and the obtained results are shown in Table 5. The experiments show that our tracker achieves the best results (precision score: 75.1%, success rate: 57.8%), with a Precision score of 1.5% higher than SiamAPN++ and a success rate of 1.8% higher than SiamAPN++. Compared to the traditional methods that have already been used for long-duration tracking, such as CFIT, our SiamMAN has made great progress in terms of Precision and success rate, with a 22.1% increase in the success rate and a significant increase in Precision of 25.9%. Thanks to the multi-phase awareness network adapted to the characteristics of different levels, SiamMAN can aggregate long-range dependency information for different level characteristics, avoiding the circumstance that the short-time loss of target features may affect the long-time target tracking and the accuracy of tracking. It can be seen that SiamMAN can obtain long-range global dependency modeling

information through the multi-phase aware mechanism and can cope well with the scenario of long-time tracking.

Table 5. UAV20L benchmark comparison results. The bold font is the best score.

NO	Metrics	SiamMAN	SiamHAS	Siam APN++	HiFT	SiamAPN	SiamFC++	SiamRPN	TCTrack	SGDViT	CFIT
1	Succ (%)	57.8	57.3	56.0	55.3	53.9	53.3	52.8	51.1	50.5	35.7
2	Pre (%)	75.1	74.5	73.6	73.6	72.1	69.5	69.6	68.6	67.3	49.2

4.3.3. DTB70 Benchmark

We compare our proposed method with seven SOTA trackers on the DTB70 benchmark, and the success and Precision plots are shown in Table 6, both of which show that our SiamMAN achieves excellent performance compared with other advanced trackers, with a success rate of 64.9% and a Precision of 83.6%. Compared with SGDViT, the success rate is improved by 1.9%, and compared with SiamAttn, the Precision is improved by 1.0%. It can be seen that the multiphase awareness network adopted by SiamMAN is more effective compared to the attention networks adopted by SiamAttn, which can obtain the context-dependent update feature expression adapted to different levels from the global perspective, achieve better expression of positional feature information at the shallow level, and achieve better expression of feature expression of semantic information at the deeper level, and better balance between deep and shallow levels of information, at the same time as achieving better performance for the aerial tracking tasks. Also, we can see that SGDViT obtains a success of 63.0%, which is mainly because it employs a large-scale transformer structure that can better measure the foreground and background properties of the target's edge pixels. However, the high computational effort makes it difficult to apply to real-time tracking. In contrast, our SiamMAN achieves a better balance between performance and computational effort.

Table 6. DTB70 benchmark comparison results. The best and the second-best results are highlighted in red and green, respectively.

NO	Metrics	SiamMAN	SiamAttn	SGDViT	TCTrack	HiFT	SiamAPN++	SE-SiamFC	Ocean
1	Succ (%)	64.9	64.5	63.0	62.2	59.4	59.4	48.7	45.6
2	Pre (%)	83.6	82.6	80.6	81.3	82.0	79.0	73.5	69.2

4.3.4. LaSOT Benchmark

To evaluate the performance of SiamMAN for long-time tracking generalization in more types of scenarios, we compare the proposed method with nine advanced trackers, as shown in Figure 5.

SiamMAN achieves the best results in both evaluation metrics of success plot and Precision plot (precision: 53.1%, success rate 52.8%), with the Precision score improving by 3.4% over ATOM and success rate improving by 6.1% over SiamMask. Compared to the Siamese family of trackers that utilize a similar architecture, such as SiamCAR, SimMAN offers a 0.7% improvement in Precision and a 1.1% improvement in success rate. It can be seen that SiamMAN shows good performance for long-time target tracking on the LaSOT benchmark, which contains more generalized scenes. The ability to achieve such enhancements on extremely challenging and large comprehensive datasets further elucidates the contribution of the proposed multiphase awareness network to the performance enhancement of the Siamese tracker. The TAN module accomplishes the information exchange in global view at a relatively shallow feature level, which makes the spatial information feature representation more accurate and enhances the model's ability to deal with the tracking tasks in the scenes of occlusion, change in view angle, and fast

motion, etc., which have a high demand for spatial information. The RCE module utilizes contextual information at a deeper level to complete the pixel-level updating of semantic information, which helps the model to more accurately measure the position of objects at the pixel level in small targets, low resolution, and other scenarios that have a high demand for semantic information. It can be said that the proposed SiamMAN completes the optimization of feature representation in multi-layer features adaptatively and has excellent generalization for the accurate tracking of objects in multiple scenes, providing a new method with excellent generalization and high efficiency in the field of target tracking.

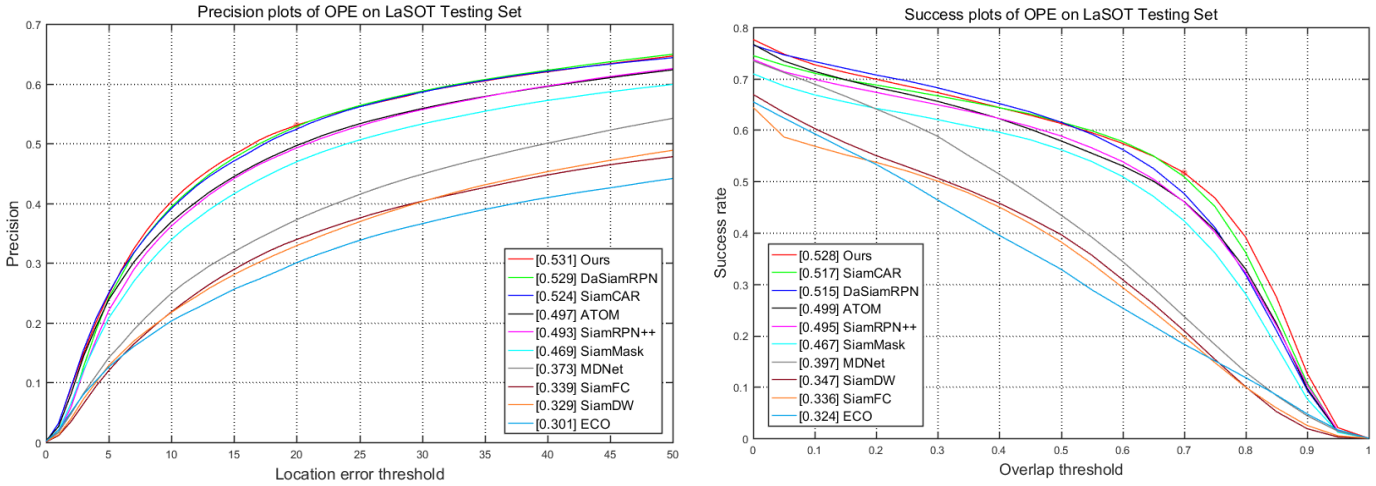

Figure 5. LaSOT benchmark comparison chart.

4.4. Heatmap Comparison Experiments

To more intuitively demonstrate the performance improvement of the proposed modules with a Siamese tracker for regions of interest in specific challenging video sequences, and further validate the performance improvement of the proposed two-stage aware network and response map context encoder in optimizing challenging scenarios, we selected an image sequence bike1 from the UAV123 benchmark for the heatmap experiments, and we added the proposed three functional modules RCE, CSE, and MCD modules to the model in turn, and the heatmaps of the three scenarios are shown in Figure 6. It can be seen that before adding the modules, the tracker's heatmap area is very fragmented, which means that the tracker's attention is easily affected by distractions rather than focusing on the target's own features. After adding the RCE module, the distraction of the tracker is significantly improved, which is mainly due to the pixel-level optimization of the RCE module for deep semantic features. After adding the CSE module, the tracker's attention area is more focused on the target itself, which is mainly due to the ability of the CSE module to obtain long-distance relevance that allows the model to focus on more levels of features of the target. After adding the MCD module, the tracker's heatmap area is more concentrated, which is mainly due to the powerful global relevance extraction ability of the MCD module, which can obtain a more accurate representation of the target's features. Meanwhile, we can see that the SiamMAN tracker containing three functional modules achieves the optimal heatmap area covering the target region, and the multiple functional modules ultimately make our SiamMAN have better accuracy and robustness.

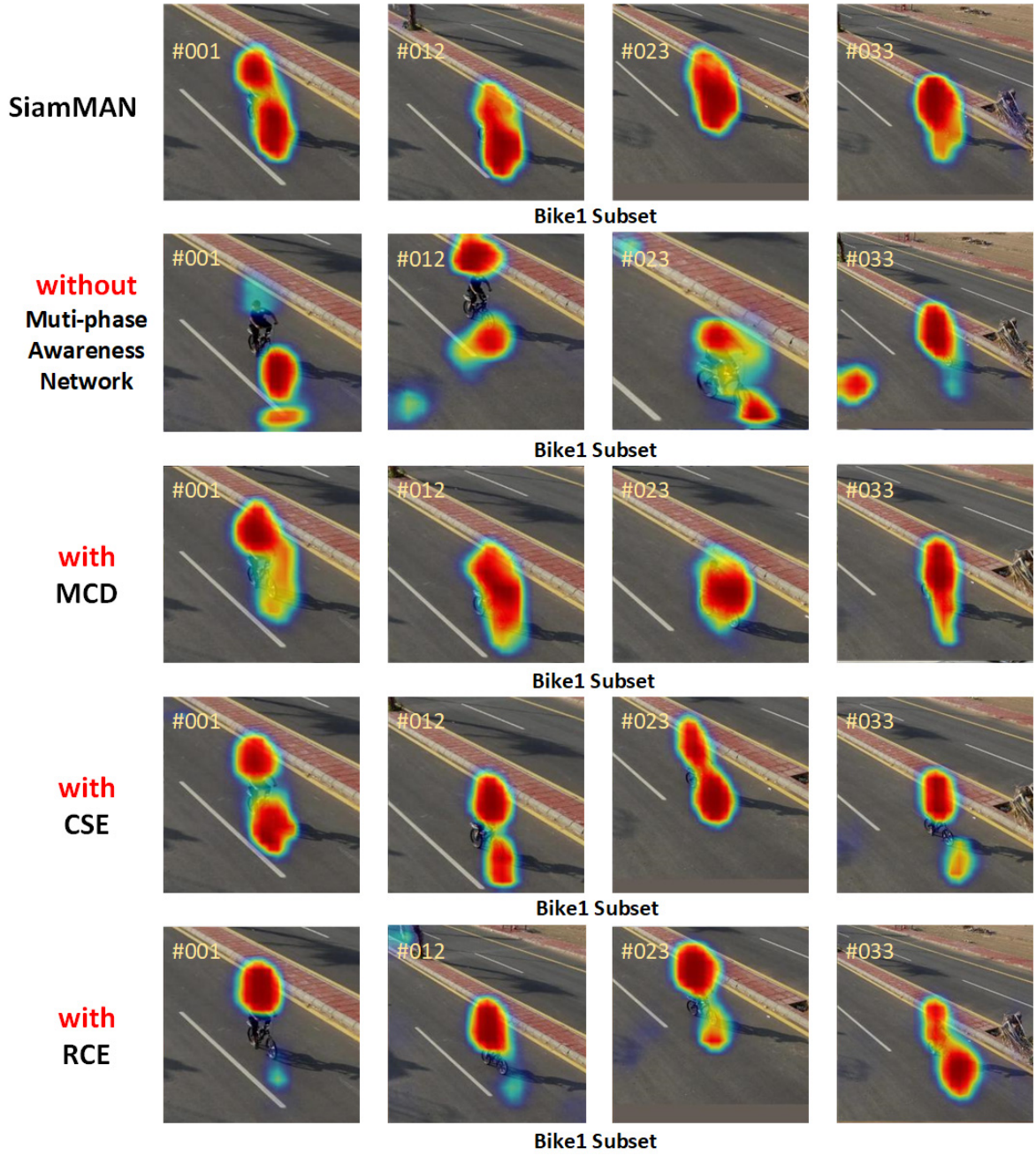

Figure 6. Heatmap comparison experiments chart.

4.5. Real-World Tests

In this section, we deploy our tracker on the UAV onboard embedded platform Jetson kits to test its practicability in real-world scenes. During the real-world tests, the utilization of the GPU and CPU is 71% and 36.8% on average. The challenging scenes in the real-world tests include scale variation, occlusion, motion blur, and low resolution. Our real-world tracking results using the UAV platform are shown in Figure 7. It can be seen that our tracker can accurately track the pedestrian by extracting global relevance when facing a complex background and small-target tracking scenarios (real-world subset1). When facing a similar object interference (real-world subset2) scenario, SiamMAN can effectively distinguish the target object. In the scenario of changing viewpoints (real-world subset3), SiamMAN can effectively perform the tracking task under different viewpoints due to its strong spatial and temporal dependency modeling capability. Finally, our tracker remains at a speed of over 20 FPS during the tests.

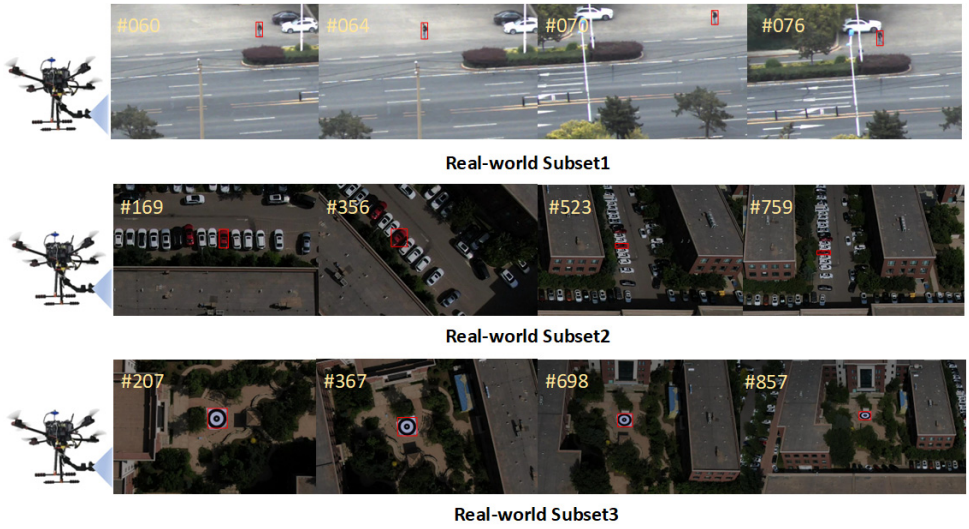

Figure 7. Results of real-world tests on the embedded platform. The tracking targets are marked with red boxes.

5. Conclusions

In this work, we propose a new multi-phase aware framework integrated into the Siamese tracker to achieve performance improvement of the algorithm in various challenging scenarios. Specifically, firstly, we propose a response map context encoder (RCE) to enable deep features to aggregate more contextual information and better balance the deep semantic information to enhance the tracker's ability to distinguish target features among deep semantic information. Secondly, we propose a two-stage aware neck which includes the multi-level contextual decoder (MCD) and cascade splitting encoder (CSE) modules to aggregate more long-range spatial–temporal information across channels to achieve global modeling and enhance the tracker's ability to cope with complex scenarios such as target occlusion and scale change. Finally, the new multi-phase aware feature-optimized functional structure is efficiently integrated into the tracker framework. Comprehensive and extensive experiments validate the effectiveness of our proposed neural network framework. Overall, we believe that our work can boost the development within the field of remote sensing, aerial tracking, and learning systems.

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

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61905240 and the National Natural Youth Science Foundation of China under Grant 62105326.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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 conflict of interest.

References

1. Zhang, J.; Zhang, X.; Huang, Z.; Cheng, X.; Feng, J.; Jiao, L. Bidirectional Multiple Object Tracking Based on Trajectory Criteria in Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5603714. [CrossRef]
2. Yan, H.; Xu, X.; Jin, G.; Hou, Q.; Geng, Z.; Wang, L.; Zhang, J.; Zhu, D. Moving Targets Detection for Video SAR Surveillance Using Multilevel Attention Network Based on Shallow Feature Module. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5200518. [CrossRef]
3. Dai, J.; Pu, W.; Yan, J.; Shi, Q.; Liu, H. Multi-UAV collaborative trajectory optimization for asynchronous 3-D passive multitarget tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5101116. [CrossRef]
4. Zhang, Y.; Wu, C.; Guo, W.; Zhang, T.; Li, W. CFANet: Efficient Detection of UAV Image Based on Cross-layer Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5608911. [CrossRef]
5. Li, B.; Fu, C.; Ding, F.; Ye, J.; Lin, F. All-day object tracking for unmanned aerial vehicle. *IEEE Trans. Mob. Comput.* **2022**, *22*, 4515–4529. [CrossRef]
6. Li, Y.; Ma, L.; Zhong, Z.; Cao, D.; Li, J. TGNet: Geometric graph CNN on 3-D point cloud segmentation. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3588–3600. [CrossRef]
7. Cao, J.; Song, C.; Song, S.; Xiao, F.; Zhang, X.; Liu, Z.; Ang, M.H., Jr. Robust object tracking algorithm for autonomous vehicles in complex scenes. *Remote Sens.* **2021**, *13*, 3234. [CrossRef]
8. Chen, Q.; Liu, J.; Wang, X.; Zuo, Y.; Liu, C. Global Multi-Scale Optimization and Prediction Head Attentional Siamese Network for Aerial Tracking. *Symmetry* **2023**, *15*, 1629. [CrossRef]
9. Song, W.; Jiao, L.; Liu, F.; Liu, X.; Li, L.; Yang, S.; Hou, B.; Zhang, W. A joint siamese attention-aware network for vehicle object tracking in satellite videos. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625617. [CrossRef]
10. Yang, J.; Pan, Z.; Wang, Z.; Lei, B.; Hu, Y. SiamMDM: An Adaptive Fusion Network with Dynamic Template for Real-time Satellite Video Single Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 3271645. [CrossRef]
11. Zeng, H.; Wu, Q.; Jin, Y.; Zheng, H.; Li, M.; Zhao, Y.; Hu, H.; Kong, W. Siam-GCAN: A Siamese graph convolutional attention network for EEG emotion recognition. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4010409. [CrossRef]
12. 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. [CrossRef] [PubMed]
13. 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*, 3–13. [CrossRef]
14. Huang, L.; Luo, R.; Liu, X.; Hao, X. Spectral imaging with deep learning. *Light Sci. Appl.* **2022**, *11*, 61. [CrossRef] [PubMed]
15. Zhang, Y.; Liu, T.; Singh, M.; Çetintaş, E.; Luo, Y.; Rivenson, Y.; Larin, K.V.; Ozcan, A. Neural network-based image reconstruction in swept-source optical coherence tomography using undersampled spectral data. *Light Sci. Appl.* **2021**, *10*, 155. [CrossRef]
16. 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, Seattle, WA, USA, 13–19 June 2020; pp. 6269–6277.
17. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese box adaptive network for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6668–6677.
18. Xing, D.; Evangeliou, N.; Tsoukalas, A.; Tzes, A. Siamese transformer pyramid networks for real-time UAV tracking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2139–2148.
19. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 15457–15466.
20. 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]
21. Huang, Y.; Li, X.; Lu, R.; Qi, N. RGB-T object tracking via sparse response-consistency discriminative correlation filters. *Infrared Phys. Technol.* **2023**, *128*, 104509. [CrossRef]
22. Zhang, J.; He, Y.; Wang, S. Learning Adaptive Sparse Spatially-Regularized Correlation Filters for Visual Tracking. *IEEE Signal Process. Lett.* **2023**, *30*, 11–15. [CrossRef]
23. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1420–1429.
24. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16 2016*; Proceedings, Part II 14 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
25. Fan, H.; Ling, H. Siamese cascaded region proposal networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7952–7961.
26. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
27. 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.

28. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
29. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12549–12556. [CrossRef]
30. Fu, C.; Cao, Z.; Li, Y.; Ye, J.; Feng, C. Siamese anchor proposal network for high-speed aerial tracking. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 510–516.
31. 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.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
33. Huang, C.; Wang, J.; Wang, S.H.; Zhang, Y.D. Applicable artificial intelligence for brain disease: A survey. *Neurocomputing* **2022**, *504*, 223–239. [CrossRef]
34. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
35. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
36. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
37. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
38. Liu, F.; Liu, J.; Chen, Q.; Wang, X.; Liu, C. SiamHAS: Siamese Tracker with Hierarchical Attention Strategy for Aerial Tracking. *Micromachines* **2023**, *14*, 893. [CrossRef] [PubMed]
39. 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, Waikoloa, HI, USA, 3–8 January 2021; pp. 2765–2774.
40. Yao, L.; Fu, C.; Li, S.; Zheng, G.; Ye, J. SGDViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. *arXiv* **2023**, arXiv:2303.04378.
41. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Proceedings, Part V 13; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
42. 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]
43. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.
44. 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, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
45. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part I 14; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 445–461.
46. Li, S.; Yeung, D.Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 1–7. [CrossRef]
47. Hu, W.; Wang, Q.; Zhang, L.; Bertinetto, L.; Torr, P.H. Siammask: A framework for fast online object tracking and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3072–3089.
48. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-aware anchor-free tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXI 16; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 771–787.
49. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669.
50. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
51. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4293–4302.
52. Yu, Y.; Xiong, Y.; Huang, W.; Scott, M.R. Deformable siamese attention networks for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 6728–6737.
53. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.

54. Chen, Y.H.; Wang, C.Y.; Yang, C.Y.; Chang, H.S.; Lin, Y.L.; Chuang, Y.Y.; Liao, H.Y.M. NeighborTrack: Improving Single Object Tracking by Bipartite Matching with Neighbor Tracklets. *arXiv* **2022**, arXiv:2211.06663.
55. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 14798–14808.
56. Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; Gong, Y. Autoregressive Visual Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 15–22 June 2023; pp. 9697–9706.
57. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13608–13618.
58. Sun, R.; Fang, L.; Gao, X.; Gao, J. A novel target-aware dual matching and compensatory segmentation tracker for aerial videos. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 3109722. [CrossRef]
59. Hu, Q.; Guo, Y.; Lin, Z.; An, W.; Cheng, H. Object tracking using multiple features and adaptive model updating. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 2882–2897. [CrossRef]