



## Article

# Slight Aware Enhancement Transformer and Multiple Matching Network for Real-Time UAV Tracking

Anping Deng <sup>1,2</sup>, Guangliang Han <sup>1,\*</sup>, Dianbin Chen <sup>1</sup>, Tianjiao Ma <sup>1</sup> and Zhichao Liu <sup>1,2</sup>

<sup>1</sup> Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun 130033, China; danganping20@mails.ucas.ac.cn (A.D.); chendb@ciomp.ac.cn (D.C.); matianjiao@ciomp.ac.cn (T.M.); liuzhichao201@mails.ucas.ac.cn (Z.L.)

<sup>2</sup> University of Chinese Academy of Sciences, Beijing 101408, China

\* Correspondence: hangl@ciomp.ac.cn

**Abstract:** Based on the versatility and effectiveness of the siamese neural network, the technology of unmanned aerial vehicle visual object tracking has found widespread application in various fields including military reconnaissance, intelligent transportation, and visual positioning. However, due to complex factors, such as occlusions, viewpoint changes, and interference from similar objects during UAV tracking, most existing siamese neural network trackers struggle to combine superior performance with efficiency. To tackle this challenge, this paper proposes a novel SiamSTM tracker that is based on Slight Aware Enhancement Transformer and Multiple matching networks for real-time UAV tracking. The SiamSTM leverages lightweight transformers to encode robust target appearance features while using the Multiple matching networks to fully perceive response map information and enhance the tracker's ability to distinguish between the target and background. The results are impressive: evaluation results based on three UAV tracking benchmarks showed superior speed and precision. Moreover, SiamSTM achieves over 35 FPS on NVIDIA Jetson AGX Xavier, which satisfies the real-time requirements in engineering.

**Keywords:** visual object tracking; UAV tracking; lightweight transformer; siamese network



**Citation:** Deng, A.; Han, G.; Chen, D.; Ma, T.; Liu, Z. Slight Aware Enhancement Transformer and Multiple Matching Network for Real-Time UAV Tracking. *Remote Sens.* **2023**, *15*, 2857. <https://doi.org/10.3390/rs15112857>

Academic Editor: Andrzej Stępczyński

Received: 18 April 2023

Revised: 25 May 2023

Accepted: 26 May 2023

Published: 31 May 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Visual object tracking is a fundamental and challenging task in the field of computer vision, aiming to continuously and steadily track the position and size of a given target based on its initial state in video frames. Due to the superior maneuverability of unmanned aerial vehicles (UAVs), UAV tracking technology has been widely applied in various fields such as aerial photography [1] and remote sensing mapping [2]. However, UAV tracking still faces several difficulties: (1) the targets are often small in size and frequently present with partial or full occlusion, rendering it challenging for the algorithm to acquire target information; (2) given that the tracking duration is prolonged and the platform experiences continuous motion, targets may undergo alterations in appearance and movement blurring, necessitating a strong degree of algorithmic robustness; (3) the limited computational capacity of UAVs requires algorithms to be lightweight to ensure real-time tracking [3]. Therefore, designing an efficient and robust tracker for unmanned aerial vehicles remains an extremely challenging task.

Current advanced tracking methods can be divided into two directions: correlation filters [4] and siamese networks [5]. Due to the limited computing power of unmanned aerial platforms, correlation filter-based methods are often employed for target tracking in practical engineering applications [6]. However, the poor performance in terms of feature representation and insufficient optimization strategies of correlation filter methods make it difficult for them to effectively cope with complex scenes that may arise during the tracking process. In contrast, trackers based on siamese neural networks have made significant progress in recent years in terms of precision and robustness, while also continuously

improving in efficiency. The deployment of lightweight siamese neural network-based trackers in real time on UAV platforms has, thus, become an area for current research activity [7]. Unfortunately, the target feature information extracted by these algorithms is insufficient to deal with complex scenarios such as changes in target appearance. Additionally, the single cross-correlation feature matching method used is unable to fully perceive the foreground and background information of the response map.

In light of these shortcomings, this article proposes a novel Slight Aware Enhancement Transformer and Multiple matching network for real-time UAV tracking. It consists mainly of a Slight Aware Enhancement Transformer (Slight-ViT) feature extraction network and a Multiple matching network, as illustrated in Figure 1. Building on the baseline, we introduce a lightweight ViT feature extraction network that effectively enhances feature salience. Furthermore, through the Multiple matching network, we strengthen the ability to discriminate between targets and the background. Testing, as depicted in Figure 1, has demonstrated the precision and robustness of our algorithm.

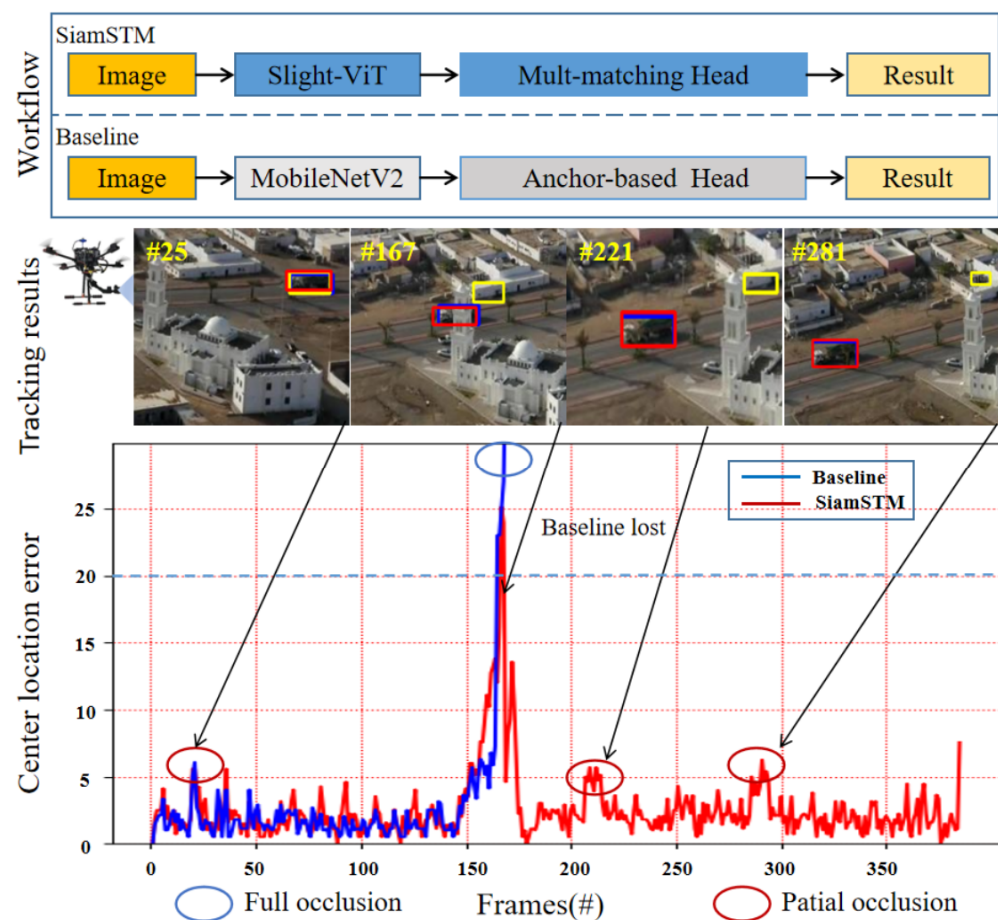


**Figure 1.** Comparison between the baseline tracker and the proposed tracker. The figures, from top to bottom, are: workflow, tracking result, and center location error comparison. When facing the constantly occluded object, our method can emphasize and aggregate the effective information selectively, avoiding interference brought by the environment. In Tracking results, the blue box represents the true value, the yellow box represents the baseline, and the red box represents ours.

The main contributions of this work can be summarized as follows:

- We propose a novel feature extraction network that adaptively aggregates the local information extracted by convolution and the global information extracted by the transformer, encoding both local and global information to enhance the algorithm's performance in terms of feature expression.

- We have developed the Multiple matching network, which separates the classification and regression branches by incorporating the focus information of the target and improving the cross-correlation method, effectively perceiving the foreground and background information of the target, and enhancing the algorithm's ability to cope with complex scenarios.
- Extensive experiments and ablation studies were conducted on three challenging UAV benchmark datasets, demonstrating the superior performance of SiamSTM in handling complex scenarios. SiamSTM achieved a high frame rate of 193 FPS on a PC, and its actual deployment on an NVIDIA Jetson AGX Xavier yielded a frame rate exceeding 35 FPS.

The rest of this article is organized as follows. Section 2 introduces the related work on visual object tracking and the real-time transformer tracking. Section 3 describes the network architecture of SiamSTM and its detailed design. Section 4 presents the experimental results on datasets and ablation study. Conclusions are drawn in Section 5.

## 2. Related Work

Previously, correlation filtering methods, such as KCF, promoted the development of UAV tracking [8], which expanded the sampling capacity using cyclic matrices while avoiding matrix pseudo-operation. In subsequent research, ARCF-HC [9] used the variation of continuous response maps to learn and suppress environmental noise. Autotracker [10] was developed to use local response graphs and global response graphs to adaptively adjust spatial and temporal weights through automatic spatiotemporal regularization. These methods, to some extent, have promoted the development of drone ground target tracking. However, correlation filtering is limited by its capacity for feature extraction and online updating strategies; its poor robustness and precision cannot meet the requirements of current practical engineering. Recently, the strong feature extraction ability of siamese neural networks has greatly improved the precision of object-tracking tasks, and siamese tracking algorithms have attracted wide attention in the field of object tracking.

Siamese neural networks consist of two independent feature extraction networks with shared weights. The template and search area are individually input into one of the two branches, extracting corresponding features. By matching the response maps, the search region with the highest similarity to the target template is identified to complete the tracking task. As a pioneering work, SINT [5] has utilized siamese neural networks to directly learn matching functions and solve tracking problems through similarity learning. SiamFC [11] introduces translational invariance, removes image size limitations, and uses five preset scale anchor boxes to regress the target box. However, the anchor boxes contain a large amount of unnecessary background information, resulting in low tracking precision. SiamRPN [12] has transformed the tracking task from a matching problem to a foreground-background classification problem by utilizing region proposal networks. By regressing on  $k$  anchor boxes of different sizes and aspect ratios at the same location, the precision and robustness of the task have been improved to some extent. Subsequent research has aimed to improve SiamRPN other properties, such as feature extraction, positive-negative sample ratio, and feature enhancement [13–16]. However, the anchor mechanism, which introduces many hyperparameters, has difficulty handling object deformation and pose changes. SiamGAT [17] uses graph neural networks for target matching, enhancing target features, and suppressing background information, thus, improving the tracking performance of the algorithm.

Transformer [18] architecture can overcome the limitation of receptive fields and can concentrate on global information. Inspired by the transformer, many methods have emerged in recent years to improve drone tracking performance [19]. TR-Siam [20] effectively enhances target tracking robustness through a transformer feature pyramid and response map enhancement method. However, this algorithm has a large computational cost and difficulty in meeting real-time requirements. HIFT [21] adopts a transformer

feature fusion method that directly integrates multi-scale feature information, but it does not provide any lightweight improvement to the transformer.

In recent years, improved lightweight algorithms have emerged specifically for the field of drone tracking: SiamAPN++ [22] employs an attentional aggregation network to enhance the robustness of processing complex scenarios; SGDViT [23] enhances spatial features by utilizing time-context data from consecutive frames; TCTrack [24] introduces temporal context information and decodes temporal knowledge to accurately adjust similarity maps; and E.T.Track [25] enables the algorithm to obtain more accurate information about the target through executive attention. While these algorithms remain constrained by the AlexNet [26] feature extraction network, there is still scope to refine their tracking precision and robustness. A baseline algorithm using MobileNetV2 [27] and an anchor box framework to regress target boxes provides a viable solution that effectively ensures robustness in addressing changes in a target's pose, while maintaining real-time speeds. However, the feature extraction abilities of this algorithm are suboptimal, and the target-matching process is relatively simplistic, indicating further opportunities for enhancement in both precision and robustness.

### 3. Proposed Method

In this section, our primary focus will revolve around elaborating the details of the Slight-ViT module and the multi-matching network, as well as exploring ways to integrate them seamlessly into the siamese network. The SiamSTM proposed in this paper is comprised of four principal components, which are illustrated in Figure 2: input, feature extraction, target localization, and output.

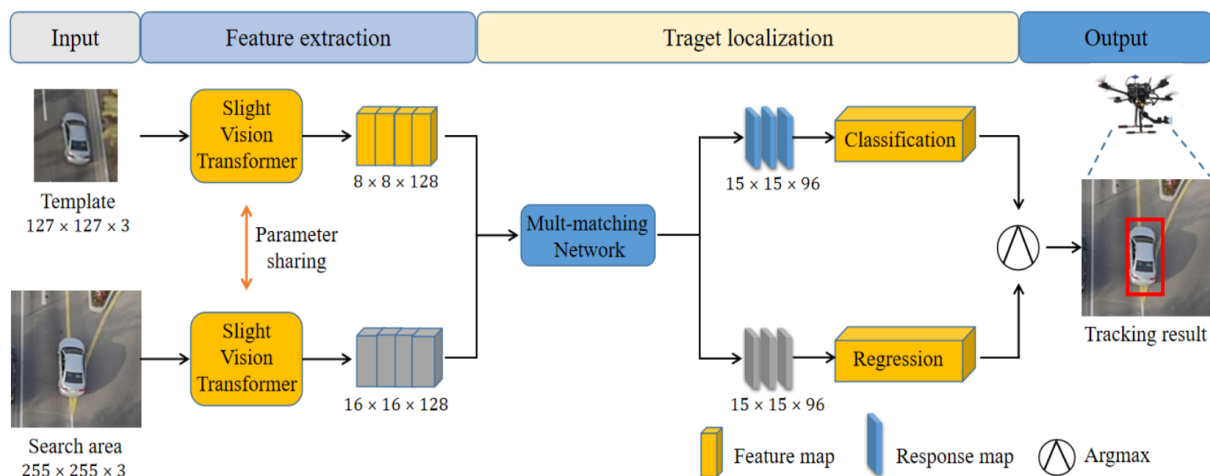


**Figure 2.** The schematic diagram of SiamSTM. Template contains target of interest. The search area is centered on the template position and is three times the size of the template in the detection frame. In the figure,  $255 \times 255 \times 3$  represents height  $\times$  width  $\times$  channel of the characteristic graph, respectively.

#### 3.1. Overall Overview

Our proposed method utilizes Slight-ViT as the backbone for the feature extraction network, with an overall stride of 16. After obtaining the feature maps for the target and search regions, they are then fed into the multi-matching network. This network utilizes the target template as a kernel to match the feature map of the search region, thereby obtaining both category response maps and position response maps. This process is instrumental in improving the tracker's ability to distinguish between the target and background. Finally, the response maps are fed into the prediction head to yield the final tracking results.

#### 3.2. Slight Aware Enhancement Transformer

It is widely acknowledged that the feature extraction network is critical to visual object tracking [13]. In terms of feature extraction algorithms, there are generally two directions:

one relies on convolution and pooling to obtain deep feature maps which represent the position and semantic information of the target [5]; the other employs a large-scale vision transformer to extract global target features and positional information [28]. As convolution neural networks generate feature information with spatial locality due to the use of locally shared convolution kernels, their ability to model long-term dependencies effectively is restricted, particularly in complex scenarios where target features are limited due to factors such as occlusion or loss of sight. Moreover, transformer structures tend to focus on global features but lack local inductive bias, which hinders their ability to model local details and multi-scale information effectively [29]. Finally, the extensive computational and parameter requirements of transformers make it difficult to run them in real time.

Building upon the limitations of existing methods, this paper proposes a lightweight transformer structure that leverages both convolution neural networks and transformers. This design incorporates the global modeling capabilities of transformers while retaining the spatial locality of convolution information, thereby improving both local detail representation and global information perception. The Slight-ViT block, as shown in Figure 3, employs the Unfold-Transformer-Folding structure for global feature modeling and uses the Fusion module to merge transformer outputs with the original input feature map. The Feed Forward Network in the figure is three fully connected layers with identical input and output channels, which are used to integrate feature sequence information. Normalization is achieved through Layer Normalization [30], which standardizes individual samples and stabilizes the distribution of feature maps, thereby avoiding problems such as gradient vanishing that can lead to model degradation. The network uses one multi-head attention layer to compute input features, with the specific formulas as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(Q \cdot K^T / \sqrt{d}\right) \cdot V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, H_2, \dots, H_N)W \quad (1)$$

$$H_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

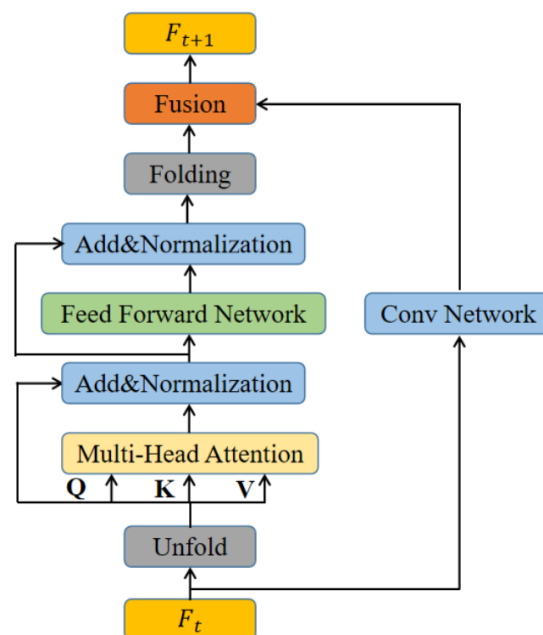


**Figure 3.** This is the schematic diagram of Slight-ViT block.  $F_t$  represents the input feature and  $F_{t+1}$  represents the output feature. Conv Network is the layer of MobileNetV2. The number of layers in Multi-Head Attention is one.

In these formulas, which utilize query (Q), key (K), and value (V) to enhance feature representation,  $d_k$  represents the number of channels,  $N$  represents the number of the attention head, and  $W_i^Q, W_i^K, W_i^V$ , and  $W$  represent the weight coefficient, which is obtained through network training. This module calculates the similarity between Q and K, multiplies V by the normalized weight, and realizes the feature enhancement of V. Additionally, removing  $\sqrt{d}$  is to ensure the stability of training.

Multi-Head Attention is designed to handle serialized data formats, which requires transforming feature maps into sequence information. However, in traditional transformer structures, images are directly fed into the attention mechanism after being flattened, resulting in excessive computations and parameters between tokens. This approach fails to meet real-time requirements for drone tracking. In light of this, this paper proposes a special Unfold and Folding method that divides the image into four categories of tokens, which accurately distinguishes each token from those around it. Each token only undergoes attention calculation with tokens of the same category, significantly reducing computation and parameter requirements while meeting the real-time demands of drone tracking, as illustrated in Figure 4.

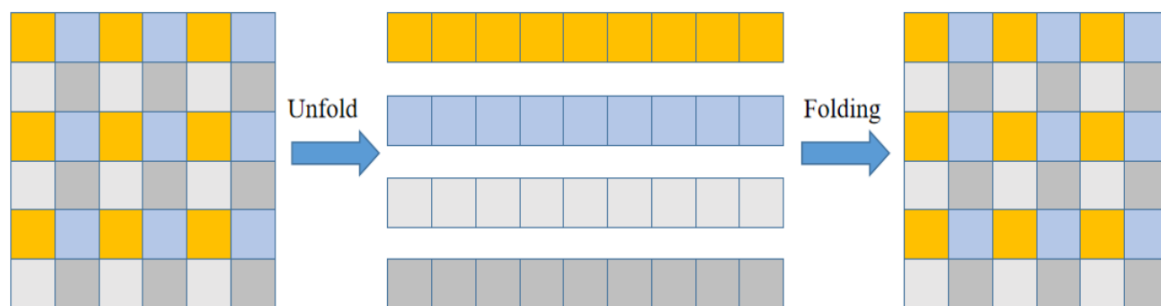


**Figure 4.** Schematic diagram of the proposed Unfold and Folding block. Split the feature map into four categories, with each one being different from the surrounding elements.

The feature extraction network used in this paper has a lower down-sampling rate, resulting in minor differences between adjacent pixel information. Additionally, in Slight-ViT, the convolution network associates image features locally with their surrounding pixels. Using traditional transformer methods would result in increased computational costs far exceeding their benefits in terms of feature salience.

Our proposed Slight-ViT method proposed in this paper successfully incorporates the transformer's exceptional global modeling capabilities while only increasing the computational requirements by 1/4 and eliminating feature map data redundancies. Following attention operations, the Folding method recovers the four classes of tokens as image feature maps and sends them to the Fusion module. This module adaptively fuses results from both the improved transformer and convolution network by channel concatenation and down-sampling, enhancing the network's ability to capture both local and global information. By inserting the designed Slight-ViT block into layers [layer3, layer4, layer5] of MobilenetV2, the Slight-ViT feature extraction network is obtained. In the Figure 5, heat maps [31] have attested to the effectiveness and robustness of this approach in providing accurate object tracking and generating feature maps.

### 3.3. Multiple Matching Network

The visual tracking task of a siamese network tracker is represented as a similarity matching problem [32]. Almost all popular siamese trackers achieve similarity learning through cross-correlation between the target branch and the search branch, sending the response map to the classification and regression networks. However, the two networks focus on different tasks, with the classification network emphasizing the similarity between the extracted features and target category, while the regression network is more concerned with the target's position and scale for bounding box parameter adjustment. Therefore,

considering the differences in content focus in classification and regression, it is necessary to use different branches for computation. In this paper, the Multiple matching network proposes a carefully designed target-matching method that fully perceives the response map information of the target. Its structure is shown in Figure 6.

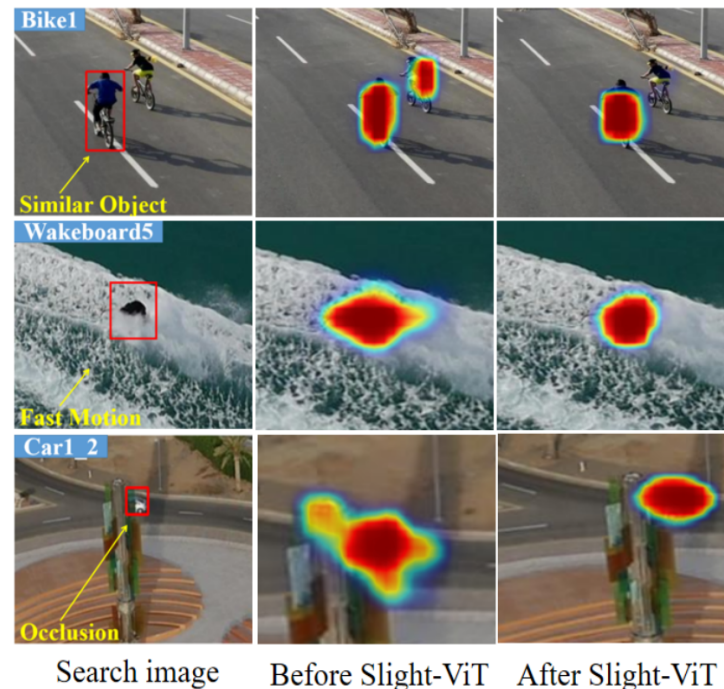


**Figure 5.** Comparison between similarity maps before Slight-ViT (second column) and after Slight-ViT (third column). It can be seen that after using Slight-ViT, the algorithm focuses more on the target itself and is less affected by various complex scene.

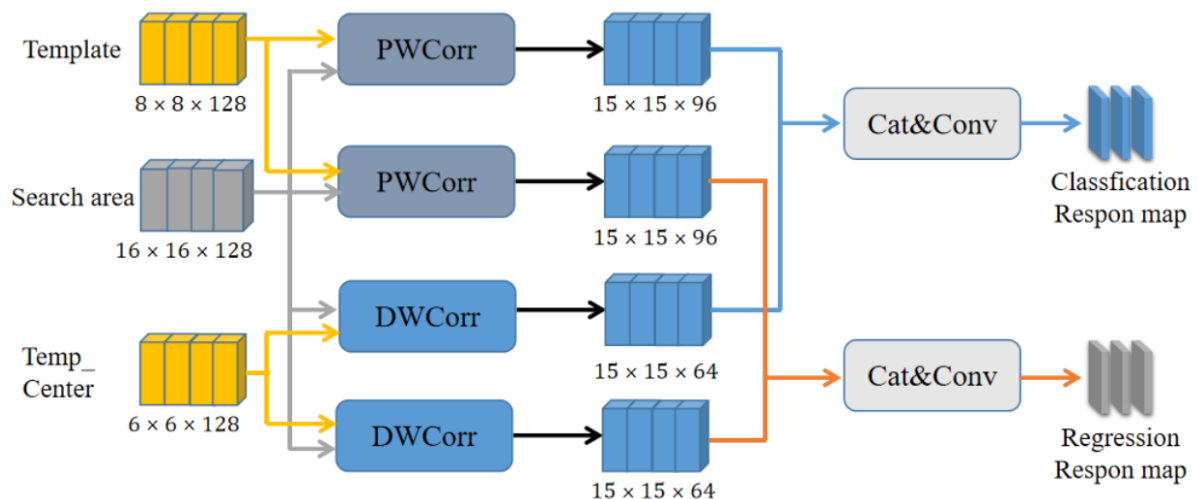


**Figure 6.** Schematic diagram of Multiple matching network structure. The number of channels in the response map obtained through PWCorr is 96, while obtained by DWCorr is 64.

Most prevalent algorithms tend to extract multi-scale features using methods such as feature pyramids to attain higher precision while significantly escalating the overall computational cost [20,21]. To surmount this obstacle, we clip the central region of the target template as a focal point for the target. The crux of the tracking task is to differentiate between the foreground and background in the search region. We opine that incorporating the focus information of the target can enhance the network's ability to discriminate foreground information and achieve equilibrium between the precision and efficiency.

In this network, we perform feature matching on Template, Temp\_Center, and the search region separately. As the Template branch contains richer feature information, we adopt PWCrr [33] to preserve the target's boundary and scale information more effectively in the response map. To generate diverse response maps while ensuring that the network is lightweight, we use DWCorr [13] to calculate cross-correlation results channel by the channel in the Temp\_Center branch, which highlights the focus information of the target and achieves efficient information association under the precondition of being lightweight. We cascade the Template branch with the Temp\_Center branch at the channel level and perform channel-wise down-sampling to adaptively fuse response map information, enhancing the network's reasoning ability for the foreground and background.

Inspired by the Decoupled head idea [34], we employ two response maps, each map dedicated to classification tasks or regression tasks, to minimize the information exchange between the two tasks as much as possible, making the learning of classification and regression networks more independent. This effectively enhances the network's perception ability for categories and discrimination ability for position and scale.

#### 4. Results

In this section, we standardized the evaluation of the performance of the SiamSTM tracker by selecting three video benchmarks: UAV123 [35], UAV20L, and UAVDT [36], to conduct precision and robustness analyses. The UAV123 benchmark consists of 123 aerial videos, while the longer video sequences in the UAV20L benchmark are better suited to real-world drone scenarios. The UAVDT benchmark is more focused on complex environments. These datasets all use precision and success rate as evaluation indicators. Precision refers to the percentage of frames with a center location error (CLE) smaller than a specific threshold in the total number of frames. The threshold is usually set to 20 pixels, and the formulas used are as follows:

$$CLE = \sqrt{(x_{pr} - x_{gt})^2 + (y_{pr} - y_{gt})^2}$$

$$f = \begin{cases} 1, & CLE < 20 \\ 0, & CLE \geq 20 \end{cases} \quad (2)$$

$$Precision = \frac{\sum_{i=1}^N f}{N}$$

In the formula,  $(x_{gt}, y_{gt})$  and  $(x_{pr}, y_{pr})$  represent the center coordinates of the ground truth box and predicted box, respectively. The center location error (CLE) refers to the Euclidean distance between the two center locations. The success rate of the tracking algorithm refers to the percentage of frames where the overlap score (S) is greater than a specific threshold in the total number of frames. The threshold is usually set to 0.5. The overlap score (S) refers to the intersection over union (IoU) ratio of the predicted box and the ground truth box. The formulas involved are as follows:

$$S = IOU \frac{|R_{pr} \cap R_{gt}|}{|R_{pr} \cup R_{gt}|}$$

$$f = \begin{cases} 1, & IOU \geq 0.5 \\ 0, & IOU < 0.5 \end{cases} \quad (3)$$

$$Success\ rate = \frac{\sum_{i=1}^N f}{N}$$

##### 4.1. Implementation Details

The algorithm in this article uses Python 3.7 and PyTorch 1.10 as the training environment, and the experimental equipment is equipped with an Intel i7-12700 CPU, NVIDIA RTX 3090 GPU, and 32 GB RAM. The GOT-10K [37], COCO [38], and VID datasets were

used as the training sets to train the network. During the training process, stochastic batch gradient descent was used as the optimization strategy with a batch size of 64. Weight decay was utilized to control the learning rate. Warm-up training was conducted for the first five iterations, during which the initial learning rate was set to 0.001 and increased by 0.001 for each iteration. After the warm-up phase, a learning rate gradient descent was used to continue training the network. The entire training process included 30 iterations and lasted for a total of 6 h.

#### 4.2. Experiments on the UAV123 Benchmark

The UAV123 dataset is currently the largest and most comprehensive drone-to-ground tracking dataset. The videos captured by drones in UAV123 have diverse scenes and targets, and the targets have various actions in the scene. The UAV123 benchmark includes twelve challenging attributes including Viewpoint Change (VC), Partial Occlusion (PO), Full Occlusion (FO), and Fast Motion (FM). In this article, our algorithm was compared with state-of-the-art algorithms, including SiamRPN++, SiamRPN, SiamAPN++, SiamAPN, SiamDW [39], as well as SiamFC and Staple [40] provided by the toolboxes.

**Overall Evaluation:** The success rate and precision curves are depicted in Figure 7. Our SiamSTM algorithm achieves a success rate of 0.618 and precision rate of 0.809, demonstrating noticeable improvements in comparison with the benchmark algorithm in terms of both the success rate and precision rate. It is noteworthy that our algorithm outperforms the tracking algorithm SiamRPN++, which employs a heavyweight ResNet-50 network for feature extraction, despite adopting a lightweight feature extraction network.

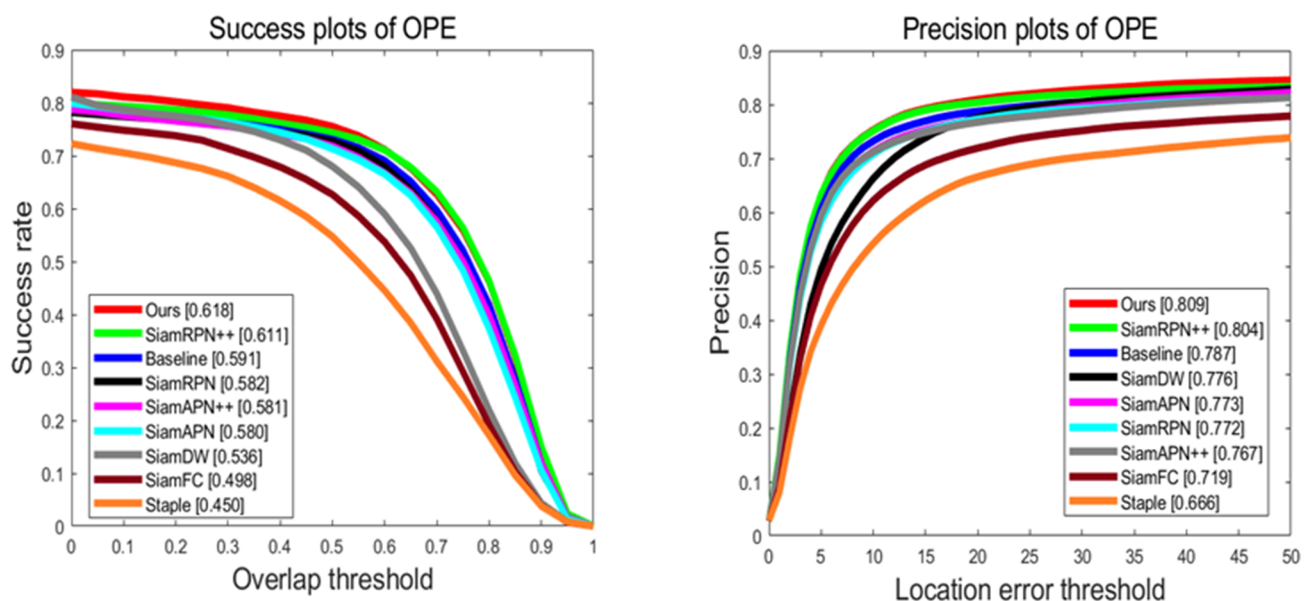


**Figure 7.** UAV123 comparison chart. The proposed algorithm SiamSTM performs favorably against state-of-art trackers. The left shows success plots; the right shows precision plots. The abscissa represents a specific threshold in the precision or success rate formula. Numbers such as 0.618 represent the area of the corresponding algorithm's curve and abscissa.

**Attribute-Based Evaluation:** We have selected challenging attributes that are frequently encountered and difficult to overcome in unmanned aerial vehicle tracking, including similar object (SO), viewpoint change (VC), scale variation (SV), partial occlusion (PO), full occlusion (FO), and fast motion (FM). As shown in Figures 8 and 9, our tracker ranks first on all of these challenging attributes and exhibits significant improvements compared to the benchmark algorithm. In summary, our algorithm can significantly enhance the accuracy and robustness of UAV tracking.

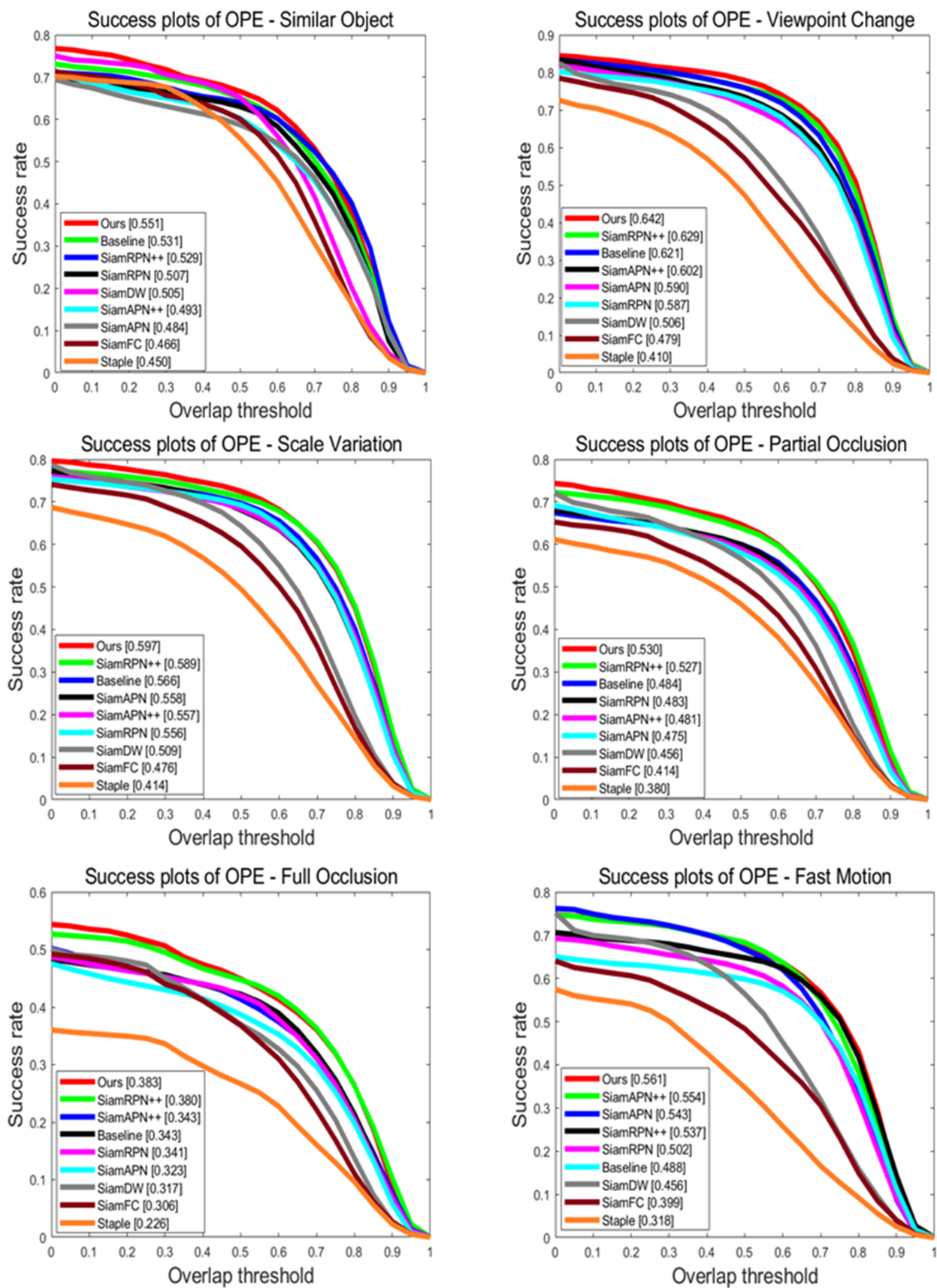


**Figure 8.** Success rates on different attributes of the UAV123 benchmark. Attribute-based performance of all trackers on six challenges. Our tracker achieves superior performance against other SOTA trackers.

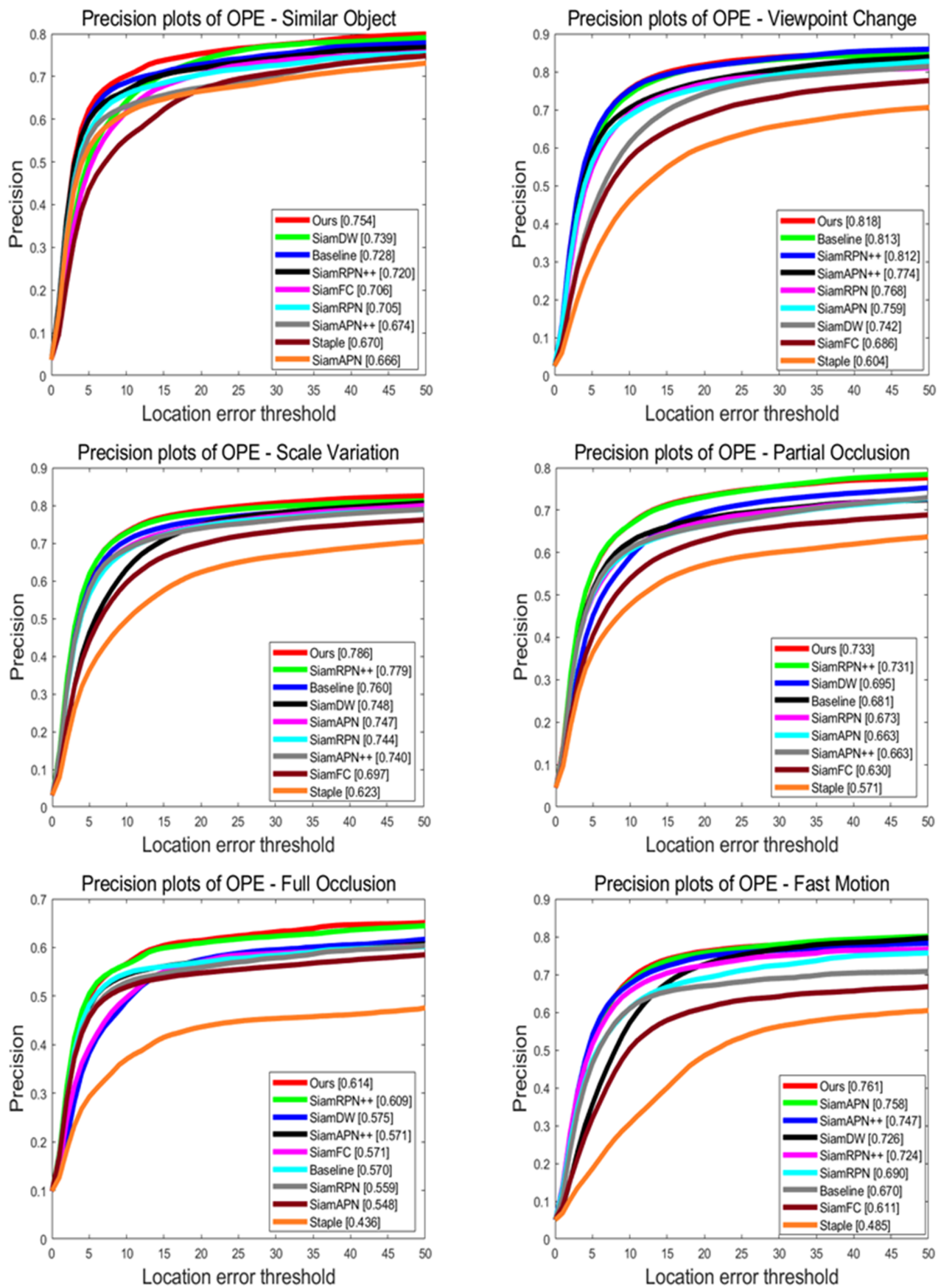


Figure 9. Precision plots on different attributes of the UAV123 benchmark.

Compared to baseline, it can be seen that SiamSTM has significantly improved its success rate and precision in scenarios with significant changes in the appearance or background interference of targets such as similar object, partial occlusion, full occlusion, and scale variation. This demonstrates that the Slight-ViT and Multiple matching networks can enhance feature saliency, effectively perceive target information, and remove background noise. Furthermore, our algorithm offers great improvements in precision as illustrated by precision plots. This indicates that our algorithm can more precisely eliminate the background and locate the center of the target. This is consistent with the experimental conclusion of the heat map reported previously [31].

*Speed Evaluation:* Our method aims to enhance the robustness of tracking algorithms to cope with complex scenarios of UAV tracking. Therefore, to comprehensively demonstrate the efficiency and performance comparison between our algorithm and other state-of-the-art (SOTA) algorithms in UAV tracking, we conducted further comparisons against the UAV123 benchmark. As shown in the Table 1, SiamSTM exhibits improved performance compared to other SOTA trackers and achieves the best computational speed on both GPU and AGX\_Xavier platforms.

**Table 1.** Speed evaluation of SiamSTM and other SOTA trackers on UAV123 benchmarks. Our tracker achieves competitive performance compared with other deeper trackers while possessing superior efficiency.

| Trackers     | Backbone   | Overall |       | Model_Size<br>(MB) | FPS_GPU<br>(FPS) | FPS_Xavier<br>(FPS) |
|--------------|------------|---------|-------|--------------------|------------------|---------------------|
|              |            | Pre.    | Suc.  |                    |                  |                     |
| SiamAPN      | AlexNet    | 0.765   | 0.579 | 118.7              | 180.4            | 34.5                |
| SiamAPN++    | AlexNet    | 0.764   | 0.579 | 187.1              | 175.2            | 34.9                |
| SGDViT       | AlexNet    | 0.766   | 0.585 | 183                | 115.3            | 23                  |
| LPAT [41]    | AlexNet    | 0.790   | 0.593 | 74.6               | 120.4            | 21                  |
| TCTrack [24] | AlexNet    | 0.800   | 0.604 | 87.1               | 125.6            | 27                  |
| Ours         | Slight-ViT | 0.809   | 0.618 | 31.1               | 193              | 36                  |

#### 4.3. Experiments on the UAV20L Benchmark

The UAV20L benchmark is comprised of 20 long sequences, with an average of nearly 3000 frames per sequence. As the frame spacing increases, the position changes of objects between frames become increasingly large and irregular, making target tracking more challenging. Therefore, the performance of long-term tracking can more intuitively reflect the performance of tracking algorithms in actual unmanned aerial vehicle ground tracking scenarios. The results of the UAV20L benchmark test are usually used to demonstrate the long-term tracking ability of the tracker. We conducted comparative experiments between our algorithm, the benchmark algorithm, and the following SOTA algorithms with publicly available results: SiamRPN++, SiamFC++ [42], SiamAPN++, SiamAPN, SESiamFC, and DaSiamRPN [43].

*Overall Evaluation:* The success rate and precision rate curves are depicted in the Figure 10. The success rate of our algorithm, SiamSTM, is 0.580, and the precision rate is 0.742. Compared to the baseline algorithm, SiamSTM exhibits a significant improvement in both the success rate (4%) and precision rate (3.1%). It is worth noting that our algorithm achieves higher precision compared to SiamRPN++ and SiamFC++, which employ large-scale networks for feature extraction; this indicates that SiamSTM can effectively improve the performance of tracking algorithms while ensuring computational speed.

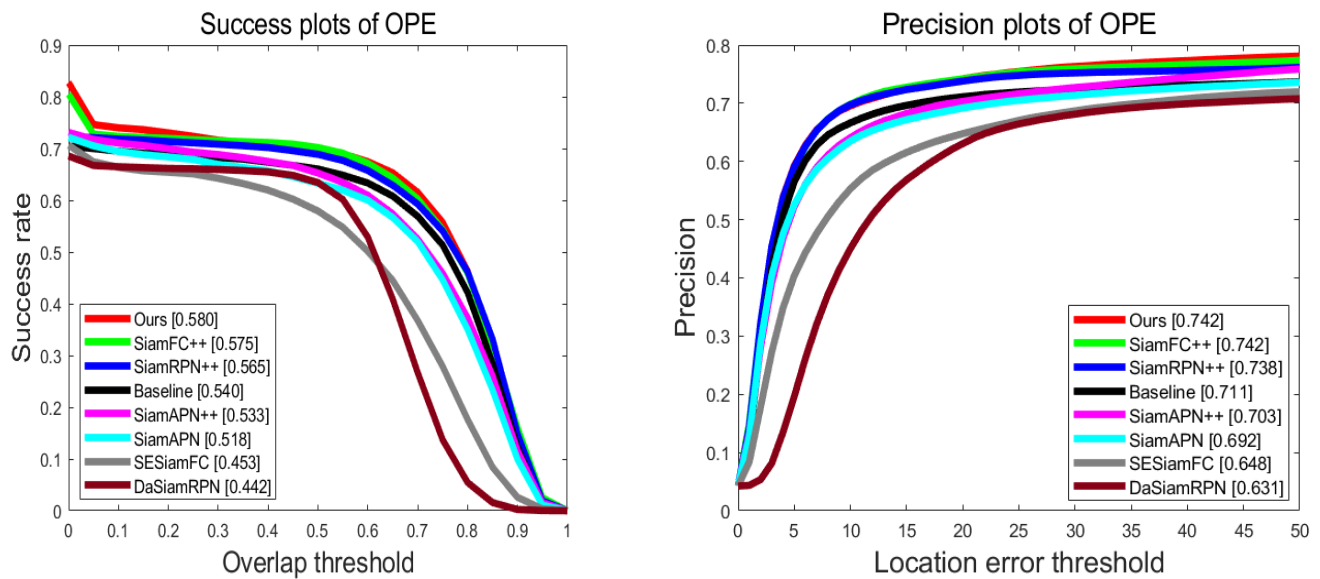


**Figure 10.** UAV20L comparison chart. Each evaluation index is the same as the previous figure. The results illustrate that our method achieves superior performance against other SOTA trackers.

*Attribute-based Evaluation:* To further demonstrate the extent of improvements achieved by our algorithm, a radar chart was used to intuitively display the precision of various algorithms under multiple challenging attributes, with the precision of SiamSTM represented by the enclosed numbers. As shown in Figure 11, our approach not only performs admirably in coping with changes in target appearance (VC, CM) but also exhibits exceptional performance in preventing model degradation (such as when the target disappears from view or is occluded). These challenges are commonplace in real-world UAV tracking scenarios, which fully demonstrates the effectiveness of our algorithm in long-term UAV tracking scenarios.

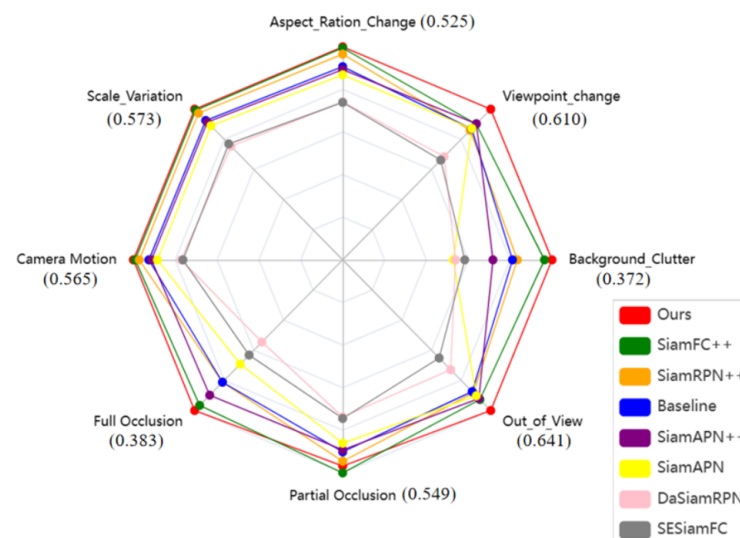


**Figure 11.** Success rates of different attributes on the UAV20L benchmark. It can be seen that our method outperforms other excellent algorithms in these complex scenarios.

It can be seen that the algorithm in this work has achieved improvement in meeting the challenge of two types of target appearance changes, namely, changes in perspective and occlusion. This indicates that the Slight-ViT and Multiple matching network proposed here can reduce the impact of environmental noise and obtain more precise target features.

#### 4.4. Experiments on the UAVDT Benchmark

The UAVDT benchmark concentrates on complex environmental conditions such as weather conditions, altitude, camera views, vehicle types, and obstructions. The benchmark is comprised of 50 sequences and 37,084 frames. In Table 2, we present the overall precision and success rate, and analyze four common attributes of UAV tracking challenges, namely, object blurring (OB), large occlusion (LO), scale variations (SV), and camera motion (CM).

**Table 2.** Attribute-based evaluation of the SiamSTM and 7 SOTA trackers on the UAVDT benchmark. The best two performances are, respectively, highlighted in red and blue.

| Trackers       | Overall |       | OB    |       | LO    |       | SV    |       | CM    |       |
|----------------|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                | Pre.    | Suc.  | Pre.  | Suc.  | Pre.  | Suc.  | Pre.  | Suc.  | Pre.  | Suc.  |
| SiamFC         | 0.681   | 0.447 | 0.738 | 0.453 | 0.479 | 0.359 | 0.604 | 0.424 | 0.616 | 0.409 |
| ECO [44]       | 0.702   | 0.451 | 0.710 | 0.437 | 0.508 | 0.360 | 0.632 | 0.431 | 0.644 | 0.422 |
| UpdateNet [45] | 0.790   | 0.487 | 0.800 | 0.467 | 0.626 | 0.420 | 0.790 | 0.498 | 0.745 | 0.452 |
| SiamAPN        | 0.710   | 0.516 | 0.777 | 0.533 | 0.546 | 0.422 | 0.698 | 0.507 | 0.619 | 0.465 |
| Ocean [46]     | 0.725   | 0.523 | 0.762 | 0.555 | 0.566 | 0.430 | 0.684 | 0.528 | 0.678 | 0.506 |
| SiamAPN++      | 0.758   | 0.549 | 0.800 | 0.560 | 0.588 | 0.451 | 0.745 | 0.536 | 0.719 | 0.519 |
| Baseline       | 0.780   | 0.571 | 0.821 | 0.593 | 0.605 | 0.463 | 0.712 | 0.559 | 0.739 | 0.548 |
| SiamSTM        | 0.802   | 0.595 | 0.866 | 0.626 | 0.647 | 0.541 | 0.771 | 0.582 | 0.763 | 0.560 |

The novel Slight-ViT and Multiple matching network used in the development of SiamSTM, achieved performance improvements in all attributes, particularly in object blurring, where the success rate and resolution improved by 4.5% and 3.3%, respectively. In large occlusion scenarios often encountered by unmanned aerial vehicles, our algorithm made significant progress with a 4.2% and 7.8% increase in the success rate and precision, respectively, compared to the baseline algorithm.

Nevertheless, our algorithm still has limitations. In the situations of scale variation, although the precision increased by 5.9% compared to the baseline algorithm, it is slightly lower than the UpdateNet algorithm, as the estimation of the algorithm's center point is affected due to the long-term accumulation of changes in the target's appearance.

#### 4.5. Ablation Study

To validate the effectiveness of the proposed Slight-ViT and Multiple matching network structures in this paper, Table 3 presents the success rate and precision in four challenges of SiamSTM and the baseline, with different components on the long-term UAV20L benchmark.

**Table 3.** Ablation study of the proposed tracker on UAVDT benchmark. The best two performances are, respectively, highlighted in red and blue. The  $\checkmark$  and  $\times$  indicate whether or not the module is in use, respectively.

| NO. | Slight-ViT   | MMH          | Overall | VC    | OV    | FO    | CM    |
|-----|--------------|--------------|---------|-------|-------|-------|-------|
| 1   | $\times$     | $\times$     | 0.711   | 0.666 | 0.705 | 0.515 | 0.693 |
| 2   | $\checkmark$ | $\times$     | 0.732   | 0.715 | 0.753 | 0.557 | 0.705 |
| 3   | $\times$     | $\checkmark$ | 0.725   | 0.690 | 0.742 | 0.533 | 0.712 |
| 4   | $\checkmark$ | $\checkmark$ | 0.744   | 0.741 | 0.784 | 0.581 | 0.728 |

Analyzing the experimental data in Table 3, it can be found the Slight-ViT can enhance feature saliency, precision, success rate, and ability to handle complex scenarios when compared to the baseline algorithm. This is because the introduction of Slight-ViT provides the proposed algorithm with global spatial contextual information and highlights the saliency of the target's features. It possesses, therefore, strong robustness when dealing with changes in objects' spatial information and appearance. Furthermore, with the addition of the Multiple matching network, the information in the response map was fully utilized,

highlighting the center position of the target and further improving the precision and robustness of SiamSTM.

#### 4.6. Qualitative Evaluation

In order to visually demonstrate the performance of our proposed tracker in handling complex environments, we conducted a qualitative analysis experiment on the UAV123 benchmark by comparing our algorithm to the baseline algorithm.

Focusing on Bike1: The primary challenges of the Bike1 subset are camera motion, scale variation, and similar object. Near frame 110, due to the rapid changes in target scale and appearance, both the baseline tracker and our algorithm experienced scale confusion. In frame 851, the benchmark algorithm appears tracking drift due to interference from similar targets. Despite some errors due to the influence of complex scenarios, SiamSTM was able to redetermine the target position and achieve impressive long-term tracking performance.

Using Bike3: as shown in Figure 12b, in the Bike3 subset, the size of the target is small, making it difficult to obtain feature information, and there may be partial or even full occlusion. When the target begins to be occluded in frame 106, all algorithms are looking for the target. When the target is completely occluded in frame 121, baseline shows tracking drift. However, the algorithm presented in this paper can still track the target correctly after the target reappears, and the CLE throughout the entire process meets the requirements.

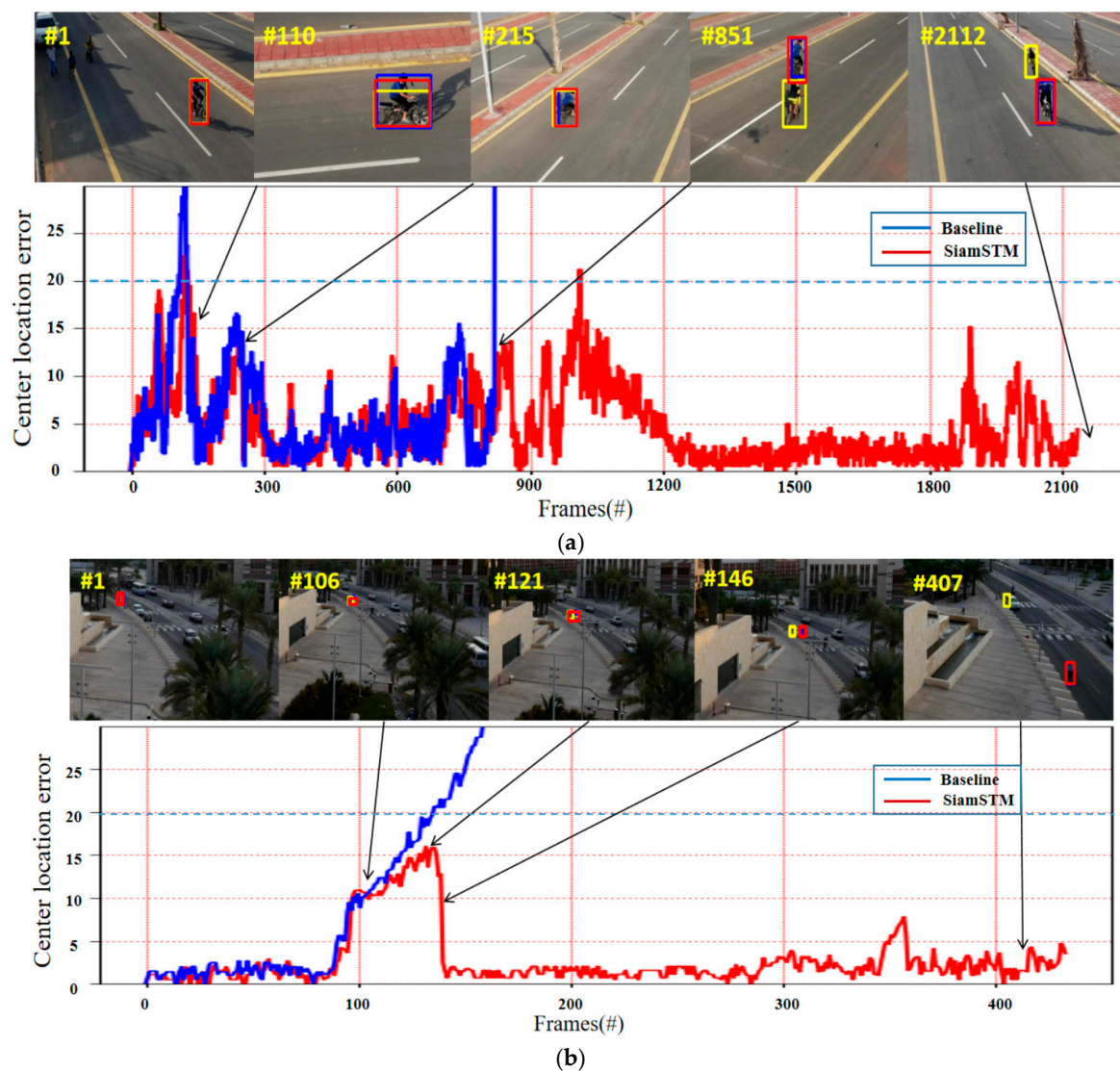


Figure 12. Cont.

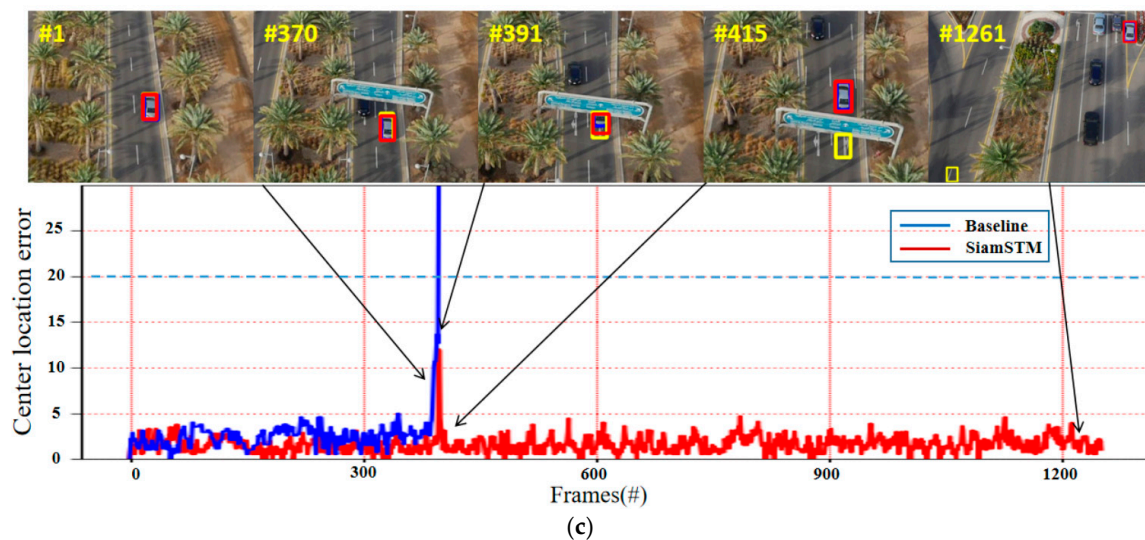


**Figure 12.** Success rates of different attributes on the UAV123 benchmark. The yellow frame represents baseline, the blue frame represents ground truth, and the red frame represents our method. The CLE below the purple dashed line is the success tracking in the test. The first frame of the sequence displays the initial state of the object. (a) Bike1 subset. (b) Bike3 subset. (c) Car4 subset.

As shown in Figure 12c, the main difficulty of the Car4 video sequence is that the car is fully occluded and is constantly moving. In frames 370 to 415, when the vehicle is completely obscured by an obstruction, the baseline algorithm has encountered tracking failure, while our proposed method can still achieve stable tracking of the target despite slight fluctuations. Through frame 1261, it can be seen that although there are similar vehicles around, the proposed algorithm can still continuously and stably track the target of interest.

The qualitative experiments mentioned above showed that with Slight-ViT, clear and accurate target features were obtained and similarity extraction ability improved and with the Multiple matching network, interference from background information was reduced. SiamSTM can, thus, effectively handle occlusion and interference from similar objects, thereby precisely and reliably obtaining the target's position and size. This proves the effectiveness of our algorithm in dealing with complex scenarios.

## 5. Conclusions

In order to meet the performance and feasibility requirements of real-time tracking for drones, this paper proposes a new type of twin neural network tracker for real-time drone tracking based on a lightweight visual perception transformer and multi-matching network called SiamSTM. To enhance the discriminative ability and robustness of the feature extraction network, Slight-ViT is proposed. Furthermore, by fully perceiving response graph information through the multi-matching network, robust and clear feature information and position scale information can be obtained. Finally, performance comparisons on multiple challenging UAV benchmarks show that the proposed method can significantly improve robustness and precision. The utility of the tracker is validated through speed tests. Therefore, we believe that our work can promote the development of UAV tracking-related applications.

**Author Contributions:** All authors participated in devising the tracking approach and made significant contributions to this work. A.D. devised the approach and performed the experiments; G.H. and D.C. provided advice for the preparation and revision of the work; T.M. and Z.L. helped with the experiments. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded in part by the Department of Science and Technology of Jilin Province under Grant 20210201132GX.

**Data Availability Statement:** The data used to support the findings of this study are available from the corresponding author upon request.

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

## References

- Li, B.; Fu, C.; Ding, F.; Ye, J.; Lin, F. All-day object tracking for unmanned aerial vehicle. *IEEE Trans. Mob. Comput.* **2022**. [\[CrossRef\]](#)
- Zhang, Z. Object Tracking based on satellite videos: A literature review. *Remote Sens.* **2022**, *14*, 3674. [\[CrossRef\]](#)
- Fu, C.; Lu, K.; Zheng, G.; Ye, J.; Cao, Z.; Li, B. Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis. *arXiv* **2022**, arXiv:2205.04281.
- Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
- Tao, R.; Gavves, E.; Smeulders, A.W.M. Siamese instance search for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
- Marvasti-Zadeh, S.M.; Cheng, L.; Ghanei-Yakhdan, H. Deep learning for visual tracking: A comprehensive survey. *IEEE Trans. Intell. Transp. Syst.* **2021**. [\[CrossRef\]](#)
- Wu, X.; Li, W.; Hong, D.; Tao, R.; Du, Q. Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey. *IEEE Geosci. Remote Sens. Mag.* **2021**, *10*, 91–124. [\[CrossRef\]](#)
- 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.
- Huang, Z.; Fu, C.; Li, Y.; Lin, F.; Lu, P. Learning aberrance repressed correlation filters for real-time UAV tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
- 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, Seattle, WA, USA, 13–19 June 2020.
- Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P. 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*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
- Bo, L.; Yan, J.; Wei, W.; Zheng, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8971–8980.
- 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.
- Zhang, Z.; Zhang, L. Domain Adaptive SiamRPN++ for Object Tracking in the Wild. *arXiv* **2021**, arXiv:2106.07862.
- Peng, J.; Jiang, Z.; Gu, Y.; Wu, Y.; Wang, Y.; Tai, Y. Siamrcr: Reciprocal classification and regression for visual object tracking. *arXiv* **2021**, arXiv:2105.11237.
- Voigtlaender, P.; Luiten, J.; Torr, P.; Leibe, B. Siam r-cnn: Visual tracking by re-detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6578–6588.
- 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, Nashville, TN, USA, 20–25 June 2021; pp. 9543–9552.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1.
- Thangavel, J.; Kokul, T.; Ramanan, A.; Fernando, S. Transformers in Single Object Tracking: An Experimental Survey. *arXiv* **2023**, arXiv:2302.11867.
- Deng, A.; Liu, J.; Chen, Q.; Wang, X.; Zuo, Y. Visual Tracking with FPN Based on Transformer and Response Map Enhancement. *Appl. Sci.* **2022**, *12*, 6551. [\[CrossRef\]](#)
- 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.
- 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.
- Yao, L.; Fu, C.; Li, S. SGdViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. *arXiv* **2023**, arXiv:2303.04378.
- 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.
- Blatter, P.; Kanakis, M.; Danelljan, M.; Gool, L.V. Efficient visual tracking with exemplar transformers. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 June 2023.
- Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)

27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
28. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
29. Xu, Y.; Zhang, Q.; Zhang, J.; Tao, D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28522–28535.
30. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
31. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
32. 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**. [[CrossRef](#)] [[PubMed](#)]
33. Yan, B.; Zhang, X.; Wang, D.; Lu, H.; Yang, X. Alpha-refine: Boosting tracking performance by precise bounding box estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5289–5298.
34. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
35. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 445–461.
36. Isaac-Medina, B.; Poyser, M.; Organisciak, D.; Willcocks, C.G.; Breckon, T.P.; Shum, H. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
37. 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](#)]
38. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Lawrence Zitnick, C.; Dollár, P. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*; Springer International Publishing: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
39. 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.
40. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H.S. Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
41. Fu, C.; Peng, W.; Li, S.; Ye, J.; Cao, Z. Local Perception-Aware Transformer for Aerial Tracking. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022.
42. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12549–12556.
43. 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.
44. Zolfaghari, M.; Singh, K.; Brox, T. Eco: Efficient convolutional network for online video understanding. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 695–712.
45. 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, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4010–4019.
46. 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, August 23–28, 2020, Proceedings, Part XXI 16*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 771–787.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.