remote sensing

MDPI

Article

Interframe Saliency Transformer and Lightweight Multidimensional Attention Network for Real-Time Unmanned Aerial Vehicle Tracking

Anping Deng [1,2], Guangliang Han [1,*], Dianbing Chen [1], Tianjiao Ma [1], Xilai Wei [1] and Zhichao Liu [1,2]

[1] Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun 130033, China; denganping20@mails.ucas.ac.cn (A.D.); chendb@ciomp.ac.cn (D.C.); matianjiao@ciomp.ac.cn (T.M.); in_weixilai@163.com (X.W.); liuzhichao201@mails.ucas.ac.cn (Z.L.)
[2] University of Chinese Academy of Sciences, Beijing 101408, China
* Correspondence: hangl@ciomp.ac.cn

Abstract: UAV visual-object-tracking technology based on Siamese neural networks has great scientific research and practical application value, and it is widely used in geological surveying, reconnaissance monitoring, and environmental monitoring. Due to the limited onboard computational resources and complex real-world environments of drones, most of the existing tracking systems based on Siamese neural networks struggle to combine excellent performance with high efficiency. Therefore, the key issue is to study how to improve the accuracy of target tracking under the challenges of real-time performance and the above factors. In response to this problem, this paper proposes a real-time UAV tracking system based on interframe saliency transformer and lightweight multidimensional attention network (SiamITL). Specifically, interframe saliency transformer is used to continuously perceive spatial and temporal information, making the network more closely related to the essence of the tracking task. Additionally, a lightweight multidimensional attention network is used to better capture changes in both target appearance and background information, improving the ability of the tracker to distinguish between the target and background. SiamITL is effective and efficient: extensive comparative experiments and ablation experiments have been conducted on multiple aerial tracking benchmarks, demonstrating that our algorithm can achieve more robust feature representation and more accurate target state estimation. Among them, SiamITL achieved success and accuracy rates of 0.625 and 0.818 in the UAV123 benchmark, respectively, demonstrating a certain level of leadership in this field. Furthermore, SiamITL demonstrates the potential for real-time operation on the embedded platform Xavier, highlighting its potential for practical application in real-world scenarios.

Keywords: visual object tracking; UAV tracking; saliency transformer; lightweight attention

check for updates

Citation: Deng, A.; Han, G.; Chen, D.; Ma, T.; Wei, X.; Liu, Z. Interframe Saliency Transformer and Lightweight Multidimensional Attention Network for Real-Time Unmanned Aerial Vehicle Tracking. *Remote Sens.* **2023**, *15*, 4249. https://doi.org/10.3390/rs15174249

Academic Editor: Francesco Nex

Received: 19 July 2023
Revised: 16 August 2023
Accepted: 18 August 2023
Published: 29 August 2023

1. Introduction

Visual target tracking for unmanned aerial vehicles (UAVs) is a fundamental task in remote sensing, whose purpose is to infer and predict the position and size of any target in consecutive aerial images given the initial state of the first frame [1]. Due to the excellent portability and flexibility of drones, UAV tracking has been widely used in various fields, such as autonomous navigation [2] and remote sensing mapping [3]. However, UAV tracking still faces significant challenges due to the following factors: (1) limited aerial perspective and target motion, which often results in occlusion and appearance changes that limit the algorithm's ability to perceive target information; (2) embedded platforms with limited computational resources, which require algorithms to be lightweight to ensure stable and real-time tracking task performance; and (3) long tracking time with temporal correlation, in which existing algorithms are still static trackers that violate the essence of spatiotemporal sequence correlation in tracking tasks, making it difficult for algorithms to effectively handle complex long-term tracking scenarios. Given these challenges, designing an accurate and robust real-time tracking algorithm for drones remains a challenging task.

Currently, the mainstream UAV tracking methods can be divided into two directions: those based on correlation filters (CF) and those based on deep learning [4,5]. Due to their excellent computational efficiency and low computational complexity in the Fourier domain, correlation filter-based trackers have been widely used in current practical engineering. However, CF-based trackers may struggle with accuracy and robustness in complex and diverse drone-ground-tracking scenarios [6]. In recent years, there have been significant advances in the accuracy and robustness of deep-learning-based methods, and the emergence of various lightweight models and hardware-level acceleration on embedded platforms have made lightweight deep-learning-based trackers deployable on UAVs in real time, making this a research hotspot [7]. However, most algorithms ignore temporal contextual information, treating tracking as a template-matching problem independent of the time dimension, which can make them struggle to effectively handle long-term and complex tracking scenarios.

In view of this, we propose SiamITL: a real-time UAV-tracking algorithm composed of the interframe saliency transformer (IST) and lightweight multidimensional attention network (LMAN), as shown in Figure 1. Building on the baseline, we introduce a lightweight multidimensional attention network to explicitly enhance feature saliency and interframe saliency transformer to implicitly improve the algorithm's target-perception ability through the introduction of time contextual information. As shown in Figure 1's test results, our algorithm demonstrates the ability to stably track the target.

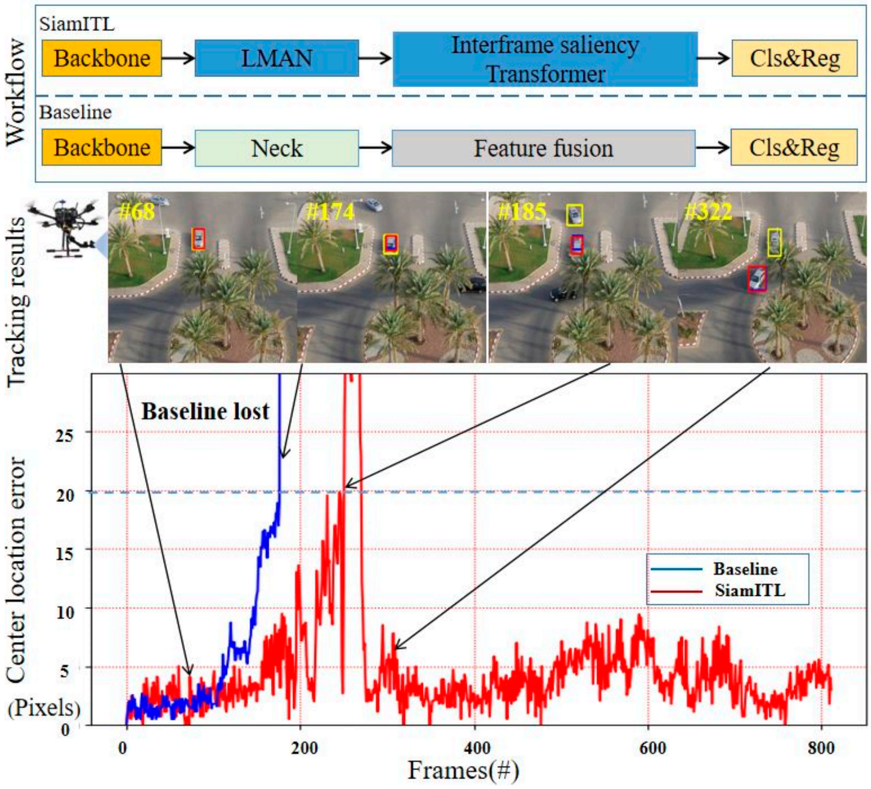


Figure 1. Comparison between the baseline tracker and SimITL. Compared to ground truth (blue), it can be seen that baseline (yellow) experiences tracking drift when encountering occlusion and similar target interference, while our method (red) can track the target robustly. And the number following # represents the current frame number. In the center location error comparison, when the value of the vertical axis is less than 20, the tracking is considered successful.

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

- We present a novel interframe saliency transformer that adaptively aggregates temporal contextual information, focusing on the dependencies between salient regions and their corresponding interframe response maps. This approach endows the algorithm with spatiotemporal correlation to enhance its ability to perceive foreground information;
- We developed the lightweight multidimensional attention network that establishes inter-dimensional dependencies with a remarkably low computational overhead, encoding both channel-wise and spatial information to enhance saliency and discriminative capability of features;
- Comprehensive evaluations on four benchmarks have validated the promising performance of SiamITL compared with other state-of-the-art (SOTA) trackers. In the speed test, SiamITL exhibits real-time performance with a speed of 32 frames per second (FPS) on real embedded platforms.

2. Related Work

The correlation filtering algorithm represented by MOSSE [8] has been widely used in the ground-target-tracking task of UAVs. This method pioneered the transformation of target tracking from the correspondence problem to the classification problem of the target and background. In subsequent studies, Su. used adaptive time regularization constraints and introduced sparse response constraints to alleviate model degradation during the training phase [9]. AutoTrack introduces the method of online adaptive-learning spatiotemporal-regularization term to alleviate the filter over-fitting problem and introduces spatial local-response mapping to improve the learning of target objects [10]. ARCF-HC [11] proposed a distortion suppression filtering algorithm that introduces response map distortion into training formulas to suppress possible environmental noise during the training phase. The above methods promoted the development of drone-ground-target tracking at that time, but calculations based on the Fourier domain made it difficult for relevant filtering algorithms to effectively cope with complex scenarios, such as severe appearance changes. In recent years, feature models based on deep learning have significant and robust feature-extraction capabilities, making significant progress in the accuracy of target-tracking tasks [12]. Among them, algorithms based on Siamese neural networks have attracted widespread attention in the field of target tracking.

Siamese neural networks have two branch inputs: target template and search region. They extract their respective features through a shared backbone network with the same weights and estimate similarity using a cross-correlation algorithm. The search region with the highest similarity is then selected as the tracking target to complete the tracking task. SINT [13] creatively transforms the target-tracking task into the foreground and background matching problem and utilizes the Siamese neural network to construct a similarity function, which is then used to solve the tracking problem through similarity learning. SiamFC [14] is the first end-to-end tracking network, transforming the tracking method from online network parameter updates to a pairwise image-similarity-evaluation problem. This is achieved by calculating the response maps of the two input images to achieve dense and efficient sliding window evaluation. SiamRPN [15] introduces the anchor mechanism, with the network directly outputting confidence scores and bounding box regression results for multiple anchors of different scales on the feature map. Subsequent research has improved SiamRPN in various aspects, such as the feature-extraction network [16], optimization of cross-correlation method [17], and instance-level segmentation. SiamPW [18] designs a classification ranking loss, which enhances recognition capability by modeling the relationship between foreground samples and background samples, thus mitigating classification and localization misalignment problems. The baseline algorithm utilizes MobileNetV2 [19] and anchor box regression on target boxes to provide robustness against similar target interference and occlusion while ensuring real-time running speed. However, this algorithm has a weaker feature discrimination capability and lacks the incorporation of temporal information; therefore, its accuracy and robustness still have room for improvement.

The transformer structure has become a research hotspot in the field of object tracking in recent years due to its excellent feature-encoding capability [20]. Inspired by the transformer, there have been numerous lightweight approaches specifically aimed at UAV tracking [21]. SiamSTM [22] improves the feature-extraction capability of the algorithm through a lightweight transformer-style feature-extraction network and a multiple matching network. LPAT [23] proposes a method that combines local modeling and global search mechanisms to enhance the algorithm's ability to distinguish local details and perceive global information. Recently, there have been several successful attempts to utilize the transformer to model temporal context information for optimization in tracking algorithms. ClimRT [24] generates highly effective potential frames between two consecutive frames to assist the algorithm in coping with complex scenes. TRTrack [25] improves the long-term spatial dependency modeling of the algorithm through trajectory-aware reconstruction of the training process. These algorithms provide a promising solution, However, their principle is somewhat similar to that of UpdateNet, as they both inject temporal context information into the algorithm by updating different dimensions of the template for target matching, ignoring the response map, which contains the most information-rich features and background clues. Therefore, there is still room for further improvement in terms of accuracy and robustness.

3. Proposed Method

In this section, we mainly introduce the details of the interframe saliency transformer and lightweight multidimensional attention network. The framework of SiamITL proposed in this article is shown in Figure 2, which can be divided into input, feature modeling, interface refinement, target localization, and output.

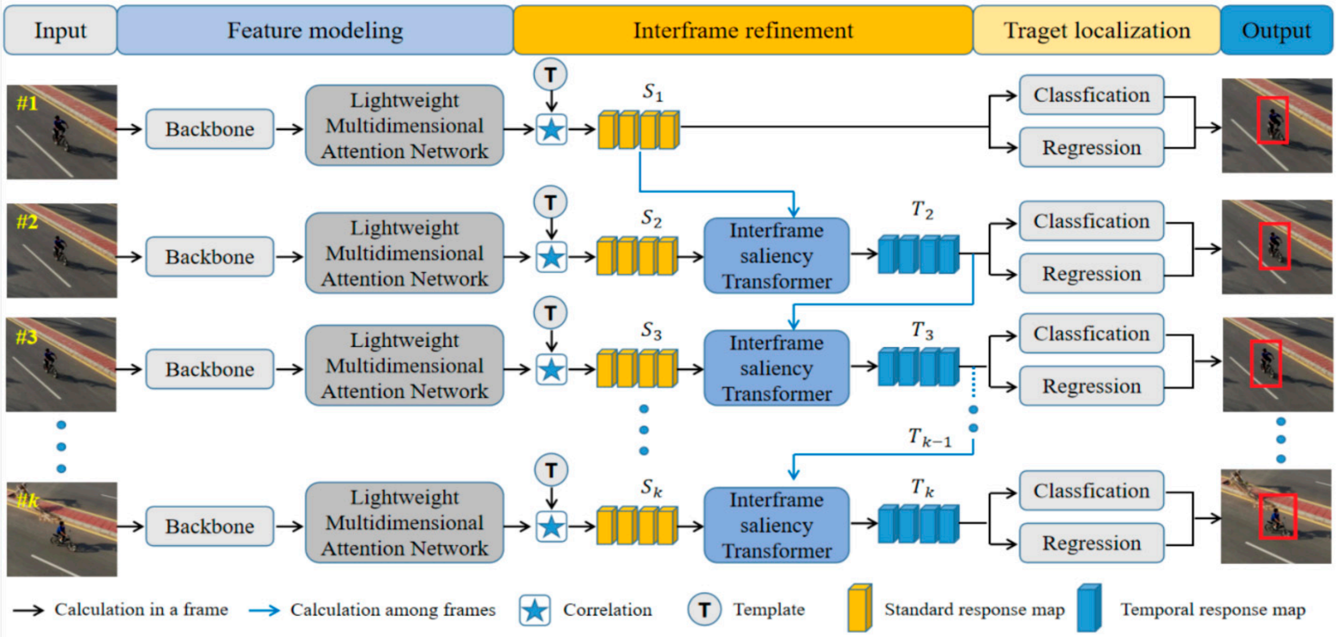


Figure 2. The overview of the SiamITL tracker. It mainly consists of the following four parts: 1. backbone for feature extraction; 2. lightweight multidimensional attention network for feature enhancement; 3. interframe saliency transformer for similarity map significance refinement, and the classification and regression function for obtaining target positions.

3.1. Overview

The method we proposed aims to achieve maximum lightest weight for embedded deployment using a lightweight backbone, MobileNet V2. It then explicitly enhances the saliency and discriminative power of the features through a lightweight multidimensional attention network. After obtaining the response map through feature matching between the search region and the template using the correlation operator, it is fed into the interframe saliency transformer for T_k, which fully utilizes the temporal context information. Finally, target location and state are obtained through a classification and regression network. In particular, when $k = 1$, it is directly input into subsequent interframe refinement and target localization.

3.2. Interframe Saliency Transformer

As the key component of our framework, the structure of the proposed method, i.e., interframe saliency transformer, to consider temporal contexts is shown in Figure 3.

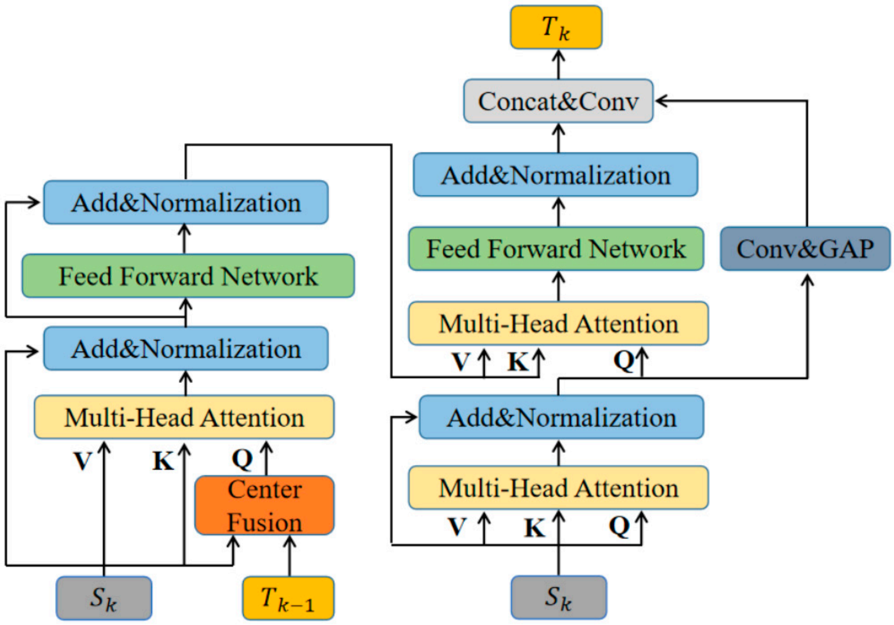


Figure 3. Structure of the interframe saliency transformer. The input is the response map of the current frame S_k, along with significant information from the time series T_{k-1}, and the final output is the enhanced response map T_k. Q, K, and V represent the query, key, and value in the transformer computation formula.

Feedforward network in the diagram consists of two fully connected layers. The first layer uses an ReLU activation function [26], while the second layer does not use an activation function, enhancing the representational power of features through nonlinear processing. Add & normalization uses residual connections and layer normalization [27] to ensure stable training and avoid gradient vanishing. Multi-head attention in the diagram serves as the basic computational component of the transformer structure.

The key to the tracking task is to distinguish the foreground from the background in the search region. The response map of each frame represents the target region of the algorithm's greatest interest, with the most interesting point serving as the central region. Therefore, we crop the central region of the temporal response map as the focus information and send it to the subsequent interframe saliency transformer. We believe that combining the focus information from the response map can improve the network's ability to recognize foreground information and achieve a balance between accuracy and efficiency by cropping away unnecessary background. Influenced by this idea, we designed the center fusion module, which is shown in Figure 4.

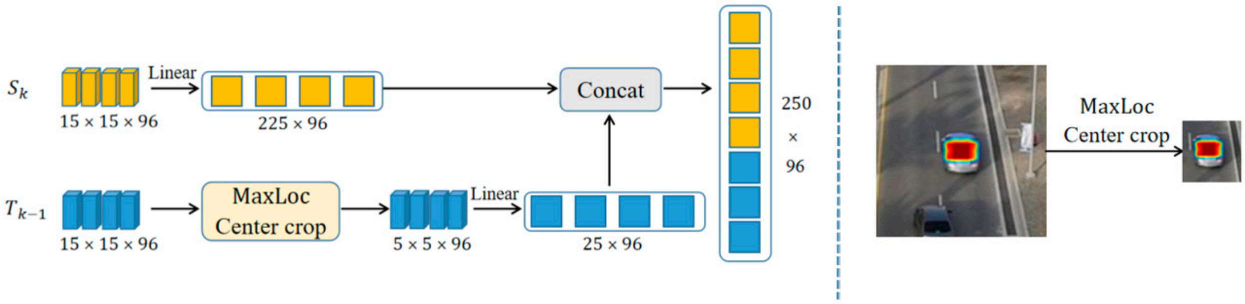


Figure 4. Schematic diagram of MaxLoc Center crop. In the figure, 15 × 15 × 96 represents height × width × channel of the characteristic graph.

The center of the response map corresponds to the predicted center of the target by the algorithm. Through the MaxLoc center crop on the right side of Figure 4, we crop the region of greatest interest to the algorithm as focal information, centered on the maximum value of the response map, to eliminate background information around the target. The focus information is linearized and then concatenated channel-wise with the linearized focus information to obtain a response map that is fused with the focus information. When it comes to how to choose the query, key, and value for the multi-head attention mechanism, we use the temporal sequence information as the query and key and use the current response map as the value to participate in the calculation. We believe that the information in the current frame is more important than the temporal sequence information, so we use this computation method to make the network put more emphasis on the current response map and represent the feature information of the current object more accurately.

During the tracking process of UAV against the ground, there may be background interference caused by similar targets and occlusion, while the transformer structure focuses more on global features and lacks a local inductive bias, which is not conducive to modeling local details. If we do not correct the transmission of temporal information, it may contain some interference information. In order to enhance the algorithm's ability to filter out unnecessary context and improve its perception of local information, we introduce an additional set of convolutions and global average pooling in the feedforward network through residual edges, allowing the network to learn the weights of residual edges to achieve adaptive filtering of temporal information.

It is worth mentioning that since both the input and output layers of the interframe saliency transformer structure are convolutional layers, the zero padding of the convolutional operation implicitly encodes spatial information, allowing for the introduction of positional information between different frames. This eliminates the need for any explicit positional Encoding operations and better adapts to scenarios with varying resolutions, while also reducing the computational complexity of the algorithm.

3.3. Lightweight Multidimensional Attention Network

Due to the limited computational resources of the drone platform, even with excellent lightweight feature-extraction networks, the feature information obtained is still insufficient to cope with complex and diverse usage scenarios. Therefore, while implicitly introducing temporal sequence information, we designed the lightweight multidimensional attention network (MDA) to establish interdependency relationships between dimensions through permute and residual transformation. This network is able to encode the channel information and spatial information of features with a very low computational overhead, improving the saliency and discriminative power of features with low computational cost. The structure is shown in Figure 5.

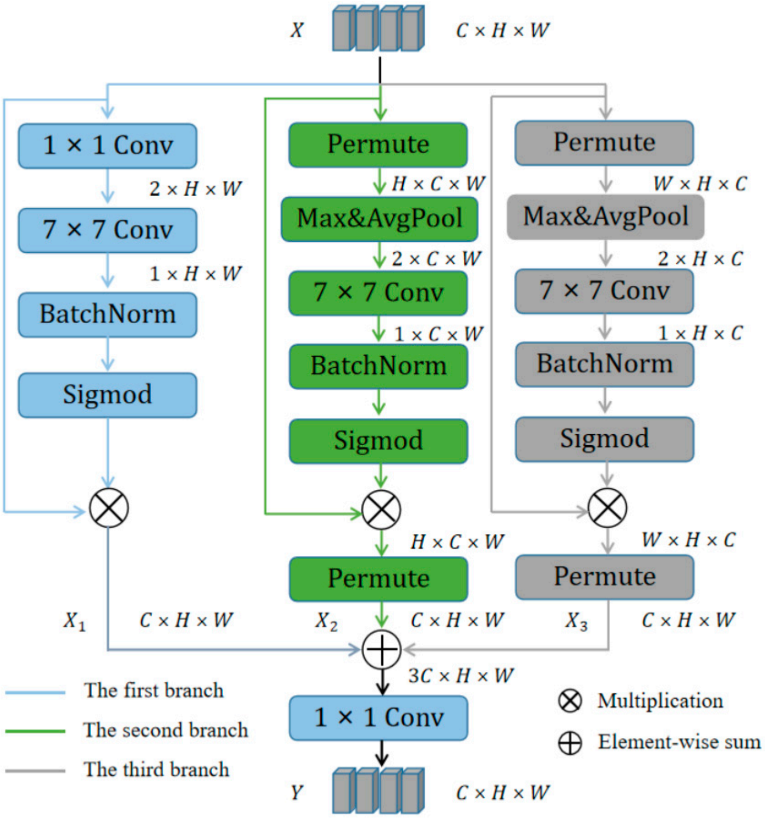


Figure 5. The schema of our lightweight multidimensional attention network. ($C \times H \times W$) represents (Channel × Height × Weight).

We have constructed three branches to enhance the input tensor from three dimensions. In the first branch, we establish information interaction between the H and W dimensions. To achieve this, the input tensor is first simplified through a 1×1 convolution operation, which reduces the number of channels to two. This simplified tensor is then passed through a large kernel convolution operation to greatly increase the receptive field and provide the network with more shape bias. Subsequently, sigmoid activation is used to generate attention weights that are applied to the input, encoding spatial information in the algorithm to obtain the output X_1. This enhances the expressive ability of the algorithm in both the H and W dimensions.

In the second branch, we establish information interaction between the C and W dimensions. To achieve this, the input tensor is first rotated through a permute operation, which reorders the dimensions of the tensor. The tensor is then simplified through a MaxPooling and AvgPooling operation, reducing the number of channels to two. This simplified tensor ($2 \times C \times W$) is then passed through the same steps as the first branch, with the output X_2 rotated back through a permute operation to maintain the same shape as the input.

In the third branch, we establish information interaction between the spatial and temporal dimensions. The input tensor is rotated through a permute operation and then simplified through a MaxPooling and AvgPooling operation with a kernel size of two, reducing the number of channels to two. This simplified tensor is then passed through the same steps as the first branch to finally obtain the output X_3.

In terms of feature discriminability, we interacted information between the spatial and channel dimensions of the features, enabling them to have a better discriminative ability for the target object against its background by encoding global spatial context information. In terms of computational complexity and parameter count, we used MaxPooling and

AvgPooling operations to reduce the number of channels, with the aim of maintaining rich feature information while reducing the depth of the layer and computational complexity.

4. Results

4.1. Implementation Details

The tracker is programmed using Python 3.7 and Pytorch 1.11 using the Ubantu 18.04 operating system and trained and validated using the PySOT library. The experimental equipment includes an Intel i7-12700 CPU, an NVIDIA RTX 3090 GPU, and 64 GB of RAM. We used VID, LaSOT [28], and GOT-10K [29] training sets for joint training, and the pre-trained model for the backbone network is obtained from MobileNet V2 trained on the ImageNet benchmark. After the initial warm-up period of five iterations, learning rate scheduling is used to gradually reduce the learning rate. A total of 20 checkpoints are performed, and the overall training time is 5 h.

4.2. Evaluation Index

In this section, we conducted extensive experiments on SiamITL in terms of accuracy and robustness, evaluating it on four standardized drone-based object-tracking datasets, including UAV123 [30], UAV20L, DTB70 [31], and UAVDT [32], using the one-pass evaluation (OPE) protocol. The tracker was initialized with the ground truth bounding box of the first frame for the entire test sequence and was not re-initialized during the evaluation. We used both accuracy and success rate as evaluation metrics. Accuracy refers to the percentage of frames with a center location error less than a specific pixel threshold. The success rate of the tracking algorithm refers to the percentage of frames where the overlap ratio between the tracking box and the real box is greater than a specific threshold in the total number of frames.

4.3. Experiments on the UAV123 Benchmark

The UAV123 dataset provides a sparse and low-altitude bird's-eye-view tracking dataset, which includes high-resolution videos captured by professional UAVs and simulated bird's-eye-view sequences generated by a simulator. In the UAV123 benchmark, as the drone moves and the camera's perspective changes, the scale and aspect ratio of the target can vary significantly, resulting in complex scenarios. The dataset includes 12 challenge attributes, including viewpoint changes (VCs), partial occlusion (PO), full occlusion (FO), and fast motion (FM). In this test, our algorithm was compared with SOTA algorithms, including SiamRPN++, SiamAPN++ [33], SiamAPN, SiamDW [34], SGDViT [35], HIFT [36] and SiamFC.

Overall Evaluation: The success and accuracy plots are shown in Figure 6. The success rate of our algorithm, SiamITL, is 0.625, and the accuracy rate is 0.818. Compared to the state-of-the-art drone-based ground tracking algorithm, SGDViT, SiamITL has a success rate and accuracy rate that are 4.3% and 5.4% higher, respectively. This indicates that our algorithm has leading accuracy and robustness.

Attribute-Based Evaluation: To analyze SiamITL's ability to cope with various complex scenarios, we conducted analyses of the six most common attributes in the UAV123 benchmark: scale variation (SV), viewpoint change (VC), partial occlusion (POC), full occlusion (FOC), out-of-view (OV), and similar objects (SOBs). The success and accuracy plots are shown in Figures 7 and 8, respectively.

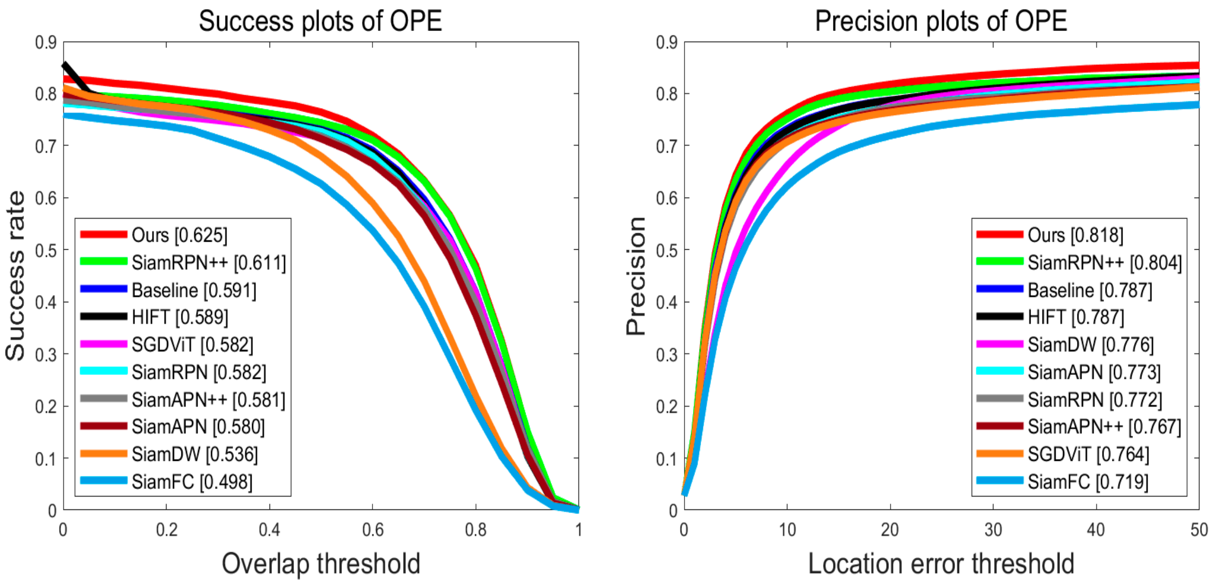


Figure 6. Overall performance of all trackers on the well-known aerial tracking benchmark UAV123. Our tracker achieves superior performance compared to other SOTA trackers.

As can be seen, in scenarios where the target's appearance changes (SV, VC), features are lost (POB, FOB, OV), or the target's information is disturbed (SOB), our algorithm, which combines the interframe saliency transformer and lightweight multidimensional attention network, achieved significant improvements compared to baseline and outperforms existing algorithms. This indicates that the introduction of time contextual information and the explicit enhancement of feature saliency in our algorithm can effectively improve the robustness of drone-based ground tracking in complex scenarios.

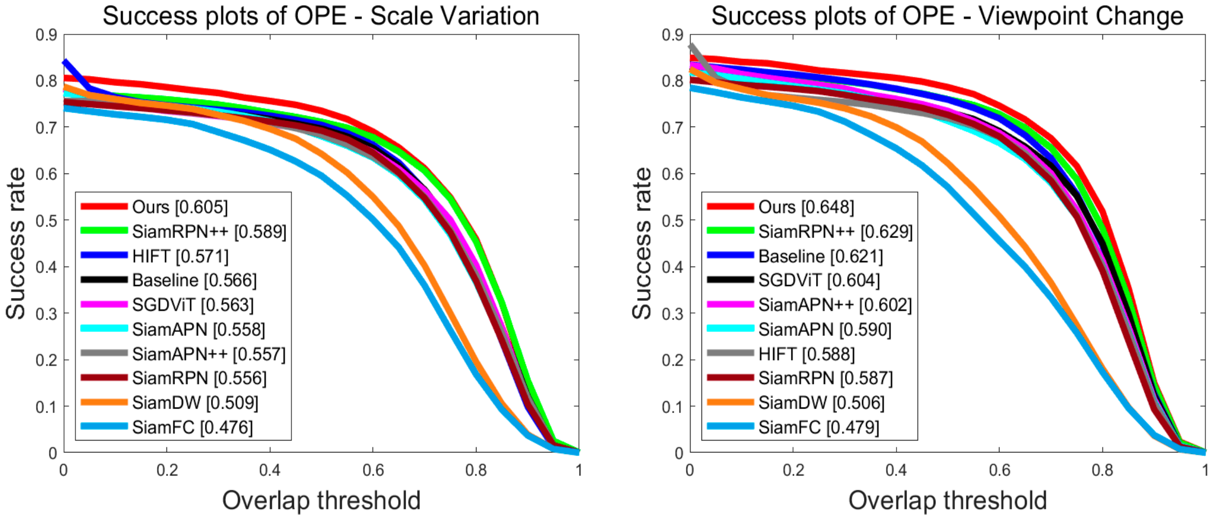


Figure 7. *Cont.*

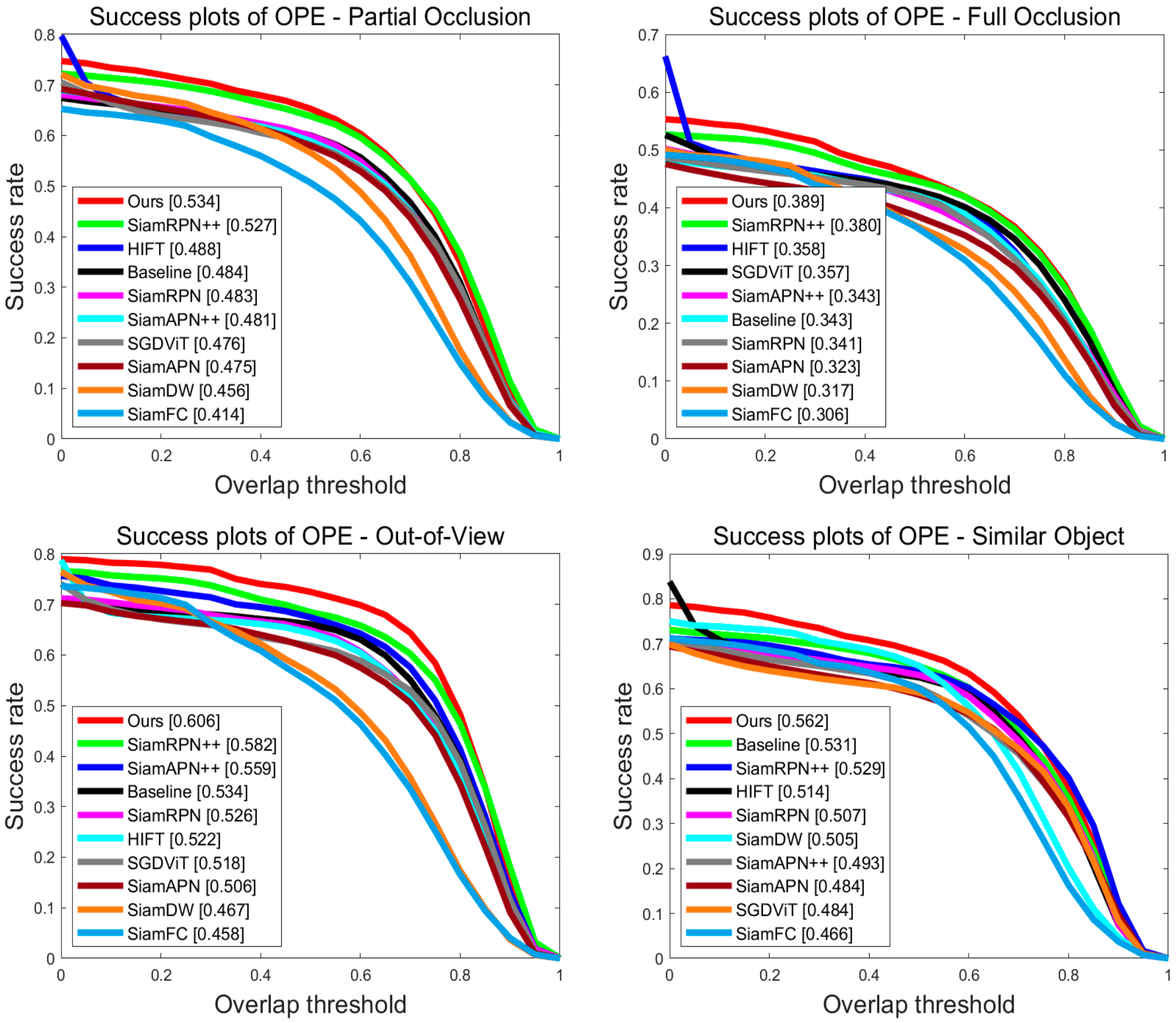


Figure 7. Success rates of different attributes of the UAV123 benchmark.

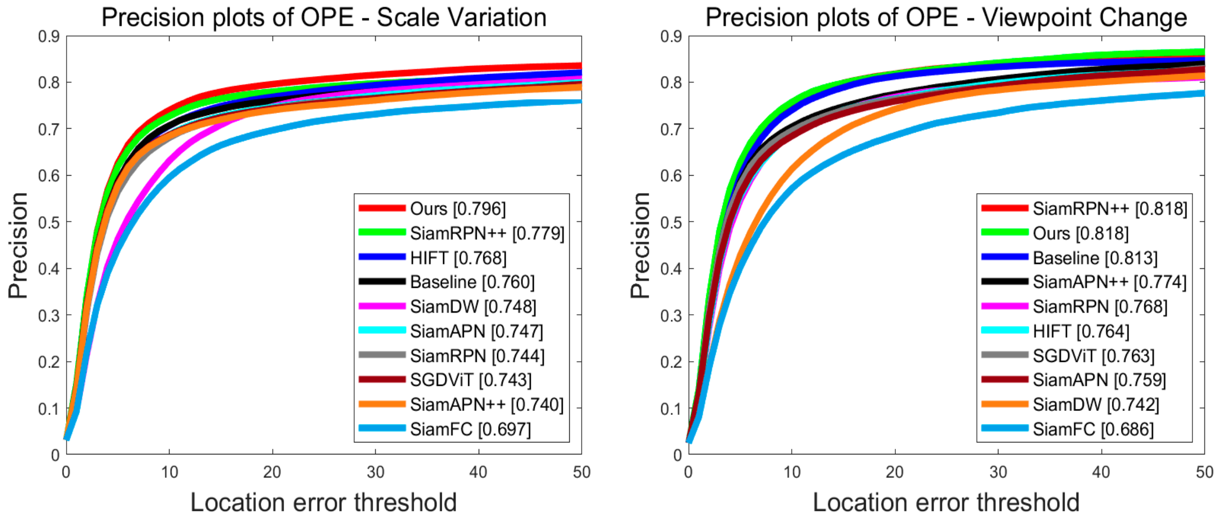


Figure 8. *Cont.*

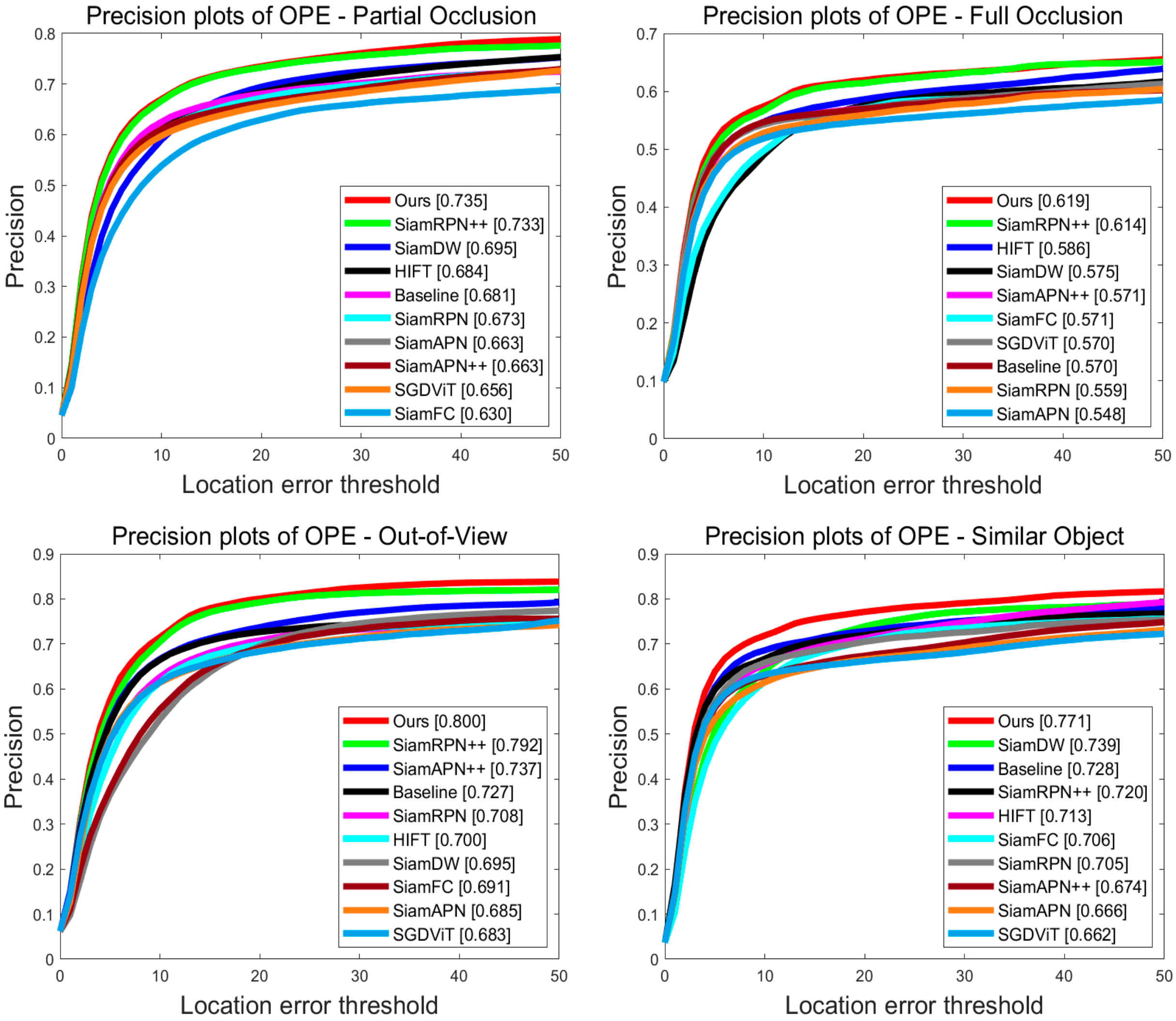


Figure 8. Precision plots of different attributes of the UAV123 benchmark.

4.4. Experiments on the UAV20L Benchmark

The UAV20L benchmark includes 20 long sequences, with an average of nearly 3000 frames per sequence. As the frame spacing increases, the position changes of the object between frames become more significant and irregular, making target tracking more challenging. Therefore, the performance of long-term tracking can more directly reflect the performance of tracking algorithms in actual drone-ground-tracking scenarios. We conducted comparative experiments on this benchmark by comparing our algorithm with the following SOTA algorithms that have publicly available dataset results: SiamFC++ [37], SiamBAN [38], SiamAPN++, SiamCAR [39], and SESiamFC.

Overall Evaluation: As can be seen from Figure 9, our algorithm achieved the best performance in terms of both success rate and accuracy in the long-term tracking scenario. Compared to the baseline, the success rate and accuracy were improved by 4.1% and 3.3%, respectively. It is worth noting that our algorithm outperformed SiamBAN, SiamCAR, and SiamFC++, which use large networks, in terms of both success rate and accuracy. This proves that our algorithm can effectively meet the requirements of long-term drone-based ground tracking scenarios while using lightweight networks.

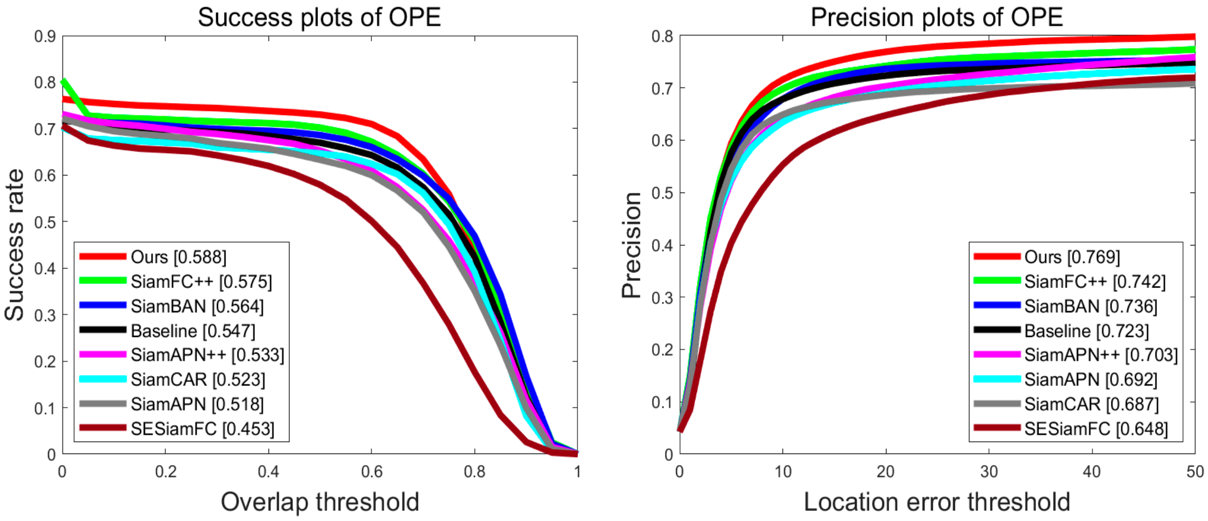


Figure 9. UAV20L comparison chart. Each evaluation index is the same as Figure 6.

Attribute-Based Evaluation: To further demonstrate the performance of our algorithm in complex scenarios, we visually presented the accuracy comparison of various algorithms under multiple challenge attributes using a radar plot. As can be seen from Figure 10, our algorithm significantly outperforms other state-of-the-art algorithms in dealing with complex scenarios, such as background interference, occlusion, out-of-view, and viewpoint changes. This fully demonstrates that our algorithm can effectively cope with various complex scenarios during the tracking process.

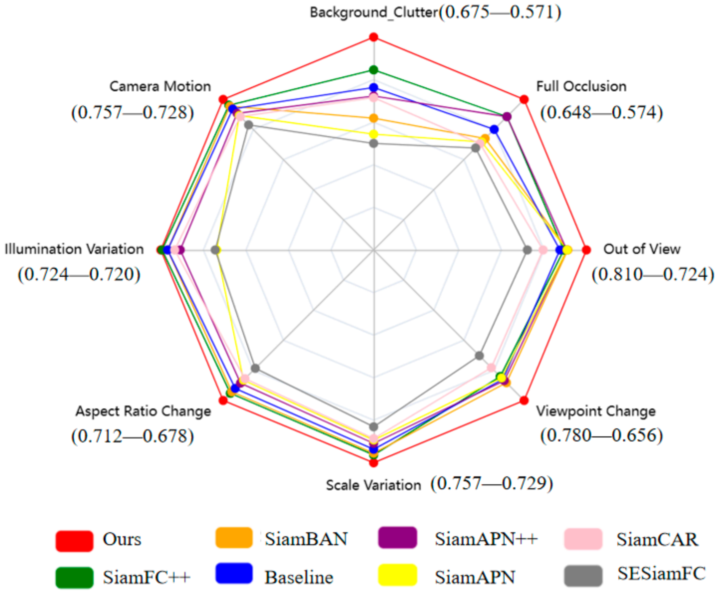


Figure 10. Success rates of different attributes on the UAV20L benchmark. The numbers in parentheses represent the accuracy of SiamITL and the second highest accuracy.

Speed Evaluation: The UAV20L benchmark is an authoritative long-term tracking dataset, and its resolution is commonly used in the field of drone-based ground tracking. Therefore, in order to comprehensively demonstrate the efficiency and performance comparison of our algorithm with other SOTA algorithms in the field of drone-based ground tracking, we conducted further comparisons using the UAV20L dataset. As shown in Table 1 below, SiamITL achieved a certain improvement in performance compared to other SOTA trackers in the field of drone-based ground tracking. Speed experiments on the embedded platform Jeston Xavier demonstrated that our algorithm can meet the real-time computational requirements of actual tracking scenarios.

Table 1. Average attribute-based evaluation of SiamITL and 7 other SOTA trackers on the UAV20L benchmarks. Due to the fact that the runtime speed of algorithms is influenced by multiple factors, such as macs, params, and the hardware platform, it is difficult to assess lightweight algorithms solely based on a single metric. Model size and frames per second (FPS) are the most representative of the algorithm's actual performance and parameter count. Therefore, we selected these two parameters for evaluation.

Trackers	**Backbone**	**Overall**		**Model_Size (MB)**	**FPS_GPU (FPS)**	**FPS_Xavier (FPS)**
		Pre.	**Suc.**			
SiamAPN	AlexNet	0.692	0.518	118.7	180.4	34.5
SGDViT	AlexNet	0.703	0.519	183.0	115.8	23.0
SiamAPN++	AlexNet	0.703	0.533	187.1	175.2	34.9
HiFT	AlexNet	0.763	0.566	82.1	127.7	31.2
SiamSTM	Slight-ViT	0.742	0.580	31.1	193.2	36.0
Ours	MobilenetV2	0.769	0.588	65.4	160.3	32.3

4.5. Experiments on the DTB70 Benchmark

The drone tracking benchmark (DTB70) consists of both short-term and long-term aerial targets, with a total of 70 video sequences. Most of the video sequences were captured by DJI drones in a real university campus at an altitude of up to 120 m. Therefore, most of the scenes in the DTB70 benchmark are low-altitude flight scenes, in which the states of the targets are flexible and the interference in the low-altitude environment is significant, making it a challenging scenario for tracking.

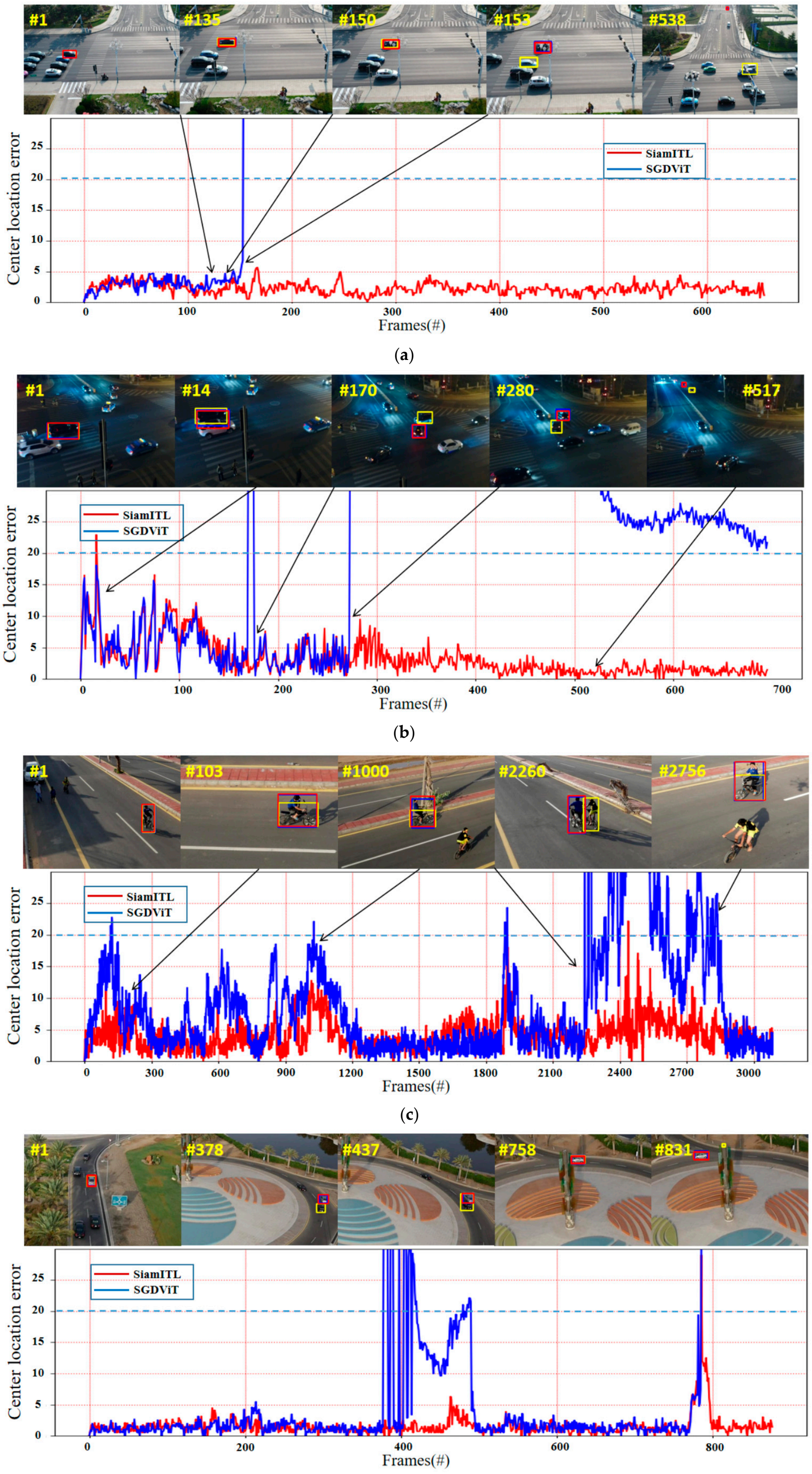
We selected four common challenges that are common in low-altitude scenes: occlusion (OCC), similar object around (SOA), out-of-view (OOV), and aspect ratio variation (ARV). Due to the interframe saliency transformer and the lightweight multidimensional attention network structure proposed in this article, SiamITL achieved good results in all attributes, as shown in Table 2. Our algorithm leads existing excellent drone-tracking algorithms in the occlusion and similar target interference scenarios that drones often encounter. However, our algorithm still has limitations. In scenarios with scale variation and out-of-view challenges, although it achieved first place in terms of success rate, its accuracy was slightly lower than that of the TCTrack algorithm. This is because even though our algorithm filters redundant information, it still difficult to effectively deal with model degradation by redundant long-term information.

Table 2. Average attribute-based evaluation of the SiamITL and 6 other SOTA UAV trackers on DTB benchmarks.

Trackers	**Overall**		**OCC**		**SOA**		**OOV**		**ARV**	
	Pre.	**Suc.**	**Pre.**	**Suc.**	**Pre.**	**Suc.**	**Pre.**	**Suc.**	**Pre.**	**Suc.**
SiamAPN	0.784	0.586	0.654	0.474	0.658	0.480	0.756	0.557	0.715	0.569
LightTrack [40]	0.761	0.587	0.617	0.438	0.642	0.466	0.768	0.586	0.755	0.603
SiamAPN++	0.790	0.594	0.713	0.517	0.682	0.495	0.772	0.590	0.728	0.575
HiFT	0.802	0.594	0.662	0.455	0.700	0.485	0.812	0.596	0.770	0.610
SGDViT	0.806	0.603	0.755	0.526	0.735	0.524	0.778	0.588	0.723	0.573
TCTrack [41]	0.813	0.626	0.751	0.540	0.728	0.535	0.894	0.641	0.769	0.619
Ours	0.824	0.629	0.782	0.552	0.746	0.542	0.883	0.663	0.765	0.633

4.6. Qualitative Evaluation

In order to visually demonstrate the tracking results of the algorithm in this article and further demonstrate its performance in dealing with complex scenarios during the long-term tracking process, we visualized the tracking results of SiamITL and the SOTA algorithm SGDViT, as shown in Figure 11.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 11. Visualization of tracking results on the UAVDT and UAV20L benchmark. The yellow frame represents SGDViT, the blue frame represents ground truth, and the red frame represents our method. The CLE below the purple dashed line (CLE < 20) is the success tracking in the test. The first frame of the sequence displays the initial state of the object. (**a**) S1606 subset. (**b**) S0301 subset. (**c**) Bike1 subset. (**d**) Car1_2 subset. The vertical coordinate. represents the Euclidean distance between the predicted box and the ground truth box, with the unit being the number of pixel points.

The first test sequence, S1606, presents major challenges, including severe similar object, occlusion, and target appearance changes. From frame 135 to frame 153, the target vehicle is partially occluded by a road lamp while a similar moving vehicle appears nearby. At this time, the SGDViT algorithm experiences tracking drift, while our algorithm is able to achieve stable tracking. Near frame 538, the target vehicle drives towards the distance. It can be seen that the yellow box representing the SGDViT algorithm is no longer able to correctly track the target, while our algorithm is able to stably track the target throughout the entire process.

By focusing on the S0301 sequence, it can be seen that this sequence is a tracking scene in a night-time scene, and the main challenges are similar target interference and target appearance changes in low-visibility conditions. Around frame 14, due to partial overlap between the similar target and the target vehicle, the tracking results of our algorithm experience a slight accuracy fluctuation, but it quickly regains correct tracking of the target vehicle. In the subsequent tracking process, the SGDViT algorithm is affected by the interference of surrounding vehicles and experiences tracking failure, while our algorithm effectively overcomes the complex scene of similar target interference and target appearance and size changes.

As shown in Figure 11c, the bike1 test is a long-term sequence of over 3000 frames, accompanied by similar target interference and repeated changes in target appearance scale. Near frames 103, 1000, and 2756, the SGDViT algorithm was unable to correctly handle the complex scene due to significant changes in target appearance scale caused by changes in the perspective of the drone. Near frame 2260, SGDViT experienced tracking drift due to similar target interference. Our algorithm demonstrated impressive long-term performance, maintaining stable and accurate performance throughout.

The Car1_2 subset presents major challenges of target occlusion and similar target interference. Near frames 378 to 437, the SGDViT algorithm experienced tracking box scale errors due to similar target interference. From frames 758 to 831, the SGDViT algorithm was unable to correctly track the target due to severe occlusion. SiamITL maintained a high CLE value compared to the ground truth box, indicating that our algorithm has a certain level of superiority in dealing with similar target interference and occlusion scenarios.

The above qualitative experiments have shown that the algorithm proposed in this article can improve its ability to recognize targets through the interframe saliency transformer, which can obtain temporal contextual information. The lightweight multidimensional attention network is used to explicitly enhance the saliency and discriminative features of the target. As a result, SiamITL can effectively handle various complex scenarios, accurately and reliably obtaining the location and size of the target. Qualitative experimental analysis demonstrated the effectiveness of our algorithm in dealing with complex scenes.

The heatmap is used to visualize the output results as loss feedback to the network structure, and the hook function is used to record the feature layers of the forward propagation and the gradients of the backward propagation in the corresponding layer structure [42]. The gradients are pooled globally and multiplied by the corresponding feature layers using the weights, which indicate how much the network structure focuses on the prediction of that feature layer. To further demonstrate the contribution of the two components of our algorithm, we use heatmaps to visualize the classification layer of the algorithm and showcase the regions of interest to the algorithm in a more prominent manner.

As shown in Figure 12, in the bike1 sequence, due to the interference of nearby similar targets, the attention of the baseline network is completely shifted to the interfering target, while the improved algorithm SiamITL can stably perceive the target. In the truck 2 sequence, it can be seen that when the truck is occluded, the attention of the baseline is diffused to the entire surrounding area, while the proposed algorithm can still accurately capture the truck's feature information. In the group 1 sequence, despite interference from nearby pedestrians, the proposed algorithm exhibits a slight shift in attention, while the baseline's attention is focused on each interfering target, burying hidden dangers for the accuracy and robustness of the algorithm.

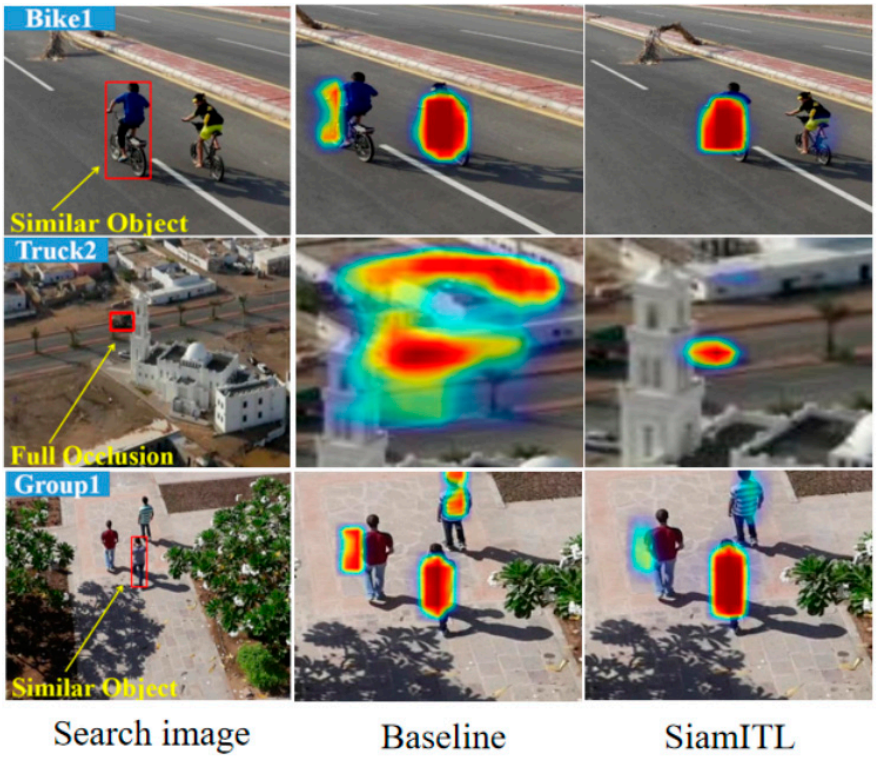


Figure 12. Comparison between similarity maps of baseline (second column) and SiamITL (third column). In order to visually demonstrate the differences in network attention, we selected occluded and similar target interference scenarios. The figure shows that the central region of the image corresponds to the target being tracked by the algorithm at that moment. Therefore, it can be seen in the bike 1 sequence that the Baseline experiences tracking drift.

From the heatmap, it can be clearly seen that the algorithm in this paper introduces the interframe saliency transformer and lightweight multidimensional attention network, and obtains temporal context information through temporal information filtering, thereby eliminating surrounding background interference and strengthening target state information. It also effectively enhances the saliency of features, making the algorithm more focused on the target itself. Therefore, when dealing with situations such as occlusion and interference from similar targets, it can minimize interfering factors and maintain the algorithm's focus on the target.

Through the aforementioned heatmap comparison experiments, it is further demonstrated that the proposed algorithm's introduction of temporal contextual information and multidimensional enhancement of feature saliency is effective in the process of UAV tracking.

4.7. Ablation Study

To verify the structural effectiveness of the interframe saliency transformer and lightweight multidimensional attention network proposed in this paper, Table 3 lists the success rates and four challenge-specific accuracy metrics of SiamITL and baseline with different components. With the addition of the lightweight multidimensional attention network to enhance feature saliency, the tracker's success rate and ability to cope with various complex scenarios have exceeded the baseline. Additionally, with the inclusion of the interframe saliency transformer structure, which adds temporal sequence information to the model, SiamITL's tracking performance and robustness have been further significantly improved.

Table 3. Ablation study of the proposed tracker on UAV20L benchmark. The $\surd$ and $\times$ indicate whether the module is or is not in use, respectively.

NO.	IST	LMAN	Overall	BC	SV	FO	OOV
1	$\times$	$\times$	0.723	0.515	0.709	0.519	0.709
2	$\surd$	$\times$	0.736	0.565	0.728	0.557	0.742
3	$\times$	$\surd$	0.752	0.640	0.743	0.615	0.781
4	$\surd$	$\surd$	0.769	0.675	0.757	0.648	0.810

5. Conclusions

To verify the structural effectiveness of the interframe saliency transformer and lightweight multidimensional attention network proposed in this paper, Table 2 lists the success rates and four challenge-specific accuracy metrics of SiamITL and Baseline with different components. With the addition of the lightweight multidimensional attention network to enhance feature saliency, the tracker's success rate and ability to cope with various complex scenarios have exceeded the baseline. Additionally, with the inclusion of the interframe saliency transformer structure, which adds temporal sequence information to the model, SiamITL's tracking performance and robustness have been further significantly improved. Real-time deployment on Xavier demonstrates the development potential of the algorithm presented in this paper. We hope that our algorithm can inspire further research in enhancing UAV performance, and in future work, we will accelerate it with TensorRT to make it more suitable for embedded chips.

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., Z.L. and X.W. 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, and by the Science & Technology Development Project of Jilin Province, Key R & D Programs No. 20210201078GX.

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

1. 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]
2. Zhang, Z.; Wang, C.; Song, J.; Xu, Y. Object Tracking Based on Satellite Videos: A Literature Review. *Remote Sens.* **2022**, *14*, 3674. [CrossRef]
3. 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]
4. Lee, D.; Kim, G.; Kim, D.; Myung, H.; Choi, H.-T. Vision-based object detection and tracking for autonomous navigation of underwater robots. *Ocean. Eng.* **2012**, *48*, 59–68. [CrossRef]
5. Marvasti-Zadeh, S.M.; Cheng, L.; Ghanei-Yakhdan, H.; Kasaei, S. Deep learning for visual tracking: A comprehensive survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 3943–3968. [CrossRef]
6. Fu, C.; Li, B.; Ding, F.; Lin, F.; Lu, G. Correlation filters for unmanned aerial vehicle-based aerial tracking: A review and experimental evaluation. *IEEE Geosci. Remote Sens. Mag.* **2021**, *10*, 125–160. [CrossRef]
7. Fu, C.; Lu, K.; Zheng, G.; Ye, J.; Cao, Z.; Li, B.; Lu, G. Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis. *arXiv* **2022**, arXiv:2205.04281. [CrossRef]
8. 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.
9. 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]

10. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11920–11929. [CrossRef]
11. 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; pp. 2891–2900.
12. Javed, S.; Danelljan, M.; Khan, F.S.; Khan, M.H.; Felsberg, M.; Matas, J. Visual object tracking with discriminative filters and siamese networks: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 6552–6574. [CrossRef] [PubMed]
13. 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.
14. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P. Fully-convolutional siamese networks for object tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–16 October 2016; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Part II 14, pp. 850–865.
15. 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.
16. 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.
17. 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.
18. Tang, F.; Ling, Q. Ranking-based siamese visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8741–8750.
19. Howard, A.; Zhmoginov, A.; Chen, L.C.; Sandler, M.; Zhu, M. Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation. 2018. Available online: https://research.google/pubs/pub48080/ (accessed on 18 July 2023).
20. 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*, I.
21. Thangavel, J.; Kokul, T.; Ramanan, A.; Fernando, S. Transformers in Single Object Tracking: An Experimental Survey. *arXiv* **2023**, arXiv:2302.11867.
22. 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. [CrossRef]
23. 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; pp. 12122–12129.
24. Fu, C.; Cai, M.; Li, S.; Lu, K.; Zuo, H.; Liu, C. Continuity-Aware Latent Interframe Information Mining for Reliable UAV Tracking. *arXiv* **2023**, arXiv:2303.04525.
25. Li, S.; Fu, C.; Lu, K.; Zuo, H.; Li, Y.; Feng, C. Boosting UAV tracking with voxel-based trajectory-aware pre-training. *IEEE Robot. Autom. Lett.* **2023**, *8*, 1133–1140. [CrossRef]
26. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2013; JMLR Workshop and Conference Proceedings. pp. 315–323.
27. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
28. 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.
29. 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]
30. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14. Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 445–461.
31. Li, S.; Yeung, D.Y. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
32. 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.
33. 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.
34. 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.

35. Yao, L.; Fu, C.; Li, S. SGDViT: Saliency-Guided Dynamic Vision Transformer for UAV Tracking. *arXiv* **2023**, arXiv:2303.04378.
36. 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.
37. 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.
38. 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.
39. 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.
40. Yan, B.; Peng, H.; Wu, K.; Wang, D.; Fu, J.; Lu, H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15180–15189.
41. Cao, Z.; Huang, Z.; Pan, L.; Zhang, S.; Liu, Z.; Fu, C. TCTrack: Temporal Contexts for Aerial Tracking. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14778–14788. [CrossRef]
42. 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.

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.