applied sciences

MDPI

Article

Double Branch Attention Block for Discriminative Representation of Siamese Trackers

Jiaqi Xi, Jin Yang, Xiaodong Chen *, Yi Wang and Huaiyu Cai

Key Laboratory of Photoelectric Information, Ministry of Education, School of Precision Instruments and Optoelectronic Engineering, Tianjin University, Tianjin 300072, China; xijiaqi@tju.edu.cn (J.X.); tiandayangjin@tju.edu.cn (J.Y.); koala_wy@tju.edu.cn (Y.W.); hycai@tju.edu.cn (H.C.)

* Correspondence: xdchen@tju.edu.cn

Abstract: Siamese trackers have achieved a good balance between accuracy and efficiency in generic object tracking. However, background distractors cause side effects to the discriminative representation of the target. To suppress the sensitivity of trackers to background distractors, we propose a Double Branch Attention (DBA) block and a Siamese tracker equipped with the DBA block named DBA-Siam. First, the DBA block concatenates channels of multiple layers from two branches of the Siamese framework to obtain rich feature representation. Second, the channel attention is applied to the two concatenated feature blocks to enhance the robust features selectively, thus enhancing the ability to distinguish the target from the complex background. Finally, the DBA block collects the contextual relevance between the Siamese branches and adaptively encodes it into the feature weight of the detection branch for information compensation. Ablation experiments show that the proposed block can enhance the discriminative representation of the target and significantly improve the tracking performance. Results on two popular benchmarks show that DBA-Siam performs favorably against its counterparts. Compared with the advanced algorithm CSTNet, DBA-Siam improves the EAO by 18.9% on VOT2016.

Keywords: object tracking; Siamese framework; self-attention mechanism

Citation: Xi, J.; Yang, J.; Chen, X.; Wang, Y.; Cai, H. Double Branch Attention Block for Discriminative Representation of Siamese Trackers. *Appl. Sci.* **2022**, *12*, 2897. https://doi.org/10.3390/app12062897

Academic Editor: Antonio Fernández-Caballero

Received: 25 December 2021
Accepted: 7 March 2022
Published: 11 March 2022

1. Introduction

Generic object tracking is a fundamental task in the field of computer vision, with a wide range of application needs in the fields of monitoring, automatic driving [1,2], surgical detection [3], posture recognition [4], and industrial measurement [5]. In recent years, many excellent achievements emerged in the visual object tracking task, but this task remains challenging due to the impact of external factors, such as target deformation, environmental illumination, and background disturbance. These adverse factors damage the feature representation of objects, making it difficult for trackers to distinguish the target from the background distractors, resulting in misdetection.

To enhance the feature representation of objects, the majority of current tracking algorithms rely on the neural network to obtain fine target features [6–8]. Benefiting from the strong information extraction capability of the neural network, visual object tracking has made rapid progress in recent years. Among them, the Siamese framework proposed by Betty [9] in 2016 has achieved a good balance between accuracy and efficiency. It uses the same set of network parameters to extract the deep features of a given target and search inputs, then locates the target by calculating the cross-correlation similarity between the two. The core of this framework is to transform the tracking problem into the template matching problem. Therefore, the template is not restricted by the training data, making the Siamese framework universal in object tracking. However, the background distractors bring side effects to the discriminative representation of the target, and the effects are hard to suppress. It is because the Siamese framework follows the principle that the target within the search region shows substantial feature similarity to the given target, while

the background shows weak feature similarity. Some background distractors may also produce features similar to the given target, thus leading the tracker to judge them as targets. Fortunately, these background distractors produce significant differences from the given target in some detail features. It would be easier to distinguish the target from distractors by performing similarity calculations for features that produce differences. However, the changeable tracking scenarios are challenging to determine which feature can produce differences in the current tracking scenario. To obtain these features and add them to the calculation of distinguishing target from background distractors, we need to expand the range of feature representation.

Many strategies have been used to obtain a rich feature representation in the Siamese framework. Some algorithms increased the amount of input information. DSiam [10] processed the characteristics of a given target, the current search image, and the previous search image in one prediction by integrating the historical target features to build a rich target template information database. It had a good treatment of the deformation problem, but the distractor problem remains. DaSiamRPN [11] added negative samples containing distractors to the training data, facilitating network learning about handling similar negative samples, but it required a large amount of additional training data from other domains. Other algorithms obtained a richer feature representation by changing the network structure. SiamRPN [12] extracted the features of the search regions of different scales and shapes through the region proposal network (RPN) and indirectly obtained the target shape information through the similarity calculation. Similar to DSiam, it only has a good effect on the target deformation problem. In addition, SiamDW [13] and SiamRCR [14] enriched feature representation by refining the current features. SiamDW obtained deeper pattern information to fit complex tracking objects by increasing the depth of the network. SiamRCR added a regression branch to adjust for features dynamically against the samples. SA-SIAM [15] used two sets of Siamese networks to obtain semantic and appearance features to enrich the variety of input features. The similar SPM [16] also described the features as two types, namely coarse and fine representations. The difference is that SPM uses a single Siamese network and a smaller number of parameters. Since these algorithms do not consider background distractors suppressing, they cannot advance the discriminative representation, though abundant features have been obtained.

To suppress the sensitivity of trackers to background distractors, we propose a Double Branch Attention (DBA) block to advance the discriminative representation for Siamese tracking. It selects the features that the current tracking scenario focuses on from rich feature representations and then selectively enhances and effectively integrates them. First, to obtain a rich feature representation, the DBA blocks concatenate channels of multiple layers of backbone. Second, the attention mechanism acts on two branches of the Siamese structure to separately compute for channel self-attention, which achieves selective enhancement of the target and detection branches. Finally, the enhanced features of the two branches are fused by cross-correlation computation. Furthermore, we associate two branches to engage the target information in the procession of feature expression in the detection branch to highlight the features related to targets. To achieve end-to-end tracking, we propose a Siamese tracker equipped with the DBA block named DBA-Siam. Experiments show that DBA-Siam achieves advanced results on UAV123 [17] and VOT [18,19] benchmarks.

2. Related Work

2.1. Siamese Trackers

In recent years, the trackers based on the Siamese framework have many advanced proposals. Full convolution trackers, such as SiamFC [9] and SiamDW [13], directly predict the target location through the similarity map and obtain a rough target bounding box with a fixed aspect ratio. Region proposal trackers, such as SiamRPN [12] and SiamRPN++ [20], feed the similarity map into the classification and regression branches and obtain a more accurate bounding box through the region proposal network. Compared with full convolution trackers, the application of the region proposal network makes trackers well fit

the shape of the target and has fine effectiveness in dealing with the target deformation problem. Mask trackers, such as SiamMask [21], add a mask branch to the region proposal tracker to provide pixel-level refinement masks, which are more accurate than bounding boxes. However, they need a lot of extra semantic segmentation training data, and the increased computational cost has a negative impact on tracking efficiency.

To achieve more accurate tracking effects at a low computational cost, we chose the region proposal tracker as the baseline tracker to carry the proposed DBA block. The baseline tracker consists of a feature extraction network and a region proposal network. The feature extraction network ϕ extracts the feature information of the target z and the search input x, respectively, and both of them share the same network parameters.

The feature information $\phi(x)$ and $\phi(z)$ are cross-correlated to obtain a similarity measure $g(x, z)$:

$$g(x, z) = \phi(x) \star \phi(z) \tag{1}$$

where $\star$ represents a cross-correlation operator. The similarity measure $g(*, *)$ expresses the similarity between the target and the proposal region of the search input, which reflects the possibility that the proposal region on the search input contains the real target. The similarity measure $g(*, *)$ is fed into the Cls_head and Loc_head to generate dense classification responses *cls* and regression responses *loc*.

$$[cls, loc] = [\text{Cls\_head}(g(x, z)), \text{Loc\_head}(g(x, z))] \tag{2}$$

The classification responses are responsible for the pixel classification to obtain the approximate location of the target, and regression responses are used to refine the size of the bounding box.

2.2. Attention Mechanism

The attention mechanism [22,23] is a common strategy to advance feature representation to suppress background distractors. Its principle is to derive feature weights from the feature information to highlight robust features dynamically. Recently, attention mechanisms have been introduced into Siamese-based trackers to improve tracking performance.

MemTrack [24] applied the attention mechanism to the target features of historical and current frames, enhancing the target representation by integrating features from previous targets. DenseSiam [25] and RASNet [26] proposed attention blocks in the target branch to refine target features, generating reliable response maps to the deformed target. These algorithms focus on the target branch of the Siamese architecture, where the role of the attention mechanism is to weight regions of the target in different channels to find and focus on regions of interest that can describe the essential characteristics of the target. The focused features can still guide the network to the target when the target is deformed in long-distance tracking.

The proposed algorithm is structurally different from the above algorithms. These algorithms were designed to find the features that describe the essential properties of the target or the tracking scene, so they do not need to pay attention to too many channels and layers. However, we aim to find features that make background distractors differ from the target, which may be distributed across partial channels at multiple layers, responding to specific properties of the object. This idea forced us to include features of channels of multiple layers as input to attention blocks to obtain more alternative features.

In the Siamese framework, search inputs contain more background regions and similar distractors than the target, so the detection branch may focus more on the feature channels that generate variability. To prevent the detection branch from losing attention to the feature reflecting the essential properties of the target during training, we compensated the detection branch with attention information from the target branch to preserve the essential information of the target.

3. Proposed Method

In this section, we introduce the details of the DBA block and how it combines with the Siamese framework. The DBA-Siam is a Siamese tracker equipped with a DBA block, whose structure is shown in Figure 1.

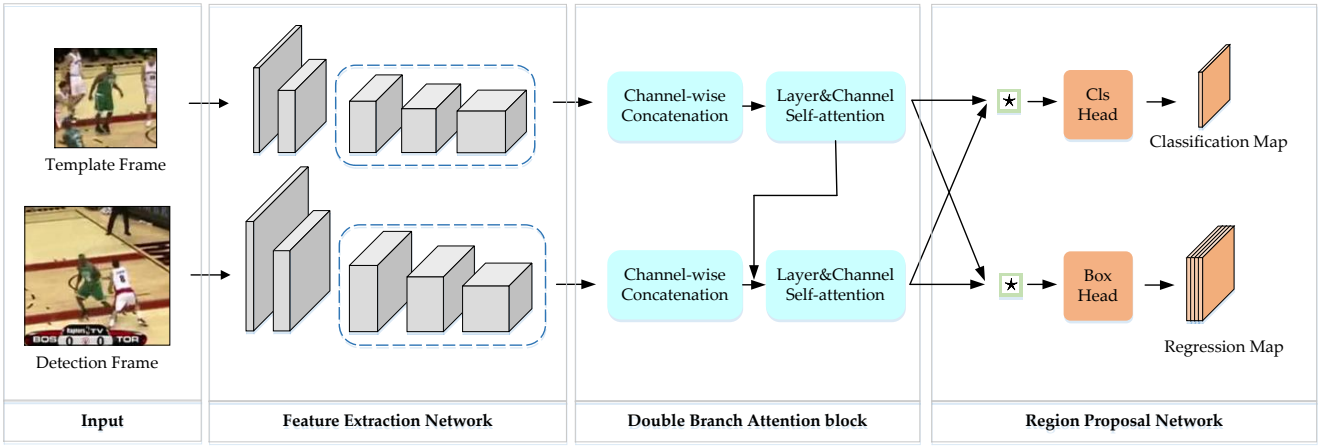

Figure 1. The schematic diagram of DBA-Siam. It includes Input, Feature Extraction Network, Double Branch Attention block, and Region Proposal Network where ⋆ stands for cross-correlation operation. The template frame contains a single target and part of the background, and the detection frame is the video frame that needs to search for the target.

3.1. Overall Overview

In the Siamese framework, the feature extraction network extracts the target and search inputs. With the increase of network layers, it calculates deeper and deeper features. We refer to the processing method of Siamban [27], which uses the features of the last three layers on the two Siamese branches as the information to be processed. The proposed DBA block performs channel concatenation of these features and weights them by attention computation to selectively enhance layers and channels. In addition, the channel self-attention matrix of the target branch is used to compensate the detection branch to ensure the tracker's ability to recognize the target.

3.2. DBA Block

For most tracking scenes, there are significant semantic gaps between target and background, so the deep features of the tracking network are peculiarly prone to be trained to pattern features of the semantics, making the tracker challenging to distinguish the target from background distractors with the same semantics. Thus, the proposed DBA blocks fuse feature information of the target branch and the detection branch at multiple layers, enabling the tracker to focus on both deep features and shallow features of the target and background distractors.

The channel attention mechanism demonstrates that equal treatment of features across all channels would hinder the network's power of representation because each channel responds to a specific feature type. For example, for deep features that can reflect object category information, each channel generally responds to a specific semantics; for shallow features that can reflect object detail information, each channel may respond to a specific shape structure. The features we need that can produce variability of the target and the distractors may exist on some channels of a certain layer, while other channels in this layer contribute less. Treating all channels equally may drown the information in these channels that produce variability in the average calculation. Thus, the proposed DBA block focuses on channels of multiple layers and hopes to assign dynamic weights to these channels through attention calculation to highlight channels that produce discriminative representation.

In summary, the proposed DBA block extends the attention block into a two-branch form combined with the Siamese framework. Its purpose is to selectively enhance features

of the two branches based on the current tracking task to enhance channels that can produce discriminative representation. In contrast to attention blocks focusing on channels or regions, we focus on three aspects of the tracker structure: layers, channels, and branches. For the different layers of each branch, we hope to determine which layer of features reflect the difference between distractors and the target according to the current tracking task; for channels of a specific layer, we hope to enhance the feature channels beneficial to target recognition while reducing the distractor of the invalid feature channel; for the two branches of the Siamese framework, we hope that the detection branch maintains the ability to recognize the target when distinguishing the target from background distractors. According to these requirements, we proposed the DBA block, as shown in Figure 2.

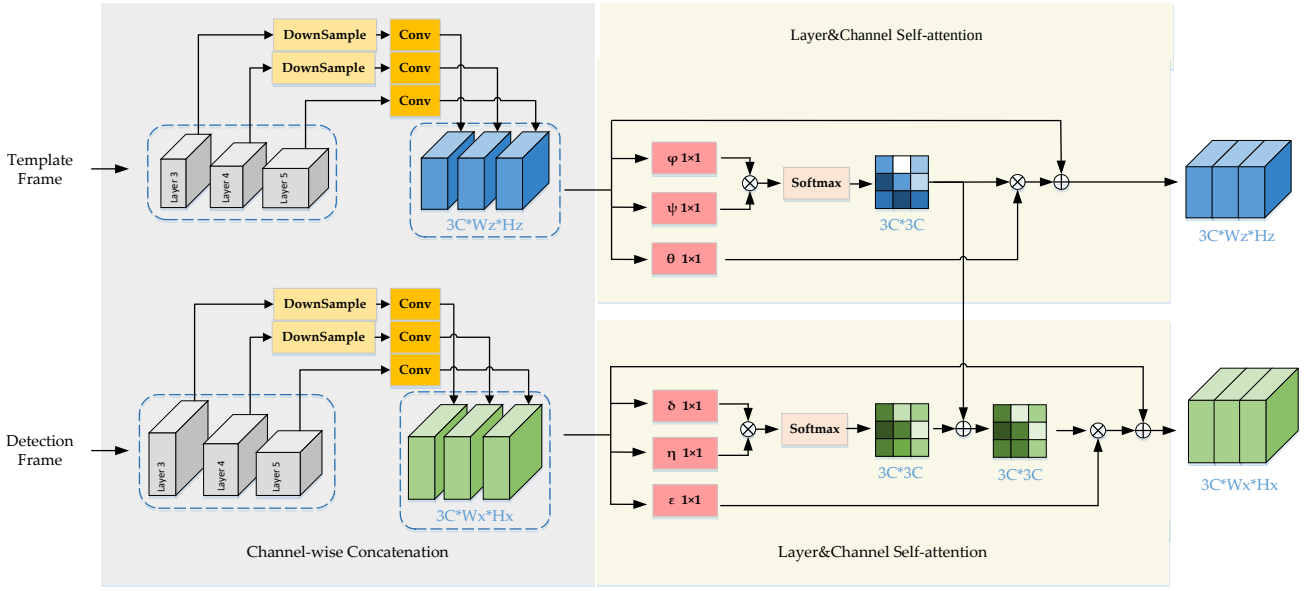

Figure 2. Schematic diagram of the proposed DBA block. The module consists of two parts: Channel-wise Concatenation and Layer-Channel Self-attention.

Specifically, the DBA block takes the feature maps of multiple layers of the two branches as input and outputs modulation features by applying the channel attention mechanism. The whole process includes channel-wise concatenation and layer-channel self-attention calculation. Among them, channel-wise concatenation gathers channels of multiple layers, so the layer-channel self-attention calculation can simultaneously calculate the attention of channels of multiple layers only by using the channel attention mechanism. To enable the detection branch to maintain the ability to identify the target's features, we integrate the channel self-attention matrix of the target branch into the detection branch. Finally, the similarity measure obtained from the cross-correlation calculation is output to the subsequent region proposal network for classification and regression.

3.2.1. Channel-Wise Concatenation

We extracted the feature maps of the last three layers of the feature extraction network, with the three layers of two branches represented as $L_3(x)$, $L_4(x)$, $L_5(x)$, $L_3(z)$, $L_4(z)$, and $L_5(z)$. We perform channel-wise concatenation on them:

$$L(x) = Cat(L_3(x), L_4(x), L_5(x)) \tag{3}$$

$$L(z) = Cat(L_3(z), L_4(z), L_5(z)) \tag{4}$$

To ensure that each layer of features has the same importance in attention calculation, $L_3(*)$, $L_4(*)$, and $L_5(*)$ are adjusted to the same thickness, respectively, include 256 channels. $L_3(*)$ and $L_4(*)$ are down-sampled to keep the same image size as $L_5(*)$.

Figure 2 shows the details of channel-wise concatenation and subsequent attention calculation.

3.2.2. Channel Self-Attention

We calculate the channel self-attention for the $L(z) \in R^{3C \times H_z \times W_z}$ and $L(x) \in R^{3C \times H_x \times W_x}$. For the target branch, we use three independent convolution layers with 1×1 kernels to generate the query feature Q_Z, the key feature K_Z, and the value feature V_Z. Then Q_Z, K_Z, and V_Z are stretched to form coded data $Q_Z{}'$, $K_Z{}'$, and $V_Z{}'$, in which Q_Z, K_Z, $V_Z \in R^{3C \times H_z \times W_z}$, $Q_Z{}'$, $K_Z{}'$, $V_Z{}' \in R^{3C \times N}$, and $N = H_z \times W_z$.

A channel self-attention matrix $A_{zsc} \in R^{3C \times 3C}$ is generated by calculating the similarity between the query feature Q_Z and the key feature K_Z. The specific method is to calculate matrix multiplication for $Q_Z{}'$ and $K_Z{}'$ and apply the softmax operation to it to get a similarity measure:

$$A_{zsc} = \text{softmax}(Q_Z{}' K_Z{}'^T) \tag{5}$$

We perform channel-wise concatenation on three layers, so A_{zsc} expresses the weight information of the channels of three layers. The channel self-attention matrix is used as weights to guide the rearrangement of the value feature V_Z, and then the new value feature is superimposed on the stretched inputs $Z \in R^{3C \times N}$ as residuals to form attention features $F_{zsc} \in R^{3C \times N}$:

$$F_{zsc} = \alpha A_{zsc} V' + Z \tag{6}$$

where α is a scalar parameter. At last, the attention feature F_{zsc} is rearranged to the original size 3C × Hz × Wz.

3.2.3. Information Compensation

In the channel self-attention mechanism mentioned above, we take the similarity between the query feature Q and the key feature K as the weight to guide the weighted summation of the value feature V. Q, K, and V are different from each other but come from the same input $L(*)$. This channel self-attention mechanism tends to enhance channels with stronger self-associations, namely the channels that remain steady when the background provides distractors. The branches of the Siamese framework share network parameters and structure, so their channels show one-to-one correspondence. Theoretically, their channel self-attention should also show a similar relationship. However, the calculations for the two branches are independent. For the template branch, the attention mechanism enhances the channel reflecting the essential attributes of the target because the template image only contains a single target. For the detection branch, the attention mechanism is dedicated to making the real target different from the background for efficient recognition, so it will enhance the channel that distinguishes the target from all the distractors, including similar objects. To avoid the detection branch losing focus on the target features in distinguishing background distractors from targets, we provide the target branch's channel self-attention matrix A_{zsc} as a clue to the detection branch, guiding the weighted superposition of V_x with the fusion of the channel self-attention maps of two branches. Specifical, the attention feature F_{xsc} of the detection branch is affected by the two branches' layer-channel self-attention A_{zsc} and A_{xsc}:

$$F_{xsc} = \alpha(A_{zsc} + A_{xsc}) V'/2 + X \tag{7}$$

where α is a scalar parameter and $X \in R^{3C \times N}$ is the stretched inputs of the detection branch.

Finally, we calculate the cross-correlation of the modulation features F_{zsc} and F_{xsc} of the two branches. The similarity measure $g(x, z)$ is input into the region proposal network for subsequent location.

$$g(x, z) = F_{zsc} \star F_{xsc} \tag{8}$$

4. Results

This section introduces the results of DBA-Siam on UAV123 and VOT benchmarks and compares them with advanced tracking algorithms. To prove the effectiveness of the proposed block, we conducted ablation experiments and analyzed the effects of the DBA block. Additionally, we visualize the tracking results and analyze the performance of DBA-Siam in some challenging video sequences.

4.1. Experimental Details

In the experiment, we use the pre-training parameters provided by SiamRPN to initialize DBA-Siam and freeze the first three backbone networks to train the parameters of the target branch of the DBA block. Then we freeze the target branch and train the detection branch on the selected challenging video sequences of the training data. Training data includes COCO [28], VID [29], and YouTube-VOS [30]. We tested DBA-Siam on an NVIDIA GTX 1660 with 6GB memory.

4.2. Algorithm Comparison

We compared the proposed tracker DBA-Siam with the advanced trackers on UAV123 and VOT benchmarks. The baseline tracker is SiamRPN [12].

4.2.1. UAV123 Benchmark

UAV123 [17] contains 123 image sequences that are collected by a low-altitude unmanned aerial vehicle (UAV). The targets in these images have a tiny size which makes them challenging for tracking. Accuracy and Precision are used to evaluate the performance of the tracker. Accuracy measures the success rate by calculating the ratio of video frames whose Intersection over Union (IoU) between the ground truth and the predicted bounding box exceeds a given threshold. The larger the Accuracy, the better the tracker. Precision measures the location precision by calculating the ratio of the video frames whose distance between the center of the ground truth and the center of the predicted bounding box exceeds a given threshold. The larger the Precision, the more accurate the tracker. Typically, the threshold distance is set to 20 pixels.

As shown in Table 1, among these advanced algorithms, DBA-Siam achieves top Accuracy. DBA-Siam achieves a Precision score of 0.792, which is close to the best score of DaSiamRPN, while outperforming the DaSiamRPN on Accuracy with a gain of 3%. Although the two methods have similar performance results, DBA-Siam needs no training negative instances from other fields, which saves training expenses. In addition, DBA-Siam surpasses the algorithm ARCF that is proposed for UAV images, with 28.5% and 18.2% improvement on Accuracy and Precision, respectively.

Table 1. Results on UAV123 benchmark. Red fonts indicate the top tracker. ↑ indicates that the larger the parameter value, the better the tracker performance.

	Accuracy ↑	**Precision** ↑
SiamFC [9]	0.447	0.681
ARCF [31]	0.470	0.670
STAPLE_CA [32]	0.425	0.597
STRCF [33]	0.457	0.627
CCOT [34]	0.409	0.659
CFNet [35]	0.428	0.680
SiamRPN [12]	0.527	0.748
ECO [36]	0.525	0.741
ECOhc [36]	0.472	0.660
DaSiamRPN [11]	0.586	0.796
Ours	0.604	0.792

4.2.2. VOT Benchmark

VOT2016 [18] and VOT2019 [19] are widely used visual object tracking benchmarks. Each tracking benchmark contains 60 sequences with various challenging factors, and part of the video sequences is identical across the two benchmarks. Each frame in the video sequences is calibrated with the bounding box of a single target object. VOT benchmark adopts Accuracy (A), Robustness (R), and Expected Average Overlap (EAO) as measurement standards. A evaluates the tracker based on the IoU of the ground truth and the predicted bounding box. The larger parameter indicates the higher accuracy of the tracker. R is used to evaluate the stability of trackers. The larger the R, the worse the stability of the tracker, the easier it is to lose the target. EAO takes the overlap rate of bounding boxes and the re-initialization time after the tracker loses the target into account. EAO evaluates the overall performance of trackers. The larger EAO parameter indicates a stronger integrated strength of the tracker.

We tested DBA-Siam on the VOT benchmark and compared it with the advanced trackers. As shown in Table 2, the proposed tracker achieves the best performance on A and EAO on VOT2016 and VOT2019. It shows that the proposed tracker has the highest accuracy for target positioning and the strongest overall performance.

Table 2. Results on VOT benchmark. Red fonts indicate the top tracker. The blank boxes mean that these algorithms did not announce the results. ↑ indicates that the larger the parameter value, the better the tracker performance. ↓ indicates that the smaller the parameter value, the better the tracker performance.

	VOT2016			**VOT2019**		
	A ↑	**R** ↓	**EAO** ↑	**A** ↑	**R** ↓	**EAO** ↑
Staple [37]	0.544	0.378	0.295			
DeepSRDCF [38]	0.528	0.326	0.276			
SiamFC [9]	0.532	0.461	0.235	0.477	0.687	0.204
SiamRPN [12]	0.560	0.260	0.344	0.517	0.552	0.224
DaSiamRPN [11]	0.610	0.220	0.411			
ECOhc [36]	0.540	1.190	0.322			
CCOT [34]	0.539	0.238	0.331	0.495	0.507	0.234
FlowTrack [39]	0.578	0.241	0.334			
MemTrack [24]	0.530	1.440	0.273			
SA-SIAM [15]	0.540	1.080	0.291	0.559	0.492	0.253
S_SiamFC	0.487	0.261	0.328	0.459	0.577	0.207
CSTNet [40]	0.571	0.219	0.349			
TADT [41]				0.516	0.677	0.207
Gasiamrpn [42]				0.548	0.522	0.247
CSRDCF [19]				0.496	0.632	0.201
Gasiamrpn [19]				0.548	0.522	0.247
SiamMsST [19]				0.575	0.552	0.252
Ours	0.634	0.242	0.415	0.587	0.602	0.261

Compared with the CSTNet and SA-Siam, DBA-Siam is inferior on R. It is because DBA-Siam uses the fixed target branch features for more efficient tracking, which leads to relocating less efficiently. Even so, our overall performance is better. Specifically, DBA-Siam outperforms SA-Siam on EAO with gains of 42.6% on VOT2016 and 3.5% on VOT2019 and outperforms CSTNet on EAO with a gain of 18.9% on VOT2016.

Thanks to the equipped DBA block, DBA-Siam surpasses the baseline SiamRPN, with 20.6% and 16.5% EAO improvement on VOT2016 and VOT2019.

4.3. Ablation Experiment

The DBA block is the core component of the proposed tracker DBA-Siam. To evaluate its effectiveness, we compared the version without this block with the version with this

block. We used ResNet50 as the backbone network to conduct ablation experiments on the DBA block. The two networks use the same pre-training parameters.

4.3.1. Data Experiments

To validate the effect of the proposed DBA block in dealing with the distractor problem, we show the number of frames that the two trackers lost targets on challenging sequences containing background distractors of the VOT dataset. We argue that the fewer lost frames indicate the stronger ability of trackers to handle background distractors.

Figure 3 shows the number of lost frames of two trackers and samples of image sequences with outstanding effects, in which the target is marked by a red bounding box. In these samples, similar background distractors exist near the target, such as players in sequence Basketball. The tracker with the DBA block loses fewer frames in these image sequences, which indicates that the DBA blocks have a positive effect on dealing with frame dropping caused by similar background distractors.

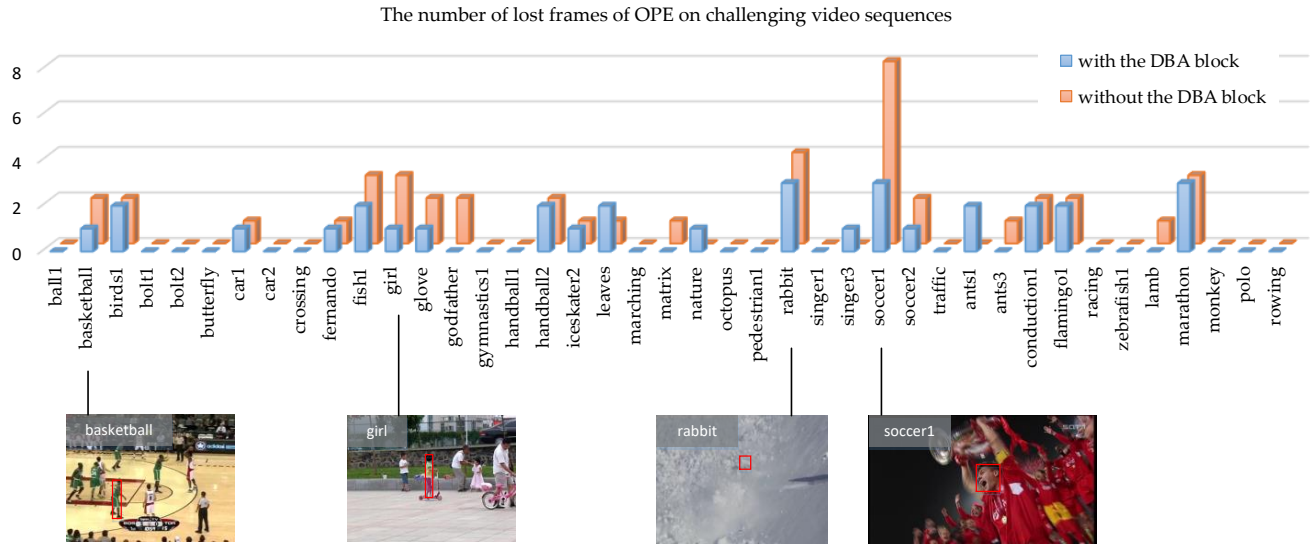

Figure 3. The number of lost frames of OPE on challenging video sequences.

Table 3 shows the comparison of evaluation results between the two trackers on three datasets, the tracker with the DBA block achieves superior tracking results on different datasets, which shows that the proposed block has a positive effect on improving tracking accuracy. Compared with VOT2016, our promotion effect on VOT2019 and UAV123 is not obvious because the proposed DBA block only aims at the background distractors problem and lacks the ability to solve other challenges in these data sets, such as tiny targets' recognition and relocation of the lost targets. Nevertheless, we still achieved a slight improvement in the effect. The detection speed shown in Table 3 is tested on a computer equipped with an NVIDIA GTX1660 graphics card. Although the proposed tracker achieved performance improvement, it increased the time cost. Nevertheless, it can still work at 30FPS in real-time.

Table 3. A comparison of evaluation results of DBA-Siam without DBA block and DBA-Siam with DBA block on multiple datasets (items with better parameters are marked in red). ↑ indicates that the larger the parameter value, the better the tracker performance. ↓ indicates that the smaller the parameter value, the better the tracker performance.

DBA Block	VOT2016-EAO ↑	VOT2019-EAO ↑	UAV123-Accuracy ↑	UAV123-Precision ↑	Speed FPS ↑
	0.344	0.260	0.582	0.772	31.6
√	0.415	0.261	0.604	0.792	30.93

4.3.2. Visualization Experiments

We selected five image samples prone to low tracking accuracy from five challenging video sequences for comparative experiments. Similar background distractors in these samples interfere with the target feature representation. Figure 4 shows the predicted bounding box and the response heat map of the compared trackers. Figure 4a–e shows the search inputs of the samples, in which the red box indicates the ground truth, the blue box indicates the prediction result of the tracker without the DBA block, and the green box indicates the prediction result of the tracker with the DBA block. Figure 4a.1–e.1 are the response heat maps of the tracker without the DBA block. Figure 4a.2–e.2 are the response heat maps of the tracker with the DBA block. The response heat maps indicate the similarity between the target and the search input. The larger the value of the point in the heat map, the more similar the target and search input are.

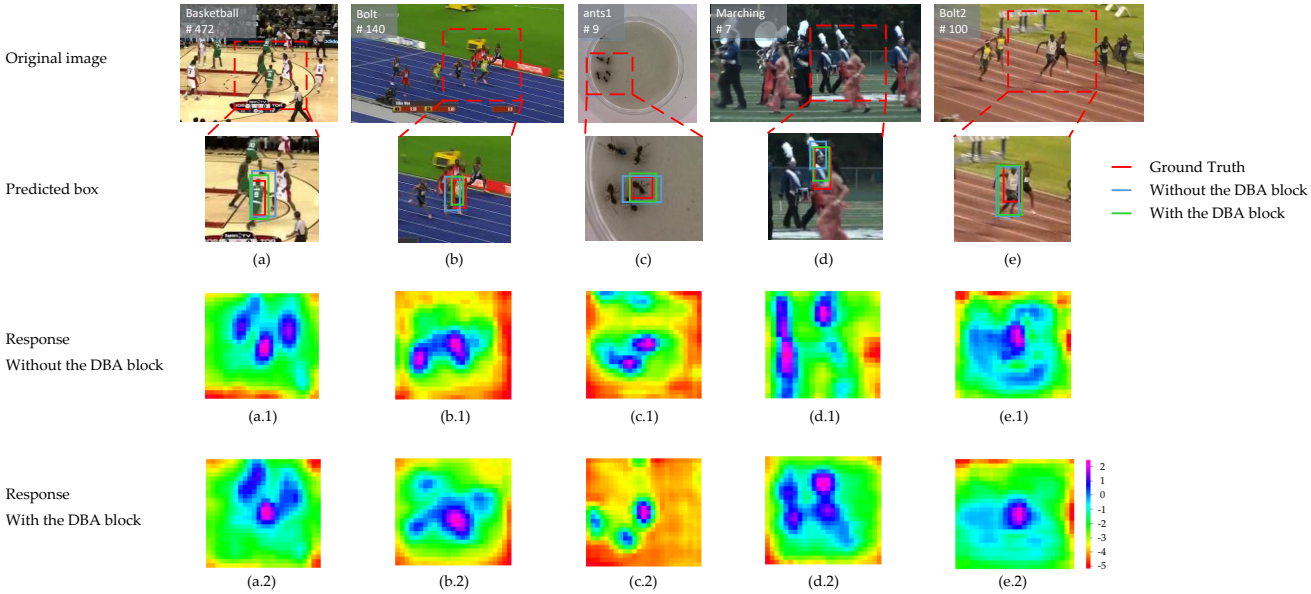

Figure 4. Response heat map of DBA-Siam without the DBA block and DBA-Siam with the DBA block in five samples. The red bounding box represents the ground truth, the blue and green bounding box respectively represents the prediction result of DBA-Siam without the DBA block and DBA-Siam with the DBA block. (**a**–**e**) are the search inputs of the five samples. (**a.1**–**e.1**) are the response heat maps of the tracker without the DBA block. (**a.2**–**e.2**) are the response heat maps of the tracker with the DBA block.

The prediction box in Figure 4a–e shows that both trackers can identify the correct target, while the predicted bounding box of the tracker with the DBA block has a larger overlap area with the ground truth, thus positioning the target more accurately. The response heat maps can reflect the reason for this result. For example, in the search input Figure 4c,c.1 contains two purple areas, which shows that the tracker without the DBA block has identified two proposal regions similar to the target ant, while Figure 4c.2 contains one purple area and less blue area, which shows that the tracker with the DBA block can distinguish the target ant from similar background ants. It can also be seen from the prediction results in Figure 4c that the target location output by the tracker without the DBA block contains the real target and a similar object, while the tracker with the DBA block can locate the real target more accurately.

5. Discussion and Conclusions

To solve the problem of background distractors, we aggregate features that may produce differences and selectively enhance features based on the current tracking task to help distinguish the target from the distractors. For trackers of the Siamese framework, we

designed the DBA block for adaptive feature fusion. It selectively enhances the features of channels of multiple layers of the two branches through attention computation. For unrestricted tracking tasks, the DBA block is adaptive to adjust the features of different channels of layers according to the current tracking target and scene. In addition, we explore how to combine the proposed DBA block with the Siamese framework to achieve end-to-end tracking. Experiments on the UAV123, VOT2016, and VOT2019 show that compared with the baseline tracker SiamRPN, DBA-Siam has superior tracking accuracy, which shows that the DBA block has a positive effect on improving the performance of the Siamese tracker.

Although the proposed block can enhance the ability of feature representation, it is limited by the fixed template input, which makes the attention mechanism not fully released. How to get richer template input and combine it with the DBA block needs further research.

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

Funding: This research was funded by [China National Key R&D Program during the 13th Five-year Plan Period], grant number 2017YFC0109901.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Lee, K.-H.; Hwang, J.-N. On-Road Pedestrian Tracking across Multiple Driving Recorders. *IEEE Trans. Multimed.* **2015**, *17*, 1429–1438. [CrossRef]
2. Li-Tian, Z.; Meng-Yin, F.; Yi, Y.; Mei-Ling, W. A framework of traffic lights detection, tracking and recognition based on motion models. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 2298–2303. [CrossRef]
3. Jia, T.; Taylor, Z.A.; Chen, X. Computerized Medical Imaging and Graphics Long term and robust 6DoF motion tracking for highly dynamic stereo endoscopy videos. *Comput. Med. Imaging Graph.* **2021**, *94*, 101995. [CrossRef] [PubMed]
4. Liu, L.; Xing, J.; Ai, H.; Ruan, X. Hand Posture Recognition Using Finger Geometric Feature. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 565–568.
5. Chang, C.-Y.; Lie, H.W. Real-Time Visual Tracking and Measurement to Control Fast Dynamics of Overhead Cranes. *IEEE Trans. Ind. Electron.* **2011**, *59*, 1640–1649. [CrossRef]
6. Nam, H.; Han, B. Learning multi-domain convolutional neural networks for visual tracking. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4293–4302.
7. Wang, L.; Ouyang, W.; Wang, X.; Lu, H. STCT: Sequentially Training Convolutional Networks for Visual Tracking. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1373–1381. [CrossRef]
8. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple People Tracking by Lifted Multicut and Person Re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 850–865.
9. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 850–865.
10. Guo, Q.; Feng, W.; Zhou, C.; Huang, R.; Wan, L.; Wang, S. Learning Dynamic Siamese Network for Visual Object Tracking. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1781–1789. [CrossRef]
11. 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.

12. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980. [CrossRef]
13. Zhang, Z.; Peng, H. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4586–4595. [CrossRef]
14. Peng, J.; Jiang, Z.; Gu, Y.; Wu, Y.; Wang, Y.; Tai, Y.; Wang, C.; Lin, W. SiamRCR: Reciprocal Classification and Regression for Visual Object Tracking. *arXiv* **2021**, arXiv:2105.11237. [CrossRef]
15. He, A.; Luo, C.; Tian, X.; Zeng, W. A Twofold Siamese Network for Real-Time Object Tracking. *arXiv* **2018**, arXiv:1802.08817. [CrossRef]
16. Wang, G.; Luo, C.; Xiong, Z.; Zeng, W. SPM-Tracker: Series-Parallel Matching for Real-Time Visual Object Tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3638–3647. [CrossRef]
17. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 445–461.
18. Kristan, M.; Fern, G.; Gupta, A.; Petrosino, A. The Visual Object Tracking VOT2017 challenge results. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2016; ISBN 9783319488813.
19. Kristan, M.; Fern, G. The Seventh Visual Object Tracking VOT2019 Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 2206–2241. [CrossRef]
20. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
21. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H.S. Fast Online Object Tracking and Segmentation: A Unifying Approach. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338. [CrossRef]
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Processing Syst.* **2017**, *30*, 5998–6008.
23. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. Available online: https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html (accessed on 24 December 2021).
24. Yang, T.; Chan, A.B. Learning Dynamic Memory Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 153–169. [CrossRef]
25. Abdelpakey, M.H.; Shehata, M.S.; Mohamed, M.M. DensSiam: End-to-End Densely-Siamese Network with Self-Attention Model for Object Tracking. In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 19–21 November 2018; Springer: Cham, Switzerland, 2018; pp. 463–473. [CrossRef]
26. Ni, Z.-L.; Bian, G.-B.; Xie, X.-L.; Hou, Z.-G.; Zhou, X.-H.; Zhou, Y.-J. RASNet: Segmentation for Tracking Surgical Instruments in Surgical Videos Using Refined Attention Segmentation Network. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 5735–5738. [CrossRef]
27. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6667–6676. [CrossRef]
28. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014, Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755.
29. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Jan, C.V.; Krause, J.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
30. Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Sep, C.V. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *European Conference on Computer Vision, Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 603–619.
31. Huang, Z.; Fu, C.; Li, Y.; Lin, F.; Lu, P. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 2891–2900. [CrossRef]
32. Mueller, M.; Smith, N.; Ghanem, B. Context-Aware Correlation Filter Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1387–1395.
33. Li, F.; Tian, C.; Zuo, W.; Zhang, L.; Yang, M.-H. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

34. Danelljan, M.; Robinson, A.; Shahbaz Khan, F.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 472–488.
35. Valmadre, J.; Bertinetto, L.; Henriques, J.; Vedaldi, A.; Torr, P.H.S. End-to-End Representation Learning for Correlation Filter Based Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5000–5008. [CrossRef]
36. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939. [CrossRef]
37. 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. [CrossRef]
38. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Learning Spatially Regularized Correlation Filters for Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318. [CrossRef]
39. Zhu, Z.; Wu, W.; Zou, W.; Yan, J. End-to-End Flow Correlation Tracking with Spatial-Temporal Attention. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 548–557. [CrossRef]
40. Yao, S.; Zhang, H.; Ren, W.; Ma, C.; Han, X.; Cao, X. Robust Online Tracking via Contrastive Spatio-Temporal Aware Network. *IEEE Trans. Image Process.* **2021**, *30*, 1989–2002. [CrossRef]
41. Li, X.; Ma, C.; Wu, B.; He, Z.; Yang, M.-H. Target-Aware Deep Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 18–20 June 2019; pp. 1369–1378. [CrossRef]
42. Kristan, M. The Sixth Visual Object Tracking VOT2018 Challenge Results. In Proceedings of the European Conference on Computer Vision (ECCV) Workshop, Munich, Germany, 8–14 September 2018; ISBN 9783030110093.