mathematics

MDPI

Article

Lightweight Target-Aware Attention Learning Network-Based Target Tracking Method

Yanchun Zhao [1,†], Jiapeng Zhang [2], Rui Duan [2], Fusheng Li [1,*,†] and Huanlong Zhang [2]

[1] School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; yczhao@uestc.edu.cn
[2] School of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 331901050046@zzuli.edu.cn (J.Z.); 331901060053@zzuli.edu.cn (R.D.); hlzhang@zzuli.edu.cn (H.Z.)
* Correspondence: lifusheng@uestc.edu.cn
† These authors contributed equally to this work.

Abstract: Siamese network trackers based on pre-trained depth features have achieved good performance in recent years. However, the pre-trained depth features are trained in advance on large-scale datasets, which contain feature information of a large number of objects. There may be a pair of interference and redundant information for a single tracking target. To learn a more accurate target feature information, this paper proposes a lightweight target-aware attention learning network to learn the most effective channel features of the target online. The lightweight network uses a designed attention learning loss function to learn a series of channel features with weights online with no complex parameters. Compared with the pre-trained features, the channel features with weights can represent the target more accurately. Finally, the lightweight target-aware attention learning network is unified into a Siamese tracking network framework to implement target tracking effectively. Experiments on several datasets demonstrate that the tracker proposed in this paper has good performance.

Keywords: target features; siamese trackers; lightweight network; target tracking

MSC: 68T45

Citation: Zhao, Y.; Zhang, J.; Duan, R.; Li, F.; Zhang, H. Lightweight Target-Aware Attention Learning Network-Based Target Tracking Method. *Mathematics* **2022**, *10*, 2299. https://doi.org/10.3390/math10132299

Academic Editor: Tao Zhou

Received: 12 May 2022
Accepted: 26 June 2022
Published: 30 June 2022

1. Introduction

Visual target tracking is a branch in the field of computer vision, and thanks to the development of deep learning techniques, especially the application of neural networks [1], target tracking has entered a new phase. In the target tracking task, the target being tracked is arbitrary, and the traditional trackers designed based on manual features [2] perform generally in target modeling. Thanks to the powerful generalization ability of depth features, which can model all kinds of targets well, depth feature-based trackers [3–5] have achieved excellent results in recent years.

Although the existing depth feature-based trackers perform well, we find that the pre-trained depth features still have some interference when modeling arbitrary targets. This is because, firstly, the targets being tracked are arbitrary, and if the dataset used to train the depth feature model does not contain such targets, that is, the depth feature model has not learned information about such targets, then when extracting the target features, it can only rely on the existing information for speculation, which often brings a lot of uncertainties and leads to more disturbances in the model. Secondly, even if the deep feature model has learned such targets, and when the general tracker uses the last layer or layers to extract the target features, it will lead to more disturbing factors in the feature model because of the huge amount of data. Finally, the existing pre-trained deep feature models are created mainly for the target recognition task, where its main task is to identify

all similar targets that appear in each frame. The target tracking task, on the other hand, is different and is to identify the same target in subsequent frames, so the tracker based on pre-trained features may be wrong in the face of interference from similar targets in the same frame.

Some trackers use the designed lightweight network as the memory module and use the target appearance information in each frame to update network parameters, to achieve good appearance memory performance. In this paper, a lightweight target-aware attention learning network is designed to learn the most effective channel features of the target online, using the target information in the first frame template to learn a series of channel features with weights, and by recombining these channel features. A compact and effective deep feature is obtained, which can better distinguish the object from the background compared to the pre-trained features. At the same time, a new attention learning loss function is developed to optimize the training of the proposed network using the Adam optimization method. Different from other methods, the lightweight network designed in this paper does not require complex parameters and is easy to implement. It only needs to learn the most salient features through the reliable information of the first frame of the target and does not need to use too much memory temporarily, which is beneficial for the efficient use of hardware resources. Finally, the lightweight target-aware attention learning network is unified into the Siamese tracking network framework to effectively achieve target tracking. Figure 1 shows that our tracker yields better tracking performance when compared with other trackers.

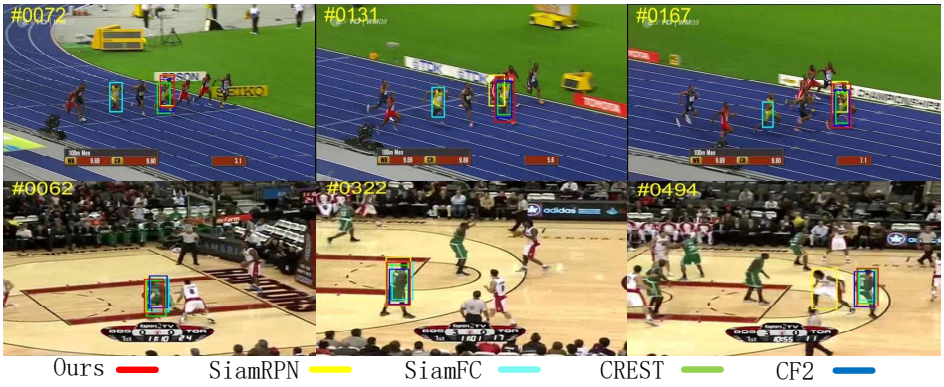

Figure 1. Comparison of our tracker with other trackers for Bolt (**top**), Basketball (**bottom**).

The main contributions of this article are described in summary as follows:

(1) A lightweight target-aware attention learning network is designed to learn the most effective channel features of the target online. The new network mines the expressiveness of different channels to the target by the first frame template.
(2) A new attention learning loss function is developed to optimize the training of the proposed network using the Adam optimization method. The loss function effectively improves the modeling capability and tracking accuracy of the network by introducing the gradient information during training.
(3) The lightweight target-aware attention learning network is unified into the Siamese tracking network framework to effectively achieve target tracking. Moreover, the proposed method performs better against other trackers.

2. Related Work

There are a large number of researchers who have made many contributions in the field of visual tracking, and many excellent trackers have been proposed. In this section, we discuss some trackers that are similar to our work.

2.1. Lightweight Network-Based Tracker

Real-time target tracking is a very relevant research element. However, when the tracking speed increases, the tracking accuracy is bound to be affected. Therefore, many researchers have researched how to increase the tracking speed without affecting the tracking accuracy. Zhao et al. [6] use a pruned convolutional neural network to construct the tracker, which is trained by a mutual learning method to further improve the localization accuracy. Cheng et al. [7] propose a real-time semantic segmentation method based on extended convolution smoothing and lightweight up-sampling on the basis of a lightweight network, which can achieve high segmentation accuracy while maintaining high-speed real-time performance. Zhao et al. [8] design a lightweight memory network, which only needs reliable target frame information to fine-tune network parameters online, so as to enhance the memory ability of the target appearance. At the same time, it can maintain good discriminant performance without a complicated update strategy. Unlike them, this paper designs a lightweight network for online learning of the most salient features of the target and achieves redundant feature channel trimming by back-propagating the weights to determine the importance of the feature channels.

2.2. Siamese Network-Based Tracker

In recent years, the combination of Siamese networks and target tracking has led target tracking to enter a new stage. Bertinetto et al. [9] propose a new structure of fully convolutional Siamese networks. In the initial offline phase, deep convolutional networks are regarded as a more general similarity learning problem, and then the simple online estimation of the problem during tracking can achieve very competitive performance, and the frame rate at runtime far exceeds the requirements of real-time performance. Li et al. [10] developed a model consisting of a Siamese network and a region proposal network, which discards the traditional multi-scale testing and online tracking, divides the network into template branches and detection branches, and uses a large amount of data for offline training to achieve a good tracking result. Gao et al. [11] propose a Siamese Attentional Key-point Network for target tracking, by designing a new Siamese lightweight hourglass network and a novel cross-attentional module to obtain more accurate target features, and propose a key-points detection approach to accurately locate target location and scale regression.

3. Proposed Method

3.1. Basic Siamese Network for Visual Tracking

Siamese networks are originally applied to template matching problems and are later introduced into object tracking. It is composed of two networks with the same structure and the same weight. These two networks are used to extract the depth feature of the target and the depth feature of the search area, and finally the cross-correlation calculation is used to find the highest response value in the search area. The position of this point is the final target position. Moreover, the whole process can be expressed by the following formula:

$$f(z, x) = \varphi(z) * \varphi(x) + b \cdot 1 \tag{1}$$

where z represents the initial frame position, x represents the position of the search region, $b \cdot 1$ denotes the deviation value, and * represents the convolution operation.

As shown in Figure 2, the proposed tracker contains a pre-trained feature extraction network, a lightweight target-aware attention learning network, and a Siamese network matching module. The VGG feature extraction network is a very deep convolutional network for image classification and achieves the state-of-the-art performance on the ImageNet challenge dataset. It is trained offline in this paper, and the proposed lightweight target-aware attention learning network is trained online by using the given first frame target information, and then the cross-correlation operation of the Siamese network is used to locate target. The attention learning loss function used to train the lightweight target-aware attention learning network is redesigned on the basis of the MSE loss function, and

the Adam optimization method is used for training, and the feature channel is determined according to the gradient value information of back propagation. The importance weight is weighted to the original depth feature to represent the target, and finally the template matching method of the Siamese network is used to locate the target. The calculation process is shown in Formula (2):

$$f_{new}(z, x) = (\varphi(z) \odot \alpha) * \varphi(x) + b \cdot 1 \tag{2}$$

where z denotes the template image, x denotes the image of the search region, $b \cdot 1$ denotes the deviation value of each, α is the channel attention weight vector of the feature channel, $\odot$ denotes the Hadamard product, $*$ denotes the convolution operation, and $f_{new}(z, x)$ denotes the response score.

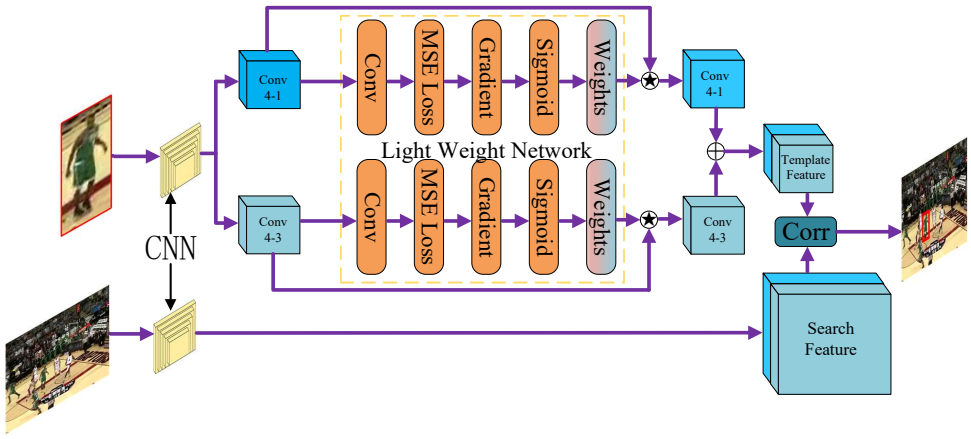

Figure 2. Overview of our network architecture for visual tracking.

3.2. Attentional Learning Loss Function

Most of the trackers based on correlation filtering use recurrent samples to train regression models, while Chen et al. [12] propose to use single-layer convolution to solve the linear regression problem and use the gradient-descent training method to solve the regression problem in target tracking, which this paper is inspired by. In the linear regression model of the work [12], the objective is to learn a linear function using the training samples $X \in R^{m \times n}$ and the corresponding regression objective $Y \in R^m$. Each element x_i in each row of the model X represents a training sample with feature dimensionality and the corresponding regression target x_i is the first element of the model Y. Then, the objective is to learn the coefficients w of a regression function by minimizing the objective function $f(x) = w^T \cdot x$ during the offline training process.

$$\arg\min_{w} \|X * w - Y\|^2 + \lambda \|w\|^2 \tag{3}$$

In Equation (3), $\|\cdot\|$ is the Euclidean parametrization, and λ is the regularization parameter to prevent overfitting.

The gradient values generated during the training of neural networks can be a good indication of the channel saliency feature information for different target classes [13], and this paper attempts to introduce this idea into a Siamese network-based tracker used for training to generate a set of weights that can represent the contribution of different feature channels to modeling, to enhance the target modeling capability of pre-trained depth features. To this end, this paper redefines its input based on Equation (3), which can be expressed by minimizing the following function:

$$\arg\min_{w} \sum_i \left((Z_i \cdot w_i') * X_i - Y_i \right)^2 + \lambda' \sum_i w_i'^2 \tag{4}$$

where $\cdot$ is the dot product operation, $*$ denotes the convolution operation, Z is the template depth feature, X is the search area depth feature; they are obtained from the same frame, and Z is located at the center of the X, λ' is regularization parameter, w' is the regression weight vector obtained by the network training, the dimension is the same as Z and X.

The comparison results of the target response maps are shown in Figure 3. Figure 3a shows the weighted features of the feature channels after learning using the attention learning loss function, and Figure 3b shows the target-specific diagnosis extracted directly using the original features.

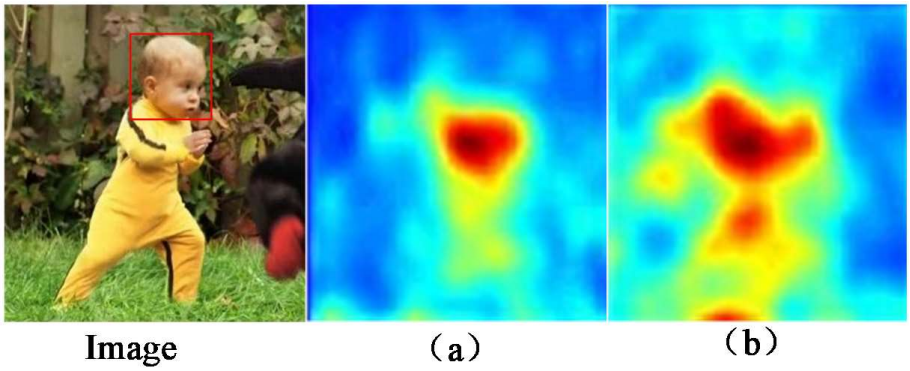

Figure 3. Comparison of the before and after learning characteristics of attentional learning loss.

Finally, the regression weights w' are mapped by the sigmoid function to obtain the channel weights corresponding to the sample images.

$$\alpha_i = {}^{1}\Big/\left(1 + e^{-w_i'}\right) \tag{5}$$

where α_i denotes the i-th value in α, and $\alpha \in [0, 1]$, w_i' denotes the i-th value in w'.

In summary, the loss function generates the gradient information by training the target information in the first frame. The gradient information is used to generate the weights of the different channels of the feature to the target information expression. The feature channel is determined according to the gradient value information of back propagation under the attentional learning loss function. The importance weight is weighted to the original depth feature to represent the target. Finally, the template matching method of the Siamese network is used to locate the target. However, the loss function is used under the assumption that the error between the model output and the groundtruth value obeys a Gaussian distribution. When this condition is not satisfied, the loss function is limited in its usefulness.

3.3. Lightweight Target-Aware Attention Learning Network

In a pre-trained deep model-based classification network, each feature channel contains a specific target feature pattern, and all feature channels together construct a feature space containing a priori information about different objects. The pre-trained network identifies object classes mainly through a subset of these feature channels, so the importance of each channel should not be calculated equally when used to track the target representation.

As shown in Figure 4, the lightweight target-aware attention learning network proposed in this paper is built on a single-layer convolutional network, which is used in the same way as a general neural network, and its kernel is set to match the size of the target template. However, to obtain better object appearance features, the lightweight target-aware attention learning network proposed in this paper only uses the given first frame object information for training and does not require complex offline training, while using the more advanced Adam Optimization method to obtain network parameters.

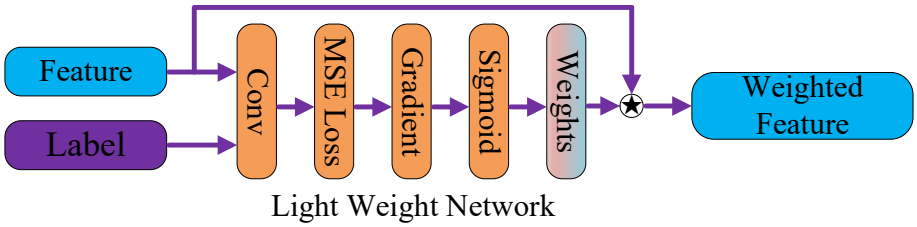

Figure 4. Lightweight target-aware attention learning network.

(1) Parameter learning process.

A search area of size X is intercepted around the given first frame target as an initial training sample, w_i' is a set of initial target feature channel weights with an initial value of 1. In the subsequent learning process, the gradient value information is calculated to update its value online according to the difference between the response values and labels of different channels. The larger the gradient value is, the smaller the contribution of the feature channel to the target model. Equation (4) is used to guide the online learning process, and the Adam optimization method is used to optimize the network by empirically setting the learning rate to, the momentum to 0.9, the weight decay to 1000, and the maximum number of iterations to 100. Compared with the traditional gradient descent (SGD) optimization method, the Adam optimization method is an improvement and extension of it, with high computational efficiency and small memory occupation. Moreover, the learning rate of the SGD optimization method is fixed, while the Adam optimization method can update the learning rate of the third training process adaptively based on the average of the first two training weights, which can improve the performance of the network on sparse gradient problems.

(2) Obvious characteristic of the lightweight target-aware attention learning network.

The network designed in this paper is implemented on a single-layer convolutional network, which learns the optimal representation of the target appearance by adjusting a certain number of feature channel weights through simple single-layer convolutional operations, using the proposed attention learning loss function to learn online, thus generating an optimal set of channel modeling parameters. This approach is computationally simple, does not require complex model computation strategies, does not take up too many valuable memory resources, and is easy to implement. Moreover, the number of parameters in the network is small, which facilitates fast computation and achieves real-time fast online tracking.

4. Experiment and Analysis

Our tracker is implemented on a PC with an i7-9700 3.0 GHz and a single NVIDIA GeForce RTX 2060 GPU with Pytorch. The algorithm proposed in this chapter uses the VGG-16 [14] neural network as the feature extraction network for the target and the search region, and the outputs of the Conv4-1 and Conv4-3 layers are used for target appearance modeling. The number of channel dimensions of the outputs is 512. Then the feature passes through the lightweight network and its feature channels are given different weights, and the number of channels is reduced to 380. Moreover, the kernel of the lightweight target-aware attention learning network is set to match the size of the target template. For the designed lightweight target-aware attention learning network, online training is performed using the attention learning loss function only in the first frame of each video sequence, setting the maximum number of iterations to 100, the momentum setting to 0.9, and the convergence loss threshold to 0.01. To handle scale variations, we also search for the object over three scales (0.957, 1, 1.047), and update the scales by scale weights (0.99, 1, 1.005). To evaluate the performance of the proposed algorithm, this section is tested on the OTB-50 [15] and OTB-100 [16] dataset, TC-128 [17] dataset, UAV123 [18] dataset set, VOT2016 [19] dataset and LaSOT dataset [20].

4.1. Ablation Studies

To better explain the validity of the proposed method, the ablation experiment of this work is analyzed on the OTB-100 dataset using one-pass evaluation. Our algorithm contains the base Siamese-based tracker and the proposed lightweight target-aware attention network. Figure 5 shows the precision and success rate of baseline without the proposed attention network and our method.

From Figure 5, we can see that when the proposed attention network is added, the accuracy and success rate of the tracking algorithm are improved. The network removes redundant and partial background information from the features to achieve superior tracking performance by online mining of different channels of the target depth features for their ability to represent the target information. The experimental results in Figure 5 show that the proposed attention network contributes to the performance of the tracking algorithm.

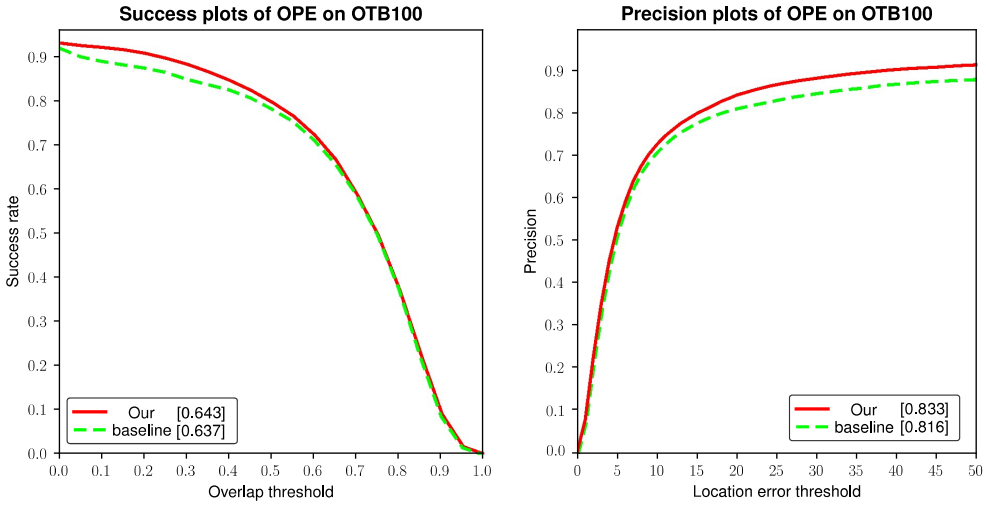

Figure 5. The ablation studies on the OTB-100 dataset.

4.2. OTB Dataset Experiments

In this paper, experiments are conducted on the popular OTB-50 and OTB-100 datasets in the field of target tracking, which consist of 50 and 100 fully annotated videos, respectively. In this paper, the accuracy maps in one-pass evaluation (OPE) are used to evaluate different trackers and are compared with 10 advanced trackers SiamFC, attention-based trackers MemTrack [21] and MemDTC [22], correlation filter-based trackers KCF [23], Staple [24], DSST [25] and SRDCF [26], deep learning and correlation filter-based tracker CF2 [27], CREST [28], and CSR-DCF [29] were compared for the results. As shown in Figures 6 and 7, the performance of the proposed tracker (Ours1) in this chapter is at the advanced level in both benchmark tests. Specifically, the proposed algorithm obtained success rate scores of 0.655 and 0.643 on OTB-50 and OTB-100, respectively, and the proposed algorithm gained 4.6% and 6.0% improvement over the Siamese network-based tracking method SiamFC, which confirms the advantages of the lightweight target-aware attention learning network and attention learning loss function proposed in this paper. CF2 algorithm uses the depth features of three layers in the VGG-16 network for target modeling to improve the discriminative power of the model, and obtains success rate scores of 0.603 and 0.562 for OTB-50 and OTB-100, respectively, and the performance of the proposed algorithm in this paper is 5.2% and 8.1% higher than that of the CF2 algorithm without using more depth features. The CREST algorithm achieves a higher success rate than the CF2 algorithm on the OTB-50 dataset and performs better than the algorithm proposed in this paper in terms of both success rate and accuracy; the reason for this is that the CREST algorithm introduces a residual network to extract the depth features of the target, and the residual network structure can be used to build a deeper network to

improve the accuracy of the features and alleviate the gradient disappearance problem caused by the deep network.

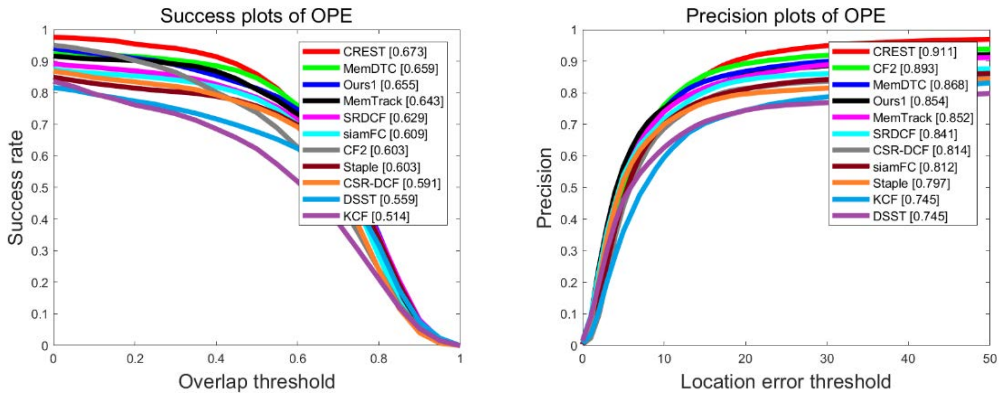

Figure 6. Success and precision rates on the OTB50 dataset.

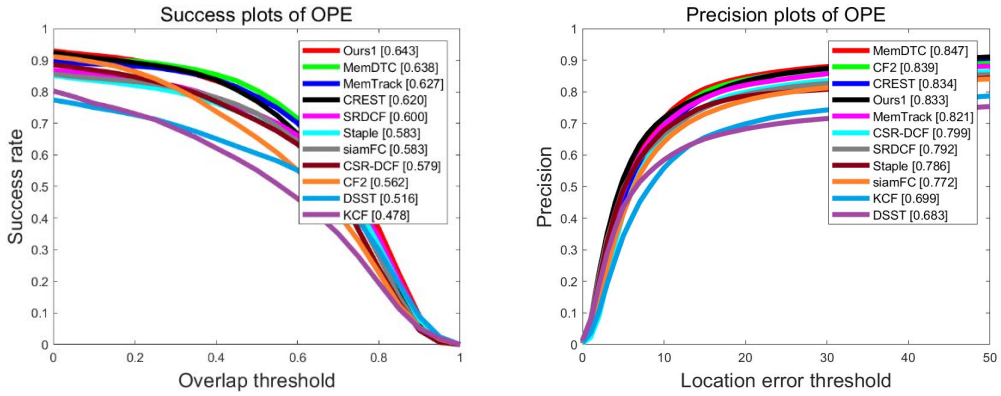

Figure 7. Success and precision rates on the OTB100 dataset.

For object-tracking algorithms, the real-time performance should also be used as one of the criteria for evaluating tracker performance. In Table 1, we compared the operational performance of some of the advanced trackers in terms of Precision score (%), Success rate (%), and Speed (FPS) on the OTB-100 dataset. Table 1 shows the results of our tracker compared with 7 advanced trackers including BaSiamIoU [30], ATOM [31], CFML [32], CREST [28], CSR-DCF [29], SRDCF [26], and SiamFC [9]. From Table 1, we can note that ATOM draws on the IoU-Net idea and proposes IoU modulation and IoU predictor to solve the scale challenge in the tracking process, achieving better tracking performance in terms of Precision score and Success rate. However, the speed performance of ATOM is not as satisfactory as our tracker. Meanwhile, although SiamFC is capable of reaching 102.3 FPS in speed, it is not able to adapt to changes in target appearance during tracking, resulting in lower tracking accuracy. Our tracker achieves 83.3% in Precision score and 64.3% in Success rate in 59 FPS. Overall, our tracker strikes a balance between Precision score, Success rate, and Speed. Therefore, for some scenes with higher requirements on tracking speed, SiamFC algorithm is a better choice, while for some scenarios where tracking accuracy is more preferred, ATOM algorithm should be chosen. Our method is more suitable for applications that require a certain degree of tracking accuracy and tracking speed.

Table 1. The real-time performance of the advanced trackers on the OTB-100 dataset. In the table, red, green and blue indicate the top three scores respectively.

Tracker	Precision Score (%)	Success Rate (%)	Speed (FPS)
Ours	83.3	64.3	59
BaSiamIoU	83.9	70.8	50
ATOM	87.9	66.7	30
CFML	85.3	64.9	32
SiamFC	77.2	58.3	102.3
CREST	83.4	62.0	1.8
CSR-DCF	79.9	57.9	8.5
SRDCF	79.2	60.0	4.2

(1) Challenge analysis of the OTB dataset

This part shows the success rate plots on the OTB-50 dataset for multiple challenge scenarios, as it contains 50 videos with relatively high tracking complexity in the OTB-100 dataset, which include: scale variation (SV), low resolution (LR), occlusion (OC), distortion (DF), motion blur (MB), fast motion (FM), in-plane rotation (IR), out-of-plane rotation (IR), out-of-field (OV), background clutter (BC), and illumination variation (IV).

More details of the performance of the proposed algorithm are shown in Figure 8. Overall, the proposed algorithm performs well in all 11 challenges. For the attributes of motion blur, distortion, and low resolution, the proposed algorithm outperforms the tracker SiamFC, which is also based on Siamese networks. The SiamRPN algorithm combines Siamese networks and region proposal network and has good tracking precision and speed, but the algorithm proposed in this paper has better performance under the background clutter challenge, indicating that the algorithm in this paper can extract the key features of the target. For exceeding the visual field, the proposed algorithm performs much better than the other nine compared trackers, which is attributed to the proposed lightweight target-aware attention learning network model and the attention learning loss function. In addition, the proposed algorithm performs better than most neural network-based trackers under the background clutter challenge, which indicates that the proposed lightweight target-aware attention learning network and attention learning loss function can effectively modify the pre-trained depth features to remove redundant information while enhancing the feature channels that are more important to the target representation, and thus it can improve the feature representation of the target. Overall, the proposed algorithm in this paper achieves good performance under several challenging attributes of the OTB-50 dataset.

(2) Qualitative experimental analysis of the OTB dataset

To qualitatively evaluate the proposed method, Figure 8 shows some tracking results of the proposed algorithm and other tracker on eleven challenging video sequences. SiamRPN is a deep learning-based algorithm, CF2 is a correlation filtering-based algorithm, where the SiamRPN algorithm also introduces region suggestion networks into the tracking, and SiamFC is a Siamese network-based algorithm, similar to the proposed algorithm in this paper. the proposed algorithm is similar.

In these six video sequences, there are many different challenges, including deformation (Bird1, MotorRolling, Skiing), occlusion (Soccer, Tiger), out-of-field (Bird1, Soccer), and background clutter (Football1, MotorRolling). SiamFC and the proposed algorithm can re-find the target after its occlusion disappears, while other trackers are unable to locate the target again due to untrustworthy samples introduced during model updates. CF2 and CREST drift rapidly in scenes where the target is out of view, and SiamFC and CF2 are unable to adapt to the challenge of scale changes in Bird1 and MotorRolling sequences. As the tracking task progresses, CREST, CF2, and SiamFC all lose targets one-by-one as the tracking drifts. In contrast, the algorithm proposed in this paper can adapt well to these challenges due to the introduction of a lightweight target-aware attention learning network

and an attention learning loss function to learn the channel weight information of the target. As in these scenarios in Figure 9, the performance of the proposed algorithm is significantly better than other trackers.

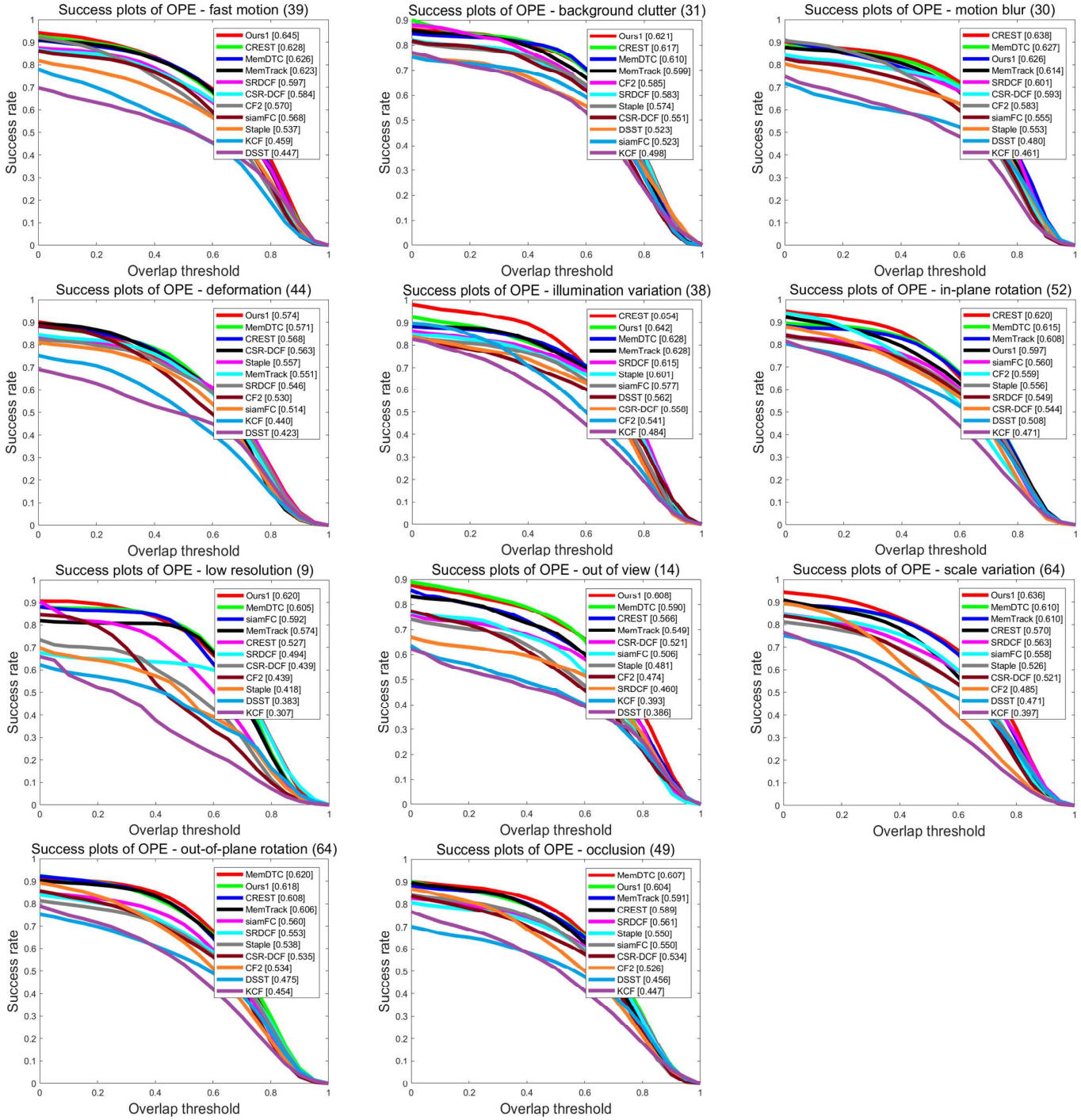

Figure 8. Comparison of 11 attribute challenge results.

Figure 9. Visualization of tracking results for focused challenge scenarios.

4.3. TC-128 Dataset Experiments

In this paper, the proposed method is evaluated on the Temple-Color (TC-128) dataset containing 128 videos. The evaluation method follows the guidelines in the OTB dataset and uses the accuracy plots in the one-time evaluation method (OPE) to compare the different trackers.

(1) Quantitative evaluation on TC-128 dataset: The proposed algorithm is compared quantitatively with 10 other trackers, including ECO [33], CREST [28], HCFTstar [34], CF2 [27], CACF [35], KCF [23], DSST [25], LOT [36], and CSK [37].

As shown in Figure 10, the proposed algorithm is in the top two positions among all trackers in terms of accuracy and success rate. Compared with the CF2 algorithm based on deep learning, the proposed algorithm achieves a higher success rate of 5.0% on TC-128, probably because CF2 uses unprocessed pre-trained deep features, while the proposed algorithm learns the most effective target channel weights through the designed lightweight target-aware attention learning network, so that the features better represent the appearance of the target. Moreover, the success rate of the proposed algorithm on TC-128 is 1.2% higher than that of CREST which learns linear regression on a single-layer convolutional network. It can also be seen that the CREST algorithm, which uses only one layer of depth features for target modeling, outperforms the CF2 algorithm, which uses multiple layers of depth features, which illustrates the great advantage of linear regression modeling on the network. The tracking robustness of the proposed algorithm is greater than that of the tracker CACF, which introduces contextual information. It can also be seen from the figure that trackers that use manual features to model targets such as KCF have significantly lower performance than other trackers that use depth features. The ECO algorithm combines color features and depth features to represent the target, and is sensitive to the color features of the target, so the performance on the TC-128 dataset designed for color features is better than the algorithm proposed in this paper. (2) Challenge analysis of TC-128 dataset: In this section, the success rate of the tracker associated with the work in this paper is tested on the TC-128 dataset for 11 challenging videos, including scale variation (SV), low-resolution (LR), occlusion (OC), distortion (DF), motion blur

(MB), fast motion (FM), in-plane rotation (IR), out-of-plane rotation (IR), out-of-field (OV), background clutter (BC), illumination variation (IV).

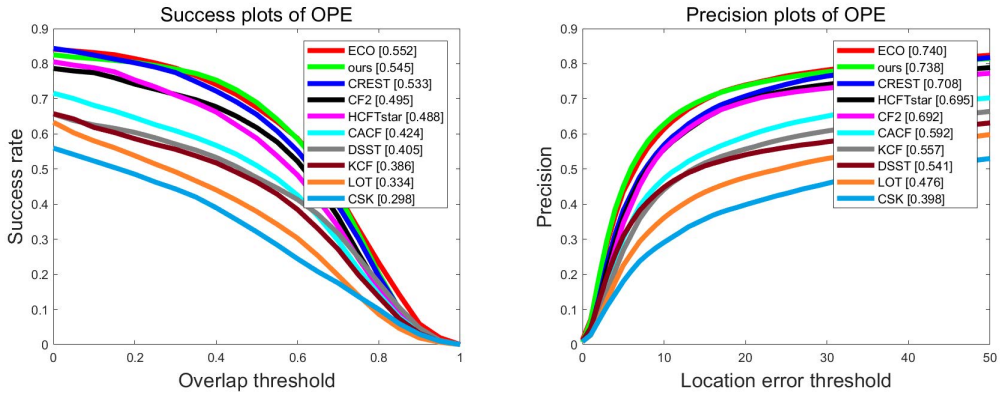

Figure 10. Success and precision rates on the TC-128 dataset.

Figure 11 shows the results of the proposed algorithm and other state-of-the-art trackers under 11 attribute challenges, and it is clear that the proposed algorithm outperforms the other trackers in overall performance. Thanks to the channel weight learning effect of the lightweight target-aware attention learning network, the proposed algorithm outperforms other trackers in the case of background clutter, motion blur, and deformation. ECO outperforms the proposed algorithm in deformation challenge scenarios due to the use of multi-feature fusion, but the proposed algorithm outperforms other trackers in several challenge scenarios with background clutter, motion blur, and out-of-field. In these scenarios, the targets often experience severe appearance changes or complex background disturbances, so the compared tracker experience tracking failures, while these compared tracker use sample update models that may contain noise, which prevents the tracker from obtaining an accurate model of the target appearance and leads to tracking failures. In contrast to these trackers, the lightweight target-aware attention learning network is introduced in this work to improve the modeling capability of depth features, allowing the tracker to adapt to target tracking tasks in complex scenes.

4.4. UAV123 Dataset Experiment

To further illustrate the performance of the proposed algorithm, the performance of the proposed algorithm is evaluated on the UAV (UAV123) dataset in this paper. Compared with typical visual object tracking datasets including OTB and TC-128, the UAV123 dataset provides low-altitude aerial video for target tracking. UAV123 is also one of the largest target tracking datasets, which contains 123 video sequences with over 110,000 images and an average sequence length of 915 frames. The UAV123 dataset has become increasingly popular due to real-life applications that are becoming increasingly popular, such as navigation, wildlife monitoring, crowd surveillance, etc. An algorithm that strikes a good balance between accuracy and real-time speed would be more practical for tracking these targets.

As shown in Figure 12, the proposed algorithm is tested on the UAV123 dataset in this paper to compare with 10 other trackers, including SRDCF [26], CREST [28], CF2 [27], SiamRPN [10], DSST [25], Struck [38], ECO [33], TADT [39], KCF [23], and CSK [37]. Thanks to the lightweight target-aware attention learning network introduced in the Siamese network framework, the proposed algorithm is higher than the TADT algorithm in terms of accuracy and success rate. Moreover, the success rate of the proposed algorithm on UAV123 is 5.8% higher than that of CREST which learns linear regression on a single convolutional layer. The performance of the CREST algorithm using only one layer of depth features outperforms that of CF2 and SRDCF using multiple layers of depth features. Trackers using manual features, such as DSST and KCF, have significantly lower performance than other trackers using depth features.

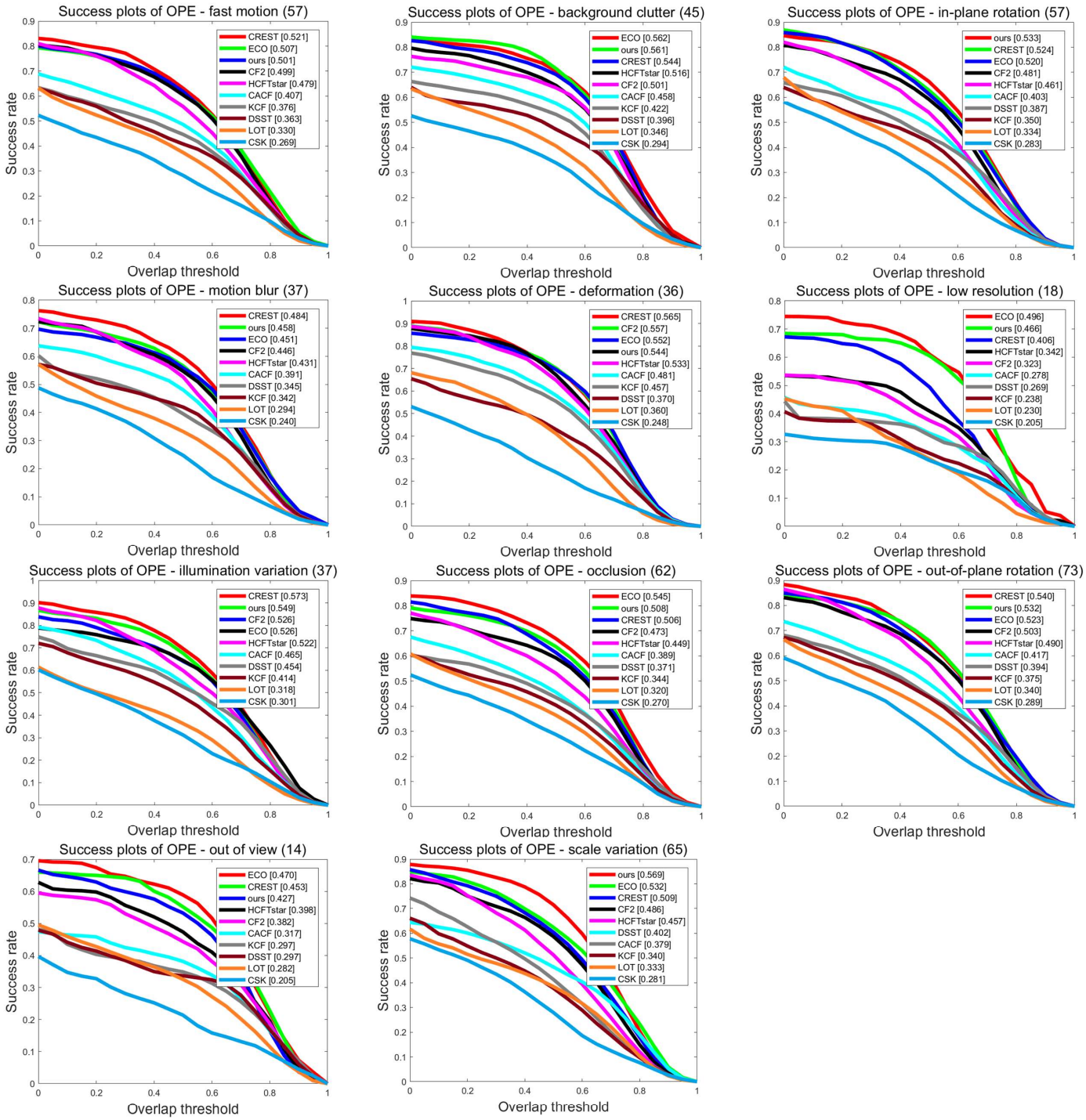

Figure 11. Comparison of 11 attribute challenge results.

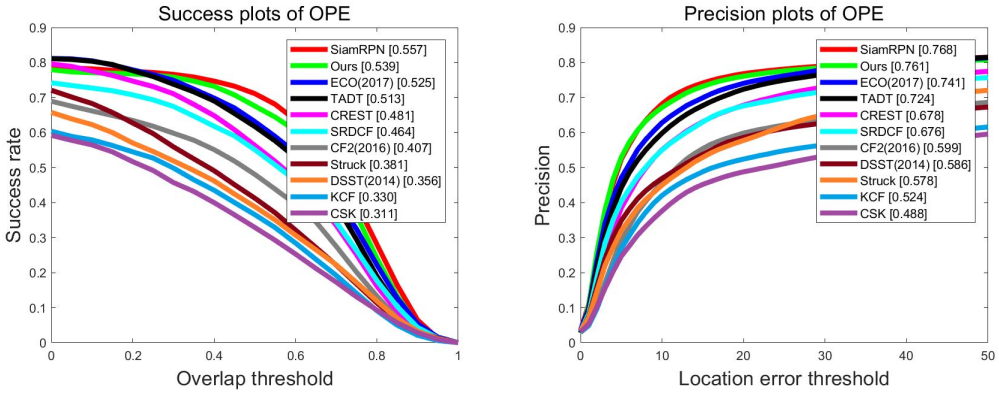

Figure 12. Success and precision rates on the UAV-123 dataset.

4.5. VOT2016 Dataset Experiment

The VOT dataset is a very popular dataset in the field of target tracking, and it uses two metrics, accuracy and robustness, to evaluate the performance of the trackers, as well as the average overlap metric (EAO) to rank the tracker. In this paper, the proposed algorithm is compared with other trackers on the VOT2016 dataset for experiments, and the compared trackers include SiamRPN++ [40], SiamRPN [10], TADT [39], DeepSRDCF [41], MDNet [42], SRDCF [26], HCF [27], DAT [43], and KCF [23]. The results of these tracker are obtained from the official results, and Figure 11 show the results of all tracker' ranking results.

As can be seen from Figure 13, thanks to the proposed lightweight target-aware attention learning network and the weight learning approach of the attention learning loss function, the proposed algorithm ranks third among all the compared trackers and performs better than the TADT algorithm that uses the regression loss function and the scale loss function for feature layer filtering. The performance of the proposed algorithm is weaker than that of SiamRPN and SiamRPN++ tracker, which also shows that SiamRPN introduces a region suggestion network to provide an accurate suggested target area and a classification regression mechanism to determine the target location and obtain a more accurate target scale through regression calculation. SiamRPN++ algorithm, on the other hand, introduces a deeper neural network to extract target features based on the SiamRPN algorithm, so it performs far ahead of the other tracker, which also shows that deep neural networks are more powerful in feature representation.

Table 2 shows some more detailed information comparing all the tracker, including the average overlap (EAO), overlap (Overlap), and failure (Failures), and the top three metrics on individual results are marked in red, green, and blue, respectively. As can be seen from the table, the proposed algorithm performs well overall in all three metrics, which reflects the ability of the proposed attention learning loss function and lightweight goal-aware attention learning network to learn reliable target features. The last column of the table shows the failure rate of the algorithm tracking, and it can be seen that the proposed algorithm ranks fourth place, which is not very far from the second-place SiamRPN and the third-place TADT, and there is still room for improvement.

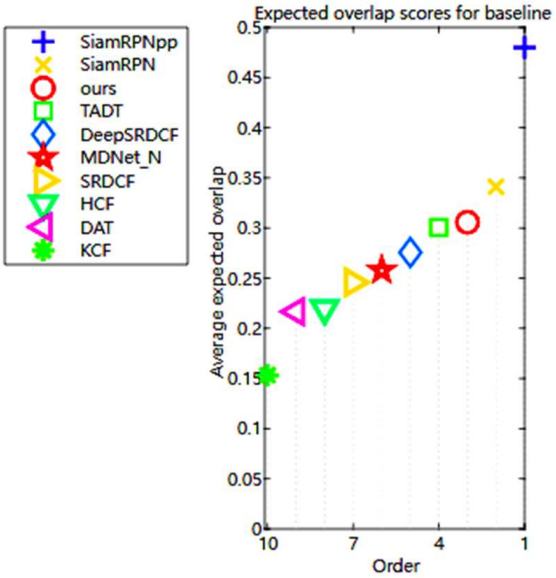

Figure 13. EAO score ranking of the compared trackers VOT2016 dataset.

Table 2. Overall performance on VOT2016 dataset, the top three trackers are marked with red, green and blue, respectively.

Tracker	EAO	Overlap	Failures
Ours	0.306	0.546	20.180
SiamRPN++	0.479	06356	11.586
SiamRPN	0.341	0.580	20.138
TADT	0.300	0.546	19.973
DeepSRDCF	0.275	0.522	20.346
MDNet	0.257	0.538	21.081
SRDCF	0.245	0.525	28.316
HCF	0.219	0.436	23.856
DAT	0.216	0.458	28.353
KCF	0.153	0.469	52.031

4.6. LaSOT Dataset Experiment

To further demonstrate the effectiveness of our method, the performance of the proposed algorithm is evaluated on the LaSOT dataset in this work. Compared with the above tracking dataset, LaSot dataset has a larger salce and more complex challenges for the tracker during the tracking process. LaSOT considers the connection between visual appearance and natural language, not only labeling the bounding box but also adding rich natural language descriptions. It contains 1400 video sequences with an average sequence length of 2500 frames and the test dataset contains 280 video sequences, with 4 videos per category.

As shown in Figure 14, our method achieved the third place in precision and success rate. Compared with the tracking algorithms based on the correlation filter, our method also obtains a good performance. However, the performance of our method is not competitive enough with the state-of-art tracking methods on the LaSOT dataset. The reason for this phenomenon is that our algorithm is not able to solve the challenge of target disappearance reproduction during long-term tracking.

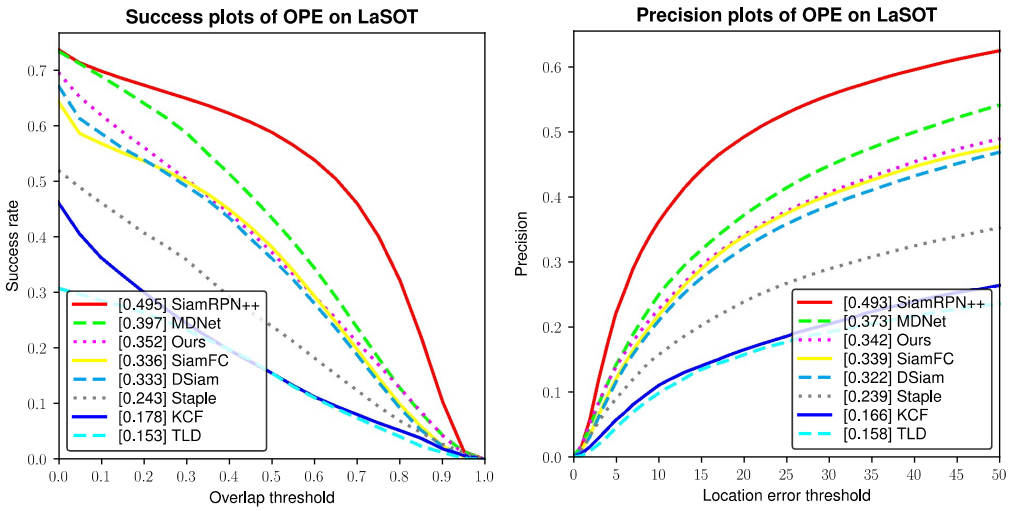

Figure 14. Success and precision plots of OPE on LaSOT dataset.

4.7. Discussions

The Siamese network tracker based on pre-trained depth features has achieved good performance in recent years. The pretrained depth features are trained in advance on large-scale datasets, and therefore contain feature information of a large number of objects. However, for a tracking video, the object being tracked is always the same, so the pre-trained features contain some redundant features. To remove redundant and interfering information from pre-trained features and learn more accurate target information, this work presents a novel tracking method with the proposed lightweight target-aware attention learning network. This lightweight target-aware attention learning network uses reliable information that the ground truth of the target is given in the first frame of each video to train the weights of the network online and obtains gradient value information by backpropagation to determine the effect of different feature channels in the target feature layer on the target, and remodel the channel of the template feature by weighting this contribution. Then the compact and effective deep feature is obtained, which can better distinguish the object from the background. The network is the single-convolutional layer network which is relatively easy to implement and compared to complex convolutional neural networks, there are fewer parameters in the network. It is worth improving that although our method can refine the target features, it does not have the ability to deal with target failure, so its performance is constrained by the target disappearance reproduction challenge in long-term tracking.

5. Conclusions

In this paper, a novel Siamese network-based target tracking method is proposed to address the problem that different feature channels often have different importance for the target representation, which enhances the feature tracking target by designing a lightweight target-aware attention learning network and using a redesigned attention learning loss learning function to learn the most effective feature channel weights for the target using the Adam optimization method representation. This lightweight target-aware attention learning network uses reliable information from the first frame of each video sequence to train the weights of the network online, and obtains gradient value information by back propagation to determine the contribution of different feature channels in the target feature layer to model the target, and re-models the target by weighting this contribution to the channels of the template features. The network is relatively easy to implement and the small number of parameters facilitates fast computation. Finally, the proposed algorithm is evaluated on OTB, TC-128, UAV123, VOT2016, and LaSOT datasets, and both quantitative and qualitative analyses show that the method achieves satisfactory performance, demonstrating the effectiveness of the proposed lightweight target-aware

attention learning network and attention learning loss function in a Siamese network framework-based tracker.

Author Contributions: Conceptualization: Y.Z., J.Z., R.D. and F.L.; methodology: Y.Z., J.Z., R.D. and F.L.; software: J.Z., R.D., H.Z.; validation: J.Z., R.D., H.Z.; analysis: Y.Z., R.D. and F.L.; investigation: H.Z.; resources: Y.Z., F.L.; writing—original draft preparation: J.Z., R.D.; writing—review and editing: Y.Z., J.Z., R.D.; visualization: R.D., H.Z.; supervision: H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (61873246, 62072416, 6167241, 61702462), Program for Science & Technology Innovation Talents in Universities of Henan Province (21HASTIT028), Natural Science Foundation of Henan (202300410495), Zhongyuan Science and Technology Innovation Leadership Program (214200510026).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: No conflict of interest exits in the submission of this manuscript, and this manuscript is approved by all authors for publication.

References

1. Farabet, C.; Couprie, C.; Laurent, N.; Yann, L. Learning Hierarchical Features for Scene Labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
2. Bousetouane, F.; Dib, L.; Snoussi, H. Improved mean shift integrating texture and color features for robust real time object tracking. *Vis. Comput.* **2012**, *29*, 155–170. [CrossRef]
3. Zhang, H.; Chen, J.; Nie, G.; Hu, S. Uncertain motion tracking based on convolutional net with semantics estimation and region proposals. *Pattern Recognit.* **2020**, *102*, 107232. [CrossRef]
4. Guo,W.; Gao, J.; Tian, Y.; Yu, F.; Feng, Z. SAFS: Object Tracking Algorithm Based on Self-Adaptive Feature Selection. *Sensors* **2021**, *21*, 4030. [CrossRef] [PubMed]
5. Cao, Z.; Fu, C.; Ye, J.; Li, B.; Li, Y. HiFT: Hierarchical Feature Transformer for Aerial Tracking. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
6. Zhao, H.; Yang, G.; Wang, D.; Lu, H. Lightweight Deep Neural Network for Real-Time Visual Tracking with Mutual Learning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019.
7. Cheng, X.; Zhao, L.; Hu, Q. Real-Time Semantic Segmentation Based on Dilated Convolution Smoothing and Lightweight Up-Sampling. *Laser Optoelectron. Prog.* **2020**, *57*, 021017. [CrossRef]
8. Zhang, H.; Chen, J.; Nie, G.; Lin, Y.; Yang, G.; Zhang,W. Light regression memory and multi-perspective object special proposals for abrupt motion tracking. *Knowl.-Based Syst.* **2021**, *226*, 107127. [CrossRef]
9. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H.S. Fully-convolutional Siamese networks for object tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
10. Li, B.; Yan, J.; Wu, W.; Zhu, 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, 18–23 June 2018.
11. Gao, P.; Yuan, R.; Wang, F.; Xiao, L.; Hamido, F.; Zhang, Y. Siamese Attentional Keypoint Network for High Performance Visual Tracking. *Knowl.-Based Syst.* **2019**, *193*. [CrossRef]
12. Chen, K.; Tao, W. Learning linear regression via single-convolutional layer for visual object tracking. *IEEE Trans. Multimed.* **2018**, *21*, 86–97. [CrossRef]
13. Ramprasaath, R.S.; Michael, C.; Abhishek, D.; Ramakrishna, V.; Devi, P.; Dhruv, B. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. Wu, Y.; Lim, J.; Yang, M.-H. Online object tracking: A benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
16. Wu, Y.; Lim, J.; Yang, M.H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
17. Liang, P.; Blasch, E.; Ling, H. Encoding color information for visual tracking: Tracker and benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5630–5644. [CrossRef] [PubMed]
18. 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, 11–14 October 2016; pp. 445–461.
19. Hadfield, S.; Bowden, R.; Lebeda, K. The visual object tracking VOT2016 challenge results. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 777–823.

20. Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Huang, M.; Liu, J.; Xu, Y.; et al. Lasot: A high-quality large-scale single object tracking benchmark. *Int. J. Comput. Vis.* **2021**, *129*, 439–461. [CrossRef]
21. Yang, T.; Chan, A.B. Learning dynamic memory networks for object tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
22. Yang, T.; Chan, A.B. Visual tracking via dynamic memory networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 360–374. [CrossRef] [PubMed]
23. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef]
24. 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.
25. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, F. Discriminative scale space tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1561–1575. [CrossRef]
26. Li, F.; Tian, C.; Zuo, W.; Zhang, L.; Yang, M.H. Learning spatial-temporal regularized correlation filters for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4904–4913.
27. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
28. Song, Y.; Ma, C.; Gong, L.; Zhang, L.; Lau, R.W.H.; Yang, M.H. Crest: Convolutional residual learning for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
29. Lukezic, A.; Vojir, T.; Zajc, L.C.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4847–4856. [CrossRef]
30. Tan, K.; Xu, T.B.; Wei, Z. Online visual tracking via background-aware Siamese networks. *Int. J. Mach. Learn. Cybern.* **2022**, 1–18. [CrossRef]
31. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4660–4669.
32. Yuan, D.; Kang, W.; He, Z. Robust visual tracking with correlation filters and metric learning. *Knowl.-Based Syst.* **2020**, *195*, 105697. [CrossRef]
33. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
34. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Robust Visual Tracking via Hierarchical Convolutional Features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2709–2723. [CrossRef]
35. Mueller, M.; Smith, N.; Ghanem, B. Context-Aware Correlation Filter Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
36. Oron, S.; Bar-Hillel, A.; Levi, D.; Avidan, S. Locally orderless tracking. *Int. J. Comput. Vis.* **2015**, *111*, 213–228. [CrossRef]
37. Henriques, J.F.; Rui, C.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the 12th European conference on Computer Vision—Volume Part IV, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012.
38. Hare, S.; Golodetz, S.; Saffari, A.; Vineet, V.; Cheng, M.; Hicks, S.L.; Torr, P.H.S. Struck: Structured output tracking with kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2096–2109. [CrossRef]
39. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 1369–1378.
40. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yang, 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.
41. Danelljan, M.; Hager, G.; Khan, F.S.; Felsberg, M. Convolutional Features for Correlation Filter Based Visual Tracking. In Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, Chile, 7–13 December 2015.
42. Nam, H.; Han, B. Learning Multi-Domain Convolutional Neural 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.
43. Pu, S.; Song, Y.; Ma, C.; Zhang, H.; Yang, M.H. Deep attentive tracking via reciprocative learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.