

Article

A Novel Target Tracking Scheme Based on Attention Mechanism in Complex Scenes

Yu Wang ¹, Zhutian Yang ^{1,*}, Wei Yang ² and Jiamin Yang ³¹ School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150080, China² Nanjing Electronic Equipment Institute, Nanjing 210013, China³ School of Information Science and Engineering, Southeast University, Nanjing 211189, China

* Correspondence: yangzhutian@hit.edu.cn; Tel.: +86-186-8685-7066

Abstract: In recent years, target tracking algorithms based on deep learning have realized significant progress, especially the Siamese neural network structure, which has a simple structure and excellent scalability. Although these methods provide excellent generalization capabilities, they fail to perform the task of learning target information discrimination smoothly due to being affected by distractors such as background clutter, occlusion, and target size. To solve this problem, in this paper we propose a newly improved Siamese network target tracking algorithm based on an attention mechanism. We introduce a channel attention module and a spatial attention module into the original network to improve the problem of insufficient semantic extraction ability of the convolutional layer of the tracking algorithm in complex environments. A channel attention mechanism enhances the feature extraction ability by using the network to learn the importance of each channel and establish the relationship between channels, while a spatial attention mechanism strengthens the feature extraction ability by establishing the importance of spatial position in locating the target or carrying out a certain degree of deformation. In this paper, the above two models are combined to improve the robustness of trackers without sacrificing tracking speed. We conduct a comprehensive experiment on the Object Tracking Benchmark dataset. The experimental results show that our algorithm outperforms other real-time trackers in both accuracy and robustness in most complex environments.

Keywords: visual object tracking; Siamese network; deep learning; full convolutional neural network; attentional mechanism



Citation: Wang, Y.; Yang, Z.; Yang, W.; Yang, J. A Novel Target Tracking Scheme Based on Attention Mechanism in Complex Scenes. *Electronics* **2022**, *11*, 3125. <https://doi.org/10.3390/electronics11193125>

Academic Editor: Dah-Jye Lee

Received: 24 August 2022

Accepted: 26 September 2022

Published: 29 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Target tracking is one of the current hot spots in computer vision, and has both a wide research range and practical significance. Its applications range from civilian medical imaging [1] and industrial detection to military missile guidance and unmanned aerial vehicle operation [2]. For target tracking in practical applications, visual tracking is particularly challenging because of the limited target information, as only objects in one frame and instantaneous changes in the appearance of objects can be employed. Despite great progress achieved by well-known tracking methods in recent years, several issues continue severely affect the performance of tracking, such as occlusion, deformation, cluttered background, low pixels, motion blur, illumination changes, and pose changes [3]. In addition, real-time tracking should be taken into account in practical applications. Therefore, the exploration of the target tracking task in complex scenes is an important research topic.

The current popular visual tracking methods focus on the Siamese network, which is characterized by a good balance of accuracy and speed [4,5]. A Siamese network is a deep learning method [6] which formulates the tracking task as a one-shot detection task, i.e., by using the bounding box in the first frame as the only exemplar. Unlike traditional network models, a Siamese network is a two-parallel Convolutional Neural Network (CNN)-based

Y-shaped deep learning framework in which one branch is utilized for the template image and another is responsible for subsequent frames [7]. It regards feature extraction in target tracking as a similarity learning problem. By convolution of the template image and the search image, a similarity measurement function is trained and the target location is completed according to the response diagram. On this basis, a number of Siamese network-based tracking algorithms, such as, SiamFC [8], SiamRPN [9], SiamRPN++ [10], and SiamDW [11], have been proposed to continuously intensify the tracking performance by improving the datasets, network architecture, network depth, and other aspects.

Despite the great success of Siamese networks, visual tracking based on Siamese networks suffers from several problems. First, due to a lack of sufficient generalization ability, shallow networks cannot properly solve the problem of background changes, which leads to limited tracking performance. Second, although the above algorithms based on Siamese networks have achieved specific achievements in performance, they face great challenges in training and high computational complexity, resulting in unsatisfactory real-time tracking performance and practical application effects. To solve these problems, this paper introduces an attention mechanism [12–14] to improve the generalization ability of the model and reduce the influence of noise in the target template. Meanwhile, due to the extremely lightweight feature of the attention module, the tracking speed is unaffected [15]. According to their different principles, attention mechanisms can be divided into two types, namely, spatial attention mechanisms and channel attention mechanisms. In this paper, the convolution structure of the Siamese network is extended and improved by combining these two attention modes through the addition of a serial channel-space attention module (CBAM). The channel attention mechanism is utilized to extract the feature types that require more attention, and the spatial attention mechanism is subsequently adopted to capture the spatial importance distribution of the features.

An extensive experiment is conducted on the Object Tracking Benchmark (OTB) [16] dataset to demonstrate the proposed tracker's effectiveness. The experimental results show that our proposed model achieves better performance compared to other state-of-the-art other trackers. In summary, the main contributions of this work are threefold:

- Based on an in-depth analysis of Siamese trackers, we propose a novel Siamese network tracking framework which introduces an attention mechanism to improve the expressive power of the neural network. The attention weights are integrated within the Siamese network without additional modules, maintaining its end-to-end property.
- We design an improved attention module which skillfully integrates a spatial attention mechanism and channel attention mechanism to systematically learn features at different levels. Recalibration of the attention maps ensures that the network attends to discriminative and robust features.
- After carrying out extensive experiments on challenging large benchmark datasets, we analyze performance under different visual attributes, then compare our proposed tracker with other state-of-the-art and real-time depth trackers. Simulation results and real data validation confirm that the proposed algorithm is efficient and performs well in cluttered environments.

The rest of this paper is organized as follows. Section 2 describes previous works closely related to the proposed methodology. Section 3 explains the proposed methodology. The experimental results are shown in Section 4, along with a thorough analysis. Our final conclusions are drawn in Section 5.

2. Related Work

In this section, we briefly introduce those basic principles of Siamese networks and attention mechanisms which are closely related to our research.

2.1. Deep Feature-Based Trackers

Object tracking is one of the most fundamental and practical tasks in computer vision. In 2013, N. Wang proposed the Direct Linear Transform (DLT) [17] algorithm, which first introduced deep learning into target tracking. This greatly improved the accuracy of trackers, and deep learning has subsequently become the mainstream approach to target tracking tasks. DLT conducts off-line unsupervised learning training to obtain network classifiers, solving the problem of insufficient training samples in target tracking and enabling online fine-tuning in the tracking process. In recent years, many target tracking algorithms based on deep learning have emerged, as shown in Table 1; these are mainly divided into generative adversarial network-based, recurrent neural network-based, and Siamese network-based target tracking.

Table 1. Summary of object tracking algorithms based on deep learning.

Method Category	Advantages	Disadvantages
Generative Adversarial Network [18]	Generating random occlusions alleviates the problem of imbalanced training samples	Struggles to converge during training
Recurrent Neural Network [19]	Makes full use of the temporal information between networks to improve the target discrimination ability	The network parameters are large, and the overall performance is poor
Siamese Network	Weights are shared, network complexity is low, and it achieves a good balance of accuracy and speed	The robustness is relatively poor in complex environments

As can be seen from the table, the three mainstream deep learning-based object tracking algorithms each possess their own characteristics. Among these approaches, Siamese networks possess outstanding ability to balance accuracy and speed thanks to their special structure, and as a result have become the primary algorithm used in target tracking. As this paper introduces a tracking algorithm based on a Siamese network, we discuss the development trend and existing shortcomings of this algorithm below.

2.2. Siamese Tracking

In machine learning, neural networks always require inputting large labeled data to obtain more ideal results [20–22]. However, traditional neural networks tend to show unsatisfactory performance when dealing with limited training samples and multiple types of samples. Unlike traditional deep learning network models, Siamese networks regard target tracking as a similarity learning problem, and have the potential to weaken the importance of labels thanks to their unique network structure, which greatly alleviates data pressure during network training. A Siamese network has two or more branches of with shared weights, and computes the distance between two sets of input vectors as a measure of difference by mapping the input data into feature vectors. The Euclidean distance and cosine distance [23] are the most commonly used functions. In this way, after the network has been adjusted, its powerful discriminative ability can be employed to make appropriate inferences from new data [24].

SiamFC [8] is one of the representative algorithms of Siamese networks for applications in target tracking, and is deemed the basis for most Siamese network improvements. The algorithm is mainly divided into three parts. The first is data input, during which the algorithm exploits two images with different fixed sizes as the template image z and search image x as the input, with the input of the former representing the image of the target to be tracked in the initial frame and the input of the latter referring to the image of the current frame. The second part is a convolutional network layer, in which the algorithm

typically selects the unfilled full convolutional network as the feature extraction tool with the aim of meeting the convenience of the input with the different sizes mentioned above. The third part is the similarity judgement. While SiamFC ensures the translation invariance of the network and balances speed and accuracy, it becomes less effective in complex environments [25]. Therefore, many improved algorithms for Siamese networks have appeared in recent years. As described in Table 2, SiamRPN introduced the Region Proposal Network (RPN) to predict the scale of the target, and SiamRPN++ further increased the depth of the network, thereby effectively improving the ability to discriminate the target. SiamDW designed new residual modules to eliminate the negative impact of padding. However, these algorithms usually have the disadvantages of high complexity and poor real-time performance. Therefore, solving such a intractable difficulties is the main focus of this paper.

Table 2. Performance comparison of different Siamese network algorithms.

The Algorithm Name	Advantages	Disadvantages
SiamFC	The tracking process is transformed into a similarity learning problem, which ensures the translation invariance of the network and balances the running speed and accuracy	The effect of target tracking in complex environments is poor
SiamRPN	The RPN network is introduced to predict the scale of the target, and the discrimination of the target is enhanced	Due to the anchor frame mechanism of RPN, there are errors in target position prediction and scale estimation
SiamRPN++	The accuracy is further improved by a multi-layer aggregation network	The improvement in accuracy is limited, and the real-time performance is reduced due to the multi-layer network
SiamDW	An internal Clipping Residual (CIR) unit is used to eliminate the impact of padding and enhance the positioning accuracy.	The outermost feature of the feature map boundary of each frame needs to be cropped, which increases the time complexity and reduces the real-time tracking performance

2.3. Attention Mechanism

Visual attention is an information processing mechanism of the human brain [26]. During the process of observing a picture, people often focus on useful information by quickly scanning the whole picture, which is an extremely efficient survival mechanism developed under long-term biological evolution. Therefore, in 2014 Bahdanau [27] and his team introduced this mechanism into a neural network for application and carried out weight distribution of data according to the focus pattern of human attention, thereby realizing much higher weights in the parts with stronger representation of the target and improving the robustness of their neural network model. Today, visual attention is extensively applied in various fields.

According to different principles, attention mechanisms can be divided into two types, namely, spatial attention mechanisms and channel attention mechanisms. In neural networks, different channels are equivalent to different filters. Channel attention increases the weight of the channel associated with the semantic feature expression of the target appearance, and reduces the weight of other irrelevant channels in order to provide direction and certainty for the selection of task filters or the consideration of features. Channel attention can change the correlation between different channels as well [28]. The channel under high-level semantics can be considered as the response of specific objects, generally accompanied by a certain correlation between each other, such as cats' ears, eyes,

nose, etc. Enhancing the correlation is beneficial in improving the extraction of feature appearance semantics. In contrast, spatial attention is capable of making feature extraction more efficient by increasing the weight of spatial positions related to important features which ought to be highly emphasized. In neural network training, an image has abundant pixel information; however, only certain specific areas are relevant to the main body of the task, and therefore need special attention. Spatial attention can establish the dependency relationship between two long-distance pixels in the feature map, which is conducive to strengthening the semantic expression of feature extraction. The addition of a spatial attention mechanism reinforces the recognition rate of key image information without incurring excessive calculation.

3. The Proposed Algorithm

Our proposed algorithm intends to adopt the overall framework of the SiamFC algorithm to tackle the unsatisfactory tracking effect in situations with complex backgrounds. After extracting features from the network, a CBAM attention mechanism module is embedded. This section introduces the proposed algorithm in three parts: its overall design, the realization of the attention mechanism, and the concrete details of its implementation.

3.1. Network Architecture

The general block diagram of the SiamFC tracking algorithm with an attention mechanism is shown in Figure 1. The overall framework applies the SiamFC algorithm to measure the similarity of two input images through a cross-convolution operation. The similarity measurement formula with the attention mechanism is expressed as follows:

$$f(z, x) = (\delta \cdot \varphi(z)) * (\varepsilon \cdot \varphi(x)) + b_1 \quad (1)$$

where z represents the value of a pixel in the target image, x represents the value of a pixel in the search image, φ represents the convolution operation function, δ represents the weight distribution of the target image obtained by the attention mechanism module, ε represents the weight of the candidate image obtained by the attention mechanism module, and b_1 represents the value at each location on the confidence map.

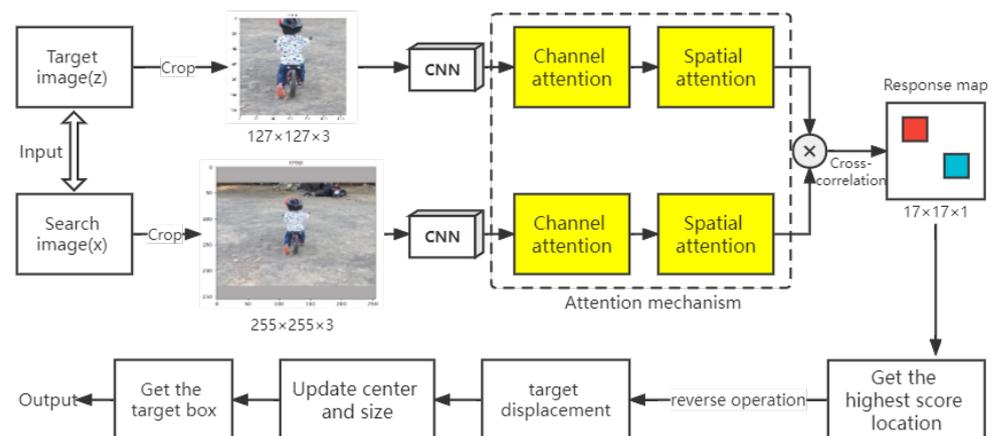


Figure 1. The overall architecture of the proposed tracker. We have highlighted the added attention mechanism module in yellow.

In the proposed algorithm, the network model first needs to be trained to input the target image and the search image, where the size of the target image is $127 \times 127 \times 3$. During the search process, the objects should be searched for in an area four times larger than the original target, so the size of the search image is $255 \times 255 \times 3$. Then, the CNN network is used to extract the features of the template image and the search image, with the feature map sizes set to $6 \times 6 \times 512$ and $22 \times 22 \times 512$, respectively. Finally, the response

map is obtained by inputting the feature map into the CBAM module for similarity measurement. CBAM is a lightweight and efficient module that takes advantage of channel and spatial attention mechanisms without requiring additional weights or computational cost. In this module, we first weighted the input data on the channel to improve the feature expression ability between different channels, then weighted the feature map in space to highlight the importance of different positions. The similarity measure refers to a cross-correlation operation performed on the features extracted from the template image and the search image. This procedure generates a 17×17 response graph.

In target tracking, the search area is cropped from the current frame image according to the estimated target position of the previous frame, then the area is upsampled to a size of 224×224 , which is input into the trained network for similarity measurement. According to the position with the highest similarity between the current frame and the first frame, the position difference can be calculated to update the target frame.

It can be seen that the main body of our algorithm is divided into two parts, model training and target tracking.

3.2. Stacked Channel–Spatial Attention

3.2.1. Channel Attention

Channel attention is a mechanism which increases the ability of neural network visual semantic extraction by changing the weight or relevance of different channels. Its structure is shown in Figure 2.

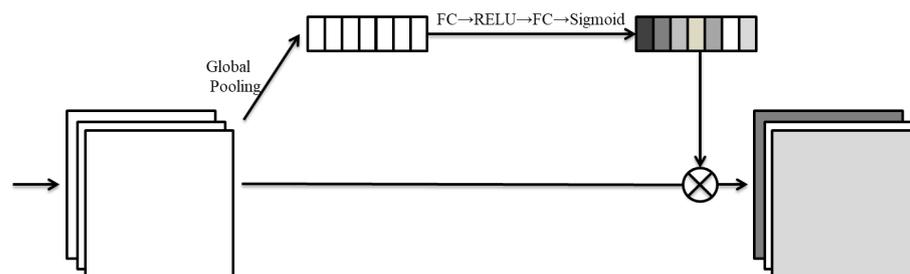


Figure 2. Channel Attention module.

The leftmost side of Figure 2 is an input set of feature channels, defined as follows:

$$A = [a_1, a_2, a_3, \dots, a_n] \tag{2}$$

where $a_k \text{---} a_k \in R^{H \times W}$, $k = 1, 2, 3, \dots, n$. Then, through the global pooling input data A , we can obtain the feature vector b , which can be expressed as

$$b = [b_1, b_2, b_3, \dots, b_n] \tag{3}$$

where $b_k \text{---} b_k \in R^{H \times W}$, $k = 1, 2, 3, \dots, n$.

After the vector b is obtained, it will be input into the Fully Connected Layer (FC), and then being activated it by Rectified Linear Units (RELU). Next, we use a full connection layer, and then activate it with sigmoid function to obtain the feature vector, which is expressed as:

$$\alpha = [\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_n] \tag{4}$$

where $\alpha_k \text{---} \alpha_k \in R^{H \times W}$, $k = 1, 2, 3, \dots, n$.

The vector α and feature map A are then multiplied to obtain the final result, namely, the channel attention feature map, which is expressed as

$$\bar{A} = \alpha \cdot A = [\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3, \dots, \bar{\alpha}_n] \tag{5}$$

where $\bar{\alpha}_k \text{---} \bar{\alpha}_k \in R^{H \times W}$, $k = 1, 2, 3, \dots, n$.

3.2.2. Spatial Attention

Spatial attention can increase the weight of spatial positions related to important features which should receive attention and establish the dependency relationship between two long-distance pixels in the feature map in order as to improve the semantic expression of feature extraction. The module structure is shown in Figure 3.

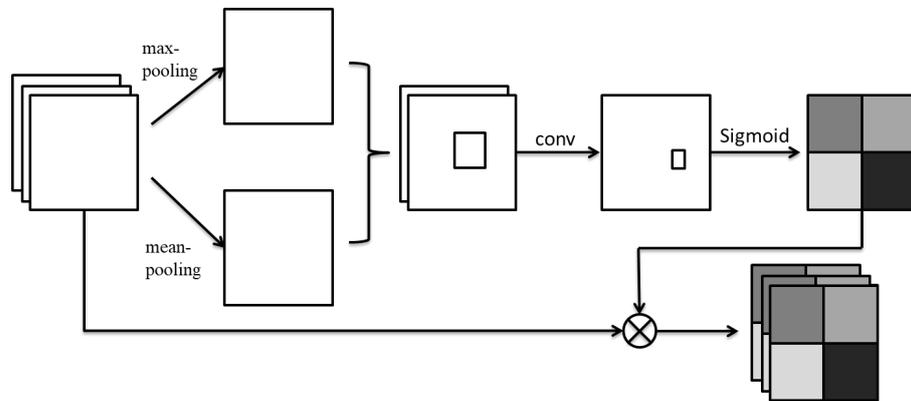


Figure 3. Spatial Attention module.

The leftmost side of Figure 3 is a set of input feature channels, described as $H \times W \times C$; the channel dimensions are the averaged pooling and maximum pooling, respectively. The results of two images with the same size and channel 1 are described as $H \times W \times 1$, then the two feature maps are spliced according to the channel. After passing through a convolution layer (usually with a size of 7×7) and activation function, the spatial attention distribution map is obtained. Finally, the attention distribution map is multiplied by the initial input [28]. The calculation process is as follows:

$$\begin{aligned}
 M_s(F) &= \sigma \left(f^{7 \times 7} ([AvgPool(F), MaxPool(F)]) \right) \\
 &= \sigma \left(f^{7 \times 7} \left(\begin{bmatrix} F_{avg}^S \\ F_{max}^S \end{bmatrix} \right) \right)
 \end{aligned}
 \tag{6}$$

3.2.3. Convolutional Block Attention Module

The combined channel attention and spatial attention mechanism can effectively improve the semantic expression of features in the tracking process, which can be arbitrarily combined in either a parallel or a serial manner. Comparisons indicate that the combined effect of the channel and spatial attention mechanisms functions better in series mode, e.g., the lightweight CBAM embedding module. Its structure is shown in Figure 4.

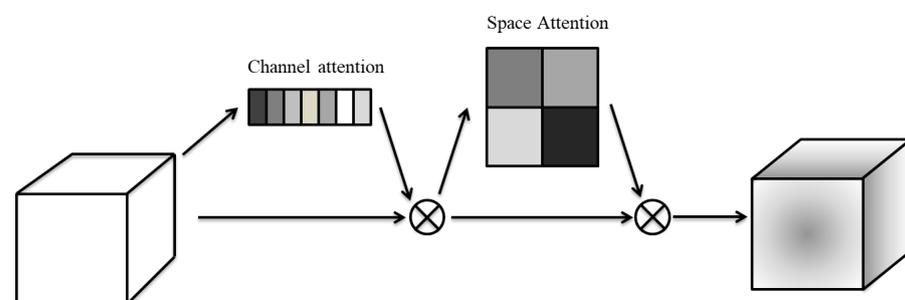


Figure 4. Convolutional block attention module. The specific process is shown below. After inputting the feature, the pooled channel passes through the channel mechanism, providing a global view. Then, convolution and activation provide the channel attention map, as shown above, which is multiplied with the original feature. On this basis, spatial attention is further allocated. The process of spatial attention map acquisition is essentially similar to that of channel acquisition, except that it is pooled in space. Finally, the output is multiplied by the result of the channel attention mechanism.

3.3. Implementation Details

3.3.1. Image Cropping

As mentioned above, the algorithm uses two images of different sizes as input, namely, the target image z and the search image x . Their sizes are generally 127×127 and 255×255 , as shown in Figure 5.

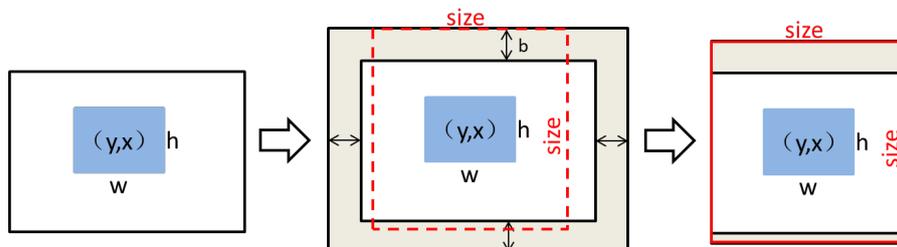


Figure 5. Process diagram of image cropping. The results in the graph are cut from the original image according to the size of the target annotation box; specifically, a certain size of boundary filling is added on the basis of the original target annotation box, then the required fixed-size image pairs are generated by scaling.

We represent the size of the target coordinate box of the input image as (w, h) and set the size of the boundary filling after clipping the template image z to p_z ; then, the size of the target in the input image is

$$S_z = (w + 2p_z) \times (h + 2p_z) \tag{7}$$

where p_z is a quarter of the sum of the length and width of the original target box, that is, $p_z = (w + h)/4$. The target image z can be obtained by zooming S_z to A^2 . The cropping of both the search image and the target image is target-centered, with the search image having less border padding.

Similarly, we set the boundary filling of search image x as p_x ; the expression is

$$p_x = \frac{(B - A) \times S_z}{A} \tag{8}$$

Thus, the size of the search image x in the unprocessed image is

$$S_x = (w_z + p_x) \times (h_z + p_x) \tag{9}$$

where w_z and h_z denote the width and height of the target image z in the unprocessed image, respectively. Similarly, the final search image x can be obtained by scaling S_x to B^2 .

3.3.2. Feature Extraction

Because the target may appear at any position in the search region, the learned feature representation for the target template should remain spatially invariant. Moreover, we determined on the basis of theory that AlexNet [29] is the only the zero-padding variant among modern deep architectures that can satisfy this spatial invariance restriction. In view of this, we base the full convolution network in this part on AlexNet [30]. We remove the last full connection layer in the original network and cancel the packet convolution to address the different input sizes. We use AlexNet to apply the ReLu nonlinear function as the activation function to replace average pooling with maximum pooling. In addition, we add Dropout [31] to the network in order to avoid overfitting. In this process, the original graph is randomly intercepted or flipped to enhance the generalization ability of the model. The overall step size of the network structure after comparison and parameter adjustment is 8; the dimensions of the parameters and activation are depicted in Table 3.

Table 3. Proposed network architecture for convolutional feature extraction.

Layer	Filter Size	Total Number of Channels	Stride	Target Image	Search Image	Channel
Input				127 × 127	255 × 255	×3
Conv1	11 × 11	96 × 3	2	59 × 59	123 × 123	×96
Pool1	3 × 3		2	29 × 29	61 × 61	×96
Conv2	5 × 5	256 × 48	1	25 × 25	57 × 57	×256
Pool2	3 × 3		2	12 × 12	28 × 28	×256
Conv3	3 × 3	384 × 256	1	10 × 10	26 × 26	×384
Conv4	3 × 3	384 × 192	1	8 × 8	24 × 24	×384
Conv5	3 × 3	256 × 192	1	6 × 6	22 × 22	×256

The final feature sizes of the target image and the search image extracted by the network are 6×6 and 22×22 , respectively, and the number of channels is 256. Based on this, we can obtain the response map.

3.3.3. Network Training

The training data used in this project consist of the Generic Object Tracking Benchmark (GOT-10k) dataset [32]. This dataset is based on the WordNet backbone and covers a majority class of 560 classes of real-world moving objects and 80 classes of motion patterns. The total amount of data is 68.8 GB; it contains more than 10,000 video segments of real-world moving objects and over 1.5 million manually labeled bounding boxes. The fair comparison of deep trackers is ensured through the agreement that all approaches use the same training data as provided by the dataset.

The parameters of the training CNN network are as follows. The training network uses Stochastic Gradient Descent (SGD) [33] as the optimization algorithm, the loss function is calculated by Binary Cross Entropy [34], and the learning rate is adjusted according to exponential decay. We set the training cycle to 50 and the batch size to 12, that is, 777 rounds per cycle. In the running process, we output the learning rate, number of rounds, and loss value in real time.

The training sample pair (z, x) can be obtained by preprocessing of the annotated video dataset. In the input search image, as long as the distance from the target is not more than R it is considered to be a positive sample; otherwise, it is a negative sample. The formula is as follows:

$$y[u] = \begin{cases} +1, & k\|u - c\| \leq R \\ -1, & \text{otherwise} \end{cases} \quad (10)$$

where k is the total step length of the network, c is the center of the target, u represents all positions of the response map, and R represents the defined radius.

When training the model, the discriminant method is used to train the positive and negative samples. The logical loss function is defined as

$$l(y, v) = \log(1 + e^{-yv}) \quad (11)$$

where y represents the true value, $y \in (+1, -1)$, and v Represents the actual score between the target image and the search image.

The loss function for model training is calculated using the average loss of all candidate locations:

$$l(y, v) = \frac{1}{D} \sum_{u \in D} l(y[u], v[u]) \quad (12)$$

where D represents the final response map and u represents all positions in the response map.

The convolution parameter θ can be obtained by minimizing the following formula by random gradient descent:

$$\operatorname{argmin}_{\theta} = EL(y, f(z, x; \theta)) \quad (13)$$

3.3.4. Testing

In target tracking, our algorithm generally applies the search image centered on the target position in the previous frame to calculate the response map and finally determines the position of the current target by multiplying the position with the maximum score by the step length. During the search process, the objects should be searched for in an area four times larger than the original target, then processed into three different scales before being introduced into the network. Next, the target position and size are updated by determining the peak point and selecting the channel with the largest response among the three response maps. During this process, the Hamming window is added to punish large displacements in order to achieve more accurate target tracking.

4. Experimental Results

In this section, the experimental evaluation of the proposed Siamese attention network is presented. This section first briefly introduces our experimental environment and evaluation metrics. Then, the algorithm is tested and evaluated in eleven different complex scenarios on the OTB-2015 dataset. Finally, the algorithm is compared with 34 other advanced trackers.

4.1. Implementation Details

4.1.1. Experimental Environment

In these experiments, we use a Linux operating system, a Compute Unified Device Architecture (CUDA) for Graphics Processing Unit (GPU), an Intel(R) Xeon(R) Silver 4114 CPU@2.20GHz, and an NVIDIA Quadro P6000 GPU with driver version 27.21.14.5239. We use the pytorch deep learning framework and Python programming language, and mainly call the following library functions: cv2, torch, os, numpy, numbers, glob, etc.

4.1.2. Evaluation Dataset and Criteria

The proposed Siamese attention network is evaluated on OTB-2015. Evaluation on OTB-2015 follows the standard protocols. OTB-2015 consists of 100 challenging sequences, including eleven challenging attributes: illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutter (BC), and low resolution (LR).

The algorithms use One-Pass Evaluation (OPE). In OPE, the ground-truth target state at the first frame of a tracking sequence is provided to a tracker, then the tracker's performance is measured in the remaining frames.

Two metrics, precision and success plots, are used to rank the trackers. Overlap success rate measures the intersection over union (IoU) between ground truth and predicted bounding boxes. The success plot shows the rate of bounding boxes with IoU scores larger than a given threshold. The precision metric indicates the percentage of frame locations within a certain threshold distance from the ground truth. The threshold distance is set to 20 pixels for all the trackers.

4.2. Experimental Results and Analysis under Different Visual Conditions

In order to test the tracking effect of the proposed algorithm in complex scenes, we divide the complex scenes involved in the OTB-2015 dataset into four categories. We select typical video sets to display, make the success plot and precision plot for each scene, and compare the performance with the 34 advanced target trackers. The first 11 trackers are displayed by default. The gray line segment is the 11th ranked tracker, and we do not show the tracker name for brevity. In the figures, our proposed algorithm is named Siamattention.

4.2.1. Target Change

In the actual target tracking process, the target itself is constantly changing, which poses certain challenges to the tracker. Common occurrences include scale variation (SV), i.e., the ratio of the bounding box of the first frame and the current frame is outside of the range t_s ($t_s = 2$), non-rigid object deformation (DEF), i.e., the target rotates in the image plane (IPR), and outside-of-plane rotation (OPR), i.e., the target rotates outside of the image plane. The tracking effect of the proposed tracker in such scenarios is shown in Figure 6. The success rate and accuracy for a one-time evaluation of the tracking results are shown in Figures 7 and 8.

4.2.2. Incomplete Target

In the actual target tracking process, sometimes the target in the picture is not complete, which poses certain challenges to the tracker. Common issues include the target being partially or fully occluded (OCC) or a portion of the target leaving the view (OV). The tracking effect of the proposed tracker for such scenarios is shown in Figure 9. The success rate and accuracy for a one-time evaluation of the tracking results are shown in Figure 10.



Figure 6. Several challenging video sequences in OTB-2015 involving target change. Pictures from top to bottom: Car4, Dancer, Bird2, Boy. Attributes: SV, DEF, IPR, OPR. In these four scenarios, even though the appearance of the target changes dramatically, the tracking recognition effect is good and there is no significant drift or tracking failure.

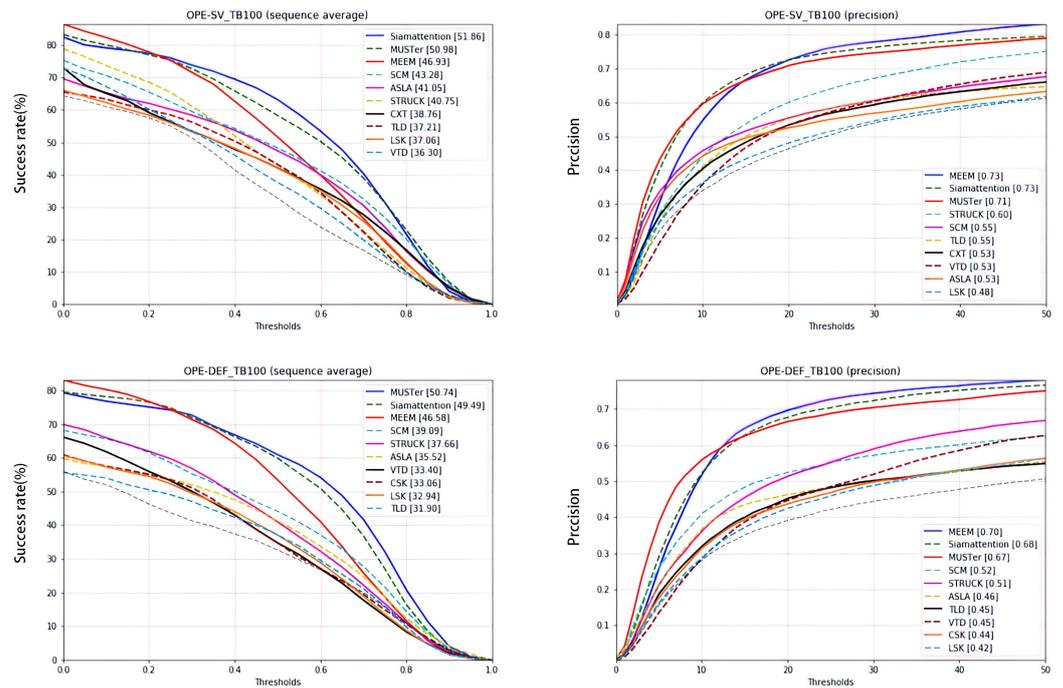


Figure 7. Success plot and precision plot of SV, DEF, IPR, and OPR under SV conditions. The proposed tracker is shown as the solid blue line in the success plot, ranking first, and the green virtual line in the precision plot, again ranking first. Under DEF conditions it is shown as the green dotted line, ranking second.

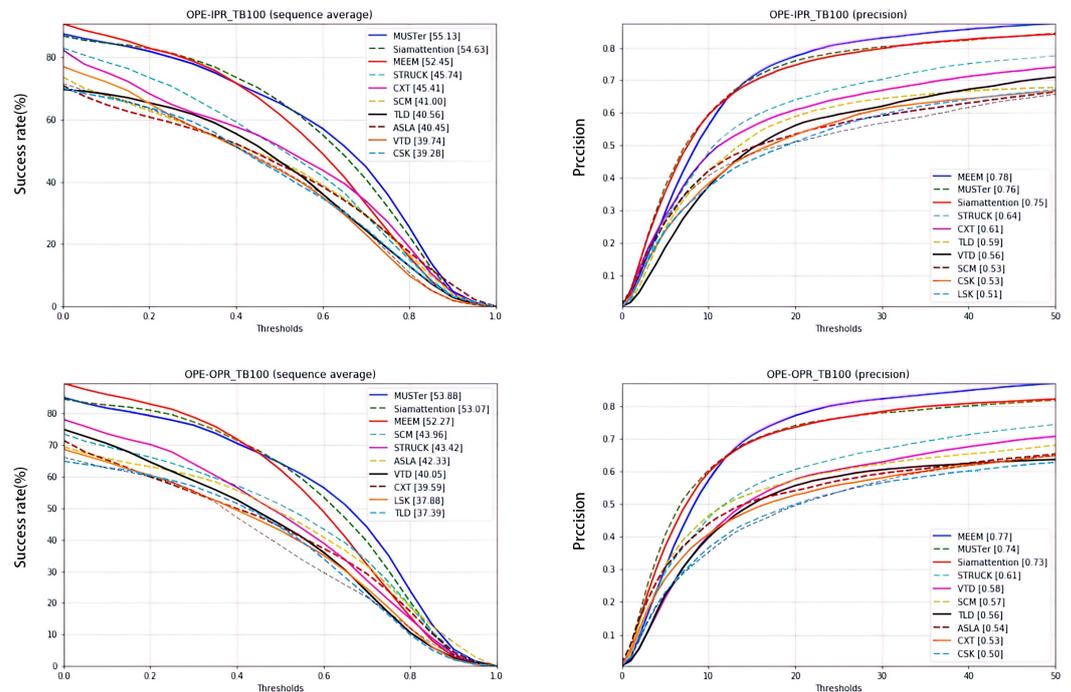


Figure 8. The success plot under IPR and OPR conditions; the proposed tracker is shown by the green dotted lines, ranking second. In the precision plot, it is shown by the red solid line, ranking third.



Figure 9. Several challenging video sequences in OTB-2015 involving incomplete targets. From top to bottom, the pictures are FaceOcc2 and Panda. The attributes are OCC and OV. In these two scenarios, despite the appearance and shape of the target being incomplete, the tracking recognition effect is good, and there is no significant drift or tracking failure.

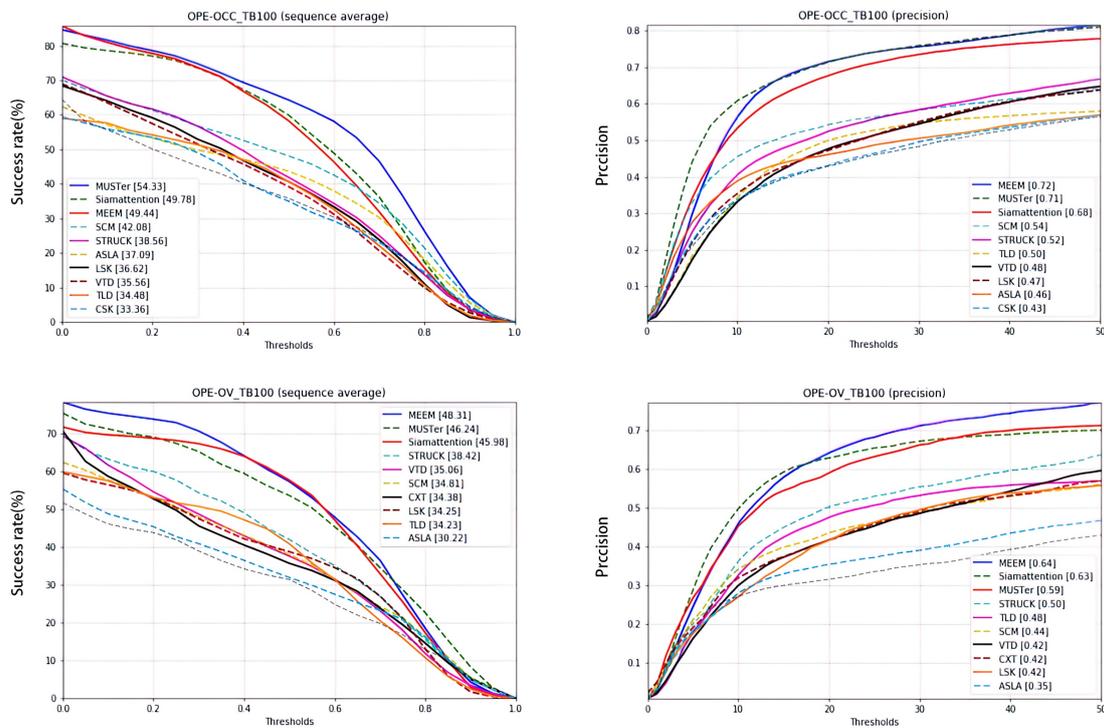


Figure 10. Success plot and precision plot of OCC and OV. Under OCC conditions, the proposed tracker is shown as the green dotted line in the success plot, ranking second, and as the red solid line in the precision plot, ranking third. Under OV conditions, it is shown as the red real line in the success plot, ranking third, and the green virtual line in the precision plot, ranking second.

4.2.3. Fuzzy Objective

In the actual target tracking process, video clarity is often non-ideal, which poses certain challenges to trackers. Common issues include the target region being blurred due to the motion of the target or camera (MB), the ground truth motion being larger than tm pixels ($tm = 20$) (FM), and the number of pixels inside the ground-truth bounding box being less than tr (LR). The tracking effect of the proposed tracker in these scenarios is shown in Figure 11. The success rate and accuracy of a one-time evaluation of the tracking results are shown in Figure 12.

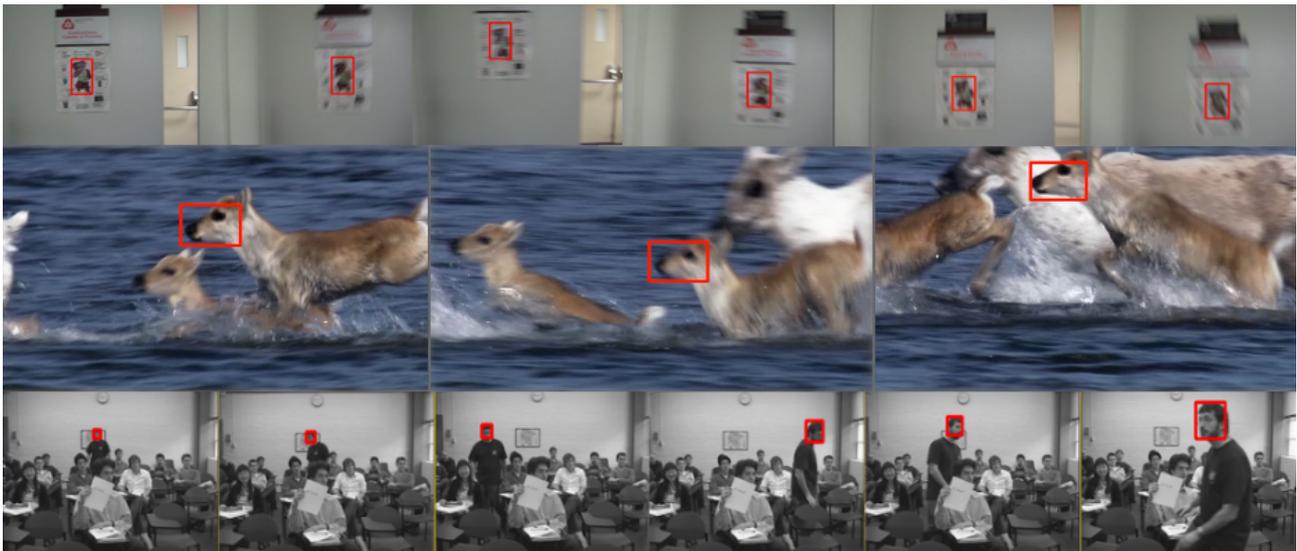


Figure 11. Several challenging video sequences in OTB-2015 involving a fuzzy objective. From top to bottom, pictures are: BlurOwl, Deer, Freeman3. The attributes are MB, FM, and LR. In these two scenarios, the tracking recognition effect is good and there is no significant drift or tracking failure.

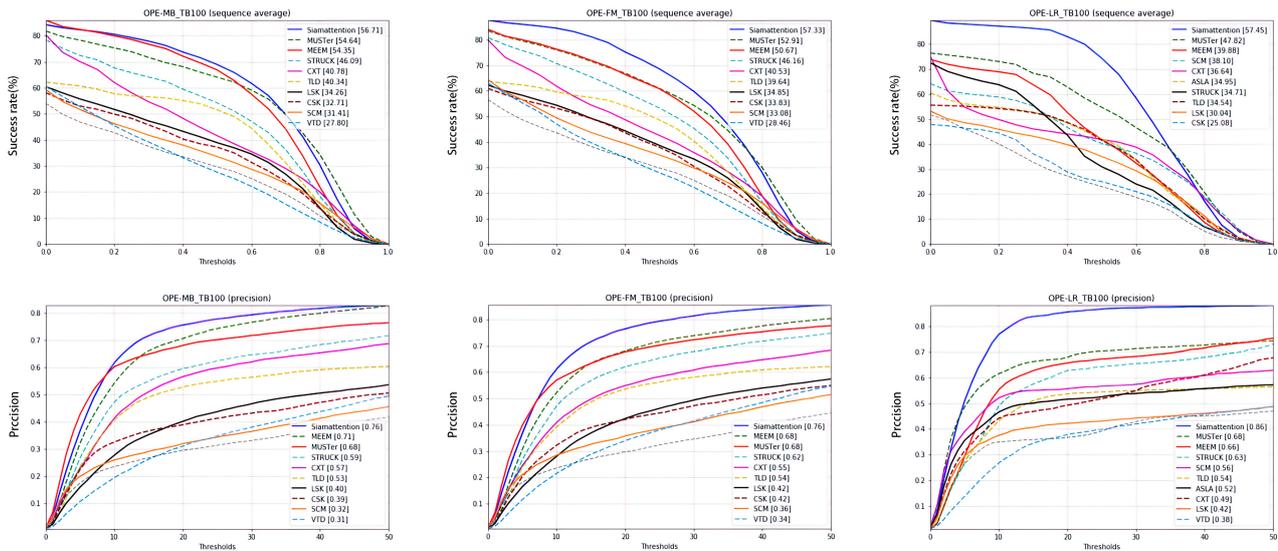


Figure 12. Success plot and precision plot of MB, FM, and LR. The proposed tracker is shown as the solid blue lines, ranking first in all instances.

4.2.4. Target Environment

In an actual target tracking process, the target environment is often uncontrollable, which poses certain challenges to the tracker. It is common that the illumination in the target region is significantly changed (IV), or that the background near the target has a similar color or texture as the target (BC). The tracking effect of the proposed tracker in such scenarios is shown in Figure 13. The success rate and accuracy for a one-time evaluation of the tracking results are shown in Figure 14.

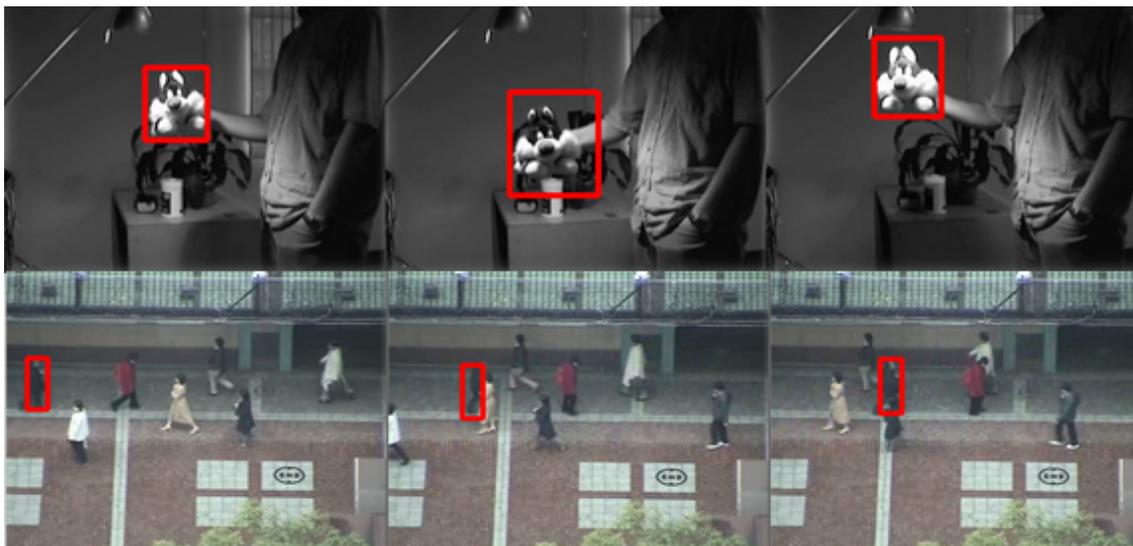


Figure 13. Two challenging video sequences in OTB-2015 involving blurry video conditions. From top to bottom, the pictures are Sylvester and Sunway, while the attributes are IV and BC. In these two scenarios, the environment in which the target is located has a great impact on the judgment of its appearance semantics. Despite this the tracker has good tracking and recognition effect, and there is no significant drift or tracking failure.

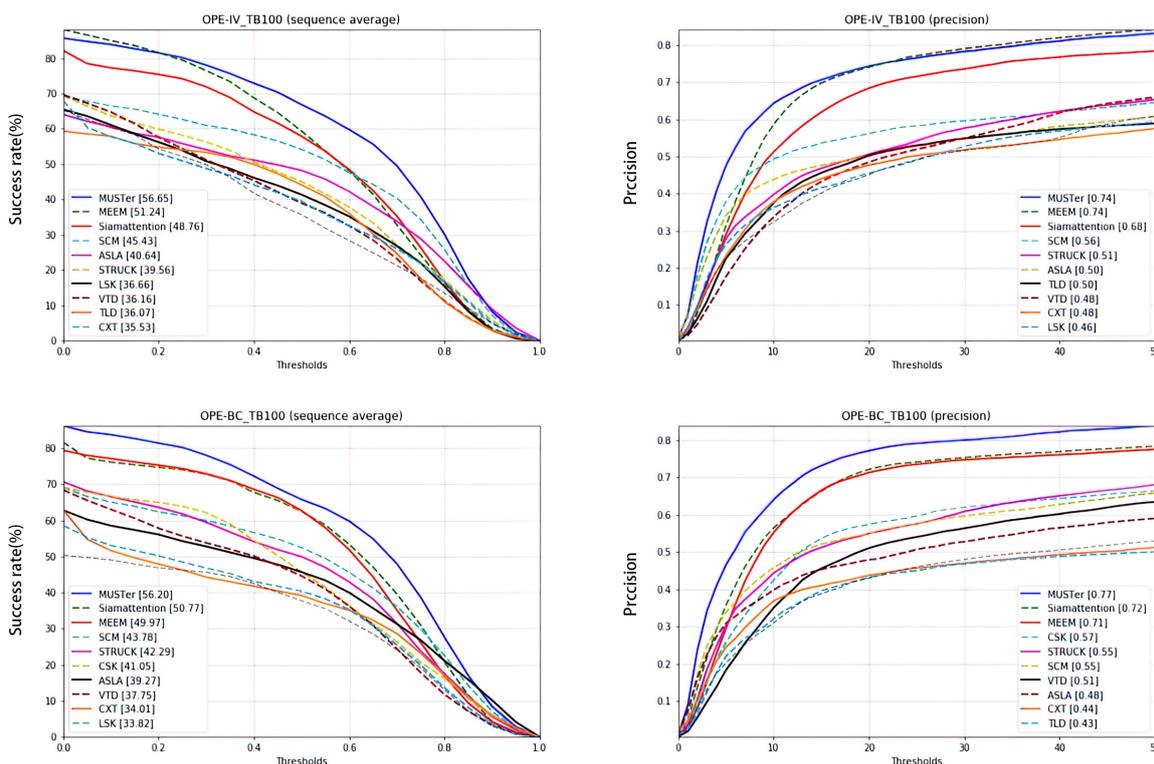


Figure 14. Success plot and precision plot of IV and BC. Under IV conditions, the proposed tracker is shown as the red real line, ranking third. Under BC conditions, it is shown as the green dotted line, ranking second.

4.3. Overall Effect Evaluation of Tracker

As mentioned in the previous section, the proposed tracker is in the top three in all visual conditions and has good adaptability. It has a high success rate and accuracy in the fast motion (FM), motion blur (MB), low resolution (LR), and scale variation (SV) scenarios, ranking first. The analysis is summarized in Tables 4 and 5.

Table 4. Attribute-based success evaluation of the proposed trackers.

Attributes	Success Rate (%)	Lower Than the First (%)	Higher Than the Next (%)	Ranking
BC	50.77	5.43	0.8	2
DEF	49.49	1.25	2.91	2
FM	57.73	0	4.43	1
IPR	54.63	0.5	2.18	2
IV	48.76	7.89	3.33	3
LR	57.45	0	9.63	1
MB	56.71	0	2.07	1
OCC	49.78	4.55	0.34	2
OPR	53.07	0.81	0.8	2
OV	45.98	2.33	7.56	3
SV	51.86	0	0.88	1

Table 5. Attribute-based precision evaluation of the proposed trackers.

Attributes	Precision (%)	Lower Than the First (%)	Higher Than the Next (%)	Ranking
BC	72	5	1	2
DEF	68	2	1	2
FM	76	0	8	1
IPR	75	3	11	3
IV	68	6	12	3
LR	86	0	18	1
MB	76	0	5	1
OCC	68	4	14	3
OPR	73	4	12	3
OV	63	1	4	2
SV	73	0	2	1

We perform overall statistics on 100 datasets of OTB-2015, using the One-Pass Evaluation (OPE) model to produce the success and accuracy plots shown in Figure 15.

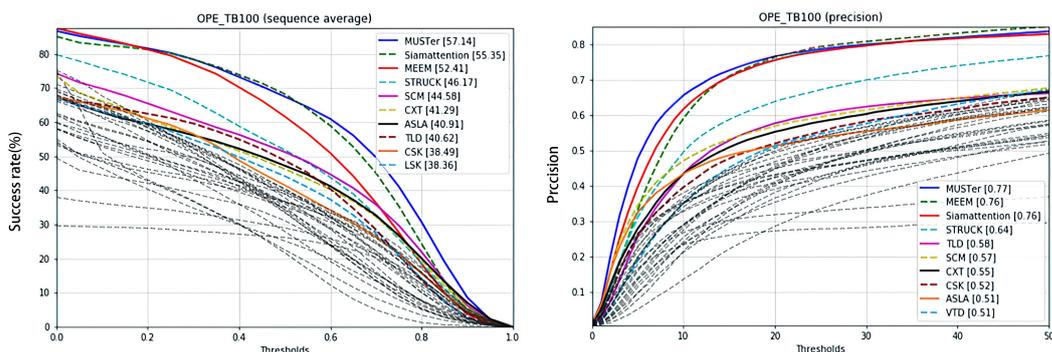


Figure 15. Success plot and precision plot of OTB100. The proposed tracker is shown in the success plot by the green dotted line, ranking second. Its average success rate is 55.35, 2.94% higher than MEEM and 1.71% lower than MUSTer. It is shown in the precision plot with a red solid line, ranking second, parallel with MEEM. The average accuracy is 0.76, 12% higher than STRUCK and 0.01% lower than MUSTer.

It is worth mentioning that we have previously trained the unimproved SiamFC algorithm with an average success rate of 52.85 and an average accuracy of 0.71. The improved Siamattention algorithm proposed in this paper increases this success rate by 2.5% and the accuracy by 5%. This shows that the attention mechanism has a significant impact on improving the appearance semantics and expression ability of feature extraction.

Among the other competitive models, the MUSTer model is mainly based on a Long Short-Term Memory network (LSTM) and its hardware requirements are extremely high, leading to problems in practical applications, while MEEM is actually a mixture of expert models. Thus, it is apparent that the tracker trained in this paper has excellent response in complex scenes and good performance, and represents a clear advance.

5. Conclusions

This paper develops a novel target tracking algorithm for Siamese networks based on an attention mechanism, which overcomes the problems of the model drifts and tracking failure in complex tracking scenes involving issues such as fast motion (FM), motion blur (MB), low resolution (LR), scale variation (SV), and more. Specifically, it highlights the beneficial part of feature extraction, inhibits useless information, improves the tracking accuracy and robustness of the algorithm, and demonstrates that a classical Siamese network and an attention mechanism can be combined with complementary results in tracking tasks. According to the results of our extensive experiments on the challenging OTB-2015 benchmark, the effectiveness and advancement of our tracker compared with other state-of-the-art methods can be verified. The proposed network architecture facilitates continued optimization of the depth and structure of the network as well as adjusting the module architecture in pursuit of higher tracking performance and a more concise network structure. We assume that our research will provide a useful reference for those who are dedicated to studying Siamese networks for visual tracking.

It should be noted that while our proposed architecture achieves better tracking results than the existing trackers on most of attributes, it is less robust on a few attributes, such as out-of-view (OV) and illumination variation (IV) scenarios. The Local Binary Pattern (LBP) operator can be used to describe the local texture features of an image. On account of its characteristics of rotation invariance and gray invariance, it is not sensitive to illumination variation. Combining LBP with the proposed Siamese network could potentially yield better performance in target tracking. In our future work, we intend to focus on these topics.

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

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the reviewers for their comments and suggestions.

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

References

1. Balta, D.; Kuo, H.; Wang, J.; Porco, I.G.; Schladen, M.; Cereatti, A.; Lum, P.S.; Della Croce, U. Infant upper body 3D kinematics estimated using a commercial RGB-D sensor and a deep neural network tracking processing tool. In Proceedings of the 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Messina, Italy, 22–24 June 2022; pp. 1–6.
2. Zhang, K.; Wang, W.; Wang, J. Robust Correlation Tracking in Unmanned Aerial Vehicle Videos via Deep Target-specific Rectification Networks. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510605. [[CrossRef](#)]
3. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual tracking: An experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1442–1468.
4. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338.
5. 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; pp. 8971–8980.
6. Heidari, A.; Navimipour, N.J.; Unal, M. Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review. *Sustain. Cities Soc.* **2022**, *85*, 104089. [[CrossRef](#)]

7. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.
8. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 850–865.
9. Wu, L.; Zhao, C.; Ding, Z.; Zhang, X.; Wang, Y.; Li, Y. A Multi-Target Tracking and Positioning Technology for UAV Based on Siamrpn Algorithm. In Proceedings of the 2022 Prognostics and Health Management Conference (PHM-2022 London), London, UK, 27–29 May 2022; pp. 456–461.
10. Wang, H.; Zhang, S.; Zhao, S.; Wang, Q.; Li, D.; Zhao, R. Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++. *Comput. Electron. Agric.* **2022**, *192*, 106512. [[CrossRef](#)]
11. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
12. Fukui, H.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. Attention branch network: Learning of attention mechanism for visual explanation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10705–10714.
13. Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; Maybank, S. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4854–4863.
14. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
15. Cao, Y.; Ji, H.; Zhang, W.; Shirani, S. Feature Aggregation Networks Based on Dual Attention Capsules for Visual Object Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 674–689. [[CrossRef](#)]
16. 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.
17. Wang, N.; Yeung, D.Y. Learning a deep compact image representation for visual tracking. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013.
18. Aggarwal, A.; Mittal, M.; Battineni, G. Generative adversarial network: An overview of theory and applications. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100004. [[CrossRef](#)]
19. Yang, T.; Chan, A.B. Visual tracking via dynamic memory networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 360–374. [[CrossRef](#)] [[PubMed](#)]
20. Fralick, M.; Campbell, K.R. The basics of machine learning. *NEJM Evid.* **2022**, *1*. [[CrossRef](#)]
21. Liu, M.; Liu, Z.; Lu, W.; Chen, Y.; Gao, X.; Zhao, N. Distributed few-shot learning for intelligent recognition of communication jamming. *IEEE J. Sel. Top. Signal Process.* **2021**, *16*, 395–405. [[CrossRef](#)]
22. Liu, M.; Wang, J.; Zhao, N.; Chen, Y.; Song, H.; Yu, F.R. Radio frequency fingerprint collaborative intelligent identification using incremental learning. *IEEE Trans. Netw. Sci. Eng.* **2021**, *9*, 3222–3233. [[CrossRef](#)]
23. Pan, C.; Huang, J.; Hao, J.; Gong, J. Towards zero-shot learning generalization via a cosine distance loss. *Neurocomputing* **2020**, *381*, 167–176. [[CrossRef](#)]
24. Wei, Z.; Yang, X. Object tracking algorithm based on fusion of SiamFC and Feature Pyramid Network. In Proceedings of the 2021 International Conference on Internet, Education and Information Technology (IEIT), Suzhou, China, 16–18 April 2021; pp. 244–248.
25. 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.
26. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *arXiv* **2022**, arXiv:2202.09741.
27. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent NN: First results. *arXiv* **2014**, arXiv:1412.1602.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
29. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J.S. Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 16–20.
30. Yuan, Z.W.; Zhang, J. Feature extraction and image retrieval based on AlexNet. In Proceedings of the Eighth International Conference on Digital Image Processing (ICDIP 2016), Chengdu, China, 20–22 May 2016; Volume 10033, p. 100330E.
31. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
32. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [[CrossRef](#)] [[PubMed](#)]
33. Ketkar, N. Stochastic gradient descent. In *Deep Learning with Python*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 113–132.
34. Ruby, U.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 5393–5397. [[CrossRef](#)]