applied sciences

MDPI

Article

Object-Aware Adaptive Convolution Kernel Attention Mechanism in Siamese Network for Visual Tracking

Dongliang Yuan [1], Qingdang Li [2], Xiaohui Yang [3], Mingyue Zhang [2] and Zhen Sun [4,*]

[1] College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China; qustydlmycr@163.com
[2] Chinesisch-Deutsche Technische Fakultat, Qingdao University of Science and Technology, Qingdao 266061, China; lqd@qust.edu.cn (Q.L.); zyy_2011@163.com (M.Z.)
[3] Faculty of Electrical Engineering and Computer Science, University of Kassel, 34001 Kassel, Germany; dekan@eecs.uni-kassel.de
[4] College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
* Correspondence: sunzhen@qust.edu.cn

Abstract: As a classic framework for visual object tracking, the Siamese convolutional neural network has received widespread attention from the research community. This method uses a convolutional neural network to obtain the object features and to match them with the search area features to achieve object tracking. In this work, we observe that the contribution of each convolution kernel in the convolutional neural network for object tracking tasks is different. We propose an object-aware convolution kernel attention mechanism. Based on the characteristics of each object, the convolution kernel features are dynamically weighted to improve the expression ability of object features. The experiments performed using OTB and VOT benchmark datasets show that the performance of the tracking method fused with the convolution kernel attention mechanism is significantly better compared with the original method. Moreover, the attention mechanism can also be integrated with other tracking frameworks as an independent module to improve the performance.

Keywords: visual object tracking; convolution kernel attention; Siamese convolutional neural network; convolutional neural network

check for updates

Citation: Yuan, D.; Li, Q.; Yang, X.; Zhang, M.; Sun, Z. Object-Aware Adaptive Convolution Kernel Attention Mechanism in Siamese Network for Visual Tracking. *Appl. Sci.* **2022**, *12*, 716. https://doi.org/10.3390/app12020716

Academic Editors: Shengzong Zhou and Yudong Zhang

Received: 3 December 2021
Accepted: 10 January 2022
Published: 12 January 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

Visual object tracking is a fundamental task in the field of computer vision. It has a wide range of application prospects in autopilot, robot navigation, drone cruise, intelligent security, human–computer interaction, etc. In the literature, researchers have proposed various visual object tracking methods. The visual tracking techniques are based on correlation filter-based methods, deep learning-based methods, and transformer-based methods. Recently, the methods based on deep learning have attracted the attention of the research community due to their advantage of end-to-end learning fashion. In particular, the object tracking methods based on Siamese convolutional neural networks have been developed rapidly due to their simple structures and efficiency. The Siamese convolutional neural network-based object tracking method uses deep convolutional features to describe objects. The deep convolutional features are extracted using multiple convolutional and activation layers. The feature maps of different layers have different characteristics. The use of features obtained using different convolution kernels to realize the effective distinction between the object and the background is a key issue for improving the effectiveness of features.

Based on the aforementioned analysis, in this work, we propose an adaptive convolution kernel attention mechanism for visual object tracking tasks. The main contributions of this paper include:

- The paper analyzes the effects of different convolution kernels on object tracking tasks and proves the necessity of convolution kernel attention.
- The paper analyzes the problems with the offline training-based attention mechanism in the tracking task and finds out the reason why this attention method is not suitable for object tracking.
- The paper presents an object-aware convolution kernel attention module. The proposed module improves the specificity of object features for different individuals based on the difference between each convolution kernel feature.

The proposed attention mechanism can be used as a general convolution kernel attention module and has good versatility. It can be used as a "plug and play" module by embedding it in various object tracking methods based on convolution feature matching, thus significantly improving the tracking performance of the original method. Finally, we combine the proposed convolution kernel attention module with classic object tracking methods. We perform experiments using OTB and VOT benchmark datasets to verify that the proposed technique significantly improves the object tracking performance.

2. Related Works

With the development of deep learning, the visual object tracking methods based on deep learning have received widespread attention from the research community. Among them, the Siamese convolutional neural network has become a research hotspot due to its simple network structure and end-to-end training ability.

The Siamese convolutional neural network was first proposed by Chopra et al. in 2005 [1] for face similarity discrimination. Bertinetto et al. [2] used this architecture for object tracking and proposed the SiamFC tracking method. Based on this method, Jack et al. [3] embedded the correlation filter as a differentiable layer in the SiamFC architecture and proposed the CFNet tracking method. Ran et al. [4] replaced the AlexNet with a better-performing VGG network and used region pooling to further improve the tracking performance. Wang et al. [5] added three attention modules to the template branch of the Siamese convolutional neural network to improve its ability to select object features. He et al. [6] proposed the SA-Siam method based on the original Siamese convolutional neural network and added a new group of Siamese networks. One group of networks tracks the object based on the appearance features of the image and the other group uses the semantic features of the template and the search area for tracking. The tracking results of the two groups of networks are merged to obtain the final tracking result, which further improves the robustness in complex scenarios. Gong et al. [7] proposed an online fast update method to reduce the computational complexity of the Siamese convolutional neural network. This method realizes the online fine-tuning of network weights. Among these derivative algorithms, the SiamRPN [8] method proposed by Li et al. combined the RPN module in Faster RCNN [9] with the Siamese convolutional neural network. After extracting the appearance features of the template image and searching the image, it is used as the input of the classification network in the RPN at the same time. By using the regression network, the tracking problem is transformed into a one-shot detection problem. This method is able to achieve excellent tracking performance. Its enhanced version, DaSiamRPN [10], aims to address the shortcoming of the original method, i.e., distinguishing objects of the same category, by adding interference training. This improves the network's ability to identify a single tracking object and avoid tracking failures caused by interference from similar objects of the same type.

With the development of tracking methods, the researchers have established that the deeper networks, such as ResNet, cannot achieve better results in the Siamese convolutional neural network tracking frameworks. In order to address this problem, Li et al. further studied the strict conversion invariance requirements implicit in the Siamese network tracking framework and proposed a method for applying deep networks in the Siamese neural network tracking frameworks [11]. Zhang et al. [12] proposed guidelines for the design of Siamese neural networks for object tracking. Wang et al. [13] proposed an unsupervised method for training Siamese neural networks. Zhou et al. [14] added a classification regression positioning branch in the original Siamese neural network, which improves the accuracy and robustness of the object tracking method. Zhang et al. [15] built a new type of Siamese neural network based on the inverse residual network to improve the performance of object tracking. Zeng et al. [16] used point regression as the main method to realize a coarse–fine classification network for visual tracking. Mu et al. [17] used a multi-layer Siamese network to consume high-level features for correcting low-level features and enhance the object response while maintaining object details to achieve real-time UAV tracking. It is evident that the object tracking method based on the Siamese network has become an important framework in the object tracking field as it has huge development potential in object tracking technology.

In the object tracking method based on the Siamese convolutional neural network, the main focus is to optimize and improve the network architecture. There are a few studies that discuss the role of different convolution kernel features in the object tracking process. In this work, we analyze the characteristics of different convolution kernels in the Siamese convolutional neural network. Moreover, we propose a new convolution kernel attention module to improve the performance of the object tracking method.

3. Convolution Kernel Feature Analysis

Usually, the convolutional neural networks have multiple convolutional layers, and each convolutional layer has multiple convolution kernels. Therefore, analyzing the response of each convolutional layer and convolution kernel can provide useful information for improving the object tracking method. First, in this work, we regard each frame in the video as an independent static image. After the image is passed through the convolutional neural network, a series of convolutional feature maps are formed. Based on the different levels of convolution feature extraction, those features are divided into low-level convolution features, middle-level convolution features, and high-level convolution features. In other words, the feature maps of different levels denote different levels of information extracted from the input image. For simplicity, in this work, we choose AlexNet in the SiameseFC tracking method to extract the convolutional features of each convolutional layer from the image. Then, we obtain the response of the object based on cross-matching, as shown in Figure 1. Since there are many channels in the high-level convolution features, only some typical features are extracted in order to highlight the key points.

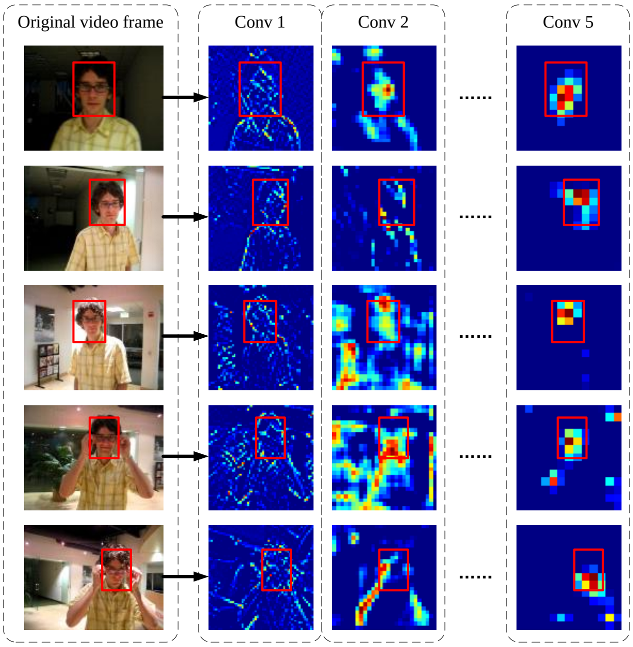

Figure 1. The response of different convolutional layers to the object in the input image.

Ma et al. [18] show that the high-level features (Conv5) mainly reflect the semantic features of the input image. By matching the semantic features, the object to be tracked can be distinguished from the background area. However, since the receptive field of a single pixel in high-level convolution features is relatively large, relying only on high-level convolution features in object tracking may provide an inaccurate position, which may reduce the tracking accuracy while ensuring robustness. At the same time, the low-level features (Conv2) retain more detailed information describing the instance level information, such as the boundary and texture of the object. Ma et al. [18], Zhu et al. [17], etc. have proposed to make full use of multi-layer features in order to improve the object features and eventually object tracking performance. Based on the feature analysis of the aforementioned convolutional layers, in this work, we further explore the role of different convolution kernels in each layer of convolutional features for specific tracking tasks.

The basic assumptions of this work are as follows.

- In the object tracking method, the objects, and scenes to be tracked are diverse.
- In different tracking scenarios, the convolution kernel activated by the object in a convolution layer may be different.
- During the tracking process, if all the convolution kernel features are used for feature matching, then the pertinence of the object is weakened, resulting in a lack of distinct characteristics of the object.

In order to verify these hypotheses, we use the "blurcar1", "coke", "david", "soccer", and "toy" video sequences in the benchmark set to perform experiments. The response of each object for different convolution kernels in the Conv5 convolution layer is shown in Figure 2.

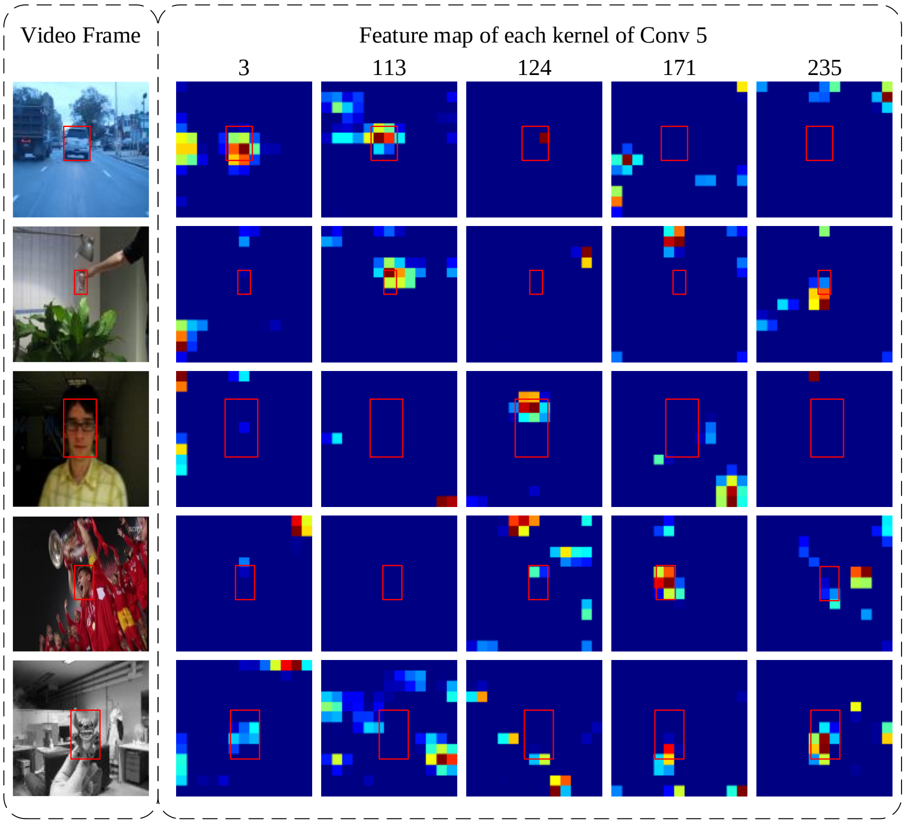

Figure 2. The response of different convolutions towards the object in the Conv5 convolution layer.

As shown in Figure 2, the objects and environments in different video sequences perform differently for each convolution kernel in the Conv5 convolutional layer. The response position of the "blurcar1" sequence in the third convolution kernel matches the position of the object contained in the image. The “coke”, “david”, “soccer”, and “toy” video sequences have better activation effects in the 113th, 124th, 171st, and 235th convolution kernels of Conv5 layer, respectively. It is evident from Figure 2 that the response maps of other convolution kernels are sparse and contain a lot of noise. Therefore, they are unable to reflect the location of the object in the image. In order to further analyze the contribution of different convolution kernels in object tracking tasks, we select two typical video sequences “soccer” and “jump” from the OTB benchmark dataset and analyze the matching response of the object in each convolution kernel. The response maps of each convolution kernel in the “soccer” video sequence are shown in Figure 3.

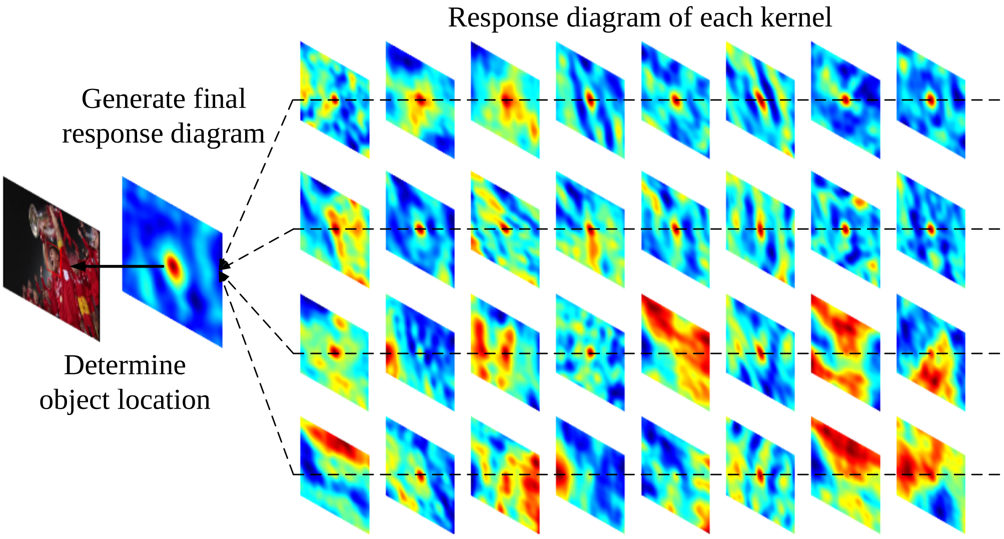
Response diagram of each kernel
Generate final response diagram
Determine object location

Figure 3. The response maps of different convolution kernels in response to the object’s position (“soccer” video sequence).

The leftmost image in Figure 3 presents the object tracking result, and the second image on the left denotes the matching response map of the object template and the search area. It is evident that the quality of the object response map is relatively good, and there is a significant peak at the real position of the object. However, it should be noted that the final response map is the result of the common matching of all convolution kernel features in the object convolution feature. If each convolution kernel feature of the object image is extracted separately and the matching response is performed, a response map of the object in each convolution kernel is obtained as shown in the figure. The figure also shows that some convolution kernels, such as 1–17, 20, 22, 26, and 30, have significant peaks in the object area (in this case, the central area). The overall distribution is similar to a two-dimensional Gaussian distribution, i.e., the convolution kernel in this part makes the primary contribution to the final object response map. At the same time, the response graphs of other convolution kernels, such as 18, 19, 21, 23–25, 27–29, 31, and 32, either have no significant peaks or have peaks at wrong positions. These convolution kernel features do not contribute to the final response map and may degrade its quality.

Similar to the “soccer” video sequence, the response maps of the “jump” video sequence for each convolution kernel are shown in Figure 4.

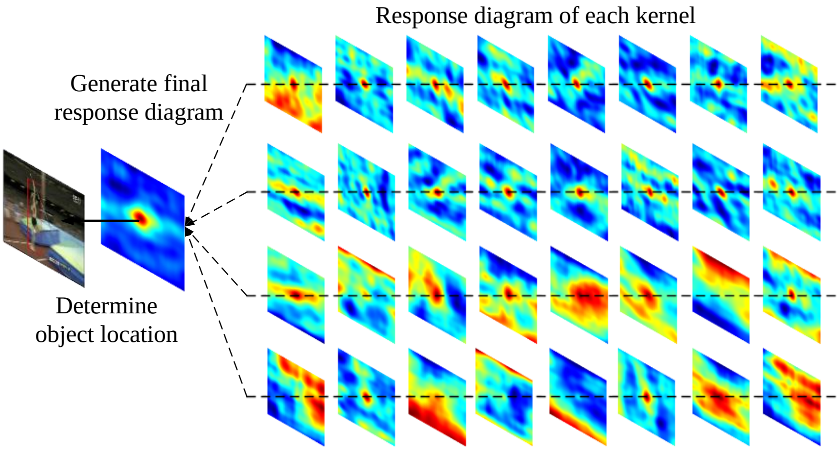
Response diagram of each kernel
Generate final response diagram
Determine object location

Figure 4. The response maps of different convolution kernels in response to the object's position ("jump" video sequence).

The response maps of the convolution kernel for the "jump" video sequence are similar to that of the "soccer" video sequence. The leftmost image in the figure denotes the result of object tracking, the second image on the left denotes the object matching response map, and the image on the right side of the figure represents the independent response map of each convolution kernel. As presented in the figure, for the "jump" video sequence, the response quality of each convolution kernel is also quite different. Among them, 2–17, 26, and 30 have better response quality and those significant peaks are in the center of the object. On the other hand, the response quality of other convolution kernels is poor. It should be noted that the convolution kernel with a good response for the "jump" video sequence is different from the "soccer" video sequence, which means that for different objects, the convolution kernels that contribute to the object response map are different.

Based on the aforementioned analysis, it can be concluded that for different tracking objects and in different tracking environments, the response levels of different convolution kernels in a convolutional layer for the same object are inconsistent. Consequently, their contribution to the tracking task is also inconsistent. Therefore, it is impossible to adapt to all objects by modifying the convolution structure of the Siamese convolutional neural network at one time. In order to achieve stable object tracking, it is necessary that the object tracking method adaptively selects different convolution kernels for different objects, makes full use of the useful features in convolution features, suppresses inefficient features, and achieves stable matching of object features. Based on the aforementioned analysis regarding the characteristics of the convolution kernel, we propose an object-aware adaptive convolution kernel attention module.

4. Object-Aware Adaptive Convolution Kernel Attention Mechanism

4.1. Analysis of the Attention Mechanism in the Pre-Training Stage

A typical attention module usually consists of a sub-neural network connected to the original neural network. The attention module and the original model are trained simultaneously to obtain the optimal parameters. In this work, we present the convolution kernel-oriented "Squeeze-and-Excitation" attention mechanism [19] (referred to as the SE attention module) based on the Siamese network object tracking framework. While the Siamese network uses a series of image pairs to train the network, the SE attention module has the ability to simultaneously learn the weight of each convolution kernel in the object tracking task and improve the neural network's feature extraction of the object sensitive area. The architecture of a typical SE attention module is shown in Figure 5.

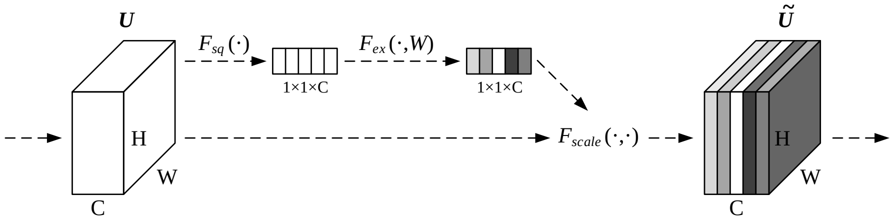

Figure 5. The architecture of the SE attention module.

where U represents the output convolution feature by a certain layer and $F_{sq}(\cdot)$ denotes a global information compression method for convolution feature U. Here, this method is implemented using global average pooling. The calculation process is expressed as follows:

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

$F_{\text{ex}}(\cdot, W)$ is a weighting process for z_c, where W represents the parameters used in the weighting process. These parameters are obtained automatically during end-to-end training of the deep learning network. When embedding the SE attention module in the Siamese network, we use ReLU and a fully connected layer to realize $F_{\text{ex}}(\cdot, W)$. Each convolution kernel $F_{scale}(\cdot,\cdot)$ uses the scalar $z_c{}'$ to multiply the original convolution feature $u_c \in \mathbb{R}^{H\times W}$ to achieve the weighting of the feature. The $F_{scale}(\cdot,\cdot)$ output weighted convolutional layer features are denoted by $\widetilde{U}$. The core of the SE attention module lies in the acquisition of the weighted parameter W. Under normal circumstances, the SE attention module should be combined and synchronized with the backbone network for training. The network architecture after embedding the SE attention module in the Siamese network object tracking framework is shown in Figure 6.

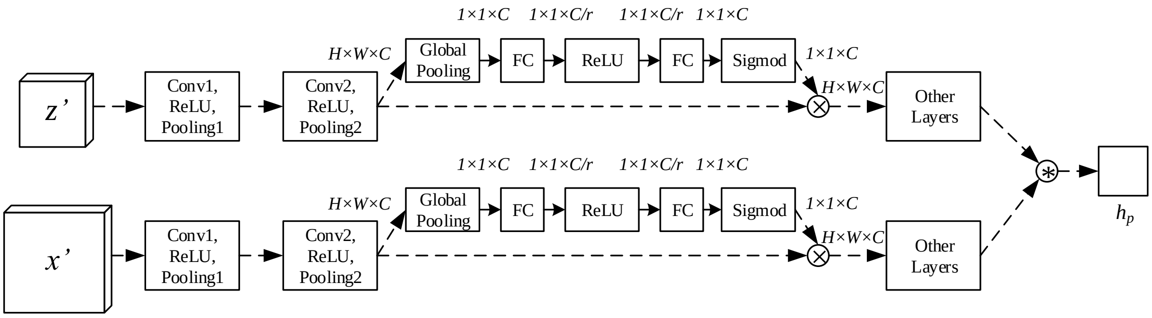

Figure 6. The structure of Siamese neural network with SE attention module.

During the training process, the SE attention module in the Siamese network learns the importance of each convolution kernel in the corresponding convolution layer based on back propagation. During the tracking process, the attention mechanism uses the importance weight of each convolution kernel to adjust the object convolution feature in order to improve the discrimination of the convolution feature and assist the object template correlation calculation for obtaining a more significant object response. It should be noted that the two branches of the Siamese network share the same parameters and that only one branch needs to be trained during the network training. In this work, we implement the aforementioned Siamese network architecture based on the Matconvnet framework [20] and embed the SE attention module in the Conv3, Conv4, and Conv5 convolutional layers of the SiamFC. We train the network using the same training process used by the classic Siamese network tracking method. After the training is completed, we use a Siamese network with SE attention mechanism to re-test the "soccer" video sequence and analyze its convolution kernel weights. The results and assigned weights of each convolution corresponding to the object are shown in Figure 7.

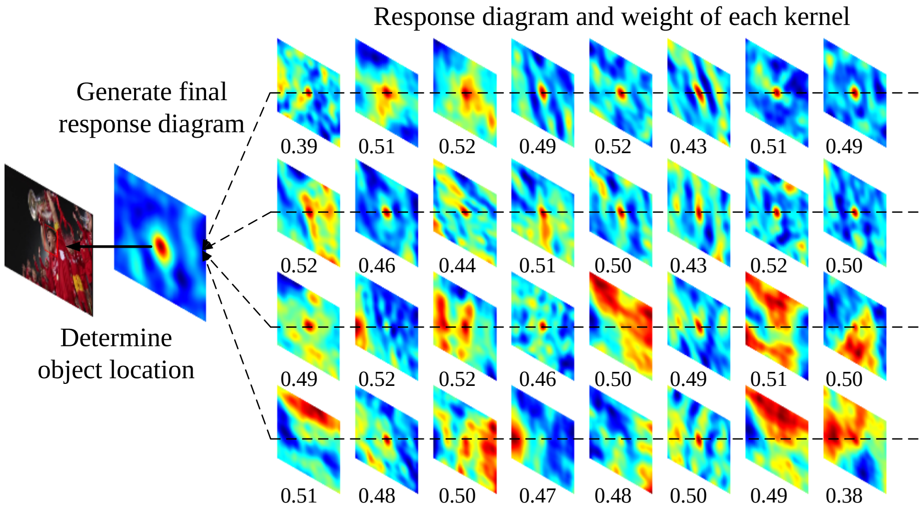

Figure 7. The results of SE attention module for each convolution kernel ("soccer" video sequence).

In Figure 7, the number below each convolution kernel response map represents the weight given by the SE attention module. It is evident that the weights of the SE attention module are not clearly distinguishable. The weights of most convolution kernels are between 0.4 and 0.6 and are not reasonably allocated according to the contribution of the convolution kernel. After analyzing the training process, we observe that the main reason for the failure of the SE attention module is that its weight acquisition method does not match the object tracking task and the Siamese network training method. Please note that the Siamese network is trained using a series of image pairs. Each pair contains the object template and the object search area. In order to obtain good generalization performance, the image pairs are acquired from various scenarios in the real world, such as pedestrians, vehicles, animals, etc. The analysis presented in Part 3 shows that the convolution kernels activated by different types of objects are different. This means that during the training process of the Siamese network, the meaning of each convolutional layer is constantly changing to correspond to the training data. Therefore, in this case, the convolution kernel weights learned by the SE attention module are repeatedly adjusted due to different training samples and are unable to converge. It is noteworthy that the attention mechanism commonly used for tasks, such as object detection, does not meet the requirements of object tracking. Therefore, a new convolution kernel weighting mechanism that meets the essential characteristics of object tracking is a dire need.

4.2. Object-Aware Convolution Kernel Attention Mechanism

Based on the analysis presented in this work, it is evident that the convolution kernels activated by different objects are different. The weighted selection of convolution kernels cannot be performed by the attention module used in ordinary end-to-end training. In order to effectively address this problem, we combine the characteristics of object tracking tasks and propose a new convolution kernel attention module for each individual object. This attention mechanism is different from the attention module traditionally applied during the training phase. The attention weight acquisition is not based on the training phase of the Siamese network. Instead, one-time training is performed at the beginning of the object tracking task. (1) At the beginning of the object tracking task, the object template is generated using the object position given in the initial frame, and the object template convolution feature is obtained by using the template branch of the Siamese network. (2) At the original object location, the search area is divided in the initial frame according to the size of the object search area set by the network. The candidate area branch of the Siamese network is used to obtain the convolution features of the candidate area. (3) The object template convolution feature and the candidate area convolution feature are both matched with each convolution kernel according to the template. At this time, the convolution kernel response maps with different response levels can be obtained. (4) Finally, based on the peak in each response map at the object position, we evaluate the support of the activated convolution kernel for the tracking task and set the convolution kernel weight used in this tracking. At the beginning of each tracking task, the above process is repeated, and different convolution kernel weight distributions are obtained based on different objects. The algorithm is shown in Table 1.

Table 1. The proposed object-aware convolution kernel attention mechanism framework.

Input	The First Frame I of the Video, the Position B of the Object in the Image.
Output	Convolution kernel weight w.
Step 1	Based on the object position B, select the object image Z from the image I. Take the location B as the center, select the image X in the search area according to the requirements of the candidate area network.
Step 2	Get the convolution feature $\varphi(Z)$ of the object image Z and the convolution feature $\varphi(X)$ of the image X in the search area. The number of channels of $\varphi(Z)$ is equal to the number of channels of $\varphi(X)$, denoted as C.
Step 3	*for* $i = 1$ to C Select the convolution feature $\varphi_\mathrm{i}(Z)$ of the i-th channel of $\varphi(Z)$ Select the convolution feature $\varphi_i(X)$ of the i-th channel of $\varphi(X)$ $S_i = \varphi_i\ (Z) * \varphi_i\ (X)$ $w_i = psr(S_i)$, where $psr()$ is defined by (2) *end*
Step 4	$w = \{w_i \mid i \in[1,C]\}$ Adjust the range of values in the set w to [0,4]
Step 5	Return the weight of the convolution kernel w

$$psr = \frac{p_{x_0,y_0} - \mu_s}{\sigma_s} \tag{2}$$

where x_0, y_0 represent the coordinates of the peak in the response map, and p_{x_0,y_0} represents the specific value of the peak. With the peak coordinate at the center, the area with radius r is called the peak area. The region outside the peak area is called the sidelobe area. μ_s and σ_s represent the average and standard deviation of all values in the sidelobe area, respectively. In this work, the value of psr is used to express the contribution of the convolution kernel response map for object tracking. It is used as the basis for the weight of the convolution kernel response map.

It should be noted that the final response map of the Siamese network is estimated based on the forward propagation of the network. It is not convenient to directly assign the weights to each convolution kernel response map. The analysis of the computations of the Siamese network shows that directly assigning the obtained weight to the object template feature is equivalent to assigning the weight to each convolution kernel response map.

The calculation process of the response map for object matching is expressed as:

$$f(Z,X) = \varphi(Z) * \varphi(X) \tag{3}$$

where Z and X represent the object image and the search area image, respectively, $\varphi(Z)$ denotes the convolution feature of the object image, $\varphi(X)$ denotes the convolution feature of the search area image, and $*$ represents the correlation operation. Essentially, the correlation operation of the multi-kernel convolution feature is the superposition of the correlation operation of each convolution kernel feature. Assuming that there are N convolution kernels, then:

$$f_i(Z,X) = \sum_{i=1}^{N} \varphi_i(Z) * \varphi_i(X) \tag{4}$$

where $\varphi_i(Z)$ and $\varphi_i(X)$ represent the i-th convolution kernel feature of the object image and the search area image, respectively. $\varphi_i(Z) * \varphi_i(X)$ denotes the correlation operation of the i-th convolution kernel feature, which is the response map of the i-th convolution kernel. Therefore, the final response map after weighting the different convolution kernel response maps is expressed as:

$$f_i(Z, X) = \sum_{i=1}^{N} w_i \times (\varphi_i(Z) * \varphi_i(X)) \tag{5}$$

where w_i denotes the weight of each convolution kernel response map obtained using the aforementioned attention module and × represents the element-wise multiplication of the weight and the corresponding response map. Since in the Siamese network, the correlation operation of the convolution feature is essentially a convolution operation, each element of the response map $f_i(Z, X)$ in the aforementioned formula is expressed as $f_i(Z, X)_{u,v}$. Now,

$$f_i(Z, X)_{u,v} = w_i \times \sum_{j,k=(u,v)}^{(u+M,v+M)} \varphi_i(Z)_{j,k} \times \varphi_i(X)_{j,k} \tag{6}$$

where (u, v) represents the coordinate of an element in the response map, (j, k) denotes the coordinate of an element in the convolution feature, M denotes the size of the object template convolution feature, $\varphi_i(Z)_{j,k}$ denotes the element in the object convolution feature, $\varphi_i(X)_{j,k}$ denotes the element in the convolution feature of the search area, and × represents multiplication by the element value. Therefore, the aforementioned expression is transformed as follows:

$$f_i(Z, X)_{u,v} = \sum_{j,k=(u,v)}^{(u+M,v+M)} \left(w_i \times \varphi_i(Z)_{j,k}\right) \times \varphi_i(X)_{j,k} \tag{7}$$

or

$$f_i(Z, X)_{u,v} = \sum_{j,k=(u,v)}^{(u+M,\ v+M)} \varphi_i(Z)_{j,k} \times \left(w_i \times \varphi_i(X)_{j,k}\right) \tag{8}$$

It is evident that (7) is equivalent to the following expression:

$$f_i(Z, X) = \sum_{i=1}^{N} (w_i \times \varphi_i(Z)) * \varphi_i(X) \tag{9}$$

or

$$f_i(Z, X) = \sum_{i=1}^{N} \varphi_i(Z) * (w_i \times \varphi_i(X)) \tag{10}$$

The aforementioned derivation shows that weighting the response maps of different convolution kernels is equivalent to weighting the features of the object template or the features of the search area. For convenience, we use the method of weighting the object template features to achieve the weighting of the convolution kernel response map.

Based on the SiamFC object tracking method, the attention method proposed above is combined with typical tracking sequences, such as “soccer” and “jump” to observe the weight assignment of each convolution kernel, as shown in Figures 8 and 9. The experimental results show that the attention module sets the weights according to the response of each convolution kernel. A high weight is assigned to a convolution kernel with a clear peak response at the object position. Similarly, a low weight is assigned to a convolution kernel without a clear object position. At the same time, since the attention module obtains weights at the beginning of the object tracking task, different convolution kernel weights can be assigned to different objects, which is more in line with the characteristics of the object tracking task.

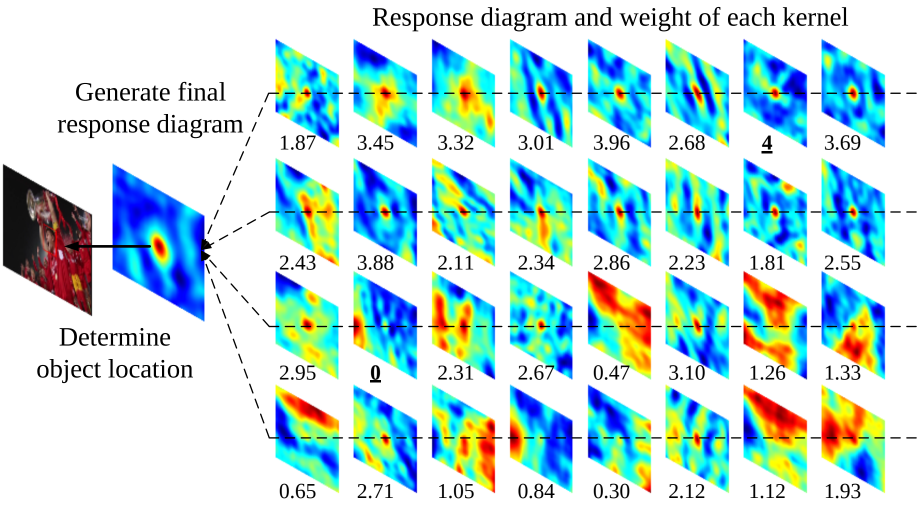

Figure 8. The proposed attention module weighting results for each convolution kernel ("soccer" video sequence).

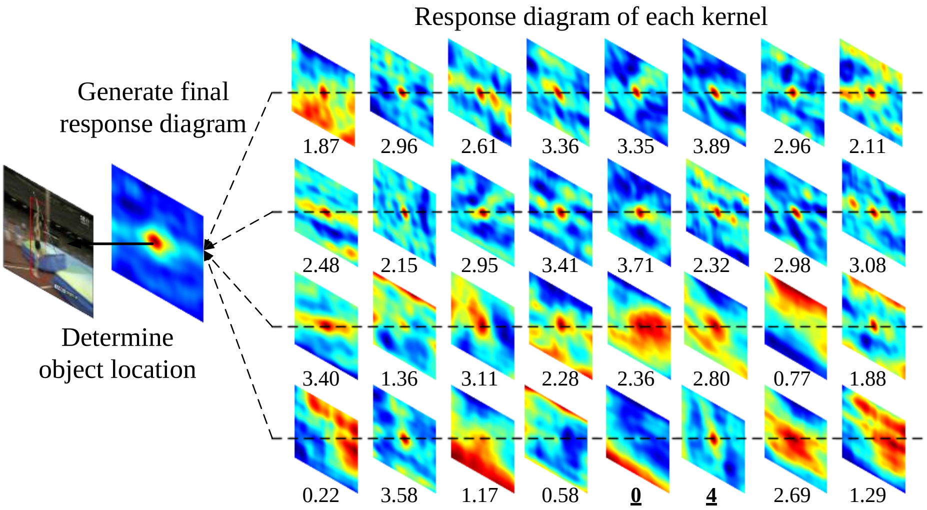

Figure 9. The proposed attention module weighting results for each convolution kernel ("jump" video sequence).

5. Experimental Verification

5.1. Experimental Design

In order to verify the effectiveness of the method proposed in this work, we use Ubuntu and Matconvnet [20] to implement the aforementioned convolution kernel attention model and combine it with the classic SiamFC tracking framework to form a new visual object tracking method, namely SiamFCCA. The proposed method is presented in Figure 10. In this work, the performance of the proposed SiamFCCA is tested by using the OTB [21,22] and VOT [23] benchmark datasets, and the results are compared with those of SiamFC and other existing object tracking methods. OTB and VOT are the two main benchmark datasets in the field of object tracking. By testing on the two benchmark datasets, the performance of the object tracking method can be reflected. The major implementation details of the model are as follows:

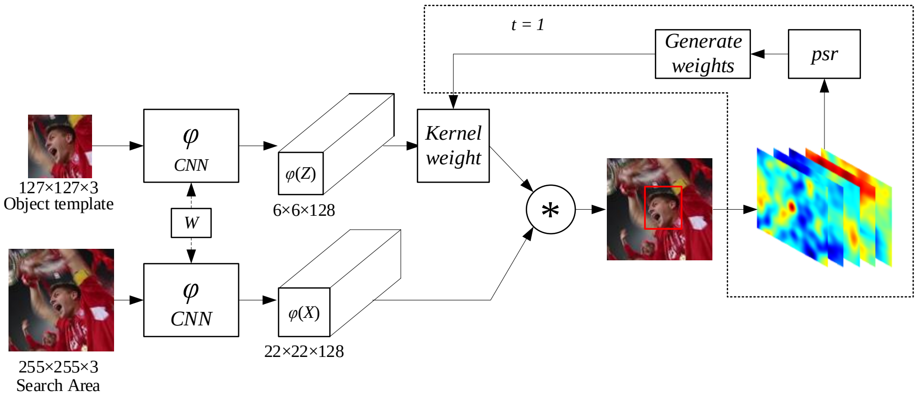

Figure 10. The SiamFCCA object tracking architecture.

(1) Backbone convolutional neural network: We use AlexNet as the basis for building the Siamese network and remove the boundary expansion operations on the pooling and convolutional layers. The resolution of the object template is 127 × 127, and the size of the candidate area is twice that of the object template.

(2) Convolution kernel attention module: In this work, we calculate the convolution kernel weights of the three convolution layers, i.e., Conv3, Conv4 and Conv5, using the initial frame. These are used as the weights of all object image features and search area features in the later stages.

(3) The Siamese network still uses a similar training method used by SiamFC. During the training process, the two branches use the same parameters. Please note that the convolution kernel attention module does not participate in the training process.

5.2. Qualitative Analysis

We analyze the performance of SiamFCCA and other tracking methods on the OTB and VOT benchmark datasets. The tracking results of some classic video sequences are shown in Figure 11.

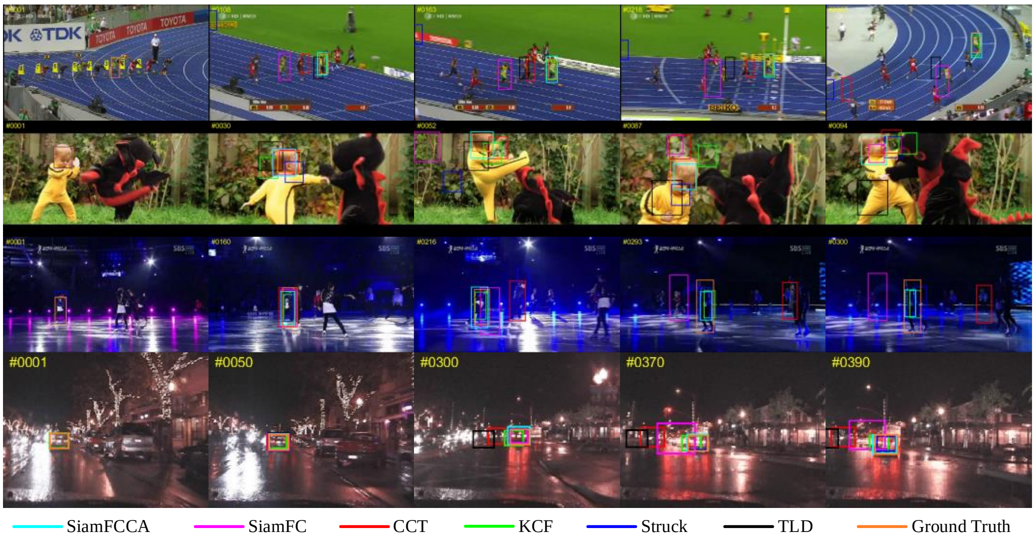

Figure 11. Qualitative comparison between SiamFCCA and other tracking methods.

The aforementioned experimental results show that the proposed SiamFCCA performs efficiently in the presence of background clutter, low resolution, and fast-moving objects compared with the original SiamFC and other tracking methods. This mainly relies on SiamFCCA's effective extraction and fusion of object appearance features, which enhance the convolution kernel features that are conducive to expressing the object characteristics and reduce the response of interfering objects in the cluttered background. At the same time, based on the suppression of other interference peaks in the response map, the object can still correctly respond to the object position, even when it is moving rapidly.

5.3. Quantitative Analysis

In order to clearly evaluate the performance of each object tracking method, we use the quantitative indicators in the OTB and VOT benchmark datasets to test the performance of each object tracking method. The indicators include the precision plots and success plots in OTB and the expected average coverage (EAO) in VOT. We test the performance of each tracking method for all the video sequences available in the OTB benchmark set. The accuracy and success maps are shown in Figure 12.

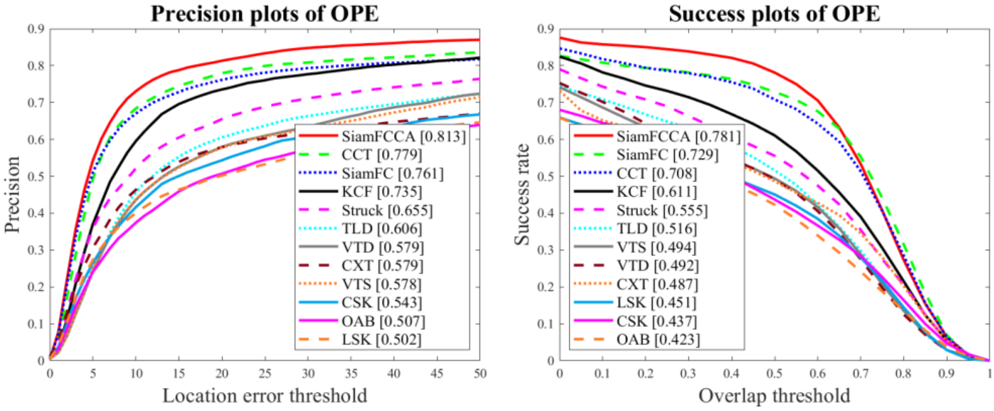

Figure 12. The precision and success plots of the proposed SiamFCCA and other tracking methods on all sequences.

Figure 12 shows that the overall performance of the proposed SiamFCCA on the OTB benchmark set is better compared with the improved SiamFC. Its accuracy and success rate are 6.8% and 7.1% better compared with SiamFC, respectively. In order to further compare the performance of each tracking method in each sub-category of the OTB benchmark set, we estimate the precision data and success plots of each sub-category, as shown in Table 2 and Figure 13.

Table 2. The comparison of precision scores of all trackers in all categories.

Tracker	All	FM	BC	MB	DE	IV	IPR	LR	OC	OPR	OV	SV
SiamFCCA	0.81	0.68	0.74	0.65	0.72	0.70	0.80	0.68	0.79	0.80	0.76	0.79
SiamFC [2]	0.76	0.68	0.72	0.60	0.67	0.69	0.72	0.56	0.72	0.74	0.69	0.76
CCT [2]	0.78	0.69	0.73	0.67	0.81	0.70	0.71	0.43	0.72	0.78	0.62	0.73
KCF [24]	0.74	0.58	0.75	0.62	0.75	0.72	0.72	0.38	0.73	0.72	0.61	0.67
Struck [25]	0.66	0.60	0.59	0.54	0.53	0.57	0.61	0.55	0.56	0.59	0.48	0.62
TLD [26]	0.61	0.54	0.43	0.50	0.52	0.54	0.58	0.35	0.56	0.59	0.53	0.60
CXT [27]	0.58	0.52	0.44	0.52	0.42	0.51	0.62	0.37	0.50	0.58	0.52	0.55
CSK [28]	0.54	0.37	0.59	0.32	0.48	0.49	0.55	0.41	0.50	0.54	0.36	0.49
VTD [29]	0.58	0.35	0.57	0.37	0.51	0.58	0.61	0.17	0.55	0.63	0.45	0.59
VTS [30]	0.58	0.35	0.58	0.37	0.50	0.59	0.58	0.19	0.54	0.61	0.44	0.57
LSK [31]	0.50	0.36	0.50	0.29	0.50	0.45	0.53	0.30	0.52	0.52	0.45	0.47
OAB [32]	0.51	0.42	0.45	0.36	0.50	0.40	0.48	0.38	0.49	0.51	0.41	0.53

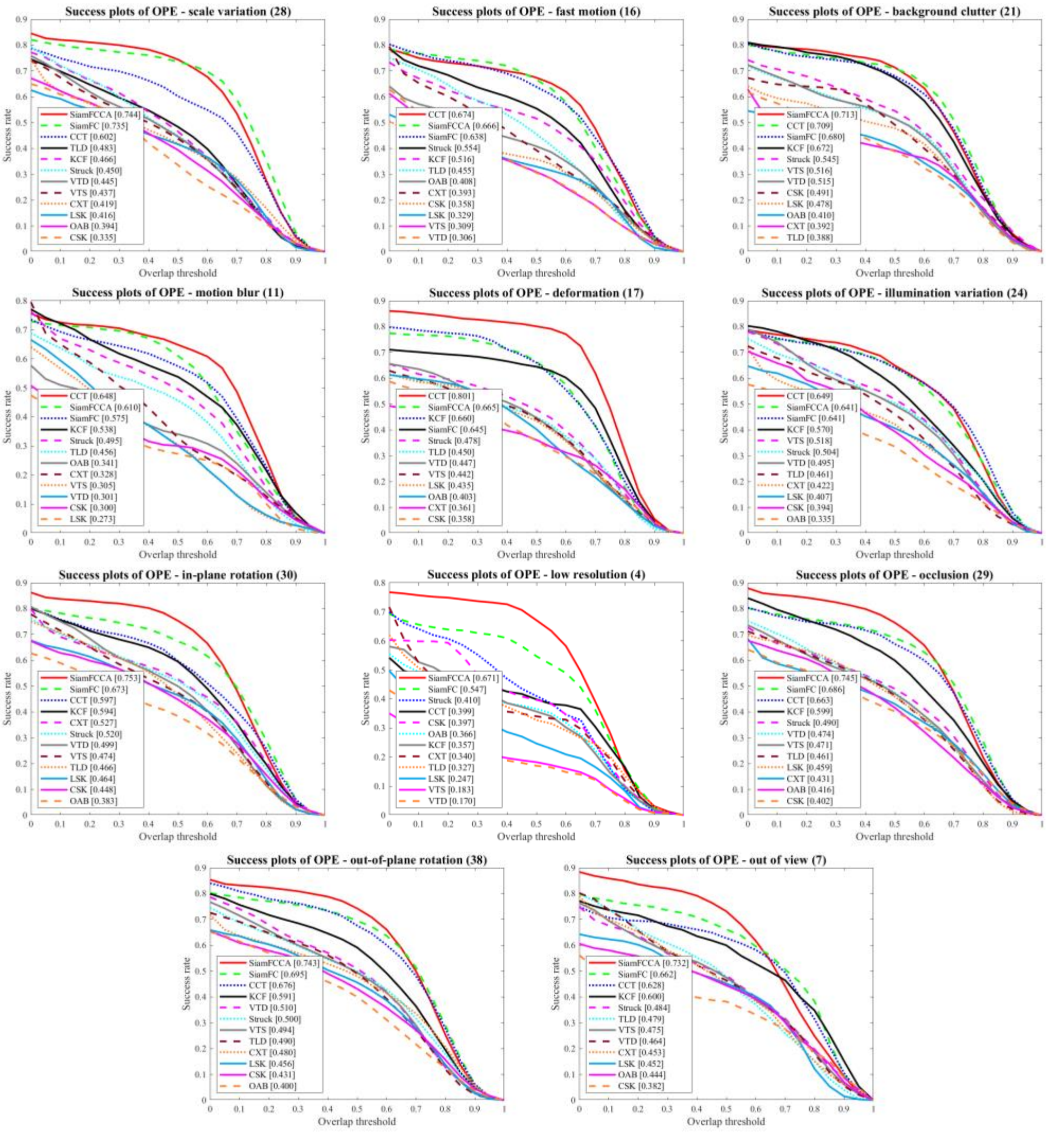

Figure 13. The success maps of the proposed SiamFCCA and other tracking methods for each sub-category.

By analyzing the success plots and precision data, it is evident that the SiamFCCA performs better compared with SiamFC in terms of most indicators. By analyzing the experiments of each tracking method performed using the OTB benchmark set, we observe that when the SiamFCCA is integrated with the convolution kernel attention module, its overall performance improves significantly compared with the SiamFC and other tracking methods. Especially in the “In-plane Rotation”, “Low Resolution” and “Out of View” sub-categories, the precision of SiamFCCA increases by 11.1%, 21.4%, and 10.1% compared with SiamFC, respectively.

We also test the proposed SiamFCCA object tracking method using the VOT benchmark set. The evaluation index is mainly reflected by the expected average coverage (EAO). The test results of each tracking benchmark set are shown in Table 3. It is evident from the test results that the tracking performance of the proposed SiamFCCA is better compared with SiamFC as the EAO score increases by 0.04 and its robustness score reduces by 14.98 compared with the original method.

Table 3. The performance comparison of trackers on the VOT benchmark.

Tracker	EAO	A	R
CCOT [33]	0.2674	0.4851	20.4138
CSRDCF [34]	0.2563	0.4849	23.5731
ECOhc [35]	0.2386	0.4896	28.7674
SiamFCCA	0.2125	0.4686	33.6019
SiamFC [2]	0.1736	0.5030	48.5851
Staple [36]	0.1688	0.5225	44.0194
SSKCF [36]	0.1661	0.5231	40.3088
MOSSEca [37]	0.1405	0.3953	50.0759
KCF [24]	0.1351	0.4445	50.0994
Struck [25]	0.0960	0.4142	80.3253
DSST [38]	0.0793	0.3913	95.5587
IVT [39]	0.0761	0.3883	104.737

The aforementioned experiments show that the SiamFCCA object tracking method proposed in this work is effective. The convolution kernel attention module adjusts the weight of convolution kernel features according to the object characteristics. Therefore, when the object resolution is relatively small or there are a large number of interfering objects, the convolution features output by the model still have a high response to the object. The tracking method does not cause tracking failure due to the response of other interference features.

6. Conclusions

In this work, based on the different contributions of different convolution kernels to the object tracking task, the attention mechanism of the convolution kernel in the Siamese network is explored. We analyze the reasons due to which the traditional convolution kernel modules are not suitable for object tracking. In this work, we propose an object-aware convolution kernel attention module. The experiments show that the weight of the convolution kernel estimated by this module meets the requirements of the object tracking task. Finally, we combine the proposed convolution kernel attention module with the classic SiamFC tracking method and present a new tracking method named SiamFCCA. The tests performed using the OTB and VOT benchmark datasets show that the proposed object tracking method has a significantly better performance compared with the original object tracking method.

Author Contributions: Methodology, Z.S.; software, Z.S.; validation, M.Z.; resources, Q.L.; data curation, X.Y.; writing—original draft preparation, D.Y.; writing—review and editing, Z.S. and Q.L.; visualization, Z.S.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tai Shan Scholar Foundation under Grant tshw201502042; the Shandong Province Key Research and Development Plan under Grant 2017CXGC0607 and 2017GGX30145; and the National Natural Science Foundation of China under Grant 61702295.

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

References

1. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a Similarity Metric Discriminatively with Application to Face Verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 539–546.
2. Bertinetto, L.; Valmadre, J.; Henriques, J.O.F.; Vedaldi, A.; Torr, P.H.S. Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 850–865.
3. Valmadre, J.; Bertinetto, L.; Henriques, J.O.F.; Vedaldi, A.; Torr, P.H.S. End-to-end representation learning for Correlation Filter based tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2805–2813.

4. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
5. 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.
6. He, A.; Luo, C.; Tian, X.; Zeng, W. A Twofold Siamese Network for Real-Time Object Tracking. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4834–4843.
7. Gong, K.; Cao, Z.; Xiao, Y.; Fang, Z. Online Update Siamese Network for Unmanned Surface Vehicle Tracking. In Proceedings of the 11th International Conference of Intelligent Robotics and Applications, Newcastle, Australia, 9–11 August 2018; pp. 159–169.
8. 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.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
10. 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.
11. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
12. Zhang, Z.; Peng, H.; Wang, Q. Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
13. Wang, N.; Song, Y.; Ma, C.; Zhou, W.; Liu, W.; Li, H. Unsupervised Deep Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1308–1317.
14. Zhou, W.; Wen, L.; Zhang, L.; Du, D.; Luo, T.; Wu, Y. SiamCAN: Real-time Visual Tracking based on Siamese Center-aware Network. *IEEE Trans. Image Processing* **2021**, *30*, 3597–3609. [CrossRef] [PubMed]
15. Zhang, F.; Qian, X.; Han, L.; Shen, Y. Inverted Residual Siamese Visual Tracking With Feature Crossing Network. *IEEE Access* **2021**, *9*, 27158–27166. [CrossRef]
16. Zeng, Y.; Zeng, B.; Yin, X.; Chen, G. SiamPCF: Siamese point regression with coarse-fine classification network for visual tracking. *Appl. Intell.* **2021**, 1–14. [CrossRef]
17. Zhu, M.; Zhang, H.; Zhang, J.; Zhuo, L. Multi-level prediction Siamese network for real-time UAV visual tracking. *Image Vis. Comput.* **2020**, *103*, 104002. [CrossRef]
18. 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 (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
20. Vedaldi, A.; Lenc, K. MatConvNet: Convolutional Neural Networks for MATLAB. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 689–692.
21. Wu, Y.; Lim, J.; Yang, M.-H. Online Object Tracking: A Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2411–2418.
22. Wu, Y.; Lim, J.; Yang, M.H. Object Tracking Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef] [PubMed]
23. Kristan, M.; Matas, J.; Leonardis, A.; Felsberg, M.; Pflugfelder, R.; Kamarainen, J.-K.; Cehovin Zajc, L.; Drbohlav, O.; Lukezic, A.; Berg, A.; et al. The seventh visual object tracking vot2019 challenge results. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
24. Henriques, J.F.; Rui, C.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *3*, 583–596. [CrossRef] [PubMed]
25. Hare, S.; Golodetz, S.; Saffari, A.; Vineet, V.; Cheng, M.M.; Hicks, S.L.; Torr, P.H. Struck: Structured Output Tracking with Kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2096–2109. [CrossRef] [PubMed]
26. Kalal, Z.; Mikolajczyk, K.; Matas, J. Tracking-Learning-Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1409–1422. [CrossRef] [PubMed]
27. Dinh, T.B.; Vo, N.; Medioni, G.E.R. Context tracker: Exploring supporters and distracters in unconstrained environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1177–1184.
28. Henriques, J.O.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In Proceedings of the European Conference on Computer Vision (ECCV), Berlin, Heidelberg, 7–13 October 2012; pp. 702–715.
29. Kwon, J.; Lee, K.M. Visual tracking decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1269–1276.
30. Kwon, J.; Lee, K.M. Tracking by sampling trackers. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1195–1202.

31. Liu, B.; Huang, J.; Yang, L.; Kulikowsk, C. Robust tracking using local sparse appearance model and k-selection. In Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1313–1320.
32. Grabner, H.; Grabner, M.; Bischof, H. Real-time tracking via on-line boosting. In Proceedings of the British Machine Vision Conference, Edinburgh, UK, 4–7 September 2006; p. 6.
33. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 472–488.
34. Lukezic, A.; Vojir, T.; Čehovin Zajc, L.; Matas, J.; Kristan, M. Discriminative correlation filter with channel and spatial reliability. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6309–6318.
35. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
36. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H. Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
37. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550.
38. Danelljan, M.; Ger, G.H.; Khan, F.S.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014.
39. Ross, D.A.; Lim, J.; Lin, R.-S.; Yang, M.-H. Incremental learning for robust visual tracking. *Int. J. Comput. Vis.* **2008**, *77*, 125–141. [CrossRef]