IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 9, 2020, accepted April 26, 2020, date of publication April 29, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991238

Attention-Based Siamese Region Proposals Network for Visual Tracking

FAN WANG[1], BO YANG[1], JINGTING LI[2], XIAOPENG HU[1], AND ZHIHANG JI[1,3]

[1]School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China
[2]Department of Financial Technology, Zhejiang Branch of China Construction Bank, Hangzhou 310016, China
[3]College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

Corresponding author: Xiaopeng Hu (huxp@dlut.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFA0704605.

ABSTRACT In this paper, we propose a multi-scale visual tracking algorithm based on attention mechanism to solve the problem that the appearance characteristic model of region proposals network has weak ability to distinguish foreground and semantic background. The method introduces attention mechanism on the basis of region proposals network to realize the self-adaptive salient characteristic expression. The attention mechanism is essentially realized by convolutional neural network. The feature optimization mainly includes spatial attention selection and channel attention selection. Specifically, the spatial attention convolutional neural network is used to learn the planar weights to enhance the foreground and suppress the interference background. The channel attention convolutional neural network is used to learn dimensional weights and discard redundant noisy feature maps to simplify appearance characteristic expression. In addition, spatial and channel attention network respectively deal with high-level and low-level features according to their structural differences to focus on the similarity appearance characteristic and semantic classification characteristic. The experimental results illustrate the outstanding performance compared with several state-of-the-art visual tracking methods on the challenging video sequences.

INDEX TERMS Attention mechanism, convolutional neural network, visual tracking.

I. INTRODUCTION

Visual tracking is one of the fundamental problems in image processing and computer vision. With the growing demand for artificial intelligence, visual tracking has been widely used in intelligent transportation [1], pavement detection [2], video monitoring [3] and other aspects. However, visual tracking still faces the challenges in complex scenarios, such as occlusion, illumination variation, background clutters and deformation.

Most tracking algorithms can be divided into two categories: generative and discriminative approaches. The generative methods describe the appearance characteristics of the target and minimize the reconstructed errors by searching the candidate target. The representative algorithms include sparse coding [4], [5], density estimation [6], principal component analysis [7] and so on. The generative methods simply focus on the target and ignore the background information. It is easily to lose the tracking target if the appearance changes drastically. The discriminative methods distinguish the target from the background by training the classifiers. This kind of methods is also called track-by-detection. The representative algorithms include multiple instance learning [8], boosting [9], structured SVM [10] and so on. The discriminative approaches are more robust than generative approaches. Nevertheless, the discriminative capability is greatly restricted because they both depend on low-level hand-crafted features.

With the continuous improvement of the computing power of modern intelligent equipment, deep learning has attracted extensive attention from researchers at home and abroad. To improve the accuracy and robustness of visual tracking task, more and more researchers start to use convolutional neural network (CNN) for visual object tracking. Several aspects including network structures [11]–[14] and updating mechanisms [15], [16] are studied. However, the online network updating and samples generation process take plenty of time, which extremely limits tracking speed. A kind of tracking algorithm based on convolutional neural network, which is named siamese network, abandons the online updating process, and it is pretrained with a large number of

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

image datasets to obtain the significant characteristic representation ability. Siamese network mitigates time-consuming problem and thus achieves real-time tracking. However, it still remains some problems. Firstly, the target template is not updated during the long-term tracking, which can easily lead to drift when target deformation and severely occlusion happen. Secondly, siamese network can only predict the target location, but cannot estimate current scale information. Besides, it is difficult to distinguish the foreground target and semantic background, and thus likely leads to drift problem.

To effectively solve the problem of multi-scale target representation, the siamese region proposals network (SiamRPN) combines siamese network with region proposals network. However, it failed to improve the ability to distinguish foreground and semantic background. In this paper, we introduce soft attention mechanism on the basis of SiamRPN to structure an adaptive appearance characteristic model, and thus improve the ability to discriminate foreground and semantic background. The attention mechanism mainly includes spatial attention and channel attention. On the one hand, the hourglass-shaped residual network is constructed to learn the plane weights and focus on the salient areas of two-dimensional feature maps. The main idea of spatial attention network is to enhance the foreground and suppress semantic background, and to assign different importance weights. On the other hand, channel attention network is constructed to learn the dimensional weights and focus on different characteristic types. The main idea of channel attention network is to eliminate redundant noisy feature maps and activate high target-relevant feature maps. As a result, the proposed method can efficiently distinguish the foreground and semantic background to avoid drift problem.

The contributions can be summarized as three folds:

1) We introduce soft attention mechanism on the basis of SiamRPN to structure an adaptive appearance characteristic model, and thus improve the ability to distinguish foreground and semantic background. The spatial attention network aims to enhance foreground and suppress semantic background to highlight characteristic otherness. At the same time, the channel attention network discards redundant information to obtain efficient characteristic expression.
2) According to the structural differences between spatial attention network and channel attention network, they deal with different characteristic gradations. Specifically, the spatial attention network deals with low-level features to learn appearance similarity characteristic. The channel attention network deals with high-level features to learn semantic classified characteristic.
3) Our experimental results demonstrate the outstanding performance of the attention-based multi-scale object tracking algorithm. The proposed method can significantly improve the ability to distinguish the foreground and background, and prevent the tracking results from quickly deviating the real target. The proposed method is compared to the state-of-the-art tracking algorithms in public benchmarks: OTB [17] and VOT [32].

The rest of the paper is organized as follows. We first review related work in Section II, and then discuss the detailed proposed method in Section III. Section IV illustrates the experimental results in public tracking benchmark.

II. RELATED WORK

A. TRACKERS BASED ON CNN

Convolutional neural network has widely been applied in object detection and recognition [18]–[23]. In recent years, there are more and more tracking algorithms based on CNN. At the beginning, Wang and Yeung [24] applies a deep model to visual tracking and acquires the characteristic of the target from the pre-trained Stacked Denoising Autoencoder (SDAE). Then they proposed SO-DLT algorithm to obtain the characteristic by using CNN, and the network achieves tracking by using long-term and short-term CNN. It is a successful application for CNN in visual tracking. To improve the accuracy of the tracking algorithms based on CNN, researchers generally study from characteristic representation, network modelling and update mechanism. For example, Nam *et al.* [14] proposed the MDNet tracker which is divided into shared layers and domain-specific layers. The tracking network is pre-trained by using the different domain-specific layers to avoid destabilizing the network. In [11]–[13], the network is pre-trained by using the ImageNet and other large-scale image datasets to obtain the efficient characteristic, such as the famous VGG-Net [25]. Nam *et al.* [14] proposed a tree structure which includes several CNN models to avoid the unreliable samples degrading the whole network. Held *et al.* [26] proposed the GOTURN tracker which performs the offline training without the online update. Its speed is fast, but it cannot adapt to the target deformation. Li *et al.* [27] update the network in a lazy way. Specifically, it is not updated until the target appearance changes a lot. In addition, there are other network types for visual tracking, such as Siamese Network [28] and Recurrent Neural Network [29].

B. TRACKERS BASED ON SIAMESE NETWORK

The essential of siamese network-based tracking algorithms is similarity comparison. This kind of trackers has the balanced accuracy and speed. Bertinetto *et al.* [28] proposed a fully convolutional siamese network to solve the similarity learning problem. The siamese network is offline pretrained using a large number of samples, and then predicts the tracking result according to the response graph obtained through cross correlation. Afterwards, CFNet [30] adds the correlation filter to the template branch to simplify original siamese network and make it more efficient. However, both of them need multi-scale samples generation which makes it time-consuming. To solve this problem, Li *et al.* [31] proposed the siamese region proposals network (SiamRPN), which uses region proposal subnetwork to generate multi-scale candidates. Benefit from the region proposal subnetwork,

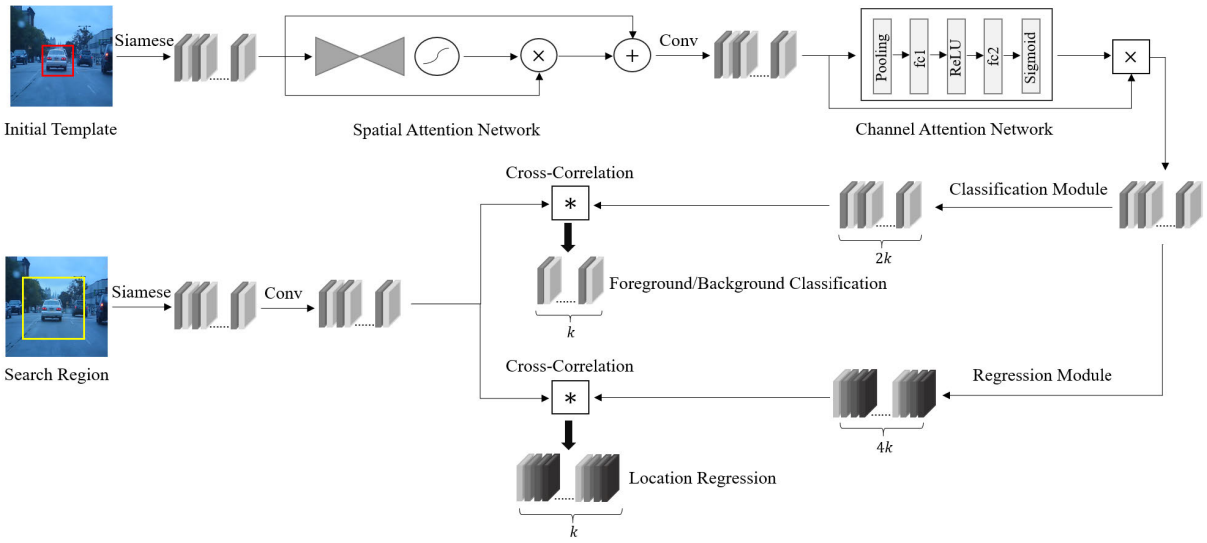


FIGURE 1. The overall framework of the proposed method.

traditional multi-scale test and online update can be discarded to achieve real-time tracking. However, it cannot improve the ability to distinguish the foreground and semantic background to effectively mitigate drift. Zhu *et al.* [38] proposed the distractor-aware siamese network, which introduces an effective sampling strategy to make the model focus on the semantic distractors, thus achieve accurate object tracking. Li *et al.* [39] proposed a simple and effective space-aware sampling strategy and successfully trained a tracker with significant performance improvements. ATOM [40] divided visual tracking into two parts: target classification and target evaluation. The former is used for coarse positioning and the latter is used for fine positioning. This two-stage tracking method can improve the accuracy of the tracker. Fan and Ling [41] proposed to concatenate a series of RPNs from high-level to low-level in a Siamese network framework to solve the positioning problem. Zhang and Peng [42] proposed to enhance the robustness and accuracy of tracking by using deeper and wider convolutional neural networks. In this paper, we introduce the soft attention mechanism on the basis of SiamRPN to structure the adaptive appearance characteristic model, which aims to enhance the foreground and suppress the semantic background.

III. ATTENTION-BASED SIAMESE REGION PROPOSALS NETWORK FOR VISUAL TRACKING

In this section, we mainly describe the proposed attention-based siamese region proposals network in detail. An overview of the proposed method is visualized in Fig. 1. The overall framework consists of the attention network and multi-scale region proposals network. The former learns the planar and dimensional weights by constructing spatial attention network and channel attention network, respectively. The latter constructs the anchor-based region proposals network to achieve multi-scale object tracking.

In the following, we first describe the overall procedure of our method. Afterwards, we elaborate the spatial attention network and channel attention network. Lastly, we describe the attention-based multi-scale tracking algorithm on the basis of the original siamese region proposals network in details.

A. OVERVIEW OF OUR APPROACH

The proposed method consists of the attention network and multi-scale region proposals subnetwork. The overall procedure of the proposed method is visualized in Fig. 2. The attention network mainly includes spatial attention network and channel attention network. The former learns the planar weights by constructing hourglass-shaped residual network to obtain the characteristic differences between the foreground and background. The latter learns the dimensional weights to eliminate redundant noisy feature maps and activate high target-relevant feature maps. Besides, the region proposals network consists of the classification module and regression module. The multi-scale region proposals network transforms the feature maps obtained from the attention network, and then the transformed target template and search region calculate the cross-correlation to obtain the response graph of classification probability and location regression.

The overall algorithm steps can be summarized as follows: Firstly, the siamese network is used to extract the features of initial target and search region. Next, the attention network is constructed to enhance the foreground and suppress the semantic background, so as to eliminate the redundant noisy feature maps and simplify appearance characteristic representation. Afterwards, the anchor-based region proposals network is used to achieve multi-scale target tracking. Finally, the search region is redetermined according to the predicted target location. The target template is fixed during the long-term tracking, and the above steps are repeated until

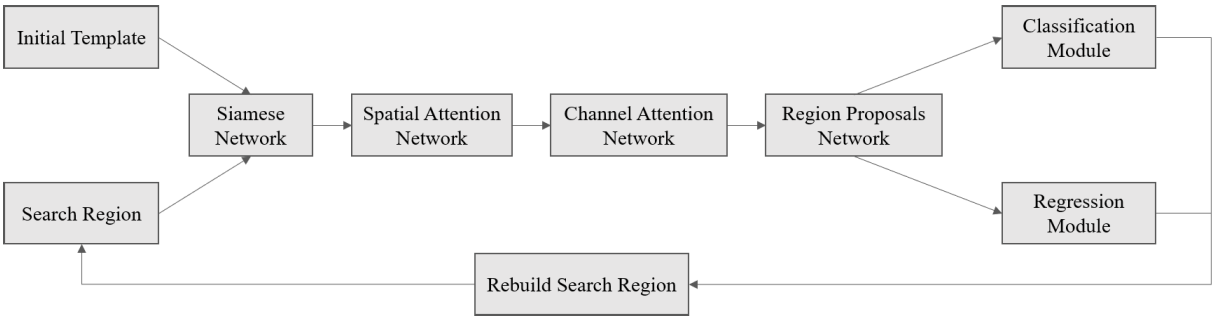


FIGURE 2. The overall procedure of the proposed method.

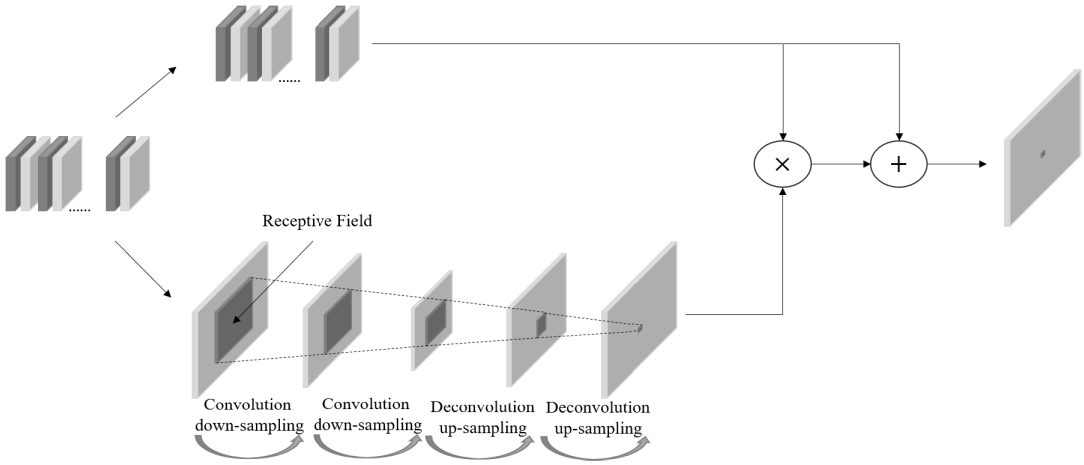


FIGURE 3. The spatial attention network structure.

the end of test sequence. The proposed method can significantly improve the ability to distinguish the foreground and background, and prevent the tracking results from quickly deviating from the real target, so as to effectively alleviate drift.

B. ATTENTION NETWORK

The essence of attention mechanism is to analyse the information obtained from vision and focus on the salient regions or objects, and then make use of the secondary information to assist in scene understanding, content recognition and other tasks. The proposed method introduces attention mechanism to focus on the difference between the foreground and semantic background, so as to improve the discrimination ability among different objects. The attention mechanism based on deep convolutional network can be divided into strong attention mechanism and soft attention mechanism. The non-differentiable nature of the strong attention mechanism makes it unsuitable for the back-propagation in deep networks. Therefore, we introduce the soft attention mechanism on the basis of original siamese region proposals network to obtain attention weights quickly. The attention network mainly includes spatial attention and channel attention network.

1) SPATIAL ATTENTION NETWORK

The spatial attention network adopts the hourglass-shaped residual network to highlight the foreground and suppress the semantic background. It reduces the size of feature maps by convolution and down-sampling to highlight the high-level semantic characteristics corresponding to the global receptive field. Afterwards, it expands the size of feature maps by convolution and up-sampling to amplify the activated salient foreground, so as to suppress the background and highlight the difference characteristics between the foreground and background. The structure of spatial attention network is shown in Fig. 3.

As Fig. 3 shows, the input feature maps extract the high-level characteristics through a series of convolution and down-sampling calculations. And then the size of feature maps is restored through deconvolution and up-sampling operations. At this time, the pixel values on the feature maps are the corresponding weights of the original feature maps. Specifically, the Sigmoid activation function is used to limit the pixel values of the weighted feature maps between 0 and 1. As a result, the weighted feature maps will not have obvious changes, and it can also suppress the interferential background information. The weighted feature maps are obtained by element-level multiplying the original feature maps and

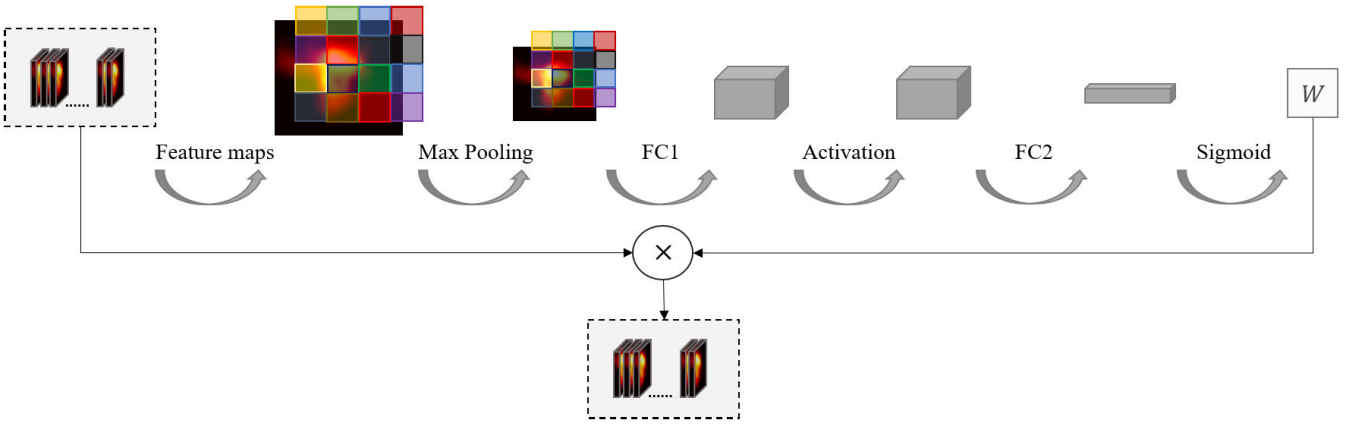


FIGURE 4. The channel attention network structure.

the corresponding weights, and its pixel values have been reduced. To avoid multiple weighting destroying data characteristics, the final spatial attention feature maps are obtained by adding the weighted feature maps and the original feature maps. Assuming that $F_o(x)$ represents the original feature maps, $F_w(x)$ represents the weighted feature maps, $F_s(x)$ represents the final spatial attention feature maps, $*$ represents the pixel-level multiplication and $+$ represents the pixel-level addition. The calculation process can be expressed as:

$$F_s(x) = F_o(x) + F_o(x) * F_w(x). \tag{1}$$

It is the extreme cases that the spatial attention feature maps are the original feature maps when the weighted feature maps $F_w(x) = 0$, which reflects the identical mapping idea of residual network. The spatial attention mechanism can enhance the foreground and suppress noisy background characteristics, so as to effectively improve the ability to distinguish the foreground and semantic background.

2) CHANNEL ATTENTION NETWORK

The channel attention network learns the dimensional weights to activate the high target-relevant characteristic types and suppress the insignificant characteristic channels, even eliminate the noisy feature maps, so as to obtain the efficient appearance characteristic representation. The high-level convolutional feature maps are essentially the semantic characteristic which is helpful for classification. The semantic characteristic is robust to deformation, but it also has the weak adaptability to the appearance changes. The channel attention network deals with the high-level characteristics can significantly improve the ability to distinguish the specific objects. The structure of channel attention network is shown in Fig. 4.

As Fig. 4 shows that the channel attention network learns the dimensional weights by the pooling and fully connection operation. It also uses the Sigmoid function to limit the weights between 0 to 1. As a result, the channel feature selection is completed by the elemental multiplication between the dimensional characteristics and the corresponding weights. The design principle of channel attention network is that the contribution of dimensional feature maps to the target characteristics representation is different, that is to say, different objects activate different channels of the feature maps. The role of channel attention network is to improve the weights which are high target-relevant, and to suppress the weights which are low target-relevant or noisy. The weights obtained from the channel attention network according to the initial target state remains fixed during tracking. Therefore, the whole network can not only enhance the characteristic difference between the foreground and background to improve the discrimination ability, but also significantly reduce the time-consumption.

C. SIAMESE REGION PROPOSALS NETWORK

The siamese region proposals network consists of the siamese network and region proposals network. The former is used to perform feature extraction, and the latter is used to generate multi-scale candidates for object tracking. The siamese region proposals network takes the tracking task as the single sample detection. It encodes the target appearance information into the correlated feature maps, that is, it extracts the candidates on the correlated feature maps of the target template and the search region to achieve multi-scale visual tracking. The region proposals network consists of the classification module and regression module. Assuming that z represents the target template, x represents the search region, $\phi(z)$ represents the feature maps of target template, $\phi(x)$ represents the feature maps of search region, $[\phi(z)]_c$ and $[\phi(x)]_c$ represents their feature maps in the classification module, $[\phi(z)]_r$ and $[\phi(x)]_r$ represents their feature maps in the regression module and the operator $*$ represents convolution. Then the correlated feature maps of classification module and regression module can be expressed as:

$$H^c = [\phi(x)]_c * [\phi(z)]_c. \tag{2}$$

$$H^r = [\phi(x)]_r * [\phi(z)]_r. \tag{3}$$

where the H^c represents the positive and negative activation of the anchor boundary boxes and H^r represents the distances

between the anchor boundary boxes and the ground-truth boundary boxes.

The siamese region proposals network learns the network weights by generating the positive and negative anchors. The positive anchor samples are labeled when the overlap rate between the anchor samples and ground-truth exceeds the ceiling threshold, and the negative anchor samples are labeled when the overlap rate between the anchor samples and ground-truth are below the floor threshold. The loss function of siamese region proposals network is composed of classification loss and regression loss. The classification loss is essentially the cross entropy, and the regression loss is the smooth L1 loss. Assuming that H_x, H_y, H_w and H_h represent the centre coordinate and scale of anchor boundary boxes, respectively. G_x, G_y, G_w and G_h represent the centre coordinate and scale of ground-truth boundary boxes, then the normalized distance can be expressed as:

$$\beta[0] = (G_x - H_x)/H_w. \tag{4}$$
$$\beta[1] = (G_y - H_y)/H_h. \tag{5}$$
$$\beta[2] = ln\,(G_w/H_w)\,. \tag{6}$$
$$\beta[3] = ln\,(G_h/H_h)\,. \tag{7}$$

The smooth L1 loss function can be expressed as:

$$smooth_{L1}\,(x,\theta) = \begin{cases} \frac{1}{2}\theta^2 x^2 & |x| < \frac{1}{\theta^2} \\ |x| - \frac{1}{2\theta^2} & |x| \geq \frac{1}{\theta^2} \end{cases} \tag{8}$$

The classification loss is known as the cross entropy, and the regression loss function can be expressed as:

$$L_r = \sum_{j=0}^{3} smooth_{L1}(\beta[j],\theta). \tag{9}$$

Then the overall loss function is weighted by the classification loss and regression loss, which can be expressed as:

$$L_t = L_c + \mu L_r. \tag{10}$$

where μ is the hyper-parameter to balance the weighted term. During the long-term tracking, the correlated feature maps of classification module and regression module can be expressed as the point collection:

$$H^c = \left\{\left(x_i^c, y_j^c, c_p^c\right)\right\}. \tag{11}$$
$$H^r = \left\{\left(x_i^r, y_j^r, d_{x_q}^r, d_{y_q}^r, d_{w_q}^r, d_{h_q}^r\right)\right\}. \tag{12}$$

where $i \in [0, w)$, $j \in [0, h)$, $p \in [0, 2k)$ in the correlated feature maps of classification module, and $i \in [0, w)$, $j \in [0, h)$, $p \in [0, k)$ in the correlated feature maps of regression module. Assuming that the siamese region proposals network needs to generate K candidates. The odd channel of classification feature maps represents the positive activation, then the K points with the highest score are retained, and the collection can be expressed as:

$$C_{cls} = \left\{(x_i^c, y_j^c, c_p^c)_{i\in I, j\in J, p\in P}\right\}. \tag{13}$$

where I, J and P represent the corresponding index collection, i, j and p represent the location and scale of anchor bounding boxes. Then the obtained anchors can be expressed as:

$$C_{anc} = \left\{(x_i^{anc}, y_j^{anc}, w_p^{anc}, h_p^{anc})_{i\in I, j\in J, p\in P}\right\}. \tag{14}$$

Similarly, the obtained bounding boxes of regression module can be expressed as:

$$C_{reg} = \left\{(x_i^r, y_j^r, d_{x_q}^r, d_{y_q}^r, d_{w_q}^r, d_{h_q}^r)_{i\in I, j\in J, p\in P}\right\}. \tag{15}$$

The obtained K candidates can be calculated by using the above anchor bounding boxes information, which can be expressed as:

$$x_i^{bb} = x_i^{anc} + d_{x_p}^r \times w_p^{anc}. \tag{16}$$
$$y_j^{bb} = y_j^{anc} + d_{y_p}^r \times h_p^{anc}. \tag{17}$$
$$w_p^{bb} = w_p^{anc} \times e^{d_{w_p}^r}. \tag{18}$$
$$h_p^{bb} = h_p^{anc} \times e^{d_{h_p}^r}. \tag{19}$$

To obtain more accurate predicted position and scale, the bounding box regression strategy is used to adjust the candidates.

D. ATTENTION-BASED MULTI-SCALE VISUAL TRACKING

The multi-scale visual tracking based on the attention network mainly consists of attention feature selection and multi-scale candidate bounding box generation. The former constructs the spatial attention network and channel attention network to deal with different planar regions and different feature types, the latter constructs the region proposals network to generate multi-scale samples. The siamese network is the pretrained AlexNet by using ImageNet dataset. At the same time, the overall network is trained offline by using ILSVRC and Youtube-BB datasets. During the training, the weights of previous layers in the siamese network are fixed, and the weights of last two layers are updated only. To give consideration to both the speed and characteristic adaptability, the network is finetuned online by using the initial target state only, and the network weights is fixed in the subsequent frames. The role of spatial attention network is to improve the adaptability of the high-level semantic characteristics to target deformation. The spatial attention network enhances the foreground and suppresses the semantic background by constructing the similar residual network. The role of channel attention network is to eliminate the redundant channels and retain the significant characteristic types. To prevent the pooling operation from filtering the useful information, the channel attention network is used to optimize the high-level semantic characteristic, so as to improve the ability to distinguish the target foreground and the semantic background.

IV. EXPERIMENT

Our method is implemented in Python based on the PyTorch framework and runs on a Titan X GPU with 6GB memory.

The proposed method is compared to the state-of-the-art tracking algorithms on the standard datasets, which are the online tracking benchmark (OTB) [17] and the visual object tracking benchmark (VOT) [32]. The experimental result shows the effectiveness and stability of the attention-based siamese region proposals network.

A. IMPLEMENTATION DETAILS

The input sizes of the template patch and search patch are $127 \times 127 \times 3$ and $255 \times 255 \times 3$. After passing through the Siamese Network, the template branch and search branch can get feature maps with dimensions of $6 \times 6 \times 256$ and $22 \times 22 \times 256$, respectively. The spatial attention network and the channel attention network are applied to achieve attention feature selection. The proposed spatial attention module applies two maxpooling with kernel sizes of 3×3 and 2×2 and then performs two up-sampling operations with output sizes of 3×3 and 6×6, followed by a ReLU activation and sigmoid activation. The proposed channel attention module applies a maxpooling with a kernel size of 3×3, and performs a fully-connected operation with an input dimension of $3 \times 3 \times 256$ and an output dimension of 256. Then the proposed channel attention module applies a ReLU activation, and performs a fully-connected operation with an input dimension of 256 and an output dimension of 256, followed by sigmoid activation. Then feature maps are input into the RPN network to obtain a classification feature map with a dimension of $17 \times 17 \times 2k$ and a regression feature map with a dimension of $17 \times 17 \times 4k$. k represents different ratios of anchor and the anchor ratios we adopted are [0.33, 0.5, 1, 2, 3]. In the offline training phase, the stochastic gradient descent (SGD) method with momentum of 0.9 is used to train the model. The initial learning rate is set to 1e-3 and the weight decay is set to 5e-5. The model is trained for 100 epochs with a maximum iteration number of 10000.

B. DATASET

We evaluate the proposed method on the public benchmark datasets OTB and VOT. The two benchmarks both contain plenty of sequences with the ground-truth labels and covers various challenging scenes, such as background clutters, motion blur, illumination variation, scale variation, occlusion, deformation and so on. Fig. 5 shows the ground-truth bounding boxes in the partial video sequences of OTB dataset.

C. EVALUATION METHODOLOGY

1) EVALUATION METHODOLOGY OF OTB

The evaluation methodology is mainly based on the precision and success plot. The precision plot essentially describes the centre location error. It is the Euclidean distance between the centre of the tracking result and ground truth bounding box, which can be expressed as:

$$\left\| E_p - E_g \right\| \leq T_p. \tag{20}$$

where E_p represents the centre position of predicted target, E_g represents the ground truth, T_p represents the threshold, $\|\cdot\|$ represents the Euclidean distance. The precision is defined as the percent of the amount of the frames whose centre location error is less than the corresponding threshold. It is changed along with the threshold. The ratio of frames with the threshold $T_p = 20$ is set to the final precision.

The success plot is used to describe the overlap ratio, we consider a frame successful if the overlap ratio is larger than the corresponding threshold. The overlap can be expressed as:

$$(S_p S_g)/(S_p S_g) \geq T_s. \tag{21}$$

where S_p represents the predicted bounding box, S_g represents the ground truth, T_s represents the threshold, the symbol represents intersection, represents union. Generally, we rank the results by the Area Under Curve (AUC) for the success plot. To accurately evaluate the proposed method, the comparative experiment employs one-pass evaluation (OPE).

2) EVALUATION METHODOLOGY OF VOT

The evaluation methodology is mainly based on the accuracy and the robustness. The accuracy is used to evaluate the accuracy of trackers. It can be expressed as:

$$\phi_t = \frac{A_t \bigcap A_{gt}}{A_t \bigcup A_{gt}} \tag{22}$$

where A_{gt} represents the ground truth of the t-th frame, and A_t represents the bounding box predicted by the tracker at the t-th frame. Assume that the tracker will run multiple times on a sequence. Define $\phi_t(i, k)$ as the accuracy of the i-th tracker on the t-th frame in the k-th repetition. Assuming the number of repetitions is N_{rep}, the accuracy at t-th frame is defined as:

$$\phi_t(i) = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} \phi_t(i, k) \tag{23}$$

The average accuracy of the i-th tracker is defined as:

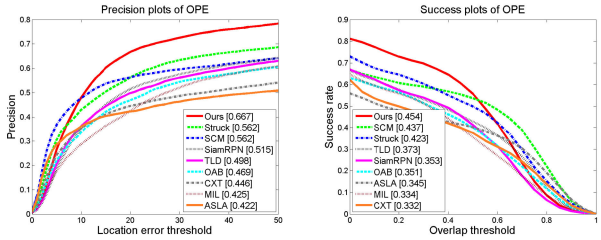
$$\rho_A(i) = \frac{1}{N_{valid}} \sum_{t=1}^{N_{valid}} \phi_t(i) \tag{24}$$

where N_{valid} is the number of valid frames. The robustness is used to evaluate the stability of the tracker. The larger the value is, the worse the stability is. Assuming that the intersection of the predicted bounding box and its ground truth in a frame is 0, it is considered to be a tracking failure. $F(i, k)$ is defined as the number of failures of the i-th tracker in the k-th repetition. The average robustness of the i-th tracker is defined as:

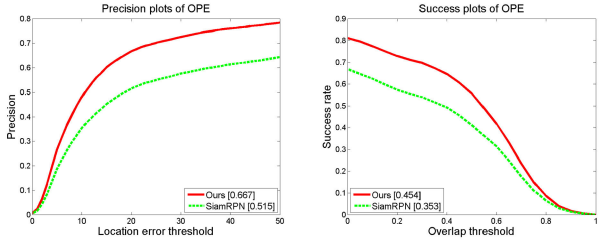
$$\rho_r(i) = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} F(i, k) \tag{25}$$

The robustness is used to evaluate the stability of the tracker. Based on the two metrics, Expected Average Overlap (EAO) is used to evaluate the performance of trackers.

FIGURE 5. The ground-truth bounding boxes in the challenging video sequences. The sequences contain various challenging scenes, such as Illumination Variation (IV), Scale Variation (SV), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Low Resolution (LR), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC).



(a) The precision and success plots of the proposed method (Ours) and other trackers.



(b) The curve comparison between the proposed method (Ours) and the benchmark SiamRPN tracker.

FIGURE 6. The precision and success plots illustrate the outstanding performance of the proposed method.

D. RESULTS ON OTB

The comparative experiment not only compares the performance between the proposed method and the benchmark SiamRPN algorithm, but also compares with the state-of-the-art trackers including TLD [33], OAB [34], MIL [8], CXT [35], Struck [10], SCM [36] and ASLA [37]. The experimental result shows the effectiveness and stability of the proposed method by drawing precision plots and success plots. Fig. 6 illustrates the precision and success plots based on centre location error and bounding box overlap ratio, respectively.

As Fig. 6 shows, the proposed method is obviously better than other comparison trackers with the 66.7% precision ratio and 45.4% success ratio. To objectively evaluate the performance of the proposed method, it is also compared with the benchmark SiamRPN to illustrate the improvement effect of introducing the attention mechanism. Compared with the state-of-the-art trackers, the proposed method makes use of the spatial attention network and channel attention network to extract the significant characteristics, so as to obtain the efficient appearance characteristic representation. The spatial attention network is used to suppress the background and highlight the difference characteristics between the foreground and background. The channel attention network is used to activate the high target-relevant characteristic types and suppress the insignificant characteristic channels, even eliminate the noisy feature maps.

In addition, the proposed method performs the evaluation under challenging attributes. The comparison curves are shown as Fig. 7 and Fig. 8. As Fig. 7 and Fig. 8 show, the proposed method has good accuracy and stability in the complex tracking scene, such as deformation (DEF), background clutter (BC) and occlusion (OCC). The reason is that the attention network can enhance foreground while suppress the semantic background to highlight characteristic otherness, so as to improve the ability to distinguish appearance characteristics.

To show the comparative experimental results more intuitively, the average precision scores and success scores are listed in Table 1 and Table 2. It clearly shows that the proposed method achieves an overall precision score of 0.667 and an overall success score of 0.454, whereas the benchmark SiamRPN are 0.515 and 0.353, respectively.

E. RESULTS ON VOT

The proposed tracker is compared on VOT with 8 the state-of-the-art trackers including OAB [34], MIL [8], CT [43], Struck [10], SiamRPN [31], MEEM [44], STC [45] and DSST [46]. Experimental results show that the proposed tracker achieves excellent performance in terms of accuracy and robustness.

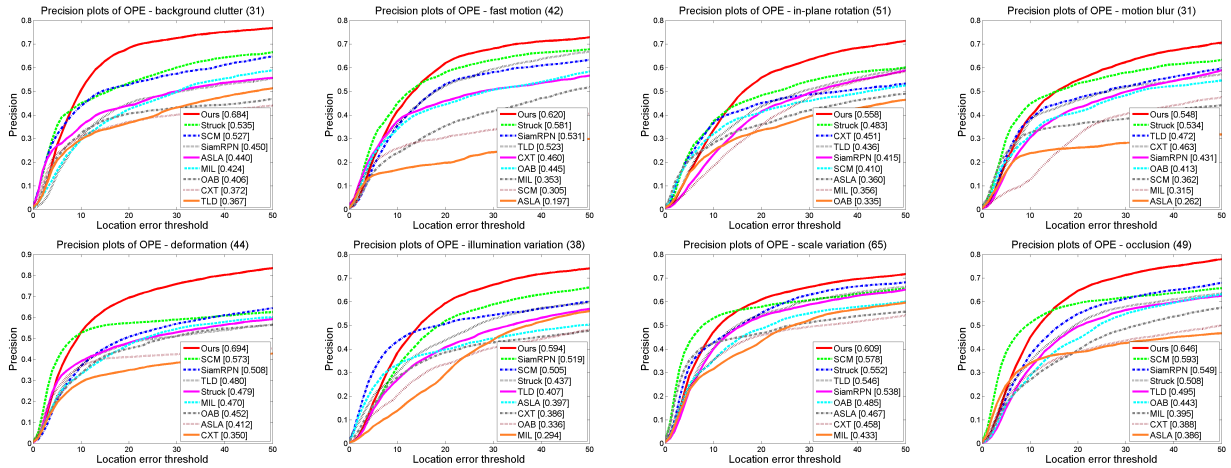


FIGURE 7. The precision plots of the proposed method under challenging attributes.

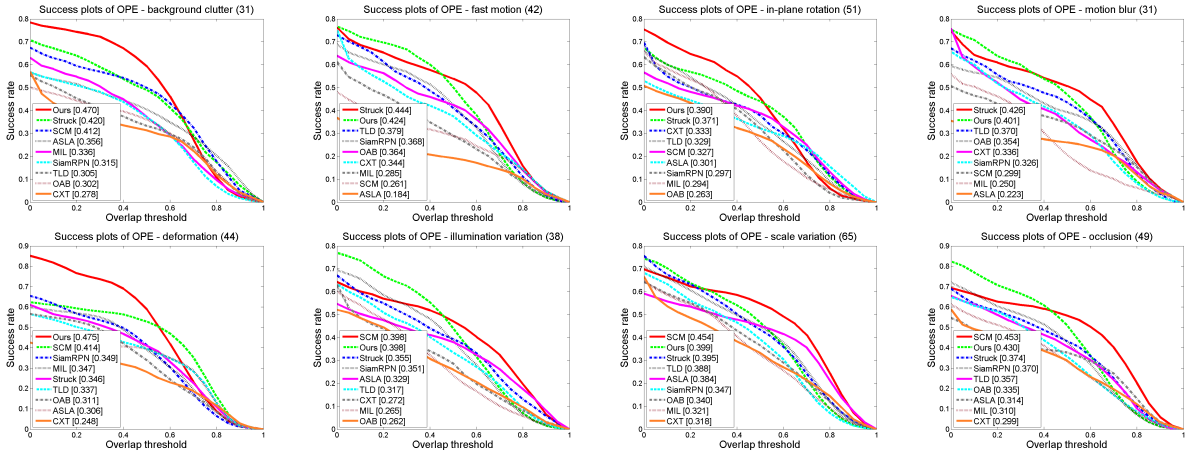


FIGURE 8. The success plots of the proposed method under challenging attributes.

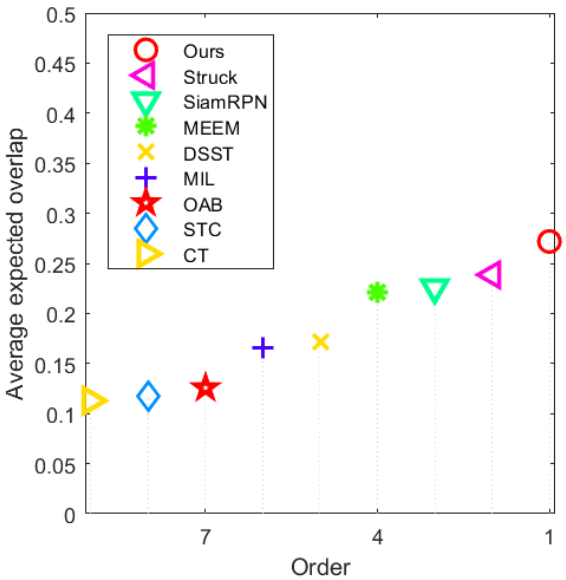


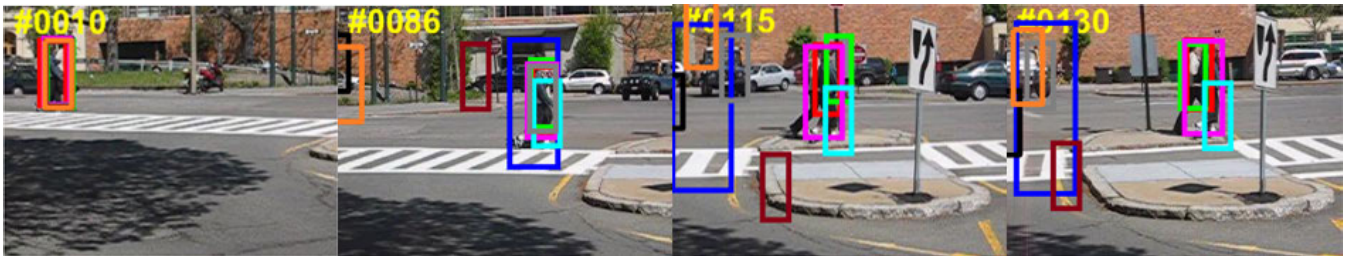
FIGURE 9. EAO ranking with trackers on the VOT dataset.

Fig. 9 shows the EAO curve evaluated on VOT dataset. To show the experimental results more intuitively, the accuracy, robustness and EAO scores are listed in Table 3. It can be seen that the proposed tracker performs well in terms of accuracy and robustness and shows a competitive EAO compared to other trackers. In baseline experiment, the proposed tracker achieves the best EAO score of 0.272. Besides, the proposed method achieved a robustness score of 0.245 and an accuracy score of 0.596, whereas the benchmark SiamRPN are 0.317 and 0.546, respectively. This demonstrates the effectiveness of the proposed attention-based Siamese region proposals network, which helps to distinguish the target foreground and the interference background.

F. QUALITATIVE RESULTS

To intuitively show the qualitative evaluation effect of the proposed method, Fig. 10 enumerates the detailed tracking results of the partial test sequence, such as CarDark, Couple, Faceocc1, Ironman and Singer2. It can be seen from the qualitative results that some trackers will be a large deviation between the predicted target and the real target under the challenging tracking scenes. For example, the CarDark sequence has the scene attribute of background clutter, which make it easily misjudge the semantic background as foreground and lead to drift. The Couple sequence

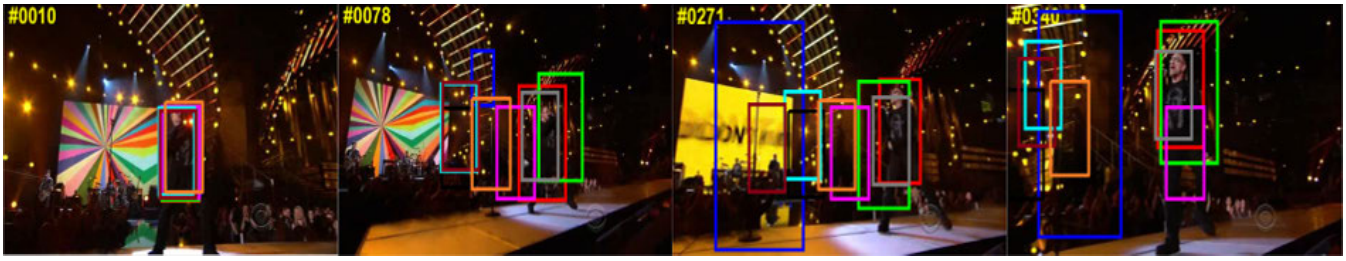
(a) CarDark



(b) Couple

(c) Faceocc1

(d) Ironman



(e) Singer2

Ours SiamRPN CXT ASLA TLD Struck MIL OAB SCM

FIGURE 10. The qualitative results of the proposed method on the OTB dataset. (CarDark, Couple, Faceocc1, Ironman and Singer2).

has the scene attribute of occlusion, and the tracking target deforms frequently, which increases the tracking difficulty. In the Faceocc1 sequence, the foreground is blocked by the semantic background for a long time, which increases the predicted error and makes it difficult to accurately estimate the location and scale information. In the Ironman sequence,

TABLE 1. The average precision scores among trackers on different attributes. The best and the second-best results are in red and green colours, respectively.

	TLD	OAB	MIL	CXT	Struck	SCM	ASLA	SiamRPN	Ours
BC	0.367	0.406	0.424	0.372	0.535	0.527	0.440	0.450	0.684
DEF	0.480	0.452	0.470	0.350	0.479	0.573	0.412	0.508	0.694
FM	0.523	0.445	0.353	0.460	0.581	0.305	0.197	0.531	0.620
IV	0.407	0.336	0.294	0.386	0.437	0.505	0.397	0.519	0.594
IPR	0.436	0.335	0.356	0.451	0.483	0.410	0.360	0.415	0.558
LR	0.349	0.376	0.171	0.371	0.545	0.305	0.156	0.292	0.469
MB	0.472	0.413	0.315	0.463	0.534	0.362	0.262	0.431	0.548
OCC	0.495	0.443	0.395	0.388	0.508	0.593	0.386	0.549	0.646
OPR	0.500	0.421	0.395	0.443	0.477	0.502	0.401	0.524	0.644
OV	0.488	0.320	0.386	0.330	0.441	0.178	0.167	0.498	0.541
SV	0.546	0.485	0.433	0.458	0.552	0.578	0.467	0.538	0.609
Overall	0.498	0.469	0.425	0.446	0.562	0.562	0.422	0.515	0.667

TABLE 2. The average success scores among trackers on different attributes. The best and the second-best results are in red and green colours, respectively.

	TLD	OAB	MIL	CXT	Struck	SCM	ASLA	SiamRPN	Ours
BC	0.305	0.302	0.336	0.278	0.420	0.412	0.356	0.315	0.470
DEF	0.337	0.311	0.347	0.248	0.346	0.414	0.306	0.349	0.475
FM	0.379	0.364	0.285	0.344	0.444	0.261	0.184	0.368	0.424
IV	0.317	0.262	0.265	0.272	0.355	0.398	0.329	0.351	0.398
IPR	0.329	0.263	0.294	0.333	0.371	0.327	0.301	0.297	0.390
LR	0.309	0.304	0.153	0.312	0.372	0.279	0.157	0.228	0.334
MB	0.370	0.354	0.250	0.336	0.426	0.299	0.223	0.326	0.401
OCC	0.357	0.335	0.310	0.299	0.374	0.453	0.314	0.370	0.430
OPR	0.365	0.305	0.312	0.314	0.363	0.386	0.319	0.351	0.428
OV	0.355	0.303	0.347	0.280	0.379	0.168	0.165	0.387	0.427
SV	0.388	0.340	0.321	0.318	0.395	0.454	0.384	0.347	0.399
Overall	0.373	0.351	0.334	0.332	0.423	0.437	0.345	0.353	0.454

TABLE 3. Experimental results on the VOT dataset. The best and the second-best results are in red and green colours, respectively.

	MEEM	DSST	OAB	MIL	CT	Struck	STC	SiamRPN	Ours
EAO	0.221	0.172	0.126	0.166	0.113	0.239	0.117	0.226	0.272
Accuracy	0.518	0.560	0.485	0.440	0.402	0.485	0.414	0.546	0.596
Robustness	0.312	0.484	0.711	0.511	0.664	0.266	0.663	0.317	0.245

the tracking target moves fast and irregular, which can result in the drift problem. The Singer2 sequence has the scene attributes of background clutter and scale variation, which make it unable to accurately predict the target state and easily cause failure.

The proposed method introduces the attention mechanism to focus on the difference between the foreground and the semantic background. The spatial attention network and channel attention network are used to obtain the salient characteristic representation of different target regions. As Fig. 10 shows, the proposed method can predict the location and scale information more accurately than other tracking methods and significantly reduce the prediction error, so as to improve the accuracy and robustness and achieve long-term object tracking.

V. CONCLUSION

In this paper, we propose a tracking method based on the attention mechanism, which focuses on the characteristic differences between the foreground and background. The method enhances the foreground and suppresses the semantic background to improve the ability to distinguish the foreground and the semantic background. The spatial attention network and channel attention network are constructed to realize salient feature selection, respectively. The former learns the planar weights by constructing the hourglass-shaped residual network, and the latter learns the dimensional weights to focus on different feature types. According to the structural differences of these two attention network, the spatial attention network deals with the low-level feature maps to focus on the appearance similarity characteristics. And the channel attention network deals with the high-level feature maps to focus on the semantic classification characteristics. The proposed method introduces the attention mechanism to simplify the characteristic representation and improve the ability to distinguish the foreground and the semantic background. The experimental result shows the outstanding performance of the proposed method compared with the benchmark SiamRPN tracker and several state-of-the-art methods on the public tracking benchmark.

The proposed tracking algorithm essentially belongs to the template matching methods. The update strategy of the target template can directly affect the performance of visual tracking. In the future, we can design the rational

and efficient template updating mechanism to improve the tracking performance.

REFERENCES

[1] A. Bhattacharyya, S. Bandyopadhyay, and A. Pal, "ITS-Light: Adaptive lightweight scheme to resource optimize intelligent transportation tracking system (ITS)—Customizing CoAP for opportunistic optimization," in *Mobile and Ubiquitous Systems, Computing, Networking, and Services*. Tokyo, Japan: Springer, 2013, pp. 1949–1955.

[2] K. Wang, Z. Huang, and Z. Zhong, "Simultaneous multi-vehicle detection and tracking framework with pavement constraints based on machine learning and particle filter algorithm," *Chin. J. Mech. Eng.*, vol. 27, no. 6, pp. 1169–1177, Nov. 2014.

[3] R. Venkatesan, P. D. A. Raja, and A. B. Ganesh, "Video Surveillance Based Tracking System," *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*. Kumaracoil, India: Springer, 2015, pp. 369–378.

[4] D. Wang, H. Lu, Z. Xiao, and M.-H. Yang, "Inverse sparse tracker with a locally weighted distance metric," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2646–2657, Sep. 2015.

[5] D. Wang, H. Lu, and M.-H. Yang, "Robust visual tracking via least soft-threshold squares," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1709–1721, Sep. 2016.

[6] L.-P. Zhang, L.-P. Wang, B. Li, and M. Zhao, "Kernel density estimation and marginalized-particle based probability hypothesis density filter for multi-target tracking," *J. Central South Univ.*, vol. 22, no. 3, pp. 956–965, Mar. 2015.

[7] Z. Y. Xiang, T. Y. Cao, P. Zhang, T. Zhu, and J. F. Pan, "Object tracking using probabilistic principal component analysis based on particle filtering framework," *Adv. Mater. Res.*, vols. 341–342, pp. 790–797, Sep. 2011.

[8] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[9] H. Grabner, C. Leistner, and H. Bischof, "Semi-supervised on-line boosting for robust tracking," in *Proc. Int. Conf. Eur. Conf. Comput. Vis.* Marseille, France: Springer-Verlag, 2008, pp. 234–247.

[10] S. Hare, S. Golodetz, A. Saffari, V. Vineet, M.-M. Cheng, S. L. Hicks, and P. H. S. Torr, "Struck: Structured output tracking with kernels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2096–2109, Oct. 2016.

[11] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3119–3127.

[12] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3074–3082.

[13] S. Hong, T. You, and S. Kwak, "Online tracking by learning discriminative saliency map with convolutional neural network," *Proc. Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 597–606.

[14] H. Nam, M. Baek, and B. Han, "Modeling and propagating CNNs in a tree structure for visual tracking," 2016, *arXiv:1608.07242*. [Online]. Available: http://arxiv.org/abs/1608.07242

[15] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4293–4302.

[16] N. Wang, S. Li, and A. Gupta, "Transferring rich feature hierarchies for robust visual tracking," 2015, *arXiv:1501.04587*. [Online]. Available: https://arxiv.org/abs/1501.04587

[17] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Jan. 2015.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Int. Conf. Neural Inf. Process. Syst.* Lake Tahoe, NV, USA: Curran Associates Inc, 2012, pp. 1097–1105.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[22] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[23] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[24] N. Wang and D. Y. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Int. Conf. Int. Conf. Neural Inf. Process. Syst.* Lake Tahoe, NV, USA: Curran Associates, 2013, pp. 809–817.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[26] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Int. Conf. In Eur. Conf. Comput. Vis.*, 2016, pp. 749–765.

[27] H. Li, Y. Li, and F. Porikli, "Robust online visual tracking with a single convolutional neural network," in *Computer Vision—ACCV*. Singapore: Springer, 2015, pp. 194–209.

[28] L. Bertinetto, J. Valmadre, and J. F. Henriques, "Fully-convolutional siamese networks for object tracking," in *Proc. Int. Conf. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 850–865.

[29] Z. Cui, S. Xiao, J. Feng, and S. Yan, "Recurrently target-attending tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1449–1458.

[30] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-End representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2805–2813.

[31] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.

[32] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, L. Cehovin, G. Fernandez, T. Vojir, G. Hager, G. Nebehay, R. Pflugfelder, A. Gupta, A. Bibi, A. Lukezic, A. Garcia-Martin, A. Saffari, A. Petrosino, and A. S. Montero, "The visual object tracking VOT2015 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 564–586.

[33] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012.

[34] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," in *Proc. Brit. Mach. Vis. Conf.*, 2006, pp. 47–56.

[35] T. B. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *Proc. CVPR*, Jun. 2011, pp. 1177–1184.

[36] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparse collaborative appearance model," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2356–2368, May 2014.

[37] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1822–1829.

[38] Z. Zhu, Q. Wang, and B. Li, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.

[39] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of siamese visual tracking with very deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4282–4291.

[40] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4660–4669.

[41] H. Fan and H. Ling, "Siamese cascaded region proposal networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7952–7961.

[42] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4591–4600.

[43] K. H. Zhang, L. Zhang, and M. H. Yang, "Real-time compressive tracking," in *Proc. Int. Conf. In Eur. Conf. Comput. Vis.*, 2012, pp. 864–877.

[44] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proc. Int. Conf. Eur. Conf. Comput. Vis.*, 2014, pp. 188–203.

[45] K. Zhang, L. Zhang, and Q. Liu, "Fast visual tracking via dense Spatio-temporal context learning," in *Proc. Int. Conf. Eur. Conf. Comput. Vis.*, 2014, pp. 127–141.

[46] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–5.

FAN WANG received the B.S. and M.S. degrees in computer science from Harbin Engineering University, in 1998 and 2001, respectively, and the Ph.D. degree in computer science from the Dalian University of Technology, China, in 2010. She is currently an Associate Professor with the School of Computer Science and Technology, Dalian University of Technology. Her research interests are wireless sensor networks, computer vision, and machine learning.

BO YANG received the B.E. degree from the Department of Computer Science and Technology, Dalian University of Technology, China, where she is currently pursuing the M.E. degree with the School of Computer Science and Technology. Her research interests include computer vision, visual tracking, and deep learning.

JINGTING LI received the B.E. degree from the Department of Computer Science and Technology, Dalian Maritime University, Liaoning, China, in 2016, and the M.E. degree from the Department of Computer Science and Technology, Dalian University of Technology, Liaoning, in 2019. She currently works with the Department of Financial Technology, Zhejiang Branch of China Construction Bank, Hangzhou, China. Her research interests include visual tracking, deep learning, and computer vision.

XIAOPENG HU received the Ph.D. degree in computer science from Imperial College London, U.K., in 2005. He is currently a Professor in computer science with the School of Computer Science and Technology, Dalian University of Technology, China. His research interests include computer vision, machine learning, and sensor fusion.

ZHIHANG JI received the B.S. and M.S. degrees in computer science and technology from the Henan University of Science and Technology, Luoyang, China, in 2003 and 2009, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Dalian University of Technology, Dalian, China. His research interests include image processing and pattern recognition.