

Received June 14, 2020, accepted July 31, 2020, date of publication August 17, 2020, date of current version September 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017179

Siamese Cascaded Region Proposal Networks With Channel-Interconnection-Spatial Attention for Visual Tracking

ZHOUJUAN CUI^{1,2}, JUNSHI AN¹, QING YE³, AND TIANSHU CUI^{1,2}

¹National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China

²University of Chinese Academy of Sciences, Beijing 100049, China

³International Education College, Henan University, Kaifeng 475001, China

Corresponding author: Zhoujuan Cui (constance669@126.com)


This work was supported by the Independent Deployment Foundation of Key Laboratory of Electronics and Information Technology for Space Systems of Chinese Academy of Sciences under Grant Y42613A32S.

ABSTRACT Trackers based on Siamese networks show great potential in tracking accuracy and speed. However, it is still challenging to adapt offline training model to online tracking. In this paper, a Siamese based tracker (SCRPN-CISA) is proposed, which integrates three attention mechanisms and a novel Cascaded Region Proposal Networks (RPN) architecture, for improving the feature extraction ability, adaptability and discrimination ability in complex scenes. Firstly, the deep network VGG-Net-D is adopted as the backbone network in the Siamese framework to increase the feature extraction capability. Then, a Channel-Interconnection-Spatial Attention module is constructed to enhance the adaptive and discriminative capability of the model. Next, a Deconvolution Adjust Block is built to fusion cross-layer features. Finally, a Three-Layer Cascaded RPN is conceived to acquire the foreground-background classification and bounding box regression by correlation calculation, and moreover, a proposal region screening strategy is presented to obtain more accurate tracking results. Experiments on OTB-2015, UAV123, VOT2016, and VOT2019 benchmarks demonstrate that, the proposed tracker (SCRPN-CISA) achieves competitive performance compared with the state-of-the-art trackers.

INDEX TERMS Visual tracking, Siamese networks, channel attention, interconnection attention, spatial attention, cascaded region proposal networks.

I. INTRODUCTION

Visual tracking is an interdisciplinary discipline that integrates theories of feature extraction, information analysis, machine learning, and computer vision. A typical scenario for visual tracking is to provide a bounding box in the first frame of the video to indicate the location of the object of interest, model the appearance and motion information of the object, and estimate the location of the object with higher possible accuracy in subsequent frames. With the rapid development of computer technology, image processing technology and artificial intelligence technology, visual tracking is widely used in visual surveillance, intelligent navigation, autopilot, human-computer interaction, military guidance, aerospace and other scenarios. As making a sea of remarkable progress,

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

visual tracking is still been recognized as an extremely challenging task for a large number of factors such as deformation, rotation, motion blur, illumination variation, background clutters, and occlusion.

As deep learning methods in image classification and object detection have made breakthrough progress, researchers have gradually introduced them into the field of visual tracking. For instance, HCF [1], C-COT [2], and ECO [3], trackers which combined with correlation filters framework and the deep convolutional features, improve the accuracy of the trackers. However, the increasingly high feature dimensions bring huge computational overhead to the online learning and update process, which directly impacts the tracking speed. The application of deep learning methods is not only limited to pre-trained deep features, but also includes an end-to-end Siamese framework. SiamFC [10] creatively transforms the tracking task into a similarity learning

problem through the Siamese networks. SiamRPN [11] accesses a Regional Proposal Network to further improve tracking performance. These trackers have achieved impressive results on all recent benchmarks [12], [13] and challenges [14]–[17].

Although these end-to-end trackers complete a balance of speed and accuracy, there are still some problems. The first problem is that the deep model learned offline does not adapt well to the online tracking process. For instance, when a category not included in the offline training datasets appears during the tracking process, the similarity learning approach is not necessarily reliable and the generalization ability is relatively weak. For example, when the object itself generates a large deformation comparing to the first frame, the stronger the offline model matching ability, the lower the possible similarity score, and the easier it is to get the wrong judgment. However, if the matching ability is weak, it is difficult to judge when similarities interfere around the object. The second problem is that the template branch operates independently of the detection branch in the Siamese networks, which makes the background information to not well utilize. Actually, it is also extremely important for the location of the object and the distinction of similarities. Therefore, we adjust the architecture of the model to acquire more efficient features, while introducing attention mechanisms into the model to generate more adaptive discriminant learning.

In particular, we design a Siamese-based tracker that devises the deep convolutional network VGG-Net-D [18] as the backbone to further enhance the feature representation by deepening the network. We explore several attention mechanisms, including channel attention, interconnection attention and spatial attention. Then, what designing a new attention block, which enhances the ability to distinguish interference sources and complexing backgrounds, makes offline training models with superior adaptability to track online. In order to obtain more accurate position, we further present a Three-Layer Cascaded RPN, to achieve the effective fusion of multi-layer features and to solve the problem of imbalanced training samples.

To summarize, the main contributions of our work are demonstrated as follows:

- We adopt the deeper convolutional network VGG-Net-D as the backbone to make the model more capable of feature representation, and then an end-to-end tracker based on Siamese Deep Network is trained successfully.
- We propose a novel Channel-Interconnection-Spatial Attention module, to capture the difference the object foreground with the semantic background, to connect the relationship between the object template and the search image, to achieve the enhancement of the object and the suppression of interference, so that online tracking possesses better adaptability and discriminating ability.
- We design a Deconvolution Adjust Block for crossing layer fusion of deep semantic features and shallow spatial features, which further optimizes the ability to distinguish complex backgrounds.

- We construct a Three-Layer Cascaded RPN, and output the classification scores and regression offsets for anchors at each RPN level. Three-layer Cascaded RPN filters negative samples carrying of little information from top to bottom. Meanwhile, multi-step regression gradually refines and adjusts the bounding box to achieve precise position.

The rest of this paper is organized as follows. The related work of trackers based on deep learning (DL) methods is introduced in Section II, and the attention mechanism also discussed in this section. The details of our work are presented in Section III. Experiments and results based on OTB-2015 [13], UAV123 [19], VOT2016 [14] and VOT2019 [17] benchmarks are manifested in Section IV. Moreover, the analysis is also discussed in this section. Finally, this paper is concluded in Section V.

II. RELATED WORK

As having been established benchmarks [12], [13], [19], [20], [21], hold annually tracking challenges [14]–[17] and improved methodologies, visual tracking has made brilliant progress.

A. DEEP FEATURE BASED TRACKERS

Algorithms based on correlation filters derived from signal processing theory become a focused area [22]–[26], which utilize correlation operations in the frequency domain, thus, both speed and accuracy accomplishing a qualitative leap. Deep learning methods demonstrate exceptional potential in computer vision tasks such as image classification. Furthermore, deep convolutional networks, with their powerful generalization ability and migration ability, are gradually introduced into visual tracking tasks for feature extraction, which greatly improve the tracking accuracy. HCF [1] utilizes pre-trained networks to extract deep and shallow features, depending on the characteristics of different layers, and respectively trains correlation filters to obtain response maps weighted fusion. C-COT [2] interpolates the feature maps of different resolutions extracted by the pre-trained networks into the continuous spatial domain, combining with multi-resolution continuous convolution filters for training and detecting, and gains more accurate position. ECO [3] optimizes from the three aspects of filter coefficients, sample division and template update strategy, while maintaining the tracking accuracy and greatly improving the tracking speed. Researchers also contribute to explore new methods which refer to object detection. They achieve excellent tracking results [4]–[8], and even in the field of Multi-Object tracking [9].

B. END-TO-END TRACKERS

Not only confine to the application of pre-trained deep networks, but also introduce the Siamese networks to train special end-to-end tracking networks. SINT [27] pioneers the introduction of the Siamese networks to transform object tracking tasks into a similarity learning problem.

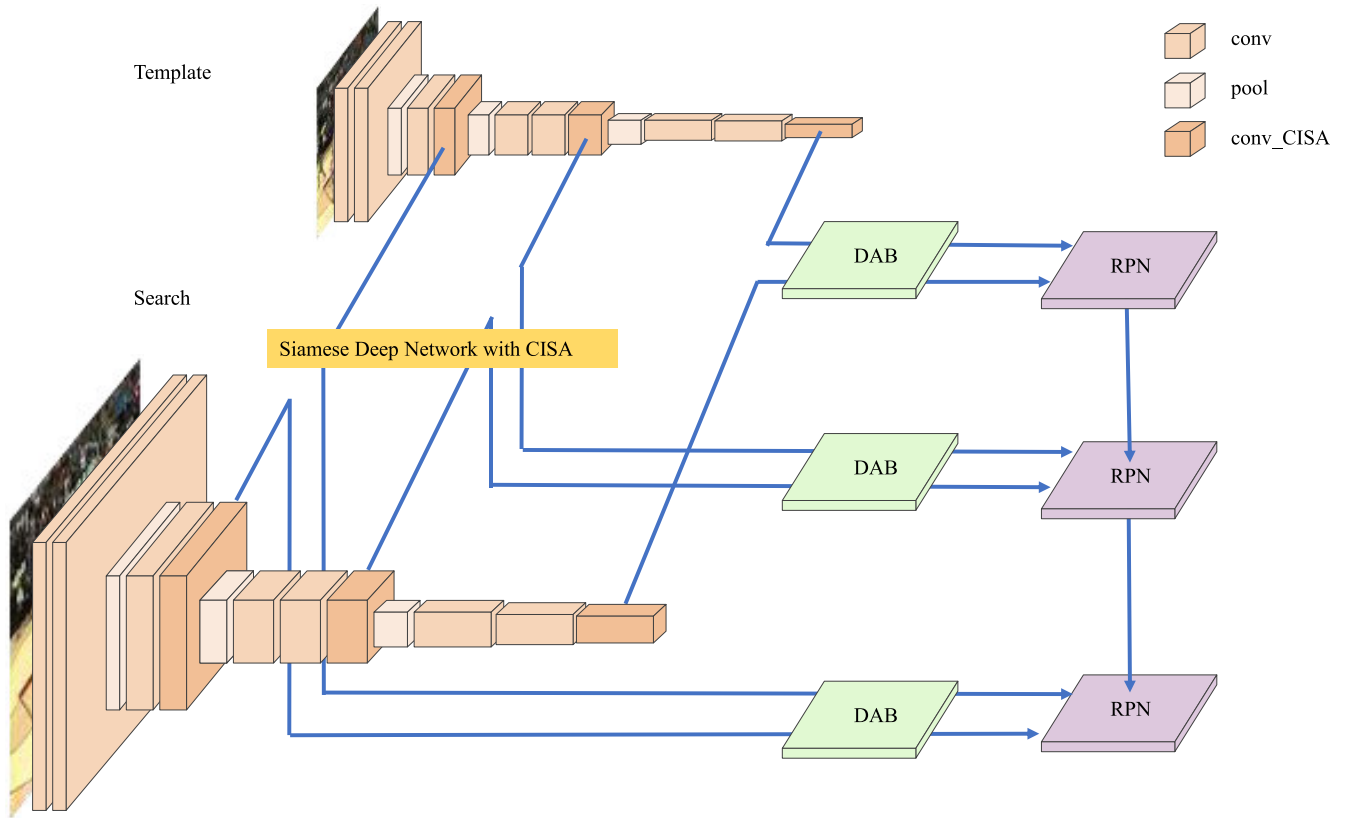


FIGURE 1. Illustration of our proposed framework. It consists of Siamese Deep Network, Channel-Interconnection-Spatial Attention module (CISA), Deconvolution Adjust Block (DAB) and Three-Layer Cascaded RPN. Given a template and search region, the networks output a dense prediction by fusion the outputs from Three-Layer Cascaded RPN.

SiamFC [10] uses a large-scale dataset to offline training a deep network, and then simply evaluates online during the tracking process. CFNet [28] converts the correlation filter into a neural network layer, combines the feature extraction network to achieve end-to-end optimization, and trains convolutional features that match the correlation filter. MDNet [29] develops a multi-domain learning framework, and then the pre-trained networks update online in the context of the sequence and adaptively learn information about Domain-specific. SiamRPN [11] introduces a Regional Proposal Network to replace multi-scale detection by bounding box regression, and connects the Regional Proposal Network that generates the candidate area to the Siamese networks for feature extraction. DasiamRPN [30] optimizes the algorithm from the aspects of imbalanced training data, adaptive model incremental learning, and long-term tracking. It performs well in occlusion and long-term tracking. Excellent algorithms are still emerging [31]–[34], refreshing records and promoting the development of the tracking field.

C. ATTENTION MECHANISM

The attention mechanism plays an important role in the human visual process, and the high-level semantic features in the image can attract human visual attention. Visual tracking

is extremely dependent on visual attention [35], [36]. Human visual tracking relies on the distinctiveness of surface features of the object[37]. The evidence displays that the role of visual attention in tracking mainly reflects in the enhancement of the object and the suppression of the interference, especially in the case of dense interference [38]. DA-VT [39] adopts discriminative spatial attention, while RTT [40] draws attention to possible objects. SA-Siam [41] designs channel attention module, which is introduced into semantic branch to improve the resolution of semantic branch and the efficiency of feature extraction. RASNet [42] explores general attention, residual attention, and channel attention three kinds of attention mechanisms, and then introduces them into the Siamese framework.

III. OUR PROPOSED METHOD

To produce effective and efficient visual tracking, a novel tracker named Siamese Cascaded Region Proposal Networks with Channel-Interconnection-Spatial Attention (SCRPN-CISA) is proposed, and the overview is as shown in Figure 1. It consists of four main components, Siamese Deep Network, Channel-Interconnection-Spatial Attention module (CISA), Deconvolution Adjust Block (DAB) and Three-Layer Cascaded RPN.

TABLE 1. Architecture of Siamese Deep Network.

Layer Name	Kernel Size	Stride	Input Channel×Map	Template Size	Search Size	Output Channel	CISA-Module
input		-		127×127	255×255	3	No
conv1_1	3×3	1	64×3	125×125	253×253	64	No
conv1_2	3×3	1	64×64	123×123	251×251	64	No
pool1	2×2	2		61×61	125×125	64	No
conv2_1	3×3	1	128×64	59×59	123×123	128	No
conv2_2	3×3	1	128×128	57×57	121×121	128	Yes
pool2	2×2	2		28×28	60×60	128	No
conv3_1	3×3	1	256×128	26×26	58×58	256	No
conv3_2	3×3	1	256×256	24×24	56×56	256	No
conv3_3	3×3	1	256×256	22×22	54×54	256	Yes
pool3	2×2	2		11×11	27×27	256	No
conv4_1	3×3	1	512×256	9×9	25×25	512	No
conv4_2	3×3	1	512×512	7×7	23×23	512	No
conv4_3	3×3	1	512×512	5×5	21×21	512	Yes

Firstly, VGG-Net-D is considered as the backbone of the Siamese Deep Network. As the layer deepens, feature extraction capability of the network also increases. At the same time, the attention module CISA is integrated into Siamese Deep Network to promote the adaptability and discrimination ability. Then, the feature maps of different layers output by the Siamese Deep Network are input to Deconvolution Adjust Block for cross-layer fusion. Finally, the above is input into Three-Layer Cascaded RPN, and the generated response maps are classified and located.

A. SIAMESE DEEP NETWORK

Feature extraction is the most critical factor that determines the performance of the tracking algorithm [43]. In order to improve the feature extraction capability of the network without introducing padding to destroy the translation invariance of the network, the network in this paper is constructed based on the more adaptive network VGG-Net-D.

VGG-Net-D is a convolutional neural network constructed by repeatedly stacking 3×3 small convolution kernels and 2×2 maximum pooling layers. Large convolution kernels are simulated by deep multiplexing of small convolution kernels to complete local perception of the image and to improve the performance of the network. The model is mainly composed of five convolutional layers, two fully connected feature layers and one fully connected classification layer.

In order to better apply VGG-Net-D to the algorithm, according to the characteristics of the Siamese networks, and taking into account cross-correlation and response map fusion operations, the VGG-Net-D is modified. The specific network structure is shown in the table 1.

As can be seen from the table, the main modifications include the following three aspects.

–First, considering the need for precise location of the object tracking task, the fifth convolutional layers conv5 are deleted, and then the total stride of the network is reduced.

–Second, the attention module CISA is integrated after conv2_2, conv3_3, conv4_3, while the size of the feature maps remains the same.

–Third, to facilitate subsequent fusion, Deconvolution Adjust Block is added to each CISA module output to adjust the number of channels.

B. CHANNEL-INTERCONNECTION-SPATIAL ATTENTION MODULE

Simply increasing the depth of the network can only relatively improve the ability to express features, and it cannot fundamentally solve the problems of the Siamese framework. Compared with trackers based on correlation filters, the trackers based on Siamese networks replace the online training by offline training in order to increase the speed. It requires the network to have two qualities at the same time. On the one hand, it can stably adapt to the changes of the object itself in various scenarios, abstracting the representative and essential characteristics of the object. On the other hand, it is able to distinguish the object from similarities sensitively and remarkably extract the differences between them.

From the aspect of feature extraction, deep convolutional networks such as VGG-Net-D trained on large-scale classification datasets have a relatively average degree of attention for each position of the image, while the tracking task needs to focus on the features of the object. Therefore, offline training networks are not fully adapted to online tracking.

From the aspect of similarity discrimination, the feature maps extracted through the Siamese Deep Network are denoted as $\varphi(\mathbf{z}) \in \mathbb{R}^{C \times H_T \times W_T}$ and $\varphi(\mathbf{x}) \in \mathbb{R}^{C \times H_D \times W_D}$, respectively. $f(\mathbf{z}, \mathbf{x}) \in \mathbb{R}^{H_f \times W_f}$ is defined as

$$f(\mathbf{z}, \mathbf{x}) = \sum_{c=0}^{C-1} \sum_{h=0}^{H_T-1} \sum_{w=0}^{W_T-1} \varphi_{c,h,w}(\mathbf{z}) \varphi_{c,H_f+h,W_f+w}(\mathbf{x}) + b, \quad (1)$$

where $H_T \leq H_D$, $W_T \leq W_D$, $H_f = H_D - H_T + 1$ and $W_f = W_D - W_T + 1$. It can be seen that the calculation process of the cross-correlation has average attention to the channel level and spatial level of the feature maps. Actually, the importance of different channels and different positions in the feature map varies greatly in the tracking task. Therefore,

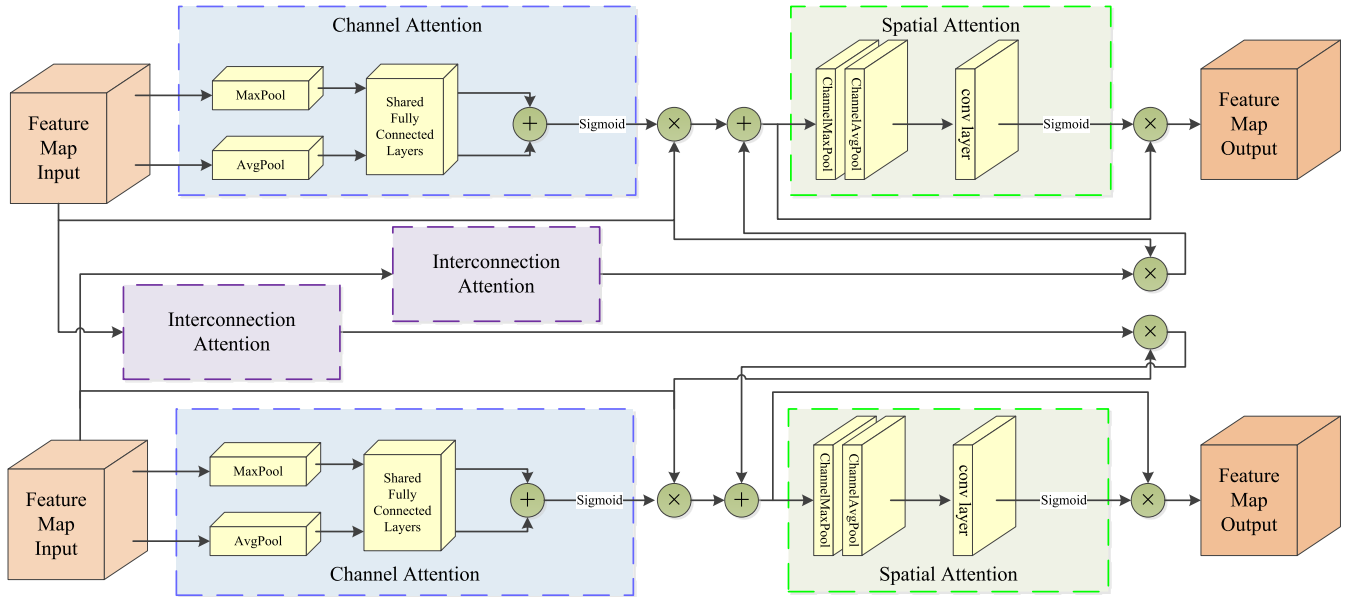


FIGURE 2. Diagram of Channel-Interconnection-Spatial Attention module. As illustrated, the Channel Attention module utilizes both MaxPool and AvgPool. The output of Interconnection Attention module is joint with the Channel Attention module. The Spatial Attention module processes them along the channel axis.

the attention modules are integrated into the feature extraction network, and the weight parameters are adjusted to highlight the important information of the object while suppressing the irrelevant detailed information, thereby improving the discriminative ability of the network.

As an important mechanism of human visual cognition, visual attention guides humans to quickly search for priority processing of the most interesting and specific areas of the screening field of view (regions of significance), selectively allocating computational resources and improving the efficiency of the visual system [44], [47]. The factors that guide attention distribution mainly include data-driven attention selection and task-driven attention selection. The former is unconsciously guided and attracts attractive viewpoints through saliency areas that are strongly different from the surroundings in the picture. The latter refers to being subjectively guided by human cognition (expected goals, empirical knowledge). The nucleus of the attentional mechanism is to suppress irrelevant detail information by adjusting the weighting parameters to highlight or screening important information about the object. Therefore, the attention mechanisms can refer to weaken the feature maps with small contribution, strengthen the feature maps with large contribution, pay attention to the difference between the object foreground and the semantic background, and achieve the enhancement of the object, the reduction of interference and the recognition between different objects, such that improve the robustness and real-time of the algorithm in complex scenarios.

Applying the attention mechanism to both channel and spatial level can be embedded in most of the current mainstream deep networks. Without significantly increasing the number of calculations and parameters, the feature

extraction capability of the network is optimized [48]. In order to differentiate the importance of different channels and different spatial positions in different tracking object feature maps, meanwhile, exploit the background information of the template image and search image, according to the structural characteristics of the Siamese networks, the Channel-Interconnect-Spatial Attention module (CISA) is constructed, as shown in Figure 2.

The feature maps output by the template branch and the detection branch are denoted as $\varphi(\mathbf{z})$ and $\varphi(\mathbf{x})$, respectively. Therefore, $\varphi_C(\mathbf{z})$ and $\varphi_C(\mathbf{x})$ can be computed as

$$\varphi_C(\mathbf{z}) = \mathbf{A}_C\{\varphi(\mathbf{z})\} \otimes \varphi(\mathbf{z}), \quad (2)$$

$$\varphi_C(\mathbf{x}) = \mathbf{A}_C\{\varphi(\mathbf{x})\} \otimes \varphi(\mathbf{x}), \quad (3)$$

where $\mathbf{A}_C\{\cdot\} \in \mathbb{R}^{C \times 1 \times 1}$ represents the attention map, and \otimes denotes element-wise multiplication.

Each channel is equivalent to a different type of feature discriminator. Channel attention optimizes and selects features through the semantic level, activates the more relevant object and deletes redundant channel features. Learn the associations between semantic features to form more cohesive and accurate features. The Channel Attention module compresses the input feature map in the spatial dimension, and uses global average pooling and global maximum pooling at the same time. Global average pooling feeds back every pixel on the feature map, while global maximum pooling supplements global average pooling.

In the Siamese networks, the template branch and the detection branch usually operate independently. Actually, encoding the respective branch into the other branch is also instructive and can make effectively use of the background

information. Therefore, the Interconnection Attention module is presented. $\varphi(\mathbf{z})$ and $\varphi(\mathbf{x})$ are passed through the Interconnection Attention module to get the attention map $\mathbf{A}_I\{\cdot\} \in \mathbb{R}^{C \times C}$. Since the dimensions of $\varphi(\mathbf{z})$ and $\varphi(\mathbf{x})$ are not same, for facilitating matrix multiplication with features, it is necessary to be dimensionally adjusted to get $\varphi_{RS}(\mathbf{z}) \in \mathbb{R}^{C \times P_T}$ and $\varphi_{RS}(\mathbf{x}) \in \mathbb{R}^{C \times P_D}$, where $P_T = H_T \times W_T$, $P_D = H_D \times W_D$. Then $\varphi_{IRS}(\mathbf{z})$ and $\varphi_{IRS}(\mathbf{x})$ are calculated as

$$\varphi_{IRS}(\mathbf{z}) = \mathbf{A}_I\{\varphi_{RS}(\mathbf{x})\} \otimes \varphi_{RS}(\mathbf{z}), \quad (4)$$

$$\varphi_{IRS}(\mathbf{x}) = \mathbf{A}_I\{\varphi_{RS}(\mathbf{z})\} \otimes \varphi_{RS}(\mathbf{x}), \quad (5)$$

After the calculation is complete, the dimensions are restored and then the interconnection feature maps $\varphi_I(\mathbf{z})$ and $\varphi_I(\mathbf{x})$ are obtained.

The channel feature maps and the interconnection feature maps of the two branches converges separately, and then the spatial feature maps $\varphi_S(\mathbf{z})$ and $\varphi_S(\mathbf{x})$ are computed as

$$\varphi_S(\mathbf{z}) = \mathbf{A}_S\{[\varphi_C(\mathbf{z}) + \varphi_I(\mathbf{z})]\} \otimes [\varphi_C(\mathbf{z}) + \varphi_I(\mathbf{z})], \quad (6)$$

$$\varphi_S(\mathbf{x}) = \mathbf{A}_S\{[\varphi_C(\mathbf{x}) + \varphi_I(\mathbf{x})]\} \otimes [\varphi_C(\mathbf{x}) + \varphi_I(\mathbf{x})], \quad (7)$$

Spatial attention is more focused on the description of the position, and complements the channel attention. By constructing the connection between different positions in the feature map, learning which parts of the feature map should have a higher response from the spatial level, according to the position weighted fusion. The Spatial Attention module uses global average pooling and maximum pooling to compress the input feature map at the channel level to obtain two 2d feature maps, which are stitched together according to the channel dimensions to obtain a feature map with two channels. Then it is convolved with a hidden layer containing a single convolution kernel to ensure that the feature map is consistent with the input feature map in the spatial dimension.

C. DECONVOLUTION ADJUST BLOCK

In different cases, the tracking effect of different convolutional features is different. In some cases, the shallow features present powerful, while in other cases, the deep features display impressively. Although the simple superposition of multi-layer feature maps can improve the expression ability of the features in object tracking to a certain extent, but the generalization ability is weak.

With the idea of residual network [49], in order to effectively utilize multi-layer features, we design Deconvolution Adjust Block(DAB) to fuse features across different layers, so that each RPN can share features and improve discrimination ability. The architecture is as shown in Figure 3, taking layer Conv4_3 as an example.

Since the deep convolutional features have more class sensitive semantic information and stronger intra class invariance, these features are very helpful to deal with the background confusion and partial occlusion in tracking, and can better infer the semantic label of the object. DAB mainly focuses on combining deep feature maps with shallow feature maps. The channels of the deep feature map Conv4_3_CISA

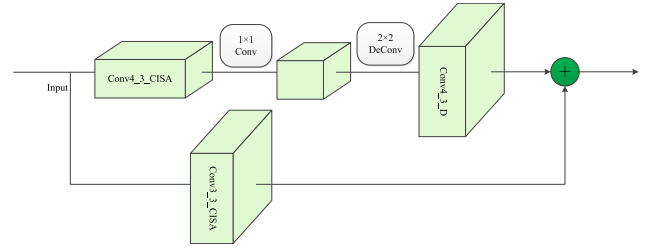


FIGURE 3. Overview of Deconvolution Adjust Block (DAB).

is reduced to 256 by 1×1 convolution. Then up-sampling is performed by deconvolution to obtain a feature map Conv4_3_D that has the same size with Conv3_3_CISA, and the elements are added to Conv3_3_CISA.

D. CASCADED REGION PROPOSAL NETWORKS

In the Three-Layer Cascaded RPN architecture, RPN includes classification branch and regression branch. The output of the template branch and the detection branch in the Siamese Deep Network are adjusted by DAB first, and then they are respectively input into the classification branch and regression branch of the Three-Layer Cascaded RPN.

In the classification branch, let the feature maps of q^{th} layer denoted as $[\varphi_{cls}^{(q)}(\mathbf{z})]$ and $[\varphi_{cls}^{(q)}(\mathbf{x})]$. In the regression branch, let the feature maps of q^{th} layer denoted as $[\varphi_{reg}^{(q)}(\mathbf{z})]$ and $[\varphi_{reg}^{(q)}(\mathbf{x})]$. Depthwise correlation operation is performed in the Three-Layer Cascaded RPN to obtain the classification branch response score maps and the regression offsets for anchors, as shown in (8) and (9).

$$\{c_{cls}^{(q)}\} = [\varphi_{cls}^{(q)}(\mathbf{x})] * [\varphi_{cls}^{(q)}(\mathbf{z})], \quad (8)$$

$$\{r_{reg}^{(q)}\} = [\varphi_{reg}^{(q)}(\mathbf{x})] * [\varphi_{reg}^{(q)}(\mathbf{z})], \quad (9)$$

The loss is the sum of softmax loss for classification and the standard smoothL1 loss for regression, following [5].

$$\ell^{(q)}(\{c_{cls}^{(q)}\}, \{r_{reg}^{(q)}\}) = L_{cls}^{(q)} + \lambda L_{reg}^{(q)}, \quad (10)$$

In the foreground-background classification of each candidate area, as the same object may exist in multiple overlapping rectangular boxes at the same time, Non-Maximum suppression (NMS) is usually used to get more accurate position. However, the problem with this strategy is that the detected bounding boxes with higher classification confidences contrarily may have smaller overlaps with the corresponding ground-truth, it is likely to cause tracking drift. Consequently, they are ameliorated as

$$s_i = \begin{cases} s_i - s_i [\text{IoU}(b_i, b_{\max})] & \text{IoU}(b_i, b_{\max}) \geq T \\ s_i & \text{otherwise.} \end{cases} \quad (11)$$

where b_i is the bounding box, s_i is the classification score of the bounding box, T denotes the threshold.

IV. EXPERIMENTS

In this part, firstly, we describe the implementation details of the proposed tracker SCRPN-CISA. Then, the attention

module is verified and analyzed. Next, we take an ablation study on OTB-2015 [13]. Finally, we evaluate our method comprehensively on the standard benchmarks OTB-2015 [13], UAV123 [19], VOT2016 [14] and VOT2019 [17].

A. IMPLEMENTATION DETAILS

1) EXPERIMENTS ENVIRONMENT

Our method is implemented using PyTorch on PC with Intel i7-9700 CPU (3.0GHz), 16GB RAM and NVIDIA GeForce RTX 2060 GPU.

2) TRAINING

We use VGG-Net-D [18], pretrained on ImageNet [50], as the backbone, and the networks are then fine-tuned on the training datasets of ImageNet VID, COCO [51] and Youtube-BB [52]. We adopt single scale images with 127 pixels for template patches and 255 pixels for searching regions. We apply stochastic gradient descent (SGD) with a momentum of 0.9 and a learning rate of 10^{-4} - 10^{-6} during training. The whole training process consists of more than 100 stages, each consisting of 6000 sample pairs.

B. ATTENTION VISUALIZATION ANALYSES

In order to analyze the role of the attention module CISA, the class-activated heat map Grad-CAM [53] is used to visualize different networks, as shown in Figure 4.

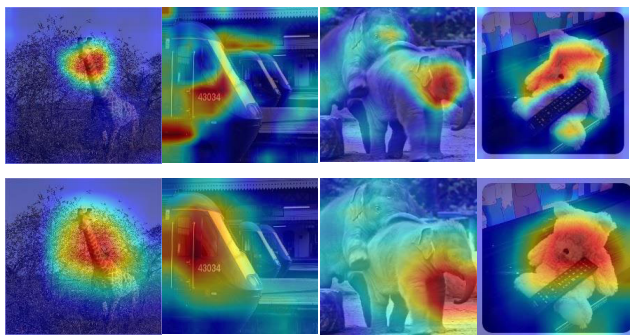


FIGURE 4. Grad-CAM networks visualization results. We compare the visualization results of the integrated networks (VGG-Net-D+CISA) with baseline (VGG-Net-D).

The more sensitive location has a higher temperature, and the less sensitive location has a lower temperature. The first row shows the results of the networks without CISA and the second row shows the results of the integrated networks with CISA. The experimental results can be seen that the networks with the CISA module have a wider range of attention and can better cover the objects to be recognized. When there are similar objects around them, they can better distinguish and focus on the objects without being disturbed by similarities. When the object is partially occluded, they can eliminate interference and cover the whole object.

C. ABLATION STUDY

The proposed tracker SCRPN-CISA is mainly composed of four modules, Siamese Deep Network, Channel-Interconnection-Spatial Attention module, Deconvolution Adjust Block and Three-Layer Cascaded RPN. In order to verify the effectiveness and better understand the contributions of various components, we conduct the comparative experiments on OTB-2013 [12] and OTB-2015 [13], the results are shown in Figure 5.

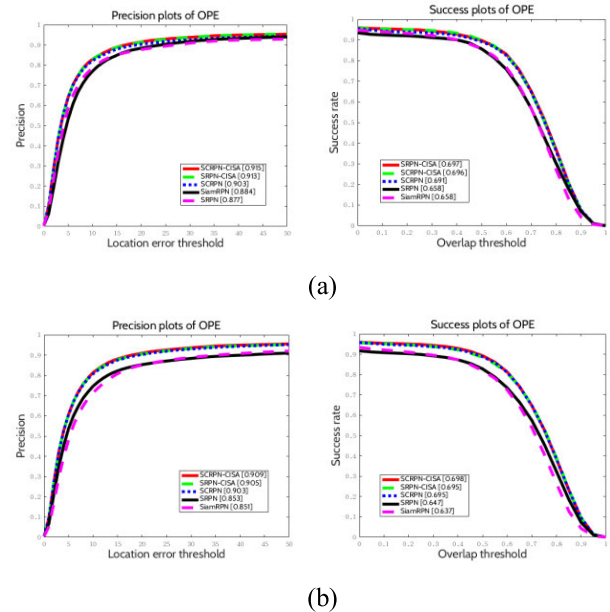


FIGURE 5. The precision plots and the success plots of ablation experiments with different models. (a) OTB-2013. (b) OTB-2015.

- SRPN: Siamese Deep Network based on VGG-Net-D;
- SRPN-CISA: Channel-Interconnection-Spatial Attention module is attached to the SRPN;
- SCRPN: Three-Layer Cascaded RPN is attached to the SRPN;

–SCRPN-CISA: the proposed tracker, which attached both CISA and Three-Layer Cascaded RPN to the SRPN.

It can be observed from the figure.

On OTB-2013, compared with SRPN,

–First, SRPN-CISA increases by 3.6% and 3.9% in the precision plot and the success plot respectively.

–Second, SCRPN increases by 2.6% and 3.3% in the precision plot and the success plot respectively.

–Third, SCRPN-CISA increases by 3.8% and 3.8% in the precision plot and the success plot respectively.

On OTB-2015, compared with SRPN,

–First, SRPN-CISA increases by 5.2% and 4.8% in the precision plot and the success plot respectively.

–Second, SCRPN increases by 5% and 4.8% in the precision plot and the success plot respectively.

–Third, SCRPN-CISA increases by 5.6% and 5.1% in the precision plot and the success plot respectively.

The above analysis shows that each part of the tracker contributes to the overall performance of the tracking. The deepening of the networks promotes the feature extraction capability. The addition of Channel-Interconnection-Spatial Attention module significantly improves the adaptability and feature extraction capability in complex scenarios. The Three-Layer Cascaded RPN enables the tracker to obtain more accurate position.

D. RESULTS ON OTB-2015

1) OVERALL PERFORMANCE

OTB-2013 [12] and OTB-2015 [13] are the general tracking benchmarks proposed by Wu Yi and others in 2013 and 2015 to evaluate the trackers. It mainly evaluates the two indicators of center position error and coverage. The precision plot refers to Euclidean distance between the predicted locations and the ground truth annotations. Mostly, the threshold distance is set as 20 pixels, which indicates the percentage of frames whose estimated location is within 20 pixels of the ground truth position. The success plot also counts the percentage of successfully tracked frames. It is set to measure the overlapping rate between the ground truth and the estimated center location which surpasses a given threshold, i.e., 0.5 [12]. Both determine whether the tracker is successful through a certain threshold.

We compare the proposal SCRPN-CISA with other state-of-the-art trackers including C-RPN [32], MDNet [29], DaSiamRPN [30], SiamRPN [11], CFNet [28], SiamFC [10], SiamDWfc [54] and DeepSRDCF [55] on OTB-2015 benchmark. The quantitative results of the nine algorithms on OTB-2015 are shown in Figure 6. In all experiments, we use the original source code or the results provided by the author to ensure a fair comparison.

We obtain a precision of 0.909 and an AUC of 0.698 which surpass that of SiamRPN [11] by 5.8% and 6.1% respectively and exceed DasiamRPN [30] by 2.9% and 4% respectively. It is proved that SCRPN-CISA extracts deeper features on the basis of the Siamese framework, and integrates the attention module simultaneously, which makes the networks with stronger adaptability, and then improves the overall accuracy and robustness of the tracker.

2) ATTRIBUTE-BASED EVALUATION

For more detailed analyses and further validation of our tracker in sequences with different challenges, such as background clutter, illumination variation and deformation, we analyze the performance in terms of different aspects of these attributes annotated in the benchmark. The precision plots and the success plots are shown in Figure 7 and Figure 8 respectively. As can be seen, our proposed tracker SCRPN-CISA performs favorably against other trackers in almost all the attributes. It is robustness with a total of 11 attributes.

3) QUALITATIVE COMPARISON

In order to better analyze the tracking performance of SCRPN-CISA, it is compared with other trackers in some

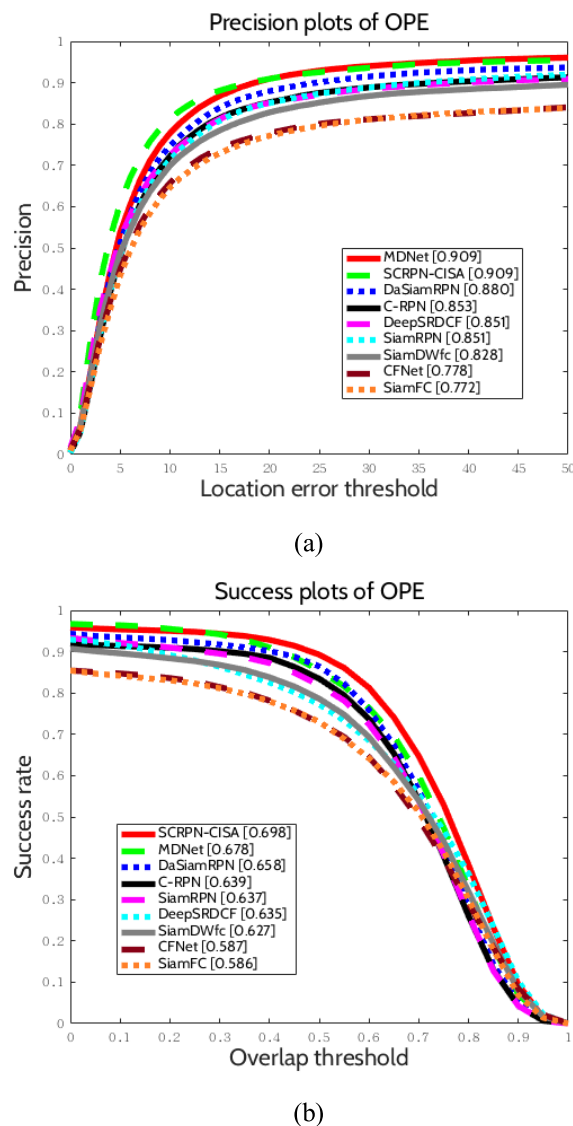


FIGURE 6. The precision plot and the success plot of OPE for 9 trackers on OTB-2015. Each tracker is ranked by the performance score. In the precision plot, the score is at error threshold of 20 pixels. In the success plot, the score is the AUC value.

sequences, and some typical results are selected for analyzing, as shown in Figure 9. Different trackers are represented by different colors, where our tracker is in red.

(a) In the Box sequence, the object moves in a chaotic environment and is occasionally blocked, sometimes with out-of-plane rotation, and the scale changes continuously. Only SCRPN-CISA and ECO have not changed the object from beginning to end.

(b) In the CarDark sequence, the object is sometimes blurred or blocked, there are similar interferences around and the lighting situation is constantly changing. The SCRPN-CISA can filter out more adaptable object features and distinguish the background, so there is never drift.

(c) In the Human4-2 sequence, the object is interfered by similar objects during the movement process, and the occlusion situation also occurs from time to time.

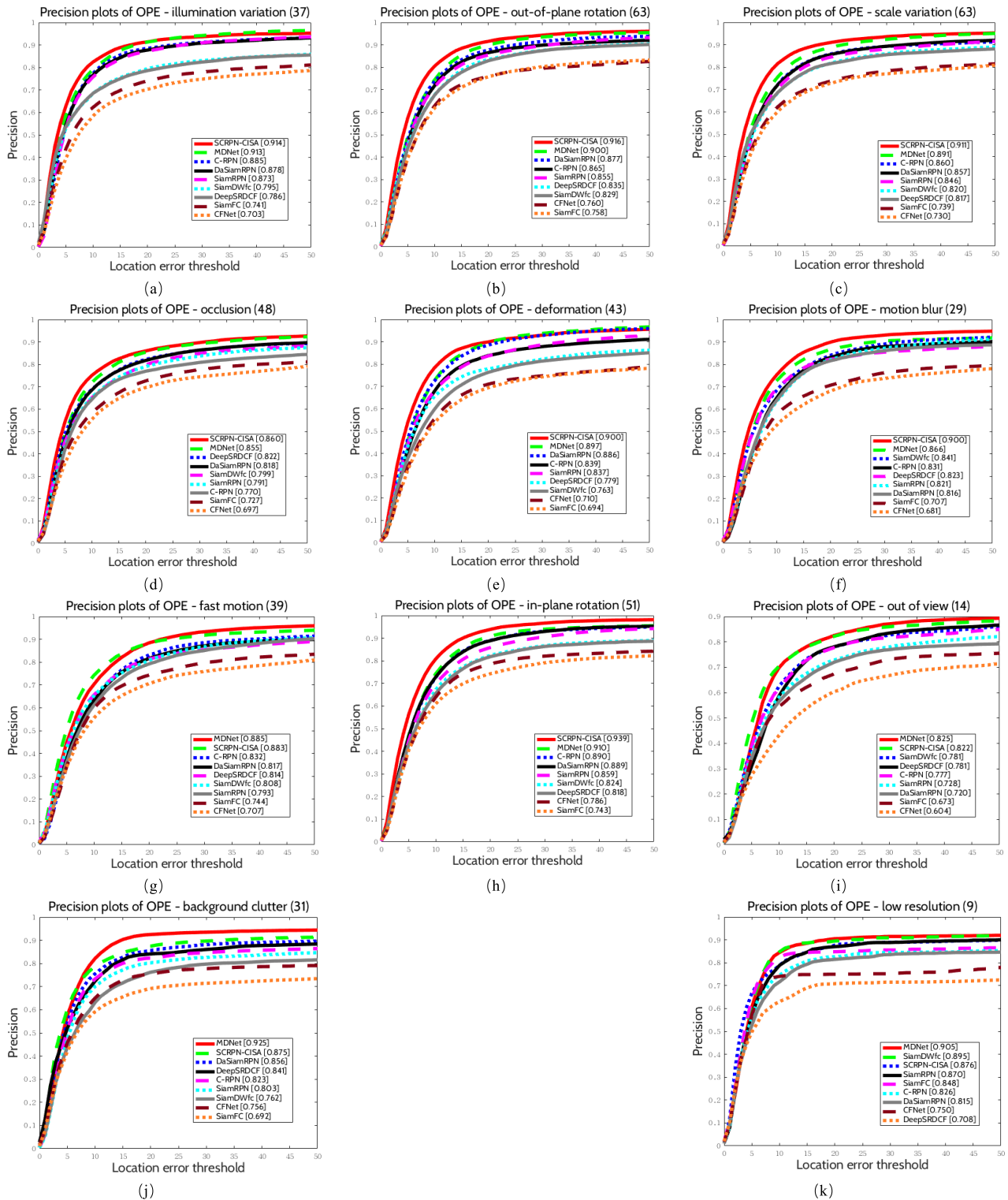


FIGURE 7. Attribute based evaluation on OTB-2015. The precision plots over eleven tracking challenges of illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-view, background clutter and low resolution.

Compared with other Siamese trackers, SCRPN-CISA develops discriminating ability of the networks and can restrain interference without drifting occurred.

(d) In the Jump sequence, the object is abruptly changed during the whole process, which is occasionally blocked, accompanied by out-of-plane rotation and

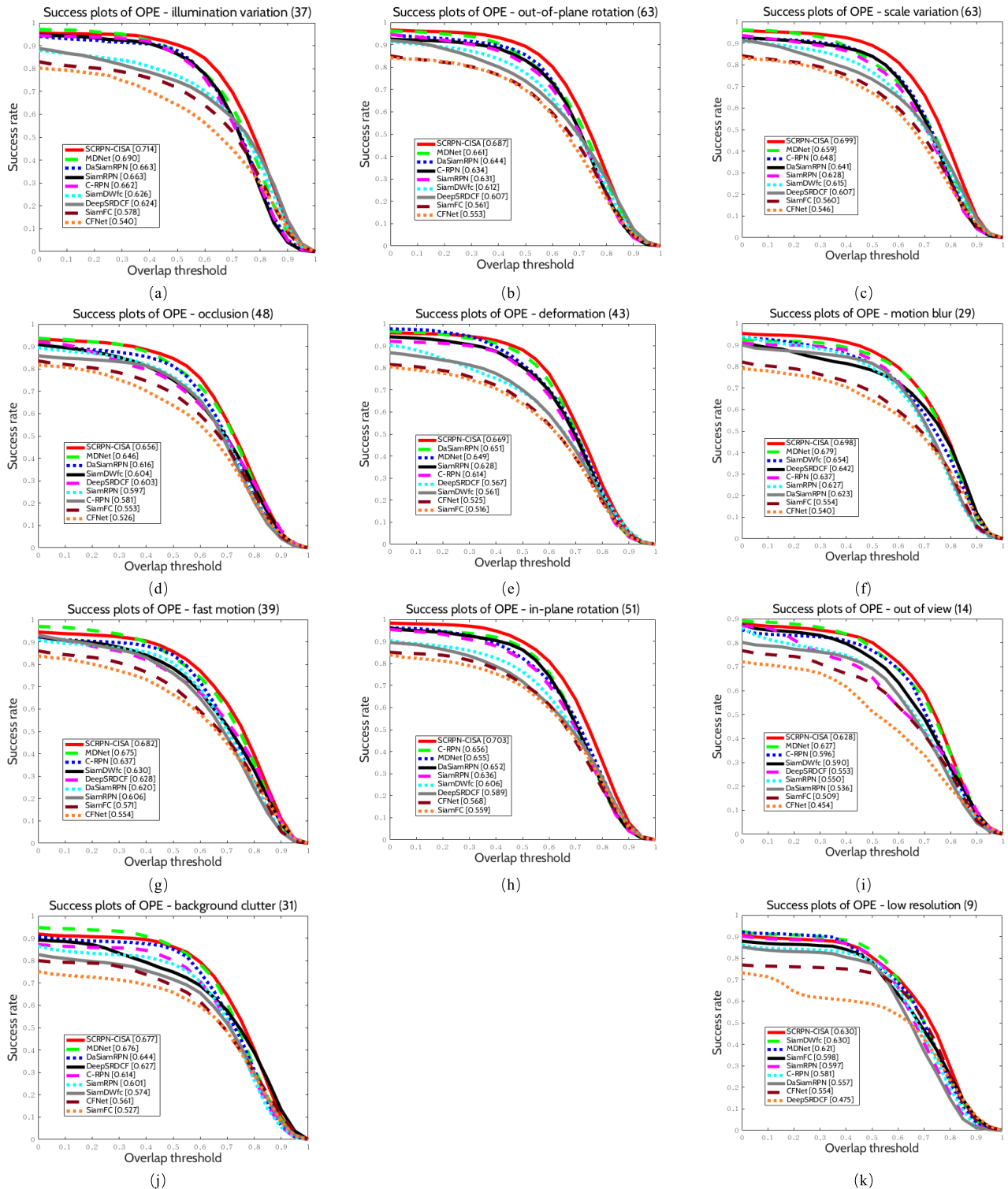


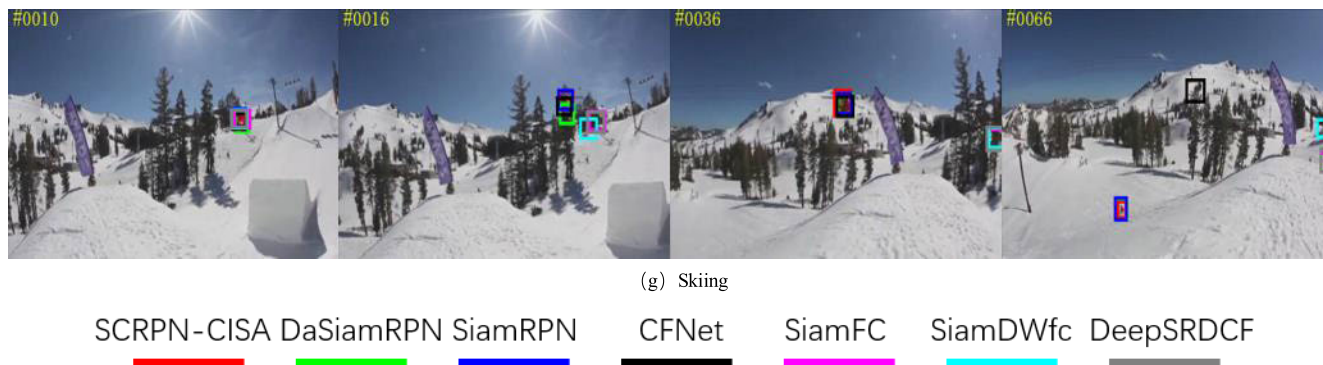
FIGURE 8. Attribute based evaluation on OTB-2015. The success plots over eleven tracking challenges of illumination variation, out-of-plane rotation, scale variation, occlusion, deformation, motion blur, fast motion, in-plane rotation, out-of-view, background clutter and low resolution. The legend contains the AUC score for each tracker.

in-plane rotation. Comparing of other Siamese based trackers, SCRPN-CISA enhances discriminating ability

of the networks, so that eliminates interference without drift.



FIGURE 9. Qualitative comparison of our tracker with the state-of-the-art trackers on the seven challenging sequences on OTB-2015.



(g) Skiing

FIGURE 9. (Continued.) Qualitative comparison of our tracker with the state-of-the-art trackers on the seven challenging sequences on OTB-2015.

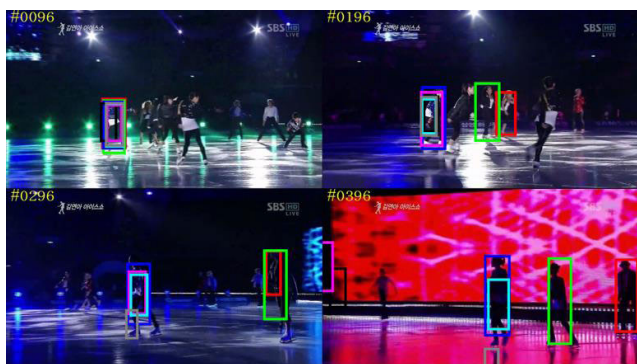


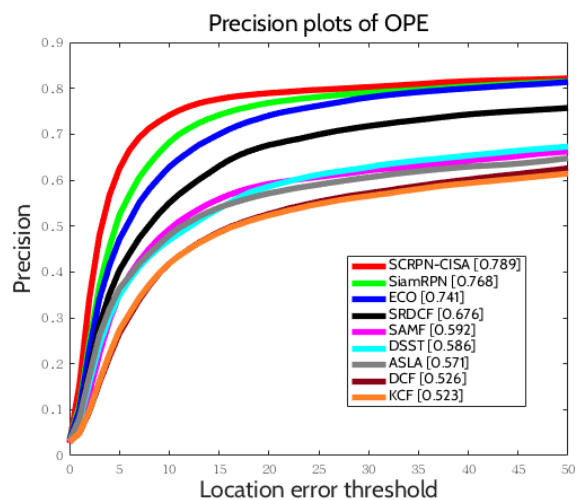
FIGURE 10. Failed sequences Skating1.

(e) In the Matrix sequence, the illuminance and angle dramatically vary resulting in a complicated background. Meanwhile, the object movement speed is relatively fast, and there are situations of in-plane rotation and out-plane rotation. Only SCRPN-CISA can always adapt and perform stable tracking.

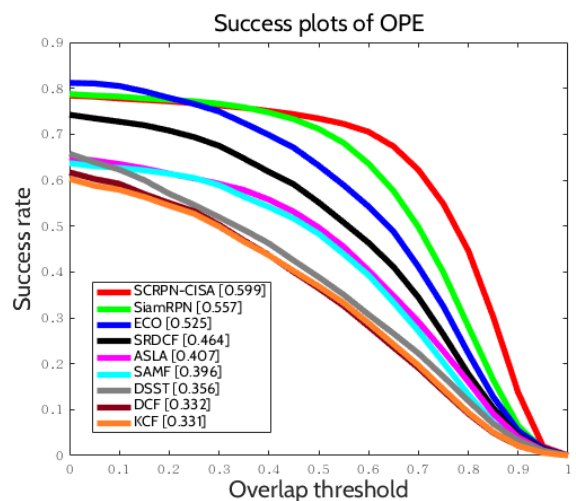
(f) MotorRolling sequence is a very challengeable video sequence. Accompanied by illumination variation, in complex background, the object sometimes becomes blur due to rapid motion, or itself rotates sometimes more than 360 degrees, which requires the tracker to have very strong feature extraction ability and adaptability, SCRPN-CISA detects object consistently and steadily, but the other algorithms drift initially in tracking.

(g) In the Skiing sequence, the object size is smaller and the speed is faster, which puts stricter requirements on feature extraction. Algorithms with shallow feature extraction ability or matching ability can no longer locate the object at the initial tracking.

By analyzing the above-mentioned video sequences with abundant attributes, our tracker has excellent performance in the situations of illumination change, scale change, rotation, similar object interference, fast motion, low resolution, and out-of-view.



(a)



(b)

FIGURE 11. The precision plot and the success plot of OPE for 9 trackers on UAV123. Each tracker is ranked by the performance score. In the precision plot, the score is at error threshold of 20 pixels. In the success plot, the score is the AUC value.

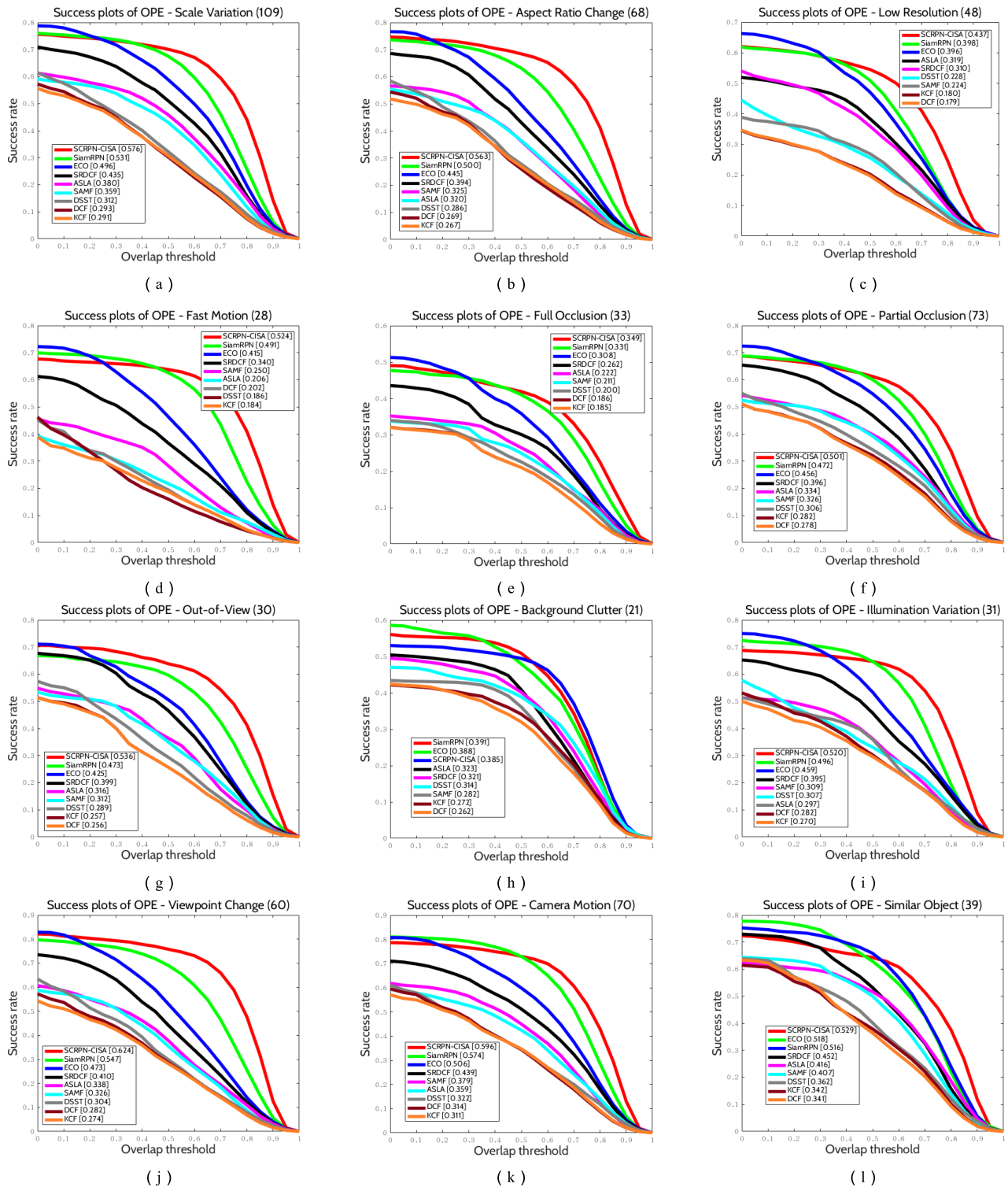


FIGURE 12. Attribute based evaluation on UAV123. The success plots over twelve tracking challenges of Scale Variation, Aspect Ratio Change, Low Resolution, Fast Motion, Full Occlusion, Out-of-View, Background Clutter, Illumination Variation, Viewpoint Change, Camera Motion, and Similar Object. The legend contains the AUC score for each tracker.

4) FAILURE

There are also individual problem sequences, as shown in Figure 10.

The Skating1 sequence is the most comprehensive video in the entire dataset. The resolution is low, and as the object moves, it continuously encounters occlusion and out-of-view.

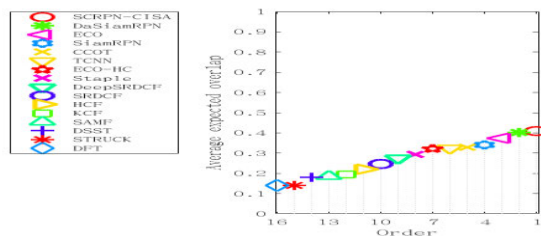


FIGURE 13. Expected averaged overlap performance on VOT2016.

The proposed tracker drifts. It depicts that the detection mechanism needs to be adjusted appropriately.

E. RESULTS ON UAV123

1) OVERALL PERFORMANCE

We compare the proposal SCRPN-CISA with other trackers including SiamRPN [11], ECO [3], SRDCF [57], ASLA [58], SAMF [59], DSST [60], DCF [26] and KCF [26] on UAV123 benchmark. The quantitative results of the nine algorithms on UAV123 are shown in Figure 11. In all experiments, we use the original source code or the results provided by the author to ensure a fair comparison. We obtain a precision of 0.789 and an AUC of 0.599 which surpass that of SiamRPN [11] by 2.1% and 4.2% respectively.

2) ATTRIBUTE-BASED EVALUATION

For more detailed analysis and further validation of our tracker in sequences with different challenges, such as background clutter, illumination variation and deformation, we analyze the performance in terms of different aspects of these attributes annotated in the benchmark. The success plots are shown in Figure 12.

As can be seen, our proposed tracker SCRPN-CISA performs favorably against other trackers in almost all the attributes. It is stable in the video sequences of UAV123 with a total of 12 attributes.

F. RESULTS ON VOT2016

As further evaluating the performance of the algorithm, tests are conducted on VOT2016. VOT2016 [14] is one of the most widely used databases in the field of visual tracking. It consists of 60 sequences and uses the accuracy (A), robustness (R), and expected average overlap (EAO) to evaluate the performance of the tracker. We compare the proposed tracker SCRPN-CISA with several excellent trackers. Figure 13 shows the EAO ranking. It can be seen from the figure that SCRPN-CISA has the best average expected overlap ratio.

The detailed comparisons are reported in Table 2. Our tracker exceeds the advanced tracker ECO 3.8% on EAO. The average speed of our tracker on the test set with GPU is 33 FPS, which is roughly 4 times as fast as ECO(8FPS)[3]. Meanwhile, comparing with SiamRPN [11], which is also based on Siamese framework, our tracker has significant improvements of 6.9% on EAO.

TABLE 2. Comparison with the state-of-the-art in terms of expected average overlap (EAO), accuracy, and robustness (failure rate) on VOT2016.

Trackers	EAO↑	Accuracy↑	Robustness↓
SCRPN-CISA	0.413	0.633	0.21
DaSiamRPN	0.411	0.61	0.22
ECO	0.375	0.55	0.20
SiamRPN	0.344	0.560	0.26
CCOT	0.331	0.539	0.238
TCNN	0.325	0.554	0.268
ECO-HC	0.325	0.54	0.30

TABLE 3. Comparison with the state-of-the-art in terms of expected average overlap (EAO), accuracy, and robustness (failure rate) on VOT2019.

Trackers	EAO↑	Accuracy↑	Robustness↓
ATOM	0.292	0.603	0.411
SiamMask	0.287	0.594	0.461
SCRPN-CISA	0.276	0.566	0.469
SA_SIAM_R	0.253	0.559	0.492
SSRCCOT	0.234	0.495	0.507
SiamRPNX	0.224	0.517	0.552
TADT	0.207	0.516	0.677

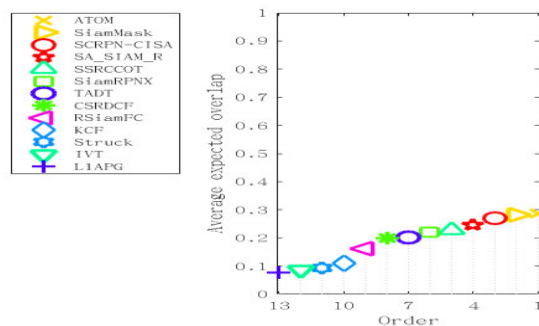


FIGURE 14. Expected averaged overlap performance on VOT2019.

G. RESULTS ON VOT2019

As further evaluating the performance of the algorithm, tests are conducted on VOT2019. VOT2019 [17] is currently the most advanced databases in the field of visual tracking. We compare the tracker SCRPN-CISA with several excellent trackers. Figure 14 illustrates that the EAO ranking.

The detailed comparisons are reported in Table 3. Our tracker is 1.6% lower than the state-of-the-art tracker ATOM [56] on EAO. Compared to SiamRPN [11], which is also based on Siamese framework, our tracker has significant improvements of 5.2% on EAO.

V. CONCLUSION

In this paper, we propose a novel Siamese based tracker SCRPN-CISA. We use the deep network VGG-Net-D as the backbone of the Siamese framework to make the model more capable of expressing features. The Channel-Interconnection-Spatial Attention module can make the model have better adaptability and discrimination ability, which achieves better object enhancement and interference suppression, and distinguish the difference between object foreground and semantic background. At the same time, we design a Deconvolution Adjust Block for cross-layer fusion of deep semantic features and shallow spatial features, thereby, further improving its ability to distinguish complex backgrounds. In addition, we construct the Three-Level Cascade RPN, and also design a bounding box adjustment strategy to achieve accurate position in particular. Experiments on OTB-2015, UAV123, VOT2016 and VOT2019 show that, comparing with current mainstream trackers, the proposed tracker SCRPN-CISA in this paper can achieve higher tracking accuracy and robust performance.

REFERENCES

- [1] 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)*, Santiago, Chile, Dec. 2015, pp. 3074–3082.
- [2] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 472–488.
- [3] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6931–6939.
- [4] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning-based searching and Coarse-to-Fine verifying," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2331–2341, May 2019.
- [5] D. Yuan, X. Zhang, J. Liu, and D. Li, "A multiple feature fused model for visual object tracking via correlation filters," *Multimedia Tools Appl.*, vol. 78, no. 19, pp. 27271–27290, Oct. 2019.
- [6] D. Yuan, N. Fan, and Z. He, "Learning target-focusing convolutional regression model for visual object tracking," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105526, doi: 10.1016/j.knsys.2020.105526.
- [7] D. Yuan, X. Li, Z. He, Q. Liu, and S. Lu, "Visual object tracking with adaptive structural convolutional network," *Knowl.-Based Syst.*, vol. 194, Apr. 2020, Art. no. 105554, doi: 10.1016/j.knsys.2020.105554.
- [8] B. Zhong, H. Yao, S. Chen, R. Ji, T.-J. Chin, and H. Wang, "Visual tracking via weakly supervised learning from multiple imperfect oracles," *Pattern Recognit.*, vol. 47, no. 3, pp. 1395–1410, Mar. 2014.
- [9] Q. Zhou, B. Zhong, Y. Zhang, J. Li, and Y. Fu, "Deep alignment network based multi-person tracking with occlusion and motion reasoning," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1183–1194, May 2019.
- [10] L. Bertinetto, J. V. Almadre, O. F. J. Henriques, A. V. Edaldi, and H. S. P. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 850–865.
- [11] 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.
- [12] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2013, pp. 2411–2418.
- [13] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
- [14] M. Kristan, A. Leonardis, and J. Matas, "The visual object tracking VOT2016 challenge results," in *Proc. IEEE ECCV Workshops*, Amsterdam, The Netherlands, Oct. 2016, pp. 777–823.
- [15] M. Kristan et al., "The visual object tracking VOT2017 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 1949–1972.
- [16] M. Kristan et al., "The sixth visual object tracking VOT2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Munich, Germany, Sep. 2018, pp. 3–53.
- [17] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J. K. Kamarainen, and A. Eldesokey, "The seventh visual object tracking VOT2019 challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1–36.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [19] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. IEEE ECCV Workshops*, Amsterdam, The Netherlands, Oct. 2016, pp. 445–461.
- [20] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "LaSOT: A high-quality benchmark for large-scale single object tracking," 2018, *arXiv:1809.07845*. [Online]. Available: <http://arxiv.org/abs/1809.07845>
- [21] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," 2018, *arXiv:1810.11981*. [Online]. Available: <http://arxiv.org/abs/1810.11981>
- [22] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2544–2550.
- [23] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. ECCV*, Firenze, Italy, Oct. 2012, pp. 702–715.
- [24] M. Danelljan, F. S. Khan, M. Felsberg, and J. V. D. Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1090–1097.
- [25] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Nottingham, Britain, 2014, pp. 1–11.
- [26] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.
- [27] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1420–1429.
- [28] 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. 5000–5008.
- [29] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 4293–4302.
- [30] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 101–117.
- [31] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1328–1338.
- [32] 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)*, Long Beach, CA, USA, Jun. 2019, pp. 7944–7953.
- [33] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6668–6677.
- [34] D. Yuan, W. Kang, and Z. He, "Robust visual tracking with correlation filters and metric learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105697, doi: 10.1016/j.knsys.2020.105697.
- [35] S. Yantis, "Multielement visual tracking: Attention and perceptual organization," *Cognit. Psychol.*, vol. 24, no. 3, pp. 295–340, Jul. 1992.

- [36] R. Allen, P. McGeorge, D. Pearson, and A. B. Milne, "Attention and expertise in multiple target tracking," *Appl. Cognit. Psychol.*, vol. 18, no. 3, pp. 337–347, Apr. 2004.
- [37] T. Makovski and Y. V. Jiang, "Feature binding in attentive tracking of distinct objects," *Vis. Cognition*, vol. 17, nos. 1–2, pp. 180–194, Jan. 2009.
- [38] M. M. Doran and J. E. Hoffman, "The role of visual attention in multiple object tracking: Evidence from ERPs," *Attention, Perception, Psychophys.*, vol. 72, no. 1, pp. 33–52, Jan. 2010.
- [39] J. Fan, Y. Wu, and S. Dai, "Discriminative spatial attention for robust tracking," in *Proc. IEEE ECCV Workshops*, Crete, Greece, Sep. 2010, pp. 480–493.
- [40] 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.
- [41] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 4834–4843.
- [42] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 4854–4863.
- [43] N. Wang, J. Shi, D.-Y. Yeung, and J. Jia, "Understanding and diagnosing visual tracking systems," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3101–3109.
- [44] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.
- [45] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Perception*, vol. 7, no. 1, pp. 1–39, 2010.
- [46] S. Frintrop, *Computational Visual Attention. Computer Analysis of Human Behavior*. London, U.K.: Springer, 2011, pp. 69–101.
- [47] A. Borji and L. Itti, "State-of-the-Art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
- [48] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [51] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 740–755.
- [52] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7464–7473.
- [53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [54] 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.
- [55] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 621–629.
- [56] 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.
- [57] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Boston, MA, USA, Dec. 2015, pp. 4310–4318.
- [58] 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.
- [59] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 254–265.
- [60] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, Nottingham, Britain, Sep. 2014, pp. 1–11.



ZHOJUAN CUI was born in Henan, China, in 1986. She is currently pursuing the Ph.D. degree with the National Space Science Center, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences, Beijing, China. Her major research interests include computer vision, deep learning, and visual tracking.



JUNSHI AN was born in Shaanxi, China, in 1969. He received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2004. He is currently a Professor and a Ph.D. Supervisor with the National Space Science Center, Chinese Academy of Sciences, Beijing, China. His research interests include deep learning, integrated space electronics technology, such as space computer hardware, software, system architecture, and data processing technology.



QING YE was born in Henan, China, in 1998. She is currently pursuing the bachelor's degree with Henan University and the bachelor's degree with the Victoria University of Australia. Her research interests include computer vision and deep learning.



TIANSHU CUI was born in Shandong, China, in 1986. He is currently pursuing the Ph.D. degree with the National Space Science Center, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences, Beijing, China. His major research interests include computer vision, deep learning, and RF fingerprinting.

...