IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2020, accepted June 19, 2020, date of publication June 29, 2020, date of current version July 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005511

Deeper Siamese Network With Stronger Feature Representation for Visual Tracking

CHAOYI ZHANG[1], HOWARD WANG[2], JIWEI WEN[1], AND LI PENG[1]

[1]Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

[2]Department of Electrical, Computer, and Software Engineering, The University of Auckland, Auckland 1010, New Zealand

Corresponding author: Jiwei Wen (wjw8143@jiangnan.edu.cn)

This work was supported in part by the National Key Research And Developement Program of China under Grant 2018YFD0400902, and in part by the National Natural Science Foundation of China under Grant 61873112.

ABSTRACT Siamese network based visual tracking has drawn considerable attention recently due to the balanced accuracy and speed. This type of method mostly trains a relatively shallow twin network offline, and measures the similarity online using cross-correlation operation between the feature maps generated by the last convolutional layer of the target and search regions to locate the object. Nevertheless, a single feature map extracted from the last layer of shallow networks is insufficient to describe target appearance, as well as sensitive to the distractors, which could mislead the similarity response map and make the tracker easily drift. To enhance the tracking accuracy and robustness while maintaining the real-time speed, based on the above tracking paradigm, three improvements including reform of backbone network, fusion of hierarchical features and utilization of channel attention mechanism, have been made in this paper. Firstly, we introduce a modified deeper VGG16 backbone network, which could extract more powerful features contributing to distinguishing the target from distractors. Secondly, we fuse diverse features extracted from deep layers and shallow layers to take advantage of both semantic and spatial information of the target. Thirdly, we incorporate a novel lightweight residual channel attention mechanism into the backbone network, which expands the weight gap between different channels and helps the network pay more attention on dominant features. Extensive experimental results on OTB100 and VOT2018 benchmarks demonstrate that our tracker performs better in accuracy and efficiency against several state-of-the-art methods in real-time scenarios.

INDEX TERMS Visual tracking, Siamese network, channel attention mechanism.

I. INTRODUCTION

Visual tracking, as one of the fundamental tasks in computer vision, is to estimate the location of an arbitrary object in a video sequence, given only target position in the first frame. It has been widely used in many applications such as video surveillance [1], autonomous driving [2] and human-computer interaction [3]. Some significant breakthroughs have been made in the past decade, it still faces many challenges like background clutters, object occlusions, and deformation.

Recently, convolutional neural networks (CNNs) have demonstrated their superior performance in various visual tasks such as image classification [4], object detection [5], and semantic segmentation [6]. The features extracted by CNNs are dependent on large-scale training data, and contain rich high-level semantic information effective for distinguishing the targets. Some of the tracking approaches integrate convolutional features into correlation filter framework [7]–[9], and achieve great improvements compared to those which utilize the traditional hand-craft features [10]–[12]. However, the CNNs utilized by the trackers based on the correlation filter are originally designed for image classification tasks, which means the extracted feature representation of the target is not fully suitable for visual tracking. Moreover, these deep trackers cannot run in real time.

To further exploit the representation capabilities of CNNs adaptive to tracking task, in recent years, trackers based on Siamese networks have achieved considerable popularity in tracking community due to their real-time speed and competitive accuracy. The typical framework of these trackers is

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

that a Siamese network is trained offline with abundant image pairs first, and the similarity between the target regions and search regions is calculated. The location with the highest response is the estimated object position. As the pioneering work for this paradigm, SiamFC [13] adopts a fully convolutional Siamese architecture followed by a cross-correlation operation between the reference patch and the candidate patches in search regions. SiamFC is highly efficient due to its lightweight network without model updating scheme during tracking.

Recently, in order to strengthen the power of the feature representation extracted by the original Siamese framework, many Siamese based tracking approaches have been developed and have achieved state-of-the-art performance. SA-Siam [14] constructs a twofold Siamese network, including an appearance branch and a semantic branch to improve the generalization capability of SiamFC. He *et al.* [15] propose a spatial masking mechanism to prevent the tracker from drifting to the salient objects in the background. Zhu *et al.* [16] improve the quality of the training data and design a distractor-aware module to enhance the discrimination of the tracker. The SPM-tracker [17] builds two modules, one is a coarse matching module aiming at strengthening the robustness of a tracker and the other is a fine matching module to enhance the discrimination power. The two blocks are connected in a series-parallel manner and the tracking performance is superior.

However, the above advanced methods still have some major limitations. Firstly, the backbone architecture adopted in these trackers is the classical AlexNet [4], a relatively shallow network, of which the output cannot capture the powerful feature representation of the target, and the tracker lacks of discrimination in the presence of similar distractors. Secondly, compared to the modern deep networks, AlexNet contains only five convolutional layers and cannot benefit from multi-layer feature fusion. In addition, only the features extracted from the last convolutional layer which focus on the semantic abstraction are utilized, in the trackers above. However, the spatial structure, which contributes to distinguishing the target from background, is ignored. Thirdly, different channels from the same convolutional layer are employed to describe the target from different aspects, which may lead to the importance gap between channels, i.e., some channels will learn more powerful features while others are less useful.

To overcome the above issues and unveil the power of the deeper network for real-time visual tracking, in this paper, we propose a simple and efficient deep architecture named HA-SaimVGG, which combines a hierarchical feature fusion strategy with a lightweight channel attention mechanism in the deeper network. Firstly, we introduce a modified deep VGG [18] network as our feature extractor, which has more powerful discriminative capability compared with the original shallow AlexNet. Secondly, we impose offline trained weights on response maps generated by deep layers and shallow layers in backbone network, making use of both semantic and spatial information of the target, to provide a high-quality response map and improve the robustness of the tracker. Thirdly, we equip the network with a lightweight residual channel attention mechanism, which could enlarge the importance gap between different channels in the same layer, and make the network pay more attention on dominant features.

In summary, The main contributions are summarized as follows:

1) A novel end-to-end Siamese architecture, which inherits the merits of the deeper VGG network, multi-layer feature fusion and attention mechanism, is proposed for high-performance real-time visual tracking.
2) A hierarchical feature fusion strategy, which takes advantage of both semantic and spatial information of the target, is utilized to enhance the discriminative capability of deep networks and improve the robustness of the tracker.
3) A residual channel attention mechanism, which is lightweight and could distinguish important channels to emphasize informative features and suppress less useful ones, is incorporated into the backbone network.
4) Extensive experiments on OTB100 and VOT2018 benchmarks are conducted to demonstrate the performance in real-time scenarios.

II. RELATED WORK

A. SIAMESE NETWORKS FOR VISUAL TRACKING

Trackers based on Siamese networks have achieved great development in recent years. Tao *et al.* [19] first introduce Siamese networks into tracking domain. They train a Siamese network to find a candidate region that best matches the original target appearance. Specifically, the initial target region will pass through one branch of the network and many candidate patches go through another, and the most similar candidate is determined by the matching function. However, their approach(SINT) is not real-time due to the utilization of optical flow. Held *et al.* [20] present GOTURN tracker, which directly regresses the locations of the target and could reach at 100 FPS on GPU, while the accuracy is not satisfied. Different from SINT, Bertinetto *et al.* propose a novel fully-convolutional Siamese network (SiamFC) [13] which trains the network offline with large-scale ILSVRC2015 (ImageNet Large Scale Visual Recognition Challenge) dataset [21] and introduce the cross-correlation layer to measure the similarity of the input pairs online. SiamFC could run at 86 FPS on GPU and maintain the competitive accuracy, which immediately attracts great concern in the tracking field. Guo *et al.* introduce a dynamic Siamese architecture (DSiam) [22] that updates the model online to adaptive to the variations of the target. Similarly, UpdateNet [23] addresses the model updating problem by training an extra network to estimate the best template for the next frame, which reduces the loss of speed. SiamRPN [24] combines a Siamese network with a region

proposal network(RPN) [5] to better handle scale variation of the target. In our approach, we take advantage of real-time speed in Siamese networks, and develop the network architecture and inference process on this basis.

B. HIERARCHICAL FEATURES FUSION

Different layers in a CNN could extract complementary information of the object. For instance, the former layers focus on the more fine-grained spatial details of the object whereas the latter layers could capture more high-level semantic information with low spatial resolution. Recently, many researches in tracking domain have exploited the way to fuse multi-layer features and obtained favorable results. Ma *et al.* [7] learn the correlation filters by utilizing the third, fourth and fifth convolutional layers in VGG19 network to extract the object description. Wang *et al.* [25] build a specific network and a general network based on the fourth layer and last layer, respectively. The heat maps generated from these two sub-networks could be adaptively fused to locate the target. The tracker in [26] combines the complementary features, the deep features help to distinguish the target from noisy background and the shallow features provide the appearance information of the object. Fan and Ling [27] cascade a sequence of RPN modules from deep high-level to shallow low-level layers in a Siamese network, which gains better performance on the basis of SiamRPN. Kuai *et al.* [28] propose a Hyper-Siamese architecture to aggregate the hierarchical feature maps with a skip-layer connection and constitute the hyper-feature representation of the target. In this paper, we are motivated by this merit to exploit multi-cues of the target for effective tracking.

C. ATTENTION MECHANISMS

As the new block of neural networks, attention models have spread to many visual tasks [29]–[31], which could reduce the irrelevant information and emphasize the significant ones. For visual tracking, Gao *et al.* [32] incorporate a novel cross-attentional module into Siamese network, and enhance both discriminative and localization capabilities of feature maps. Li *et al.* [33] introduce a encoder-decoder attention module which squeezes the feature maps first and builds the relationships between each channel in a Siamese network, realizing the filter for different features. Wang *et al.* [34] strengthen the feature representation for the Siamese architecture by integrating different kinds of attention mechanisms including general attention, residual attention and channel attention, resulting in the great alleviation of over-fitting. Gao *et al.* [35] utilize a novel hierarchical attentional module with long short-term memory and multi-layer perceptrons to effectively facilitate visual pattern emphasis, and learn the reinforced attentional representation for accurate target object discrimination and localization. In terms of these insights, this paper proposes a lightweight residual channel attention module to help the backbone extract more discriminative features.

III. PROPOSED APPROACH

In this section, three improvements including deeper feature extraction network, feature fusion strategy and channel attention mechanism in our tracker are introduced.

A. OVERALL ARCHITECTURE

Our HA-SiamVGG takes SiamFC as a basic framework, which formulates tracking as a template matching task between the exemplar image z and the candidate patch x in a search region using cross-correlation as

$$f_\theta(z, x) = \varphi_\theta(z) \star \varphi_\theta(x) + b \cdot 1 \tag{1}$$

where $\varphi_\theta(\cdot)$ denotes a convolutional feature embedding with parameters θ, $b \cdot 1$ denotes a bias term and $\star$ is the cross-correlation operation. The output of Equation 1 reflects the similarity between the input image pairs and the maximal value matches with the estimated target position.

In our work, we take full advantage of the deeper network VGG to learn a more discriminative feature representation, equipped with a hierarchical feature fusion strategy and a lightweight residual channel attention mechanism for real-time tracking. The overall network architecture is shown in Fig.1.

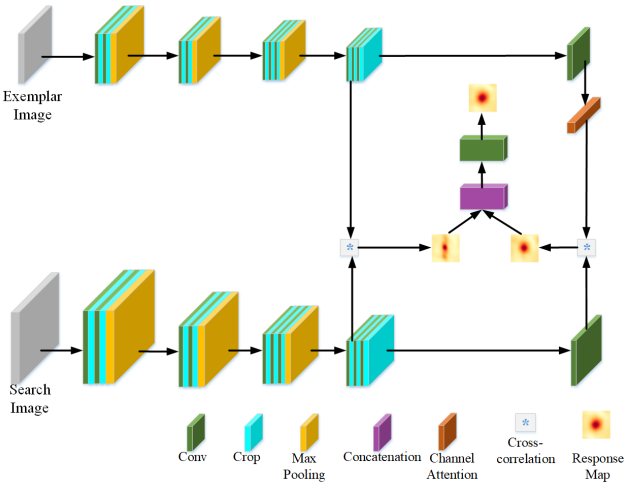


FIGURE 1. Overall network architecture of the proposed HA-SiamVGG tracker. Different components of the modified VGG16 are illustrated in different colors. We first utilize hierarchical features extracted from the last layer and the first crop layer in the 4^{th} block to generate two response maps, then, we fuse these two maps by a convolutional operation to provide a more robust response map. Meanwhile, the last layer is equipped with a lightweight channel attention module to further strengthen the dominant features.

B. DEEPER BACKBONE

As we know, the increasing depth of networks is beneficial for elevating model capability [36]. However, a straightforward replacement with original VGG cannot bring improvement and even lead to substantially drops. We analyze the Siamese framework and identify that the padding operation and the network stride are the two dominant factors of performance degradation.

Padding operation will bring potential position bias. Especially, when the target moves to the edge of the image. If the network contains padding operation, as shown in the Fig.2, the embedded features of the exemplar will include both the original target patch and additional padding regions. Whereas for the candidates in the search image, some of them are the features only extracted from themselves(blue regions in Fig.2), and some of them could contain padding regions plus themselves(orange regions in Fig.2). Thus, it leads to an inconsistency between the embedding features of the target in different search regions, and results in an inaccurate similarity reflection of the final output between input pairs. To address padding interference, we crop out the outermost features affected by padding operation of the feature map [37].

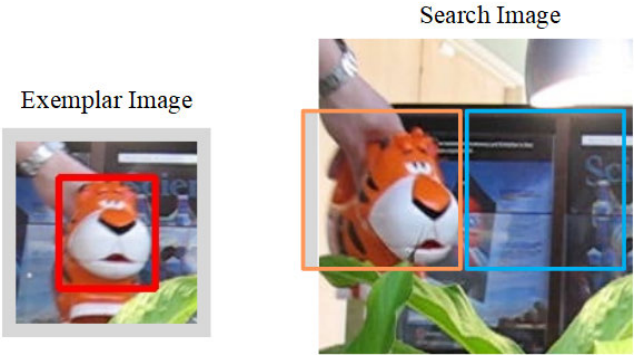


FIGURE 2. Padding influence. In the search image, when the target move to image borders, some candidate regions only contain themselves(blue regions) whereas some contain both themselves and outer padding regions(orange regions), which lead to an inconsistency between the embedding features of the target at different positions.

The network stride of the original VGG16 which has 5 max pooling layers is set to 32, whereas tracking task aims to localize the object precisely instead of classification, such a big stride could result in a low spatial resolution of the last layer, which is insufficient to locate the target accurately. Thus, we reserve the first three max pooling layers, i.e., we narrow the stride to 8, with regards to accuracy and efficiency.

C. FEATURE FUSION STRATEGY

In SiamFC framework, the quality of the response map plays a critical role in position estimation. However, due to the limited information extracted by the AlexNet which contains only five convolutional layers, shallow network based trackers could hardly benefit from hierarchical feature fusion strategy. Thus, the response map is not discriminative and easily distracted. In this paper, we replace AlexNet with a deeper modified VGG16 network and integrate multi-layer cues to improve the quality of the final response map. As illustrated in Fig.1, we utilize features extracted from the last layer and the first crop layer in the 4^{th} block and generate two response maps. Furthermore, these two response maps including both semantic and spatial information can measure the similarity between input pairs from more perspectives. Then, we concatenate these two maps and fuse them through

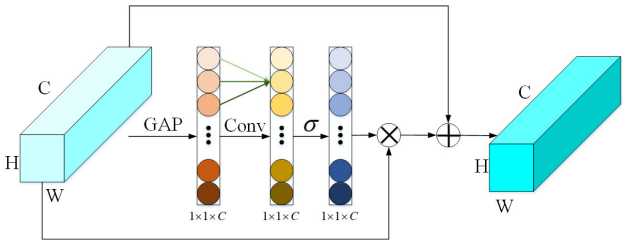


FIGURE 3. Channel attention module. H, W, C represents the height, width and number of channels for features, respectively. GAP means Global Average Pooling layer to provide a channel-wise descriptor. Conv denotes the one-dimensional convolution to build connections between adjacent channels. σ is the sigmoid activation.

a convolutional operation with kernel size set to one. Thus, different from some trackers combining response maps in an empirical manner [7], [38], that is to say, their weighting coefficients are defined by manual adjustment, our approach is data-driven and end-to-end. Our expectation is that the network can learn the robust weighting coefficients through training phase, instead of manually fine tuning on small test datasets.

D. CHANNEL ATTENTION MECHANISM

In order to further efficiently strengthen the robustness of extracted features in the complex scenario during visual tracking, for instance, the background with many noises could lead to a drifted tracker, a novel lightweight residual channel attention mechanism is proposed. Instead of building sophisticated relationships for one channel with all other channels, which uses fully connected layers leading to higher model complexity and computational burden [29], [30], we focus on the interaction between a single channel and its neighbors. As shown in Fig.3, the extracted features are firstly passed through a global average pooling layer to provide a layer-wise descriptor, followed by an one-dimensional convolution (1D Conv) aiming to build connections between adjacent channels, then, a gating unit with a sigmoid activation is employed to calculate weights for different channels,

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

where Equation 2 represents the sigmoid function, x denotes the features embedded after 1D Conv with size of $1 \times 1 \times C$, C is the number of channels. Then, the calculated weights are imposed on the input features in a channel-wise manner. Lastly, we adopt the residual architecture by adding the weighted features to the original input,

$$\tilde{F}_i = w_i \times F_i + F_i, \quad i = 1, 2, \ldots C \tag{3}$$

where w_i represents the weight for each channel, F_i denotes the original features extracted by each channel, and $\tilde{F}_i$ denotes the features equipped with attention mechanism. The number of parameters for our attention module is k, which equals to the kernel size of 1D Conv, however, the number in [29] and [30] is larger than C^2. Thus, our method could significantly reduce the number of parameters. It is worth noting

TABLE 1. Detailed backbone configuration of HA-SiamVGG.The first column represents the name of each layer, followed by four columns representing the parameters setting of the convolution or maxpooling operation. The last two columns represent the size(height,width and number of channels) of exemplar and search region after its corresponding layer operation. This network configuration corresponds to Fig.1, the layer name represents the order of this operation. For instance, Conv1-1 denotes that this layer is the first convolutional layer in the first block, Crop1-1 denotes that this layer is the first crop operation in the first block, MaxPool1 denotes that this is the first maxpooling operation.

Layer	Kernel Size	Padding	Output Channels	Stride	Exemplar Size	Search Size
Input			3		127×127×3	255×255×3
Conv1-1	3	1	64	1	127×127×64	255×255×64
Crop1-1					125×125×64	253×253×64
Conv1-2	3	1	64	1	125×125×64	253×253×64
Crop1-2					123×123×64	251×251×64
MaxPool1	2	0		2	61×61×64	125×125×64
Conv2-1	3	1	128	1	61×61×128	125×125×128
Crop2-1					59×59×128	123×123×128
Conv2-2	3	1	128	1	59×59×128	123×123×128
Crop2-2					57×57×128	121×121×128
MaxPool2	2	0		2	28×28×128	60×60×128
Conv3-1	3	1	256	1	28×28×256	60×60×256
Crop3-1					26×26×256	58×58×256
Conv3-2	3	1	256	1	26×26×256	58×58×256
Crop3-2					24×24×256	56×56×256
Conv3-3	3	1	256	1	24×24×256	56×56×256
Crop3-3					22×22×256	54×54×256
MaxPool3	2	0		2	11×11×256	27×27×256
Conv4-1	3	1	512	1	11×11×512	27×27×512
Crop4-1					9×9×512	25×25×512
Conv4-2	3	1	512	1	9×9×512	25×25×512
Crop4-2					7×7×512	23×23×512
Conv4-3	3	1	512	1	7×7×512	23×23×512
Crop4-3					5×5×512	21×21×512
Conv5-1	1	0	256	1	5×5×256	21×21×256

that our channel attention mechanism resembles the ECA-Net [39], can not only widen the activation gap between different channels, but also retain the capacity of the original features. In HA-SiamVGG, we utilize our attention block on the features extracted from the last convolutional layer, which further strengthens the semantic level of features.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we first provide the implementation details including the detailed backbone configuration of HA-SiamVGG, and then evaluate the performance of our tracker on OTB100 [40] and VOT2018 [41] benchmarks, lastly, we carry out ablative studies to analyze the effectiveness of the proposed feature fusion strategy and channel attention mechanism.

A. IMPLEMENTATION DETAILS

The detailed backbone configuration of HA-SiamVGG is presented in Table 1.

The proposed tracker is trained on the dataset of Got-10K [42] which contains about 10000 video sequences with 1.5 million annotated axis-aligned bounding boxes. In the training phase, we apply the stochastic gradient descent with momentum of 0.9 and the weight decay of 0.0005 to train the network. The loss function is defined as

$$l(y, v) = \log(1 + e^{-yv}) \tag{4}$$

where v is the estimated score which represents the similarity of each exemplar-candidate pair, y is the ground-truth label and $y \in \{-1, +1\}$. Different candidate regions have different scores, and constitute a map of scores D. We define the loss of a score map to be the mean of the individual losses,

$$L(y, v) = \frac{1}{D} \sum_{u \in D} (l(y(u), v(u))) \tag{5}$$

where $y(u) \in \{-1, +1\}$ denotes a true label for each position $u \in D$ in the score map. The convolutional layer parameters of our network are initialized with the weights of VGG16 pre-trained on ImageNet. The model is trained for 50 epochs in total using mini-batches of 8 and the learning rate exponentially decays from 10^{-2} to 10^{-5}. The kernel size of 1D Conv in channel attention module is set to 3.

For online tracking, as shown in Table 1, the size of template and search region are 127 × 127×3 and 255 × 255×3, respectively. The input image pair is fed into each branch of the network and the size of score map is 17, where the position of maximal value represents the center of the target. The channel attention mechanism is used only in the first frame to extract more powerful features of the exemplar. Scale variation of the target is estimated by processing the search image at three scales with a fixed aspect ratio that is set to 1.0375.

All the experiments are implemented in Pytorch1.1.0 with an Intel i7-8700 CPU, a NVIDIA GeForce GTX 1070 GPU and the RAM is 16g.

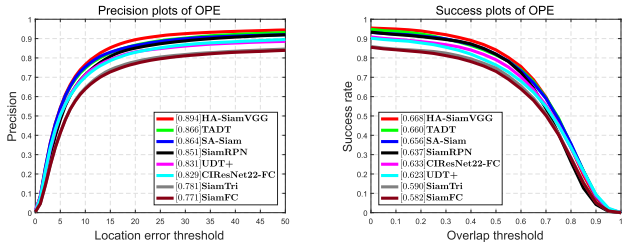
B. RESULTS ON OTB100

OTB100, a widely-used tracking benchmark, contains 100 fully annotated sequences and utilizes precision score and AUC score as two standard evaluation metrics [40]. Precision score calculates the Euclidean distance error between the center positions of tracking results and ground-truth. AUC

score measures the overlap rate, i.e., the intersection and union of the tracking results and ground-truth. Following the OTB100 benchmark settings, we compare our tracker with the other seven state-of-the-art trackers using a one-pass evaluation(OPE) strategy.

1) OVERALL PERFORMANCE ANALYSIS

We select 7 recent Siamese based tracking approaches including SiamFC, SiamTri [43], UDT+ [44], CIRseNet22-FC [37], SiamRPN [24], SA-Siam [14], TADT [45] to compare with our HA-SiamVGG. Fig.4 illustrates that our tracker achieves the best performance in both precision plot and success plot. Compared with original SiamFC, HA-SiamVGG obtains relative improvements of 12.3% in precision score and 8.6% in AUC score. Furthermore, our method surpasses the SiamRPN tracker by 4.3% and 3.1% in precision plot and success plot, respectively. Different from SA-Siam which builds semantic branch and appearance branch, our architecture is more concise and powerful, the feature fusion strategy could consider appearance and semantic information of the target simultaneously and the attention mechanism develops the discriminability of the semantic features, thus, our algorithm achieves better results. It is worth noting that our tracker has a 3.5% improvement in AUC score compared with CIResNet22-FC, which is a recent novel method that uses deep ResNet [36] as its backbone equipped with CIR unit. The achievements can be attributed to the deeper network combined with hierarchical features and attention mechanism, making HA-SiamVGG robust to challenging scenarios. Moreover, our tracker can run at a speed of 40 FPS on GPU on this benchmark.



(a) (b)

FIGURE 4. The overall precision plot(a) and success plot(b) with 8 trackers on OTB100 benchmark. The digit in the legend denotes the precision score and AUC score, respectively. The proposed HA-SiamVGG performs best.

2) ATTRIBUTE ANALYSIS

In order to comprehensively compare our tracking algorithm with others, we evaluate trackers using 11 annotated attributes on OTB100 benchmark. Fig.5 and Fig.6 present precision plot and success plot on each challenging attribute, respectively. In Fig.6, our HA-SiamVGG performs best in 9 challenges including fast motion, motion blur, deformation, illumination variation, in-plane rotation, low resolution, out-of-plane rotation, out of view and scale variation. Compared with other Siamese based trackers, the advantage of our tracker is obvious. Especially, our method improves the baseline algorithm SiamFC in a large margin on every challenging factor. In Fig.5, our approach obtains 0.996 in precision score under the challenge of low resolution, which means almost every frame regarding with this attribute can be accurately tracked. Moreover, it can be seen from both Fig.5 and Fig.6 that combining feature fusion strategy and channel attention mechanism helps improve the robustness of the tracker, the CIRseNet22-FC tracker which only replaces the backbone with deep ResNet does not perform as well as ours on all attributes.

On the other hand, we also find that our HA-SiamVGG could not perform well for scenes with occlusion, especially when the target is occluded by the same kind objects. The main reason is that our tracker has more power to distinguish the target from different categories, however, as the same kind of objects will produce similar response compared to the target in both attention mechanism and feature fusion module, it is difficult to identify the target accurately when the similar objects coincide with it. As illustrated in Fig.7, when the target(man) is occluded by a little boy, due to the similar appearance and the same category(human), the tracker could not distinguish the target from such a strong distractor, leading to drift.

3) QUALITATIVE ANALYSIS

In order to visualize tracking process, we give the qualitative comparison of our proposed algorithm against other Siamese based trackers in Fig.8. It can be seen that HA-SiamVGG is robust to most difficult scenarios. In the sequence of Box and Bolt, the target is under the complex surrounding with similar objects, even occlusions. Both original SiamFC and SiamTri tracker could not distinguish the target from other distractors and directly lose the target, but ours could follow the target in each frame. In the sequence of DragonBaby and MotorRolling, the target moves fast and lead to motion blur. The recent TADT and SiamRPN tracker could not accurately estimate the position of the target in some hard frames, for instance, in the 40^{th} frame of MotorRolling, the results of TADT and SiamRPN capture the bottom or top part of the target, respectively, whereas ours always precisely locate the motorcycle rider. In the sequence of Singer2, due to the serious background clutter, i.e., the singer and the background are both black, only our approach and SiamRPN could find the target and track successfully, this can be attributed to the proposed channel attention module that makes our tracker more discriminative to distinguish the target from background. Moreover, the tracking result in the 185^{th} frame illustrates that our tracker is robust to illumination variation. In the sequence of Skiing, the area of the target is relatively small, our HA-SiamVGG equipped with hierarchical feature fusion strategy could utilize the spatial information to accurately track, whereas CIRseNet22-FC tracker which only uses deeper semantic features could not follow the target.

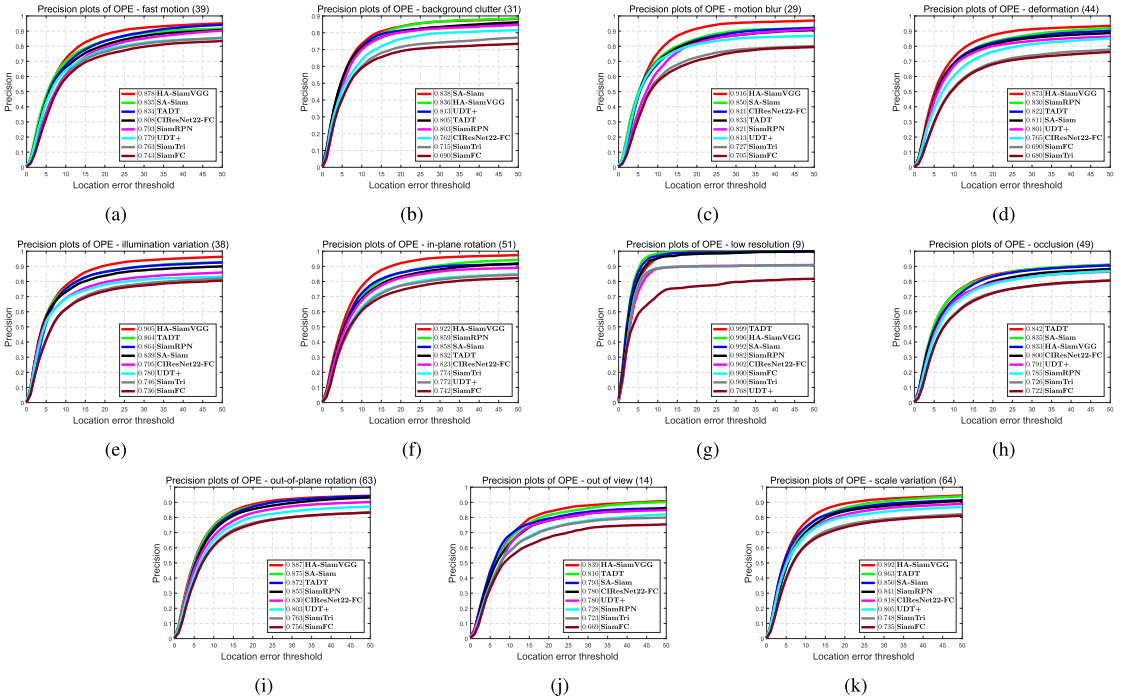


FIGURE 5. Attribute-based precision plots with 8 trackers on OTB100 benchmark. The digit in the legend of each sub-picture reflects the precision score of each approach under the corresponding attribute. The digit in the title of each sub-picture means the number of videos with the corresponding attribute. The proposed HA-SiamVGG performs best under 8 attributes against other state-of-the-art tarckers.

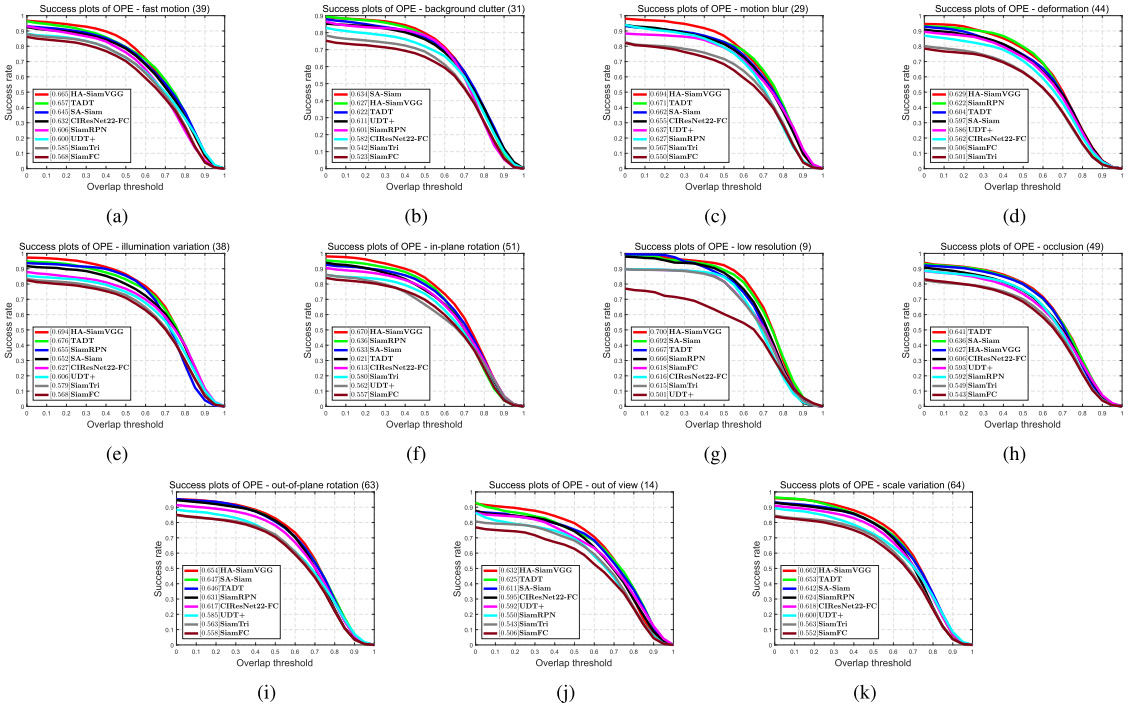


FIGURE 6. Attribute-based success plots with 8 trackers on OTB100 benchmark. The digit in the legend of each sub-picture reflects the AUC score of each approach under the corresponding attribute. The digit in the title of each sub-picture means the number of videos with the corresponding attribute. The proposed HA-SiamVGG performs best under 9 attributes against other state-of-the-art trackers.

C. RESULTS ON VOT2018

To further validate the generality of the proposed tracker, we conduct experiments on VOT2018 benchmark, which is a more challenging dataset with 60 colored sequences labeled by rotated bounding rectangle. The main evaluation measurement used to rank the trackers is expected average overlap(EAO) which combines accuracy(A) and robustness(R).

FIGURE 7. Failure case on the Human3 sequence in an occlusion scenario on OTB100 benchmark. When the target is occluded by a little boy, due to the similar appearance and the same category(human), the tracker could not distinguish the target from such a strong distractor, leading to drift.

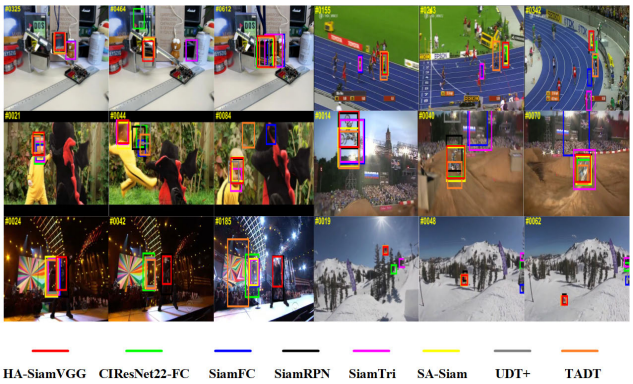


FIGURE 8. Qualitative evaluations of 8 trackers on 6 challenging image sequences(from left to right and from top to bottom are Box, Bolt, DragonBaby, MotorRolling, Singer2 and Skiing, respectively) on OTB100 benchmark.

TABLE 2. Baseline experimental results for trackers on VOT2018. The red fonts, blue fonts and green fonts indicate the best, the second best and the third best performance. The A, R and EAO evaluate the accuracy, robustness and expected average overlap of a tracker, respectively.

Tracker	A	R	EAO
HA-SiamVGG	0.537	0.309	0.313
SiamVGG [46]	0.531	0.318	0.286
SA-Siam [14]	0.533	0.337	0.286
ECO [47]	0.484	0.276	0.280
MCCT [38]	0.532	0.318	0.274
SiamDW [37]	0.538	0.398	0.270
SiamFC [13]	0.503	0.585	0.187

We select 6 recent outstanding tracking approaches including 4 Siamese based methods and 2 correlation filter based algorithms, including SiamVGG [46], SA-Siam [14], SiamDW [37], SiamFC [13], ECO [47] and MCCT [38]. Table 2 presents the baseline experimental results on VOT2018 benchmark. Our HA-SiamVGG still surpasses SiamFC in a large margin, over 12 percents in the EAO criteria. Compared with SiamVGG and SiamDW which only utilize the deeper network, HA-SiamVGG equipped with feature fusion strategy and attention module could achieve better performance. Compared with two correlation filter based tracking approaches, HA-SiamVGG obtains higher accuracy and competitive robustness.

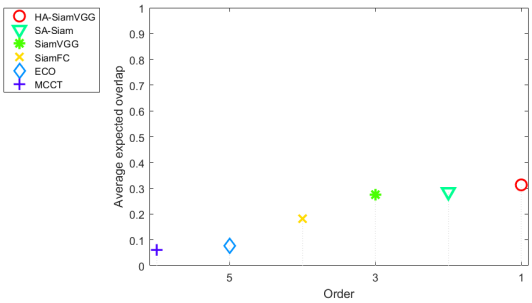


FIGURE 9. Real-time experiment with 6 trackers on the VOT2018 benchmark. The proposed HA-SiamVGG could perform best.

TABLE 3. Components analysis of HA-SiamVGG on OTB100 benchmark. Each component could improve the tracking accuracy of the baseline(SiamFC), and the proposed HA-SiamVGG integrating all of the components could achieve the best result.

Component	SiamFC	HA-SiamVGG			
Modified VGG16?		√	√	√	√
Channel attention?			√		√
Hierarchical features?				√	√
Precision score	0.771	0.852	0.873	0.879	0.894
AUC score	0.582	0.644	0.650	0.661	0.668

To verify the real-time performance of our tracker, we also conduct the real-time experiment on the VOT2018 benchmark. As illustrated in Fig.9, the proposed HA-SiamVGG still performs best, and could meet real-time requirements.

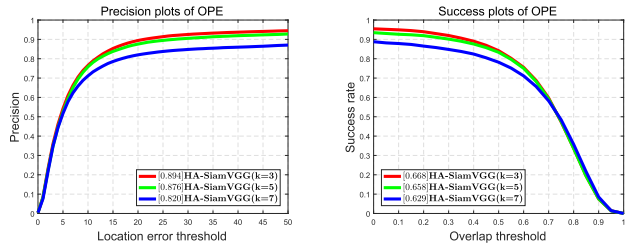
D. ABLATION STUDY

1) PARAMETERS ANALYSIS

To analyze the influence of the kernel size(k) in our channel attention mechanism, we compare three different parameter settings on OTB100 benchmark. Fig.10 shows that as we increase the value of k, the tracking performance drops rapidly, and the best parameter setting is 3. We infer that the last layer with 256 channels which is equipped with the attention module does not need long-range interaction between the channels, the short-range interaction could provide enough importance differences to help the network pay more attention on the dominant features.

2) COMPONENTS ANALYSIS

To verify the contributions of each component in the proposed HA-SiamVGG, we implement and evaluate three variations



(a) (b)

FIGURE 10. The overall precision plot(a) and success plot(b) for the analysis of the parameter k in attention module on OTB100 benchmark. The digit in the legend denotes the precision score and AUC score, respectively. The best value of k is 3.

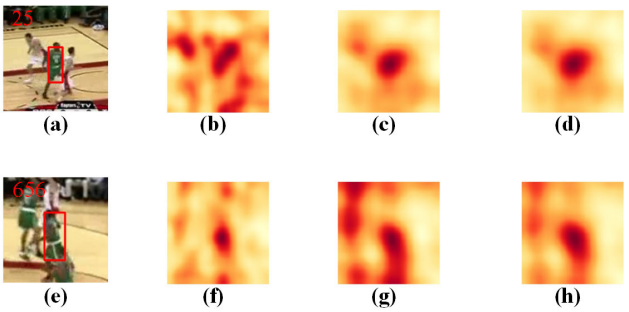


FIGURE 11. The first column(a),(e) are snapshots of the used sequence Basketball on OTB100 benchmark. The target is in the red bounding box. The second column(b),(f) and the third column(c),(g) show the response map generated by Crop4-1 and Conv5-1, respectively. And the last column(d),(h) show the final fused response map.

of our approach. We use the OTB100 benchmark for the ablation analysis.

As shown in Table 3, SiamFC is our baseline algorithm. The improvement is huge when the backbone(AlexNet) is replaced by the modified VGG network. Furthermore, the channel attention mechanism and hierarchical feature fusion strategy both help promote performance. Finally, the proposed HA-SiamVGG integrating all of the components achieves the best result.

3) FEATURE FUSION VISUALIZATION

We visualize the response maps generated by two utilized layers to qualitatively analyze that our hierarchical feature fusion strategy can improve the final result. Fig.11 illustrates that it is unreliable to depend only on one response map when distractors appear and easily lead tracker to drift. For instance, in the 25^{th} frame, response map from the last layer is ideal, whereas in the 656^{th} frame, response map from the last layer is noisy but the response map from the former layer is concentrated. Thus, response maps from different layers are complemental, and the proposed fusion strategy can make use of both spatial and semantic information to generate a high-quality response map, to alleviate mentioned problems and make the tracker more robust.

4) CHANNEL ACTIVATION VISUALIZATION

In order to analyze the effect of the proposed channel attention mechanism, we first visualize the activation of the last layer with this module or not. Similarly, we use the first frame of the Basketball sequence on OTB100 benchmark. Fig.12(a) illustrates that the proposed residual attention mechanism can expand the influence of important channels while maintaining the capacity of the less important features. Thus, the learnt feature representation is more discriminative and help the tracker discriminate the target from background.

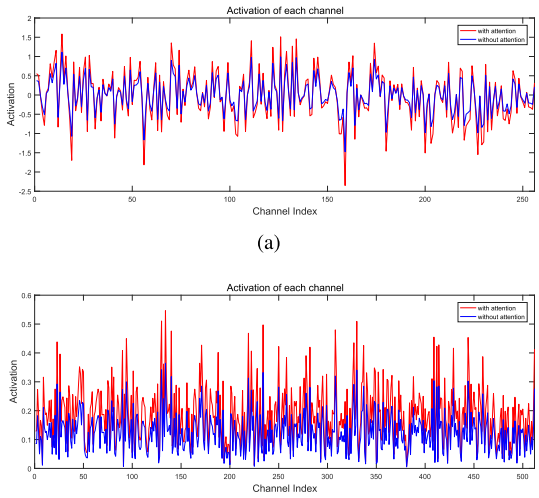


FIGURE 12. The top (a) shows the last layer(Conv5-1) activation of each channel with attention mechanism or not, and bottom (b) shows the former layer(Crop4-1) activation of each channel with attention mechanism or not.

From our experiments, it is better to employ only the last layer with the attention module. In other words, the former convolutional layer cannot benefit from the attention block even we fuse the features from two layers. We also provide the visualization of activation of the former layer in Fig.12(b). Compared to the last layer, the response values of the former layer are restricted within a small range. Furthermore, even if we add the attention block to this layer, the activations are quite small and the gap between different channels still cannot be enlarged, which makes it difficult to define the importance of each channel.

V. CONCLUSION

In this paper, we present a real-time tracking approach that can unveil the power of deep network. The proposed HA-SiamVGG first utilizes a modified VGG16 as the backbone to extract more discriminative features of the target. Then a hierarchical feature fusion strategy in a data-driven manner is introduced to improve the quality of the final response map. Finally, a lightweight residual channel attention mechanism is adopted to make the network focus on dominant features. Extensive experiments on OTB100 and VOT2018 benchmarks demonstrate the effectiveness of the

proposed tracker with a favorable performance against the state-of-the-art tracking approaches.

REFERENCES

[1] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3539–3548.

[2] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 2722–2730.

[3] G. Wu and W. Kang, "Vision-based fingertip tracking utilizing curvature points clustering and hash model representation," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1730–1741, Aug. 2017.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Honolulu, LT, USA, Dec. 2012, pp. 3539–3548.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2015, pp. 91–99.

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

[7] 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.

[8] Y. Qi, S. Zhang, L. Qin, Q. Huang, H. Yao, J. Lim, and M.-H. Yang, "Hedging deep features for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1116–1130, May 2019.

[9] G. Bhat, J. Johnander, M. Danelljan, K. Shahbaz, and M. Felsberg, "Unveiling the power of deep tracking," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 483–498.

[10] 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.

[11] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van De Weijer, "Adaptive color attributes for real-time visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, FL, USA, Jun. 2014, pp. 1090–1097.

[12] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1401–1409.

[13] M. Cen and C. Jung, "Fully convolutional siamese fusion networks for object tracking," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Amsterdam, The Netherlands, Oct. 2018, pp. 850–865.

[14] 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, Utah, USA, Jun. 2018, pp. 4834–4843.

[15] A. He, C. Luo, X. Tian, and W. Zeng, "Towards a better match in siamese network based visual object tracker," in *Proc. Eur. Conf. Comput. Vis. Workshop*, Munich, Germany, Sep. 2018, pp. 132–147.

[16] 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.*, Munich, Germany, Sep. 2018, pp. 8971–8980.

[17] G. Wang, C. Luo, Z. Xiong, and W. Zeng, "SPM-tracker: Series-parallel matching for real-time visual object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, NV, USA, Jun. 2019, pp. 3643–3652.

[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] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1420–1429.

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

[21] O. Russakovsky, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[22] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1763–1771.

[23] L. Zhang, A. Gonzalez-Garcia, J. V. D. Weijer, M. Danelljan, and F. S. Khan, "Learning the model update for siamese trackers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 4010–4019.

[24] 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.*, Salt Lake City, Utah, USA, Jun. 2018, pp. 8971–8980.

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

[26] K. Chen and W. Tao, "Once for all: A two-flow convolutional neural network for visual tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 12, pp. 3377–3386, Dec. 2018.

[27] 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. 7952–7961.

[28] Y. Kuai, G. Wen, and D. Li, "Hyper-siamese network for robust visual tracking," *Signal, Image Video Process.*, vol. 13, no. 1, pp. 35–42, Jul. 2018.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, Jun. 2018, pp. 7132–7141.

[30] S. Woo, J. Park, J. Lee, and I. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.

[31] W. Du, Y. Wang, and Y. Qiao, "RPAN: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3725–3734.

[32] P. Gao, R. Yuan, F. Wang, L. Xiao, H. Fujita, and Y. Zhang, "Siamese attentional keypoint network for high performance visual tracking," *Knowl.-Based Syst.*, vol. 193, Apr. 2020, Art. no. 105448.

[33] D. Li, G. Wen, Y. Kuai, and F. Porikli, "End-to-end feature integration for correlation filter tracking with channel attention," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1815–1819, Dec. 2018.

[34] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, Jun. 2018, pp. 4854–4863.

[35] P. Gao, Q. Zhang, F. Wang, L. Xiao, H. Fujita, and Y. Zhang, "Learning reinforced attentional representation for end-to-end visual tracking," *Inf. Sci.*, vol. 517, pp. 52–67, May 2020.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[37] 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)*, Long Beach, NV, USA, Jun. 2019, pp. 4591–4600.

[38] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, Jun. 2018, pp. 4844–4853.

[39] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-net: Efficient channel attention for deep convolutional neural networks," 2019, *arXiv:1910.03151*. [Online]. Available: http://arxiv.org/abs/1910.03151

[40] 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.

[41] M. Kristan, "The sixth visual object tracking vot2018 challenge results," in *Proc. Eur. Conf. Comput. Vis. Workshop*, Munich, Germany, Sep. 2018, pp. 3–53.

[42] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 4, 2019, doi: 10.1109/TPAMI.2019.2957464.

[43] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 459–474.

[44] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1308–1317.

[45] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1369–1378.

[46] Y. Li and X. Zhang, "SiamVGG: Visual tracking using deeper siamese networks," 2019, *arXiv:1902.02804*. [Online]. Available: http://arxiv.org/abs/1902.02804
[47] 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.

CHAOYI ZHANG received the B.E. degree in automation from Jiangnan University, Wuxi, China, in 2018, where he is currently pursuing the master's degree in control science and engineering. His current research interests include deep learning, computer vision, and object tracking.

HOWARD WANG received the bachelor's and master's degrees from the Department of Automation, Zhejiang University, China, in 2001 and 2004, respectively, and the Ph.D. degree from The University of Auckland, New Zealand. His interests include but not limited to image processing, CCTV video intelligent analysis and processing, large scale video surveillance systems, and deep learning in computer vision.

JIWEI WEN received the Ph.D. degree in control science and control engineering from Jiangnan University, Wuxi, China, in 2011. From 2015 to 2016, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, The University of Auckland. He is currently an Associate Professor with the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. His research interests include stochastic switched systems, model predictive control, and T-S fuzzy modeling and control. He serves as an Associate Editor of *International Journal of Sensors, Wireless Communications and Control.*

LI PENG received the Ph.D. degree from the School of Information Engineering, University of Science and Technology Beijing, in 2002. He is currently a Professor with the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. He is a member of the Chinese Computer Association, and also Chinese Artificial Intelligent Association. His research interests are computer simulation, intelligent control, and visual wireless sensor networks.