IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Siamese Network Using Adaptive Background Superposition Initialization for Real-Time Object Tracking

JUNAN ZHU[1,2], TAO CHEN[1,*], and JINGTAI CAO[1]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China

[*]Corresponding author: Tao Chen (e-mail: chent@ciomp.ac.cn).

This work was supported in part by the National Natural Science Foundation of China under Grant 61601195 and 61605199.

ABSTRACT Object tracking has become widespread in many fields, such as autonomous vehicles, video surveillance and robotics. However, it is far from the requirements for real-world applications. Recently, Siamese network based trackers have attracted high attention by balancing accuracy and speed. Because these trackers only learn a similarity measurement model via off-line training, the exemplar branch has insufficient discriminant information to adapt to the constantly changing appearance of the target in subsequent frames. We propose a Siamese network based tracker that improves upon tracking performance as follows. First, an adaptive background superposition initialization is proposed and used in the exemplar branch to make full use of the limited prior information in the first frame. Second, a light-weight convolutional neural network is proposed and applied as the tracker's backbone; it compresses the dimensions of the feature to ensure speed and accuracy. Third, the channel attention module is introduced into our tracker and integrated with adaptive background superposition initialization. The feature map of the original exemplar image and its background changed image are adjusted by a channel attention model and fused to enhance the representation of the exemplar image. The GOT-10k dataset is applied to train our tracker. Finally, experiments on the object tracking benchmark (OTB) and visual object tracking (VOT) demonstrate the effectiveness of our proposed approach compared with state-of-the-art trackers.

INDEX TERMS Adaptive background superposition initialization, channel attention module, object tracking, Siamese network.

I. INTRODUCTION

Visual object tracking is a vital and fundamental research topic in computer vision as it provides a basis for further semantic level analysis (action recognition, scene recognition, etc.). In a model-free object tracking task, given the initial bounding box of the target in the first frame, the objective is to localize the same target in subsequent frames. However, lack of prior information of the arbitrary target makes it difficult to design trackers; furthermore, there are many challenges in object tracking, such as illumination variations, occlusions, deformations, etc. Object tracking has many important real-world applications, such as, autonomous vehicles, video surveillance, and robotics. Therefore, it is necessary to build a tracking algorithm that can balance speed and accuracy.

Recently, tracking algorithms based on correlation filters [1]-[8] have been widely used. These trackers significantly reduce computational complexity and enable the algorithm to achieve real-time requirements by using the properties of the circulant matrix that can be diagonalized in Fourier space. However, their tracking results are inaccurate. Subsequently, convolutional neural networks (CNNs) have been proven as excellent for target detection, image classification, and other computer vision tasks. CNNs have incredible performance in feature representation. Thus, they have been introduced into correlation filters as feature extractors; trackers with deep features [9]-[14] have achieved state-of-the-art results on object tracking benchmark (OTB) [15], [16] and visual object tracking (VOT) [17]. However, the extractor for deep features is computationally complex, and it is difficult to achieve an acceptable real-time speed even if GPU, FPGA, and other hardware are used to accelerate the computing speed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

Siamese networks are another solution to solve the problem of object tracking [18]-[24] as they have achieved good results in face recognition, image matching, and other computer vision tasks. Object tracking can be expressed as a similarity learning problem; the similarity between an exemplar image and a candidate image is evaluated by learning a similarity measure function. Because these Siamese trackers only learn a similarity measurement model via off-line training, the exemplar branch has insufficient discriminant information to adapt to the constantly changing appearances of the targets in subsequent frames. Furthermore, AlexNet [8] is still the backbone of most Siamese trackers, which is a relatively shallow network (only five convolutional layers and two max pooling layers). Thus, this simple network cannot fully exploit the capabilities of deep neural networks.

In this work, we propose a novel tracking algorithm based on Siamese networks. The main contributions are as follows:

1) We propose an adaptive background superposition initialization strategy. The background color of the target in the exemplar image is adaptively changed; then, the feature maps of the original exemplar image and its background changed image are fused to enhance the feature representation of the target.
2) We design an efficient light-weight network that contains thirteen convolutional layers and three max pooling layers. This network can compress the dimensions of the feature to ensure speed and accuracy and it is used as the backbone of our tracker.
3) We introduce a channel attention module [25], which is an improvement of SEblock in SENet [26], to our tracker in the exemplar branch. This module is used to re-weight the channels of the feature maps to improve feature extraction. Meanwhile, the channel attention module can be integrated with the adaptive background superposition initialization.
4) Extensive experiments conducted on OTB benchmarks [15], [16] and VOT challenge [17] demonstrate that our algorithm achieves competitive results compared with those of state-of-the-art trackers.

The remainder of this article is organized as follows. Closely related works are presented in Section II. The details of out tracker are provided in Section III. Experiments are conducted on multi benchmarks and the results are shown in Section IV. Finally, Section V concludes this work.

II. RELATED WORK

In this section, we introduce three aspects of work related to our study: (i) deep features for correlation filter based tracking; (ii) Siamese network based tracking; (iii) attention mechanisms.

A. DEEP FEATURES FOR CORRELATION FILTER BASED TRACKING

Because the accuracy of image classification has been improved by nearly 10% from 74.2% to 83.6% by using AlexNet at the Large Scale Visual Recognition Challenge (ILSVRC) 2012 [27]. CNNs have gained unprecedented attention in computer vision. The close integration of correlation filter trackers and CNN has promoted the development of object tracking in recent years. DeepSRDCF [12] applies deep features from a pre-trained VGGNet [28] and principal component analysis (PCA) is used to reduce the data dimensionality. C-COT [11] applies features from three convolutional layers by using a pre-trained VGGNet and learns a discriminative continuous convolution operator to improve performance. ECO [10] combines deep features with handcrafted features (Histogram of Oriented Gradients (HOG) [29] and Color Names (CN) [30]) and factorizes the convolution operator to reduce the number of parameters. UPDT [9] introduced the adaptive fusion of deep and shallow features to fully utilize the capabilities of CNNs. However, the speeds of these trackers struggle to meet the requirements of real-life applications.

B. SIAMESE NETWORK BASED TRACKING

To overcome the low speed of trackers that use pre-trained networks as feature extractor. Recently, the Siamese network has drawn great attention in the field of object tracking. GOTURN [31] introduced the Siamese network into the field of object tracking for the first time and uses a simple feed-forward network without online training. SINT [23] matches the given target in the initial frame with the candidates in next frame, and then returns the most similar target as determined by the learnt matching function. Re3 [32] uses a recurrent network to obtain better features generated by the exemplar branches. Inspired by the correlation methods. SiamFC [22] first introduced a correlation layer to the Siamese network. CFNet [24] improves SiamFC by adding a correlation filter to the exemplar branch to update the template model, making the Siamese network more robust to appearance changes. DSiam [20] learns the appearance variation of the target and background suppression from previous frames online. Dong *et al.* [33] introduced triplet loss to improve the performance of SiamFC and CFNet. Kuai *et al.* [34] introduced target objectness model and the target template model to improve the performance of SiamFC. SiamRPN [35] introduced the region proposal network (RPN) to generate the bounding boxes of targets. However, the backbone networks utilized in most Siamese trackers, such as AlexNet, are still shallow, so they cannot fully exploit the capabilities of deep neural networks.

C. ATTENTION MECHANISMS

Attention modules have been used widely in the field of deep learning. In 2015, Bengio *et al.* [36] first introduced an attention module into natural language processing (NLP). Then, attention module was implemented in computer vision applications, such as image classification [37], object detection [38], and object tracking [18], [19], [21]. Li *et al.* [39] introduced channel attention to adaptively impose channel-wise weight on the integrated features in correlation filter

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

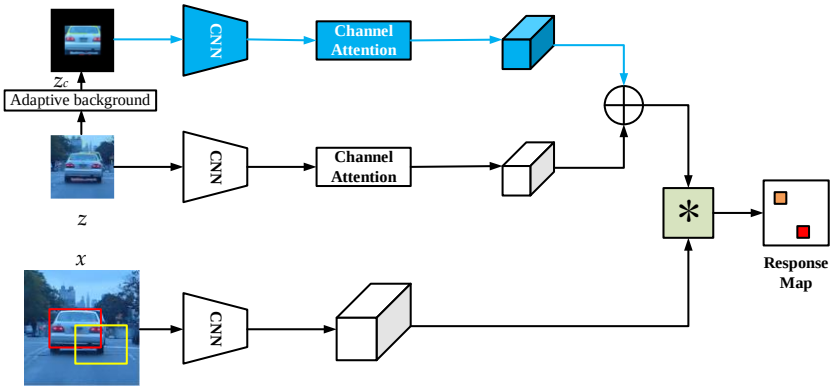


FIGURE 1. Overall architecture of our tracker. It contains two asymmetric branches, namely the exemplar and search branches. The blue branch is the adaptive background superposition initialization branch, which is part of the exemplar branch.

tracking. RASNet [40] introduced general attention, residual attention, and channel attention to Siamese trackers. DensSiam [19] applies self-attention to make the network focus on non-local features. SA-Siam [21] applies channel attention module in its semantic branch to improve the distinguishing ability of the semantic branch. In this work, we apply a channel attention module to improve feature extraction.

III. PROPOSED ALGORITHM

The overall of our algorithm is introduced in Section III.A. Then, the adaptive background superposition initialization, convolutional network architecture, and channel attention module are presented in detail in Sections III.B, III.C, and III.D. Finally, the details of how to train this network are described in Section III.E.

A. ALGORITHMIC OVERVIEW

The overall architecture of our tracker is shown in Fig. 1. This Siamese network is made up of two asymmetric branches, namely the exemplar and search branches. The exemplar branch consists of a main branch and an adaptive background branch; they each contain a CNN and a channel attention module. However, the search branch only uses a convolutional network. Finally, the exemplar and search branches are joined by a convolution layer, and the convolution kernel is the feature map of the exemplar image. The final output is a response map that represents the similarity between the search and exemplar images.

After training, the tracking process can be described as a cross-correlated operation using a pre-trained Siamese network:

$$f(z,x) = \varphi(x) * (\alpha(\varphi(z)) + \lambda\alpha(\varphi(z_c))) + b \qquad (1)$$

where $\varphi(\bullet)$ denotes the convolutional network, $\alpha(\bullet)$ denotes the channel attention module, λ is the merge parameter, and b denotes a bias. The workflow in Fig. 1 shows that two inputs are required for our Siamese network, namely a target image patch z and a candidate search region x. z_c is the background color changed image of z, and it is calculated via the background color adaptive selection algorithm. A response map that represents the similarity between the search image patch and the target image patch will return. Then, we determine the maximal value of the response and map it onto the original frame to acquire the position of our target in the following frame.

B. ADAPTIVE BACKGROUND SUPERPOSITION INITIALIZATION

In a Siamese network based tracker, if we update the model, it will introduce drift over time due to a miss-match of the target. Thus, the model is not updated in almost all Siamese network based trackers. This makes the feature representation of the target obtained from the initial frame crucial. Thus, we propose an adaptive background superposition initialization strategy to enhance the feature representation of the targets.

FIGURE 2. The top images are the original exemplar images and the bottom images are their background changed images. The sequences from left to right are *Jumping*, *Car25*, and *Walking2* from the OTB benchmarks.

According to visual common sense, the greater the difference between the target and the background, the easier it is to distinguish the target. Therefore, we design a background color adaptive selection algorithm to increase the contrast between the target and the background so as to improve the distinguishing ability. As shown in Fig. 2, the exemplar

TABLE I
BACKBONE ARCHITECTURE OF OUR TRACKER. ALL THE CONVOLUTIONAL LAYERS ARE INTEGRATED WITH RELU EXCEPT THE LAST ONE. 'CONV*-BN' STANDS FOR CONVOLUTIONAL LAYER WITH BATCH NORMALIZATION LAYER. 'CONV*'STANDS FOR CONVOLUTIONAL LAYER. 'MP*' STANDS FOR THE MAX POOLING LAYER.

				Activation Size		
Layer	Kernel Size	Out and In Chan.	Stride	For Exemplar	For Search	Chan.
				135 × 135	263 × 263	×3
CONV1-BN	3 × 3	32 × 3	1	133 × 133	261 × 261	×32
CONV2-BN	3 × 3	64 × 32	1	131 × 131	259 × 259	×64
MP1	2 × 2		2	65 × 65	129 × 129	×64
CONV3-BN	3 × 3	128 × 64	1	63 × 63	127 × 127	×128
CONV4-BN	1 × 1	64 × 128	1	63 × 63	127 × 127	×64
CONV5-BN	3 × 3	128 × 64	1	61 × 61	125 × 125	×128
MP2	2 × 2		2	30 × 30	62 × 62	×128
CONV6-BN	3 × 3	256 × 128	1	28 × 28	60 × 60	×256
CONV7-BN	1 × 1	128 × 256	1	28 × 28	60 × 60	×128
CONV8-BN	3 × 3	256 × 128	1	26 × 26	58 × 58	×256
MP3	2 × 2		2	13 × 13	29 × 29	×256
CONV9-BN	3 × 3	512 × 256	1	11 × 11	27 × 27	×512
CONV10-BN	1 × 1	256 × 512	1	11 × 11	27 × 27	×256
CONV11-BN	3 × 3	512 × 256	1	9 × 9	25 × 25	×512
CONV12-BN	1 × 1	256 × 512	1	9 × 9	25 × 25	×256
CONV13	3 × 3	256 × 256	1	7 × 7	23 × 23	×256

images comprise a target and its background. If the pixel mean of the target and its background is within a certain range and the mean of the target is more than 127, we fill the background with black; if the pixel mean of the target and its background is within a certain range and the mean of the target is less than 127, we fill the background with white; otherwise, we do not change the color of background. The color of the background area is determined by

$$B_{color} = \begin{cases} black & if \ \|M_t - M_b\| \le T \ and \ M_t \ge 127 \\ white & if \ \|M_t - M_b\| \le T \ and \ M_t < 127 \\ remain\ the\ same & otherwise \end{cases}, \quad (2)$$

where M_t and M_b stand for the means of pixels in the area of the target and its background, respectively. T is the threshold of the means of pixels between the area of the target and its background.

The feature map of the original exemplar image and its background changed image are adjusted by a channel attention model and then fused to enhance the representation of the exemplar image.

C. CONVOLUTIONAL NETWORK ARCHITECTURE

We apply a CNN as the backbone of our tracker. Our backbone network has thirteen convolutional layers and three max pooling layers. The convolutional filters are 3 × 3 filters and 1 × 1 filters, the number of channels is doubled after every max pooling operation, and 1 × 1 filters are placed between 3 × 3 convolutions. Batch normalization layers are also used after convolutional layers to accelerate the training and regularize the model. Details of our convolutional network are listed in Table 1.

The significant feature of our network is that 1 × 1 filters are used many times as compared with AlexNet and VGGNet. 1 × 1 filters can compress the number of channels, which reduces the decline in speed when using a deeper convolutional network. 1 × 1 filters reduce the number of parameters and allow our tracker to run on some small GPU memory devices. Furthermore, 1 × 1 filters can increase the non-linearity and mix cross-channel information to improve the generalization capability of the network.

D. CHANNEL ATTENTION MODULE

The feature map $U \in R^{W \times H \times C}$ contains multiple channels with different types of visual patterns. The information of some channels are more useful than that of others. Channel attention module is used to re-weight the channels and it has been validated as effective for some computer vision tasks. Thus, the channel attention module in CBAM [25] is introduced to our network to improve performance. This channel attention module comprises two pooling layers, a shared multi-layer perceptron (MLP), and two sigmoid activation layers, whereas the MLP consists of two fully connected layers. The overall architecture of this module is shown in Fig. 3.

This channel attention module is an improvement of SEblock in SENet [21]. SEblock only uses a global average pooling layer. Whereas global average pooling has feedback for every value on the feature map, global max pooling only has feedback where the value is the maximum in a sub-region,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

which can be used as a complementary to global average pooling. Consequently, we introduce this improved channel attention module to our tracker. Given the input $U \in R^{W \times H \times C}$, the feature map of the final output $V \in R^{W \times H \times C}$ can be calculated by

$$V = Sigmoid(MLP(AvgPool(U)) + MLP(MaxPool(U))) \otimes U, \quad (3)$$

where $\otimes$ signifies a channel-wise multiplication function.

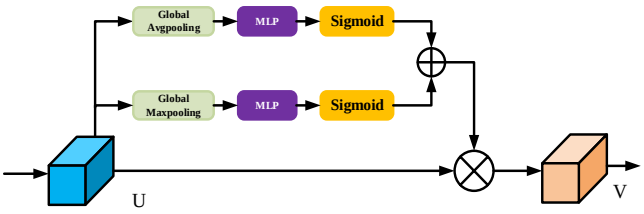


FIGURE 3. **Overall architecture of the channel attention module.**

E. TRAINING THE NETWORK

To train our Siamese network, we select GOT-10k [41], which contains more than 10000 video sequences and more than 1.5 million annotated bounding boxes, as the training dataset. This dataset has a wide range of scenarios and objects. Therefore, it is very suitable for training deep trackers.

Generation of Training Samples. Considering the speed of image transformation in training, each video frame of the dataset is cropped and scaled offline in advance to obtain a 127×127 exemplar image and a 255×255 search image. This processing of images is similar to that in SiamFC [22] and CFNet [24]. Before cropping and scaling the exemplar image *z*, the center position of the target and the target size (*w*, *h*) are obtained according to the labeled information of the target. The square exemplar image is cropped from the original frame and the center is the same as the target. The scale factor *s* can be formulated as

$$s(w+2c) \times s(h+2c) = 127 \times 127, \quad (4)$$

where $c=(w+h)/4$ is the context of our target image patch. Each element $u \in D$ is divided into positive or negative samples according to the following equation:

$$y[u] = \begin{cases} +1 & if \ k\|u - p_c\| \le R \\ -1 & otherwise \end{cases}, \quad (5)$$

The total stride of the network is *k*, p_c denotes the center of the score map, and *R* is the radius used to distinguish between positive and negative samples.

Loss function. The logic loss function is similar to that in SiamFC [22] and it can be formulated as

$$L(y,v) = \frac{1}{|D|} \sum_{u \in D} l(y[u], v[u]), \quad (6)$$

where *v* denotes the returned score value of the inputted target-candidate image pair and $y \in [+1,-1]$ is the ground truth label. Then, the parameters θ of the Siamese network can be calculated by minimizing the loss function:

$$\arg\min_{\theta} \mathbb{E}_{z,x,y} L(y, f(z,x;\theta)), \quad (7)$$

After stochastic gradient descent (SGD) is applied, (7) can be solved.

IV. EXPERIMENTS

In this section, our algorithm is compared with other state-of-the-art algorithms. Experiments are conducted on OTB benchmarks and VOT challenge. We train and evaluate our tracker using Python and Pytorch on an Intel i7-8700K CPU with a GTX 1080 Ti GPU card. Implementation details are shown in Section IV.A, and the experiments on OTB benchmarks and VOT challenge are described in Sections IV.B and IV.C. Ablative studies and further discussion about the proposed light-weight network, adaptive background superposition initialization, and channel attention module are described in Section IV.D. The effects of different training datasets on tracking performance are shown in section IV.E. Finally, the analysis of merge parameter is shown in Section IV.F.

A. IMPLEMENTATION DETAILS

Training. Our proposed tracker are trained on GOT-10k [41] by solving (7) with SGD. We set the radius *R* to 8 pixels and the batch size to 8. Kaiming Normal Initialization [42] is used to initialize the parameters of our network. The learning rate exponentially decays from 10^{-2} to 10^{-5}. The weight decay is set to 0.0005 and the momentum is 0.9.

Tracking. To further improve the performance of our algorithm, we follow the SiamDW [43] to use larger size image in tracking phase. The sizes of the exemplar image and the search image are set to 135×135 and 263×263. The merge parameter λ is set to 0.21. *T* is set to 60. Scale change often accompanies tracking; to obtain a better performance when dealing with scale variation, three fixed scales {0.96;1;1.04} are used to search the target. Linear interpolation is used to update the scale, and the scale factor is set to 0.59. The cosine window function is introduced and applied in the score map to penalize large displacements, and the window influence is set to 0.27. The tracking algorithm can run at an average speed of 50 frames per second (fps).

B. EXPERIMENTS ON THE OTB DATASET

OTB benchmarks [15], [16] are widely used tracking benchmarks in recent years, which consist of three datasets, namely OTB-2013, OTB-50 and OTB-100. One-pass evaluation (OPE) is applied to evaluate the performance of trackers in OTB benchmarks. Overlap ratio and center location error are applied to obtain their own success and precision plots, respectively. Detailed information on these evaluation indicators can be found in [15], [16].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

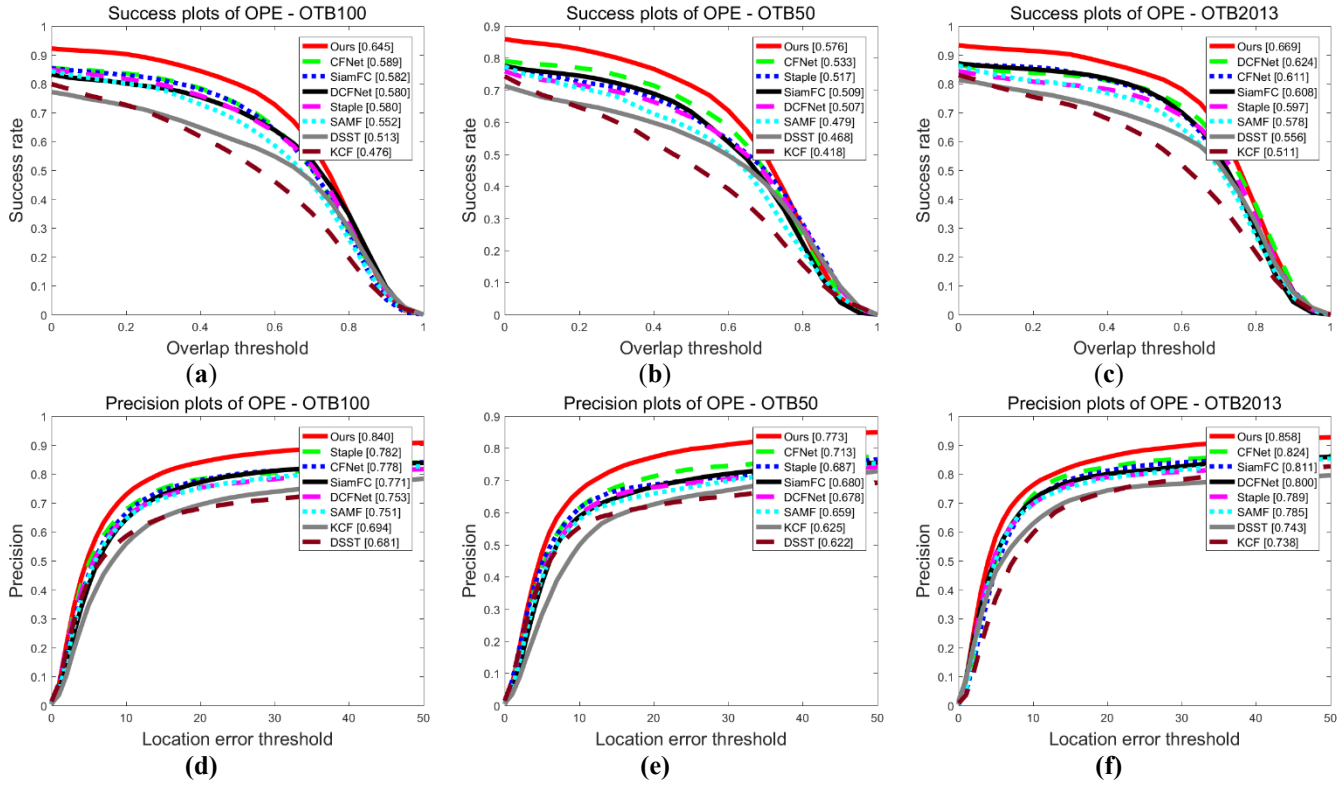


FIGURE 4. Success and precision plots of OPE on OTB-100, OTB-50, and OTB-2013, respectively

1) COMPARISON WITH STATE-OF-THE-ART TRACKERS

Our trackers are compared with two types of algorithms, namely correlation filter-based trackers (DSST [2], Staple [3], KCF [4], and SAMF [5]) and Siamese network based trackers (DCFNet [12], SiamFC [16], and CFNet [18]). The experimental results of the eight trackers are shown in Fig. 4. The threshold for distance precision is 0 to 50 pixels and the threshold for overlap precision is 0 to 1.

Fig. 4 shows that our proposed tracker achieves the best overall performance with AUC scores of 64.5%, 57.6%, and 66.9% on OTB-100, OTB-50, and OTB-2013, respectively.

2) ATTRIBUTE-BASED EVALUATION

In OTB benchmarks, all sequences are annotated with eleven attributes, and each sequence contains several challenges. To measure the performance of our algorithms on different attributes, the experiment is conducted on OTB-100. The success plots of different attributes are shown in Fig. 5.

The success rate of challenge attributes is applied to measure the performances of tracking algorithms in handling specific challenging situations. As shown in Fig. 5, our tracker ranks first except for low resolution and out-of-view, which are slightly lower than those of other trackers. Fig. 5 demonstrates that our tracker has good robustness and can handle almost all challenging situations.

3) QUALITATIVE ANALYSIS

We select five representative image sequences (*Bird1*, *DargonBaby*, *Jump*, *MotorRolling*, and *Soccer1*) from the OTB-100 dataset. Almost all challenges can be found in these sequences. The bounding boxes of the targets predicted by different tracking algorithms are shown in Fig. 6. To better present the results of different algorithms, we only show the three best trackers (CFNet, SiamFC, and DCFNet) along with our tracker.

As shown in Fig. 6, our tracker has high practical performance in handling various challenges. In the *Jump* and *MotorRolling* sequences, our tracker can always track the targets, but other trackers fail; in the *Bird1*, *DargonBaby*, and *Soccer1* sequences, other trackers can also follow the targets, but our bounding boxes of the targets are more accurate than those of other trackers.

C. EXPERIMENTS ON THE VOT DATASET

TABLE 2
COMPARISON OF TRACKERS ON VOT2018

	EAO	Accuracy	Robustness
Ours	0.2397	0.5297	0.501
DSiam	0.1963	0.5123	0.646
SiamFC	0.1880	0.5029	0.585
DCFNet	0.1825	0.4702	0.543
DensSiam	0.1740	0.4621	0.688
Staple	0.1694	0.5296	0.688
KCF	0.1349	0.4472	0.773
DSST	0.0788	0.3947	1.452

The VOT challenge is a competition of model-free object tracking algorithms and it has been held every year since 2013. VOT2018 [17] is made up of 60 annotated sequences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

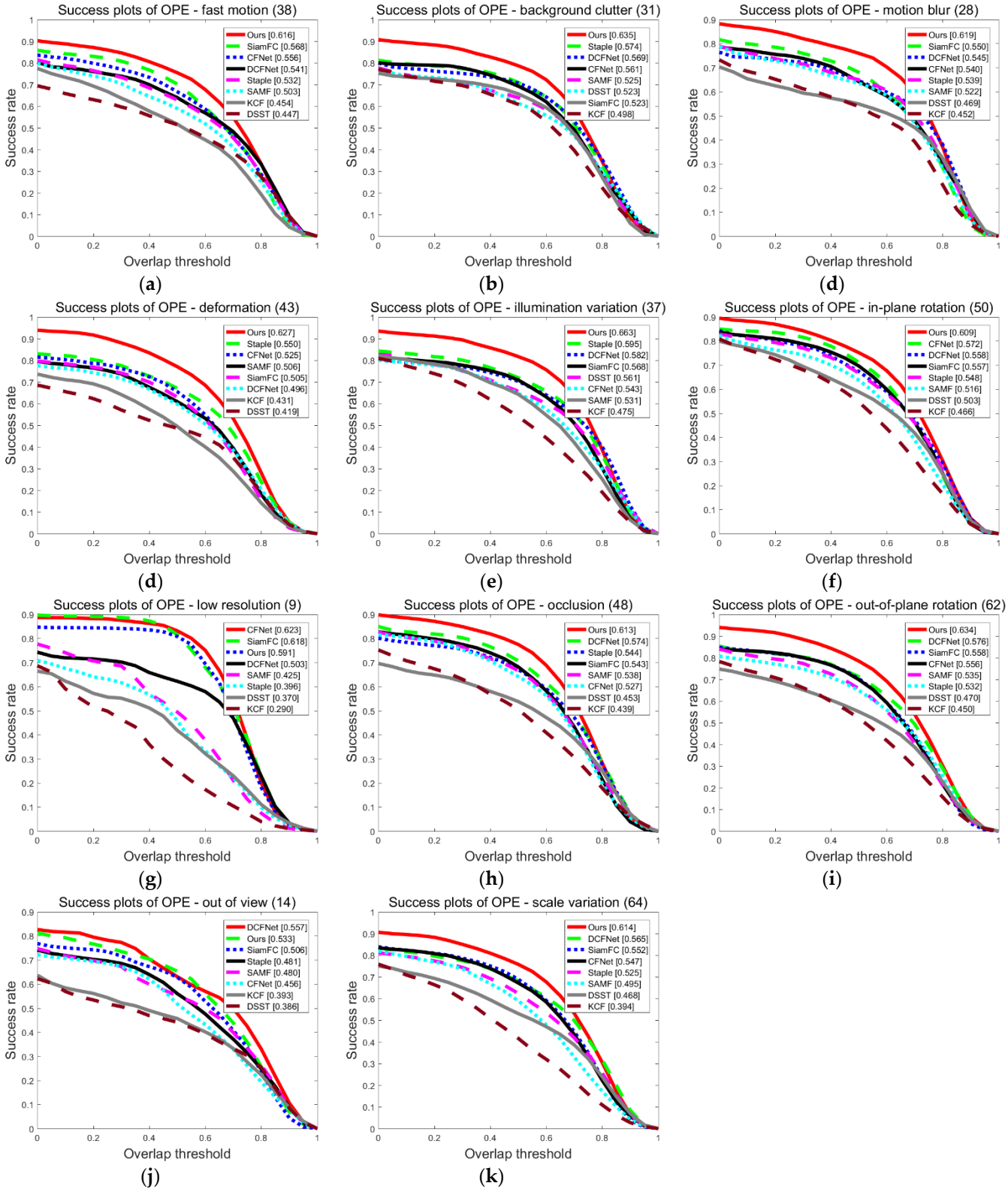


FIGURE 5. Success and precision plots of OPE on OTB-100, OTB-50, and OTB-2013, respectively. The values in brackets in success and precision plots of OPE stands for the area under curve (AUC) value and the scores at local error threshold of 20 pixels.

Expected average overlap(EAO), accuracy and robustness are used to evaluate the performance of difference trackers. Detailed information on these evaluation indicators can be found in [17]. Our tracker are compared with DSST [2], Staple

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

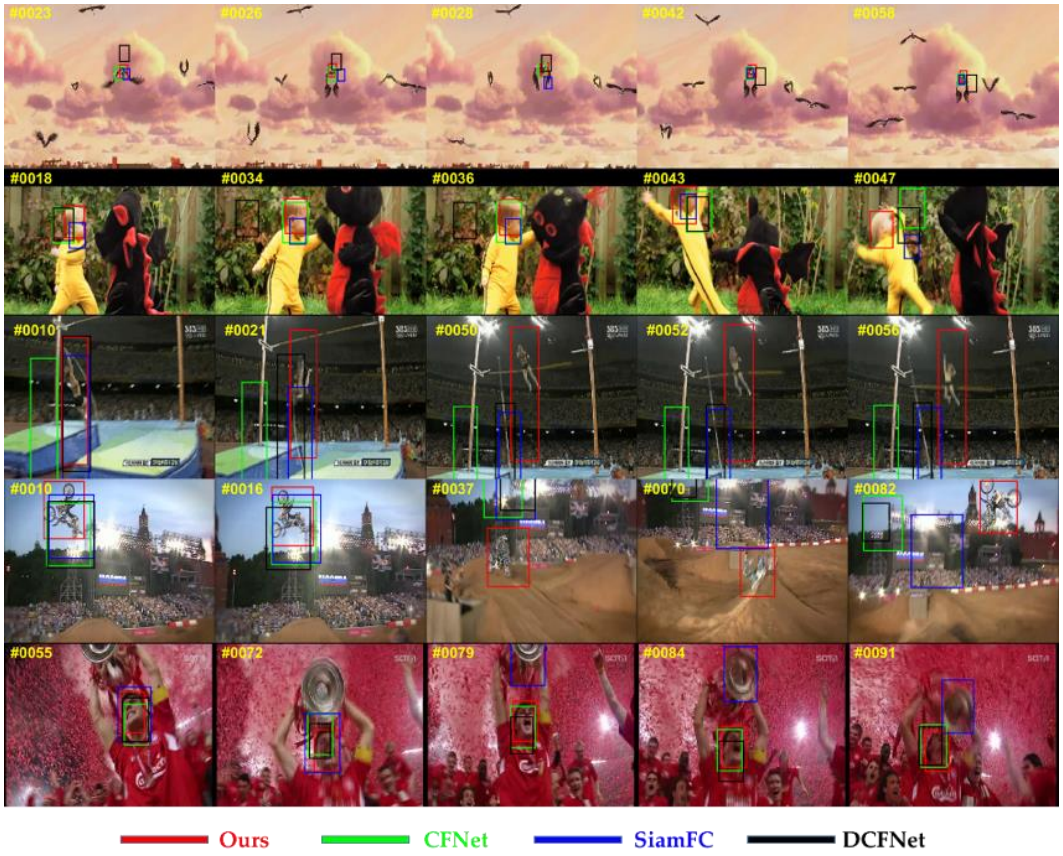


FIGURE 6. Qualitative evaluation of the proposed algorithm compared with CFNet, SiamFC, and DCFNet on five challenging sequences (from top to bottom: *Bird1*, *DargonBaby*, *Jump*, *MotorRolling*, and *Soccer1*).

[3], KCF [4], DCFNet [12], SiamFC [16], DensSiam [19], and DSiam [20], the raw results of these trackers are from http://votchallenge.net/vot2018/results.html. Our tracker is compared with these state-of-the-art trackers and the results are shown in Table 2.

As shown in Table 2, the best three results are marked in red, blue and green. Our tracker ranks first in the three evaluation metrics. EAO is the most important metric in VOT challenge. The EAO score of our method gains relative improvement about 27.5% compared with our baseline tracker SiamFC and it also improves DSiam by 22.1%.

D. ABLATION ANALYSIS

To validate the effectiveness of the proposed light-weight network, adaptive background superposition initialization module, and channel attention module in our tracker, we design and train two other algorithms related to our algorithm. We apply SiamFC as our baseline and compare these trackers to evaluate the impact of the three module on OTB-100. The results are presented in Fig. 7. Among them, the Siam+LWN method is similar to SiamFC, but its backbone is replaced by our proposed light-weight network. Siam+LWN+CA method stands for the tracker with our proposed light-weight network and channel attention module. Siam+LWN+CA+ABSI method is our final tracker that includes light-weight network, channel attention and adaptive background superposition initialization modules.

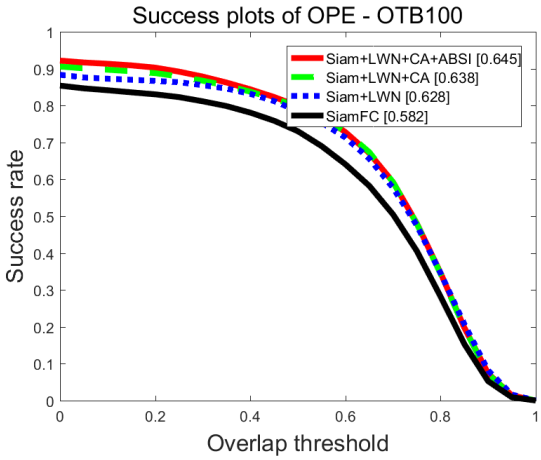


FIGURE 7. Ablation study of our tracker on OTB-100.

1) ROLE OF THE PROPOSED LIGHT-WEIGHT NETWORK

Our algorithm is based on SiamFC. The difference between the Siam+LWN method and SiamFC is that our Siam+LWN method replaces the backbone of SiamFC with our proposed light-weight network. To validate the effectiveness of our

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2937166, IEEE Access

proposed light-weight network, we compare the two trackers. As shown in Fig. 7, the AUC scores of the Siam+LWN method improve SiamFC by 7.90% on OTB-100. This is because we use a deeper light-weight network to replace the shallow backbone in SiamFC, and this network can improve the feature extraction. Simultaneously, our proposed network applies many 1×1 convolutions to compress the number of parameters; thus, our algorithm maintains high accuracy and high speed, which indicates the superior performance of our proposed network.

2) ROLE OF THE CHANNEL ATTENTION MODULE

To validate the effectiveness of our channel attention module, we compare Siam+LWN method and Siam+LWN+CA method. As shown in Fig. 7, the Siam+LWN+CA method improves the AUC scores of the Siam+LWN method by 1.59% on OTB-100. In our algorithm, the channel attention module is learned along with our proposed light-weight network via off-line training using GOT-10k, which makes it more efficient in object tracking tasks. This module can automatically acquire the importance of different channels and utilize them to re-weight the channels of the feature map. This operation can improve feature extraction and thus improve the performance of our proposed tracker.

3) ROLE OF ADAPTIVE BACKGROUND SUPERPOSITION INITIALIZATION

In object tracking, the only prior information of the target is the initial bounding box in the initial frame. To fully utilize this limited prior information, we design an adaptive background superposition initialization strategy. As shown in Fig. 7, our Siam+LWN+CA+ABSI method improves the AUC scores of the Siam+LWN+CA method by 1.10% on OTB-100. In the initialization phase, we first apply a background color adaptive selection algorithm to change the background color of the exemplar image to increase the contrast between the target and the background so as to improve the discrimination ability. The feature map of the original exemplar image and its background changed image are adjusted by a channel attention model and then fused to enhance the feature representation of the target. Furthermore, our final tracker, which includes the three modules, can improve the AUC scores of the Siam+LWN method by 2.70% on OTB-100, which indicates that integrating the adaptive background superposition initialization module with the channel attention module can provide optimal results.

E. TRAINING DATASET ANALYSIS

At present, the training dataset used in most Siamese trackers is ImageNet-VID [44], which only contains 30 object classes with 5.4 thousand videos. The object class of ImageNet-VID is too few and it may cause overfitting in training Siamese tracker. Consequently, Huang *et al.* proposes a large scale dataset called GOT-10k, which has 563 object classes with 10 thousand videos. The statistical comparison of GOT-10k and ImageNet-VID is presented in Table 3. As shown in Table 3, GOT-10k is a richer dataset and more suitable for tracking Siamese tracker. So we apply GOT-10k dataset as our training dataset. The evaluation result of our proposed tracker trained on VID and GOT-10k is shown in Fig. 8.

TABLE 3
COMPARISON OF GOT-10K AND IMAGENET-VID

	GOT-10k	ImageNet-VID
Classes	563	30
Videos	10k	5.4k
BBoxes	1.5M	1.03M

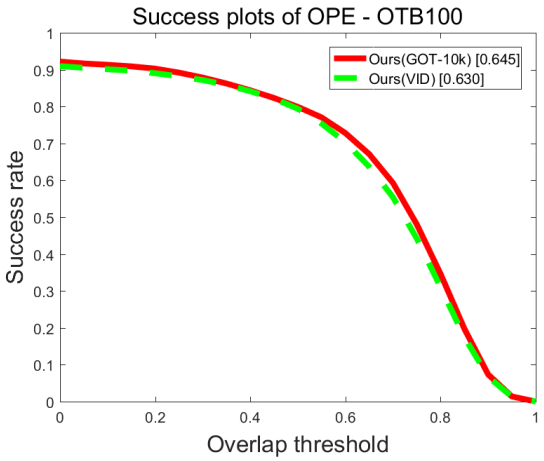


FIGURE 8. Success plot result on OTB-100 of our proposed tracker trained on ImageNet-VID and GOT-10k.

As shown in Fig. 8, Our tracker trained on GOT-10k can improve the AUC scores of trained on ImageNet-VID by 2.38% on OTB-100. This experiment indicates that larger scale training dataset can improve the performance of Siamese tracker.

F. MERGE PARAMETER ANALYSIS

To find the optimal merge parameter λ, different values of λ are used to evaluate our tracker on OTB-100. The results are shown in Fig. 9, the best performance is achieved around λ=0.21. When we increase or decrease the value of λ, AUC score of our tracker will deteriorate. So the merge parameter λ is set to 0.21 in our final tracker.

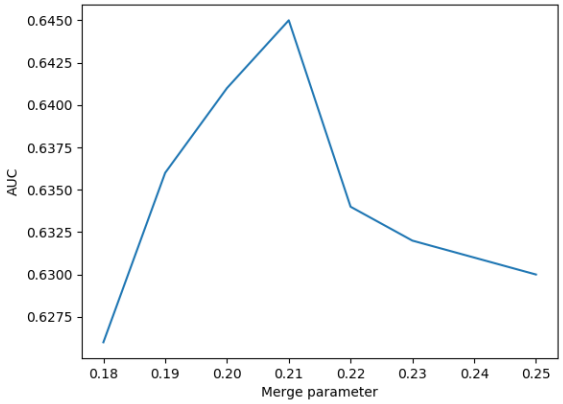


FIGURE 9. AUC scores on OTB-100 using different merge parameter λ.

V. CONCLUSIONS

In this paper, we propose a real-time model-free object tracking algorithm based on Siamese networks. An adaptive background superposition initialization strategy is proposed and integrated with channel attention module to fully utilize the initial frame and enhance the feature representation of the target. Additionally, a novel light-weight network is designed and applied as the backbone of our tracker. This network can compress the dimensions of features to reduce computational complexity. Our tracker is evaluated on OTB-100, OTB-50, OTB-2013, and VOT2018 datasets. Experiments demonstrate that our algorithm achieves competitive results compared with the state-of-the-art trackers.

REFERENCES

[1] C. Rui, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Florence, Italy, 7–13 October 2012, pp. 702-715.

[2] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, p. 1561, 2016. doi: 10.1109/TPAMI.2016.2609928.

[3] 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, 27–30 June 2016, pp. 1401-1409.

[4] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583-596, 2015. doi: 10.1109/TPAMI.2014.2345390.

[5] Y. Li and J. Zhu, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, 8–11 September 2014, pp. 254-265.

[6] R. Z. Han, Q. Guo, and F. Wei, "Content-Related Spatial Regularization for Visual Object Tracking," in *proc. ICME*, San Diego, USA, 23-27 July, 2018.

[7] W. Feng, R. Han, Q. Guo, J. Zhu, and S. Wang, "Dynamic Saliency-Aware Regularization for Correlation Filter-Based Object Tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3232-3245, 2019. doi: 10.1109/TIP.2019.2895411.

[8] P. Zhang, Q. Guo, and W. Feng, "Fast and object-adaptive spatial regularization for correlation filters based tracking," *Neurocomputing*, vol. 337, pp. 129-143, 2019.doi: 10.1016/j.neucom.2019.01.060.

[9] G. Bhat, J. Johnander, M. Danelljan, F. S. Khan, and M. Felsberg, "Unveiling the Power of Deep Tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8-14 September 2018, pp. 483-498.

[10] 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, 21–26 July 2017, pp. 6638-6646.

[11] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016, pp. 472-488.

[12] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Convolutional Features for Correlation Filter Based Visual Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston Massachusetts,USA, 7-12 June 2015, pp. 621-629.

[13] Y. Qi *et al.*, "Hedging Deep Features for Visual Tracking," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 1-1, 2018. doi: 10.1109/TPAMI.2018.2828817.

[14] S. Zhang, Y. Qi, F. Jiang, X. Lan, P. C. Yuen, and H. Zhou, "Point-to-Set Distance Metric Learning on Deep Representations for Visual Tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 187-198, 2018. doi: 10.1109/TITS.2017.2766093.

[15] Y. Wu, J. Lim, and M. H. Yang, "Online Object Tracking: A Benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, 23–28 June 2013, pp. 2411-2418.

[16] Y. Wu, J. Lim, and M. H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, p. 1834, 2015. doi: 10.1109/TPAMI.2014.2388226.

[17] M. Kristan *et al.*, "The Sixth Visual Object Tracking VOT2018 Challenge Results," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8-14 September 2018, pp. 3-53.

[18] W. Qiang, G. Jin, J. Xing, M. Zhang, and W. Hu. (2017) "DCFNet: Discriminant Correlation Filters Network for Visual Tracking," [Online]. Available: https://arxiv.org/abs/1704.04057

[19] M. H. Abdelpakey, M. S. Shehata, and M. M. Mohamed, "DensSiam: End-to-End Densely-Siamese Network with Self-Attention Model for Object Tracking," in *Proc. ISVC*, 2018, pp. 463-473.

[20] Q. Guo, F. Wei, C. Zhou, H. Rui, W. Liang, and W. Song, "Learning Dynamic Siamese Network for Visual Object Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 22–29 October 2017, vol. 1, pp. 1781-1789.

[21] A. He, C. Luo, X. Tian, W. J. c. v. Zeng, and p. recognition, "A Twofold Siamese Network for Real-Time Object Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City Utah, USA, 18-22 June 2018, pp. 4834-4843.

[22] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 8–10 October 2016, pp. 850-865.

[23] 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, 27–30 June 2016, pp. 1420-1429.

[24] 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)*, Honolulu, HI, USA, 21–26 July 2017, pp. 5000-5008.

[25] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8-14 September 2018, pp. 3-19.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City Utah, USA, 18-22 June 2018, pp. 7132-7141.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. NIPS*, Lake Tahoe, NV, USA, 3–6 December 2012, pp. 1097-1105.

[28] K. Simonyan and A. Zisserman. (2015) "Very Deep Convolutional Networks for Large-Scale Image Recognition," [Online]. Available: https://arxiv.org/abs/1409.1556

[29] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, 21–23 September 2005, pp. 886-893.

[30] A. D. Bagdanov, "Color attributes for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 23–28 June 2012, pp. 3306-3313.

[31] D. Held, S. Thrun, and S. J. e. c. o. c. v. Savarese, "Learning to Track at 100 FPS with Deep Regression Networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, 8–10 October 2016, pp. 749-765.

[32] D. Gordon, A. Farhadi, and D. Fox. (2017) "Re3 : Real-Time Recurrent Regression Networks for Object Tracking," [Online]. Available: https://arxiv.org/abs/1705.06368

[33] X. Dong and J. Shen, "Triplet Loss in Siamese Network for Object Tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8-14 September 2018, pp. 472-488.

[34] Y. Kuai, G. Wen, and D. Li, "Masked and Dynamic Siamese Network for Robust Visual Tracking," *Information Sciences*, vol. PP, no. 99, pp. 169-182, 2019. doi: 10.1016/j.ins.2019.07.004.

[35] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City Utah, USA, 18-22 June 2018, pp. 8971-8980.

[36] D. Bahdanau, K. Cho, and Y. J. i. c. o. l. r. Bengio. (2015) "Neural Machine Translation by Jointly Learning to Align and Translate," [Online]. Available: https://arxiv.org/abs/1409.0473

[37] W. Fei *et al.*, "Residual Attention Network for Image Classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu,HI, USA, 21–26 July 2017, pp. 6450-6458.

[38] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person Search via a Mask-Guided Two-Stream CNN Model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 8-14 September 2018, pp. 764-781.

[39] D. Li, G. Wen, Y. Kuai, and F. Porikli, "End-to-End Feature Integration for Correlation Filter Tracking With Channel Attention," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1815-1819, 2018. doi: 10.1109/LSP.2018.2877008.

[40] Q. Wang, Z. Teng, J. Xing, J. Gao, 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 Utah, USA, 18-22 June 2018, pp. 4854-4863.

[41] L. Huang, X. Zhao, and K. Huang. (2019) "GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild," [Online]. Available: https://arxiv.org/abs/1810.11981

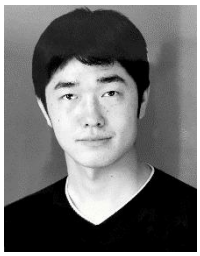
[42] K. He, X. Zhang, S. Ren, and S. Jian, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 7–13 December 2015, pp. 1026-1034.

[43] Z. Zhang, H. Peng, and Q. Wang, "Deeper and Wider Siamese Networks for Real-Time Visual Tracking," [Online]. Available: https://arxiv.org/abs/1901.01660

[44] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *the International Journal of Computer Vision*, vol. 115, no. 3, pp. 211-252, 2015. doi: 10.1007/s11263-015-0816-y.

JUNAN ZHU received the B.E. degree from Northeastern University, Shenyang, China, in 2015. He is currently pursuing the Ph.D. degree in Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. His research interests include object detection, tracking, and recognition.

TAO CHEN is currently a Research Fellow and a Supervisor of Ph.D. Candidates of the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He received his Ph.D. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China, in 2007, His research interests include digital image processing and photoelectric measurement.



JINGTAI CAO is currently an Associate Research Fellow of the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. He received his M.S. degree and Ph.D. degree from Jilin University, Changchun, China, in 2009 and 2017, respectively. His current research interests include digital image processing, object detection and tracking.