Dual-Template Siamese Network with Attention Feature Fusion for Object Tracking

*Minhua LIU, Jiantong SHI, Yu WANG**

Beijing Key Laboratory of Big Data Technology for Food Safety, School of Artificial Intelligence, Beijing Technology and Business University, 100048, Beijing, China

liuminhua@btbu.edu.cn, sjt1525@126.com, wangyu@btbu.edu.cn

Submitted October 11, 2022 / Accepted February 28, 2023 / Online first July 17, 2023

Abstract. *In order to alleviate the adverse effects resulted from complex scenes for object tracking, such as fast movement, mottled background, interference of similar objects, and occlusion etc., an algorithm using dual-template Siamese network with attention feature fusion, named SiamDT, is proposed in this paper. The main idea include that the original ResNet-50 network is improved to extract deep semantic information and shallow spatial information, which are effectively fused using the attention mechanism to achieve accurate feature representation of objects. In addition, a template branch is added to the traditional Siamese network in which a dynamic template is generated together with the first frame image to solve the problems of template failure and model drift. Experimental results on OTB100 dataset and VOT2018 dataset show that the proposed approach obtains the excellent performance compared with the state-of-the-art tracking algorithms, which verifies the feasibility and effectiveness of the proposed approach.*

Keywords

Object tracking, Siamese network, feature extraction, feature fusion, attention mechanism

1. Introduction

Visual object tracking is an important content in the field of computer vision, which can be widely used in security monitoring, human-computer interaction, traffic navigation and so on [1]. Although the object tracking technology has been developed rapidly, many problems still exist in this subject, such as the change of shape, scale, speed, and illumination of the object, motion blur, occlusion and disappearance under the influence of complex scenes [2]. Therefore, it is of great significance to ensure that the above difficulties in the object tracking algorithm can be resolved or relieved, and accurate and robust tracking performance can be achieved.

Traditional object tracking algorithms based on machine learning, such as filtering algorithms, have good tracking accuracy and speed, but their capabilities on feature extraction are weak, which makes them difficult cope with tracking tasks in complex scenes.

With the continuous development of deep learning, object tracking algorithms based on them have become the mainstream due to their excellent performance, and they can be divided into filtering algorithms combined with deep features and Siamese network based algorithms [3].

In recent years, object tracking algorithms based on Siamese network can better balance the accuracy and the speed, so they have been paid more attention to. Tao et al. [4] proposed Siamese instance search for tracking (SINT) method which pioneered to change the problem of object tracking into the one of object matching. Subsequently, Bertinetto et al. [5] proposed the SiamFC algorithm in which a fully convolutional Siamese network with end-to-end training was designed and implemented.

Based on SiamFC algorithm, many improved versions are proposed. SiamRPN [6] algorithm was proposed by the terms of the region proposal network in the Faster R-CNN object detection algorithm, and the multi-scale prediction was replaced by the bounding box regression, so as to obtain a more accurate object boundary, and to further improve the tracking success rate. In SiamRPN++ [7] algorithm a spatially aware sampling strategy was proposed based on the SiamRPN algorithm which enables the Siamese network to break through the limitation of spatial invariance, to increase the depth of the network, and to further improve the tracking performance. SiamMask [8] algorithm adds a mask branch to the fully convolutional Siamese network, which can track and segment the object in real time only by relying on a bounding box of the initial frame, and can improve the segmentation accuracy by enhancing the network loss.

Although the object tracking algorithms based on Siamese network are developing continuously, many deficiencies still exist. First, in most of these methods a shallow convolutional neural network model is used for feature extraction, which makes the extracted features weak in representation, and cannot accurately describe the object to be tracked. Second, only deep semantic features are used

DOI: 10.13164/re.2023.0371

for feature matching, while the spatial information contained in shallow features are ignored. In addition, during the course of long-time tracking, the object template is easy to fail due to the large change of the object. Therefore, only relying on the first frame image of the object as the template will lead to the failure of object tracking.

Inspired by the above ideas, an object tracking algorithm using dual-template Siamese network with adaptive attention feature fusion is proposed in this paper which is built under the SiamFC architecture. Our contributions can be concluded as follows:

(1) In order to fully utilize the rich feature extraction capabilities of deep networks, the originally shallow backbone AlexNet is replaced with a deeper ResNet-50. And the ResNet-50 network is improved including the step size of network, receptive field etc. to make the network more suitable for feature extraction in complex scenes.

(2) The features output from the last three residual blocks of improved ResNet-50 are adaptively fused by attention mechanism to fully utilize the deep semantic information and shallow spatial information, so as to obtain features with stronger representation ability.

(3) In order to resolve the problem that the object template degrades due to large changes of objects during course of the long-term tracking, the first frame image and the previous one of search frame of the object are together used as templates to match with the search frame image.

(4) The proposed approach obtains the excellent performance on both OTB100 dataset and VOT2018 dataset.

The rest of this paper is organized as follows. Firstly, the basic Siamese network architecture is described in Sec. 2. Secondly, the overall framework of the proposed approach and the specific improvements of each part are described in Sec. 3. Then in Sec. 4, extensive experiments and comparisons with other algorithms on OTB100 dataset and VOT2018 dataset are presented to demonstrate the feasibility and effectiveness of the proposed approach. Finally, a conclusion is drawn in Sec. 5.

2. Fully-Convolutional Siamese Network

In fully convolutional Siamese networks for object tracking (SiamFC) a cross-correlation problem is defined, and the similarity of objects from the depth network of Siamese architecture with two branches are learned as shown in Fig. 1. Specifically, one branch is used to learn the features of the object, and the other one is used to learn the features of the search area.

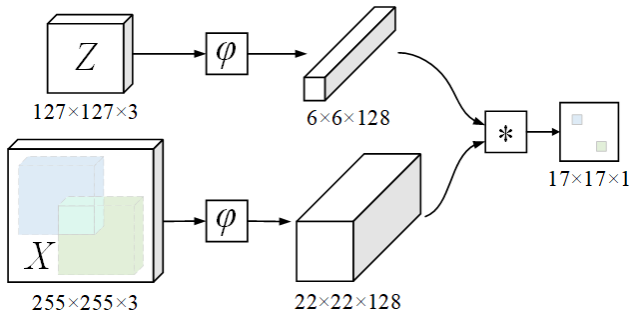


Fig. 1. The architecture of fully convolutional Siamese network.

The object is usually given in the first frame of the video sequence which can be regarded as a template Z, and the task is to find the most similar position between the subsequent search frame image X and template Z in the semantic embedding space. The calculation method on similarity is shown by

$$f(Z,X)=\varphi(Z)*\varphi(X)+b \tag{1}$$

where $\varphi(Z)$ and $\varphi(X)$ respectively represents the result of feature extraction of template Z and subsequent search frame image X, and b is used to simulate the offset of similarity value.

3. Methodology

3.1 Overall Framework of the Proposed Approach

Based on SiamFC algorithm, the overall framework on the proposed approach is composed of feature extraction network, attention feature fusion module and dual-template strategy, as shown in Fig. 2. Among them, the dual-template strategy means that additional template branch is added to the original two convolutional neural networks with shared weights, so that the previous frame image of search frame and the first frame image are together used as templates to build a Siamese architecture consisting of three convolutional neural networks with shared weights. In each branch, the original AlexNet is no longer used in the feature extraction network, but an improved ResNet-50 is used to extract deep and shallow features of template images and search frame images. The attention feature fusion module adaptively fuses the features extracted by the last three residual blocks in the improved ResNet-50 using the channel attention mechanism and the spatial attention mechanism. Then, the fused template features and the subsequent frame image features are matched to calculate the similarity, so as to generate two score maps. Finally, the two generated score maps are weighted to predict the object location and scale, so as to complete the object tracking task.

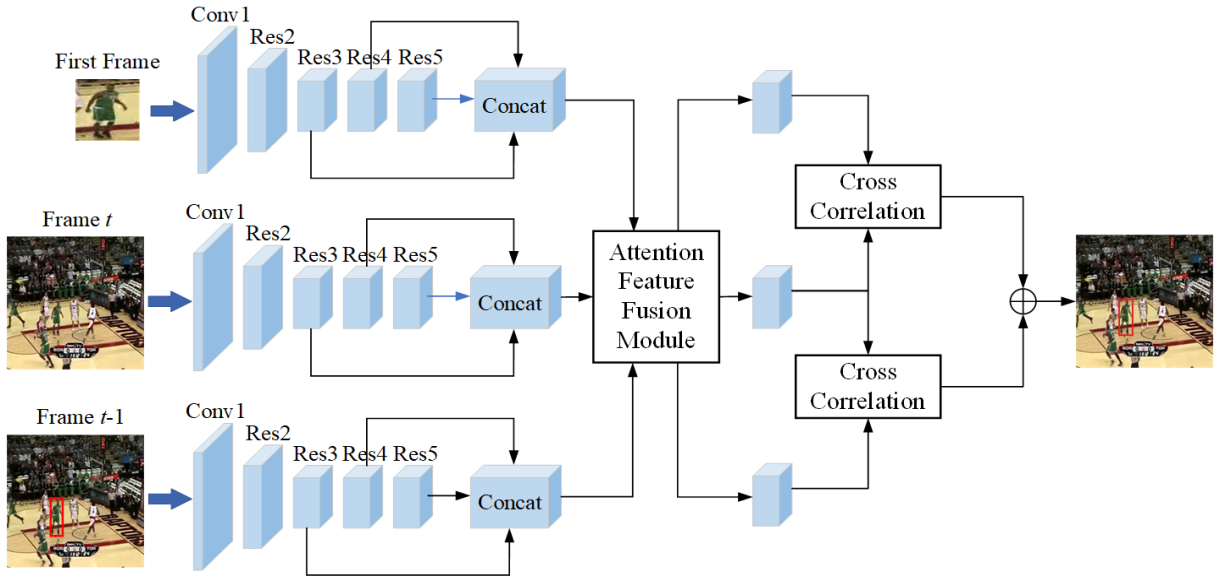


Fig. 2. Overall framework of the proposed approach. Compared to the baseline model SiamFC, we first added a template branch Frame *t*–1 to generate dynamic dual-template. Then we improved the ResNet-50 to replace the original feature extraction network AlexNet, and we use attention feature fusion module to fuse features extracted by the last three residual blocks. Finally, the features of dual-template and search frame image are compared to generate two score maps which are weighted to predict the object location and scale.

3.2 Dual-Template Strategy

Object tracking algorithm based on Siamese network generally only relies on the first frame of the object image as a template, and the template is not updated as the subsequent tracking process. When tracking for a long time, and the situation on object changes greatly such as scale, speed, illumination, etc., the template is so easy to fail, further resulting in the failure of object tracking.

In order to solve the above problems, a new branch is added to the original two convolutional neural networks of the Siamese network. The new branch also shares weights with the other two branches, and its role is to extract the features of the previous frame of search frame as a template, and to match this template with the search frame, as shown in the Frame *t*–1 branch in Fig. 2.

3.3 Deeper Feature Extraction Network

Because of the advent of the SiamFC algorithm, a large number of algorithms for object tracking based on Siamese network appeared, but in these algorithms shallow convolutional neural networks such as AlexNet are still used for feature extraction. In the article [9] extensive ablation experiments on AlexNet, VGG, ResNet and other networks were conducted, and the results verified that a deeper convolutional neural network model can significantly improve the tracking performance. Therefore, a deeper ResNet-50 is considered as the backbone for feature extraction in this paper. However, simply using the ResNet-50 to replace the original AlexNet will not only fail to improve the tracking performance, but also will lead to a decrease on the accuracy of the tracking algorithm because padding in deep networks will destroy the strict translation in variance [7]. In order to solve the above problem, in the literature [7] the spatial aware sampling strategy is adopted, combined with the ResNet-50 structure to achieve deeper feature information and to improve the tracking performance.

By the above analysis, in this paper the improved ResNet-50 is used as the backbone for feature extraction. The original ResNet-50 has a large stride of 32 pixels, but for the object tracking task, the gap between the object in the front and back frames may be small, and the positioning accuracy of the object will be reduced due to the large network stride. Therefore, the stride of the last two residual blocks of ResNet-50 are reduced, so as to reduce the overall stride to 8 pixels, and dilated convolution [10] to keep the receptive field size unchanged are added. At the same time, in order to reduce the computational complexity, an additional 1×1 convolutional layer is added behind each residual block, which can reduce the number of channels to 256, and the 7×7 center region cropped from the original 15×15 template feature space is considered as template feature. The feature processed in this way can still represent the entire object area [11]. The structure of improved ResNet-50 is shown in Fig. 3.

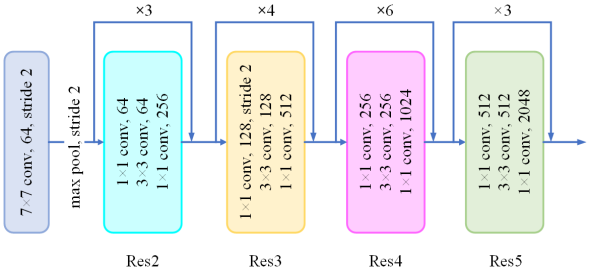


Fig. 3. The structure of improved ResNet-50.

Details	Structure	Bottleneck in Res4			Bottleneck in Res5		
		1×1 Conv	3×3 Conv	1×1 Conv	1×1 Conv	3×3 Conv	1×1 Conv
stride	Original	1	2	1	1	2	1
	Improved	1	1	1	1	1	1
padding	Original	0	1	0	0	1	0
	Improved	0	2	0	0	4	0
dilation	Original	1	1	1	1	1	1
	Improved	1	2	1	1	4	1

Tab. 1. Comparison between the original ResNet-50 and the improved ResNet-50.

The detailed comparison between the original ResNet-50 and the improved ResNet-50 is shown in Tab. 1.

3.4 Attention Feature Fusion Module

During the course of object tracking, rich and accurate description of object features is the primary condition. Therefore, strong feature expression plays a crucial role for the accuracy of tracking algorithms. For a deep network, the extracted features contain more semantic information. In addition, they are less affected by the object illumination, deformation and so on, and have stronger discrimination ability. But deep features may lose a lot of spatial details. On the contrary, shallow features contain more fine-grained spatial information, and they can better perceive the position of the object. Therefore, a way for richer and more robust feature expression can be obtained by combining the deep features and shallow features. Attention mechanism can help to find attention regions in scenes with dense objects, accordingly enhancing the extracted features [12].

In order to better fuse deep features and shallow features, we apply channel attention mechanism and spatial attention mechanism to adaptively fuse the features extracted by the last three residual blocks of the improved ResNet-50. Based on the above improvement of ResNet-50 structure, the feature resolution output from the last three residual blocks are the same. Therefore, the features extracted from the last three residual blocks can be concatenated directly to obtain the preliminarily fused feature map F. Then, the initially fused feature map F is weighted by the channel attention mechanism to get the feature map F', and finally the feature map F' is weighted by the spatial attention mechanism to get the feature map F".

(1) Channel Attention Mechanism

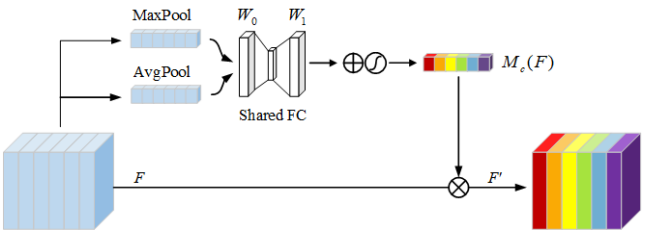


Fig. 4. Channel attention mechanism.

Using channel attention mechanism [13], different weights can be assigned to each channel by judging the importance of each feature channel during the course of object tracking as shown in Fig. 4.

First, global max pooling and global average pooling are performed on the input feature map F to achieve spatial aggregation, and to obtain their respective one-dimensional vectors. After that, two one-dimensional vectors are respectively input into the multilayer perceptron (MLP). Then element-wise addition and sigmoid activation on the features output by MLP are performed to generate the channel attention weight vector. Finally, the original input feature map F is weighted using the channel attention weight vector to obtain the final channel attention feature map F'. The channel attention weight vector $\mathbf{M}_c(\mathbf{F})$ is calculated by

$$\begin{aligned}\mathbf{M}_c(\mathbf{F}) &= \sigma\left(\mathrm{MLP}(\mathrm{Maxpool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{Avgpool}(\mathbf{F}))\right) \\ &= \sigma\left(W_1(W_0(\mathbf{F}^c_{\max})) + W_1(W_0(\mathbf{F}^c_{\mathrm{avg}}))\right)\end{aligned} \tag{2}$$

where $\sigma(\cdot)$ represents sigmoid operation. W_0 and W_1 denote multi-layer perceptron transformation. $\mathbf{F}^c_{\max}$ and $\mathbf{F}^c_{\mathrm{avg}}$ are the vectors after global max pooling and global average pooling respectively.

(2) Spatial Attention Mechanism

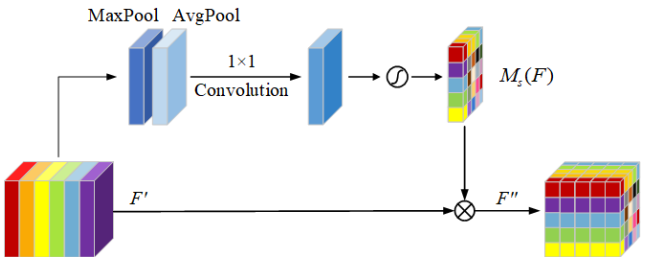
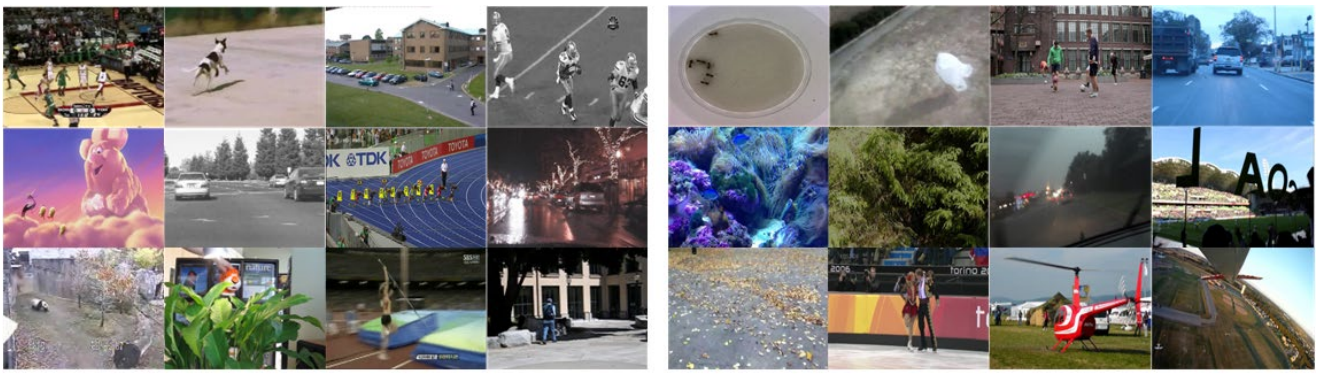


Fig. 5. Spatial attention mechanism.

Similarly, using spatial attention mechanism [14], different weights are assigned to each region by judging the importance of different regions during the course of object tracking as shown in Fig. 5.

Different from the channel attention mechanism, on the spatial attention mechanism max pooling and average pooling of the input feature map F' on the channel first are carried out to obtain two-layer feature map. After that, 1 × 1 convolution is used to reduce the dimensionality of the two-layer feature map into one of a single-layer feature map. Then the single-layer feature map is activated by sigmoid to generate a spatial attention weight matrix. Finally, the original input feature map is weighted using the spatial attention weight matrix to obtain the final spatial attention feature map. The spatial attention weight matrix $\mathbf{M}_s(\mathbf{F})$ is calculated by

$$\begin{aligned}\mathbf{M}_s(\mathbf{F}) &= \sigma\left(\mathrm{Concat}\left(\mathrm{Maxpool}(\mathbf{F}'),\mathrm{Avgpool}(\mathbf{F}')\right)\right) \\ &= \sigma\left(f^{1\times1}\left(\mathbf{F}'^s_{\max},\mathbf{F}'^s_{\mathrm{avg}}\right)\right)\end{aligned} \tag{3}$$



(a) (b)

Fig. 6. Some examples of (a) OTB100 dataset and (b) VOT2018 dataset.

where $f^{1\times1}(\cdot)$ stands for 1 × 1 convolution. $\mathbf{F'}^{s}_{max}$ and $\mathbf{F'}^{s}_{avg}$ represent the feature map after max pooling and average pooling respectively.

4. Experimental Results and Analysis

4.1 Implementation Details and Datasets

The hardware conditions of our experiments are Intel Core i7-10875H 2.30 GHz CPU, NVIDIA GeForce RTX2060 GPU, and 16 GB RAM. The compiling environment is configured with Python 3.7, CUDA 10.1, and Pytorch 1.2.0. End-to-end training was performed using stochastic gradient descent, with an initial learning rate of 10^{-2} which gradually decayed to 10^{-5} during training. The model was trained for 50 iterations, and the batch size was set to 32.

Our approach was trained on the ILSVRC 2015-VID database which contains more than 4000 video sequences with over 1 million frames annotated, and was tested on the OTB100 dataset and VOT2018 dataset. OTB100 dataset contains 100 video sequences involving many complex situations such as illumination change, scale change, deformation, in-plane rotation, out-of-plane rotation, occlusion, disappearance, rapid motion, and motion blur of the object, etc. Compared with OTB100 dataset, VOT2018 dataset has higher resolution and more complex object changes in tracking sequence, and now has become a mainstream standard for evaluating tracking algorithms. Therefore, the performance of object tracking algorithm can be well evaluated on OTB100 dataset and VOT2018 dataset. Some samples of the dataset are shown in Fig. 6.

4.2 Experimental Results on OTB100 Dataset

(1) Objective criteria

In the experiments on OTB100 dataset, the tracking precision and success rate are often used to evaluate the performance of algorithms. The tracking precision is calculated by the percentage of the number of frames on which the error ρ between the center position of the object prediction box (x_p, y_p) and the center position of ground truth (x_{gt}, y_{gt}) is within the threshold to the number of total frames in the video, as shown in (4) and (5).

$$Precision = \frac{frames(\rho < threshold)}{frames(all)}, \tag{4}$$

$$\rho = \sqrt{(x_p - x_{gt})^2 + (y_p - y_{gt})^2}\ . \tag{5}$$

The tracking precision is a real number between 0 and 1. The larger the value is, the better the result is. The value of precision in this paper adopts the value when the threshold is 20.

The tracking success rate is calculated by the percentage of the number of frames on which the overlap rate *IOU* between the object prediction area A_p and the ground truth area A_{gt} is greater than the threshold to the number of total frames in the video, as shown in (6) and (7).

$$SuccessRate = \frac{frames(IOU > threshold)}{frames(all)}, \tag{6}$$

$$IOU = \frac{A_p \bigcap A_{gt}}{A_p \bigcup A_{gt}}. \tag{7}$$

Similarly, the tracking success rate is also a real number between 0 and 1. The larger the value is, the better the result is. The value of success rate in this paper adopts the area under the curve of the success plots.

(2) Ablation Experiments

In order to verify the effectiveness of each improvement of the proposed approach, ablation experiments are performed on the OTB100 dataset, and the results are shown in Tab. 2.

It can be seen from Tab. 2 that the deeper feature extraction network (Improved ResNet-50) contributes 4.4% to the improvement of tracking precision, thus more accurate results are provided for object positioning. But the contribution in the success rate is small, only increased by 0.6%. The reason is that the rich spatial details contained in the shallow features are not fully utilized. Therefore, after adding the attention feature fusion module, the fused features from deep and shallow features contain more com-

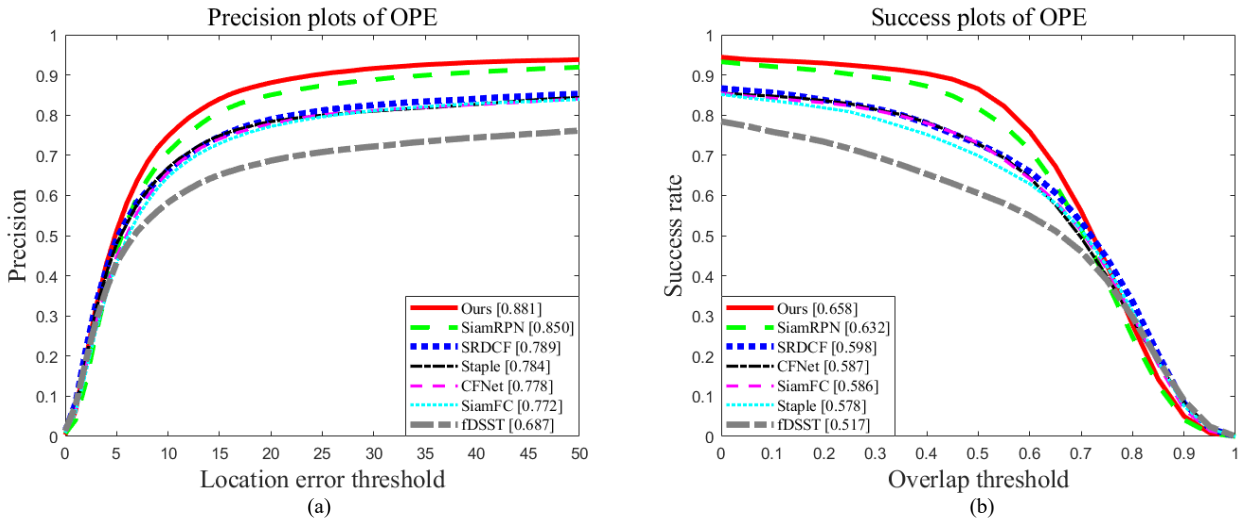


Fig. 7. Comparison results of (a) tracking precision and (b) success rate on OTB100 dataset.

Methods	Improved ResNet-50	Attention feature fusion module	Dual-template strategy	Precision	Success rate
SiamFC	×	×	×	0.772	0.586
Proposed Method	√	×	×	0.816	0.592
	√	√	×	0.843	0.645
	√	√	√	0.881	0.658

Tab. 2. Performance evaluation of improvement ideas of the proposed approach. √ represents the module has been added, × on the contrary.

prehensive information, which significantly contributes 7.1% and 5.9% to the improvement of tracking precision and success rate, respectively. In addition, the added dual-template strategy makes full use of the information of historical frames, alleviates the risk of tracking drift when the object changes greatly, and makes the performance of the model more robust. Thus the tracking precision and the success rate are further improved by 3.8% and 1.3%, respectively.

It is noteworthy that using the improved ResNet-50 as the feature extraction network reduces the amount of parameters to a certain extent. However, due to the addition of attention mechanism and dual-template strategy, the final amount of parameters continues to increase, reaching 24.9 M, and the FLOPs (Floating Point Operations) reaches 19.5 G. The overall computational complexity of the algorithm is increased, but the performance improvement brought by this is worthwhile.

(3) Contrast Experiments

In order to verify the effectiveness of our approach, and to better evaluate the tracking performance, we selected three mainstream tracking algorithms based on Siamese network and three ones based on correlation filtering for comparison. Among them, tracking algorithms based on Siamese network include the baseline SiamFC [5], the improved algorithm CFNet [11] based on SiamFC, and SiamRPN [6] which has performed well in recent years and has gradually become an essential comparative algorithm. Tracking algorithms based on correlation filtering include SRDCF [16], fDSST [19] which used a single histogram of oriented gradient (HOG) feature, and Staple [15] which was combined HOG features with color features. For fair comparison, the experimental results of each algorithm are obtained by optimizing parameters under the same hardware conditions and database settings. Figure 7 shows the comparison results of the overall tracking precision and success rate of our approach and the above six algorithms with one-pass evaluation (OPE) on the OTB100 dataset.

Figure 7(a) shows the comparison results of tracking precision. The abscissa is the threshold of the center position error. It can be seen that our approach has the highest curve slope, which shows that it also has high tracking precision compared with other algorithms at a lower threshold, and can always be optimal.

Figure 7(b) shows the comparison results of tracking success rate. The abscissa is the threshold of the overlap rate between predicted box and the actual object box. It can be seen that the success rate curve of our approach has the slowest decline, which indicates that under the condition of more stringent requirements for tracking objects, our approach can ensure that the success rate decreases less sharply and the tracking performance is stable.

In general, our approach performed the best in both tracking precision and success rate. Compared with the baseline algorithm SiamFC, the tracking precision increased by 10.9%, and the success rate increased by 7.2%. Compared with the well-performing algorithm SiamRPN, the tracking accuracy increased by 3.1%, and the success rate increased by 2.6%. This shows that our approach has excellent tracking performance

For further illustrating the ability of the proposed approach to cope with complex scenes during the tracking process, the experimental results on 11 complex scenes of OTB100 database are given respectively in Tab. 3, Tab. 4

and Tab. 5, besides the above overall tracking precision and success rate.

It can be seen from Tab. 3, Tab. 4 and Tab. 5 that our approach achieved the most advanced results in nine scenes such as fast motion, background clutter, deformation etc. except for low resolution and out of view scenes in which the results still rank second and third respectively, which proves the adaptability of our approach in complex scenes.

In order to further verify the ability of our approach to deal with complex scenes, we visually compare our approach with SiamFC and SiamRPN on OTB100 dataset. Figure 8 shows the visual comparison results of four groups of video sequences with high tracking difficulty. In

Methods	**Out of view**		**Deformation**		**In-plane rotation**		**Out-of-plane rotation**	
	Precision	**Success rate**	**Precision**	**Success rate**	**Precision**	**Success rate**	**Precision**	**Success rate**
fDSST	0.478	0.386	0.541	0.422	0.698	0.505	0.655	0.477
SiamFC	0.673	0.509	0.694	0.516	0.743	0.559	0.758	0.561
CFNet	0.604	0.454	0.710	0.525	0.786	0.568	0.760	0.553
Staple	0.668	0.475	0.747	0.548	0.768	0.548	0.738	0.533
SRDCF	0.597	0.460	0.728	0.540	0.745	0.544	0.742	0.550
SiamRPN	0.728	0.544	0.837	0.626	0.857	0.631	0.854	0.628
Ours	0.721	0.537	0.887	0.652	0.889	0.652	0.878	0.644

Tab. 3. Comparison results in scenes of out of view, deformation, in-plane rotation and out-of-plane rotation.

Methods	**Scale variation**		**Fast motion**		**Motion blur**		**Background clutter**	
	Precision	Success rate	Precision	Success rate	Precision	Success rate	Precision	Success rate
fDSST	0.644	0.473	0.571	0.458	0.568	0.469	0.704	0.523
SiamFC	0.739	0.560	0.744	0.571	0.707	0.554	0.692	0.527
CFNet	0.730	0.546	0.707	0.554	0.681	0.540	0.756	0.561
Staple	0.724	0.519	0.709	0.540	0.700	0.541	0.749	0.560
SRDCF	0.741	0.559	0.769	0.597	0.767	0.594	0.775	0.583
SiamRPN	0.846	0.622	0.793	0.602	0.820	0.625	0.803	0.594
Ours	0.858	0.641	0.82	0.621	0.820	0.625	0.856	0.642

Tab. 4. Comparison results in scenes of scale variation, fast motion, motion blur and background clutter.

Methods	**Illumination variation**		**Low resolution**		**Occlusion**	
	Precision	**Success rate**	**Precision**	**Success rate**	**Precision**	**Success rate**
fDSST	0.716	0.556	0.602	0.395	0.596	0.456
SiamFC	0.741	0.578	0.848	0.598	0.727	0.553
CFNet	0.703	0.540	0.750	0.554	0.697	0.526
Staple	0.778	0.590	0.610	0.400	0.723	0.541
SRDCF	0.786	0.61	0.655	0.494	0.730	0.556
SiamRPN	0.872	0.660	0.870	0.580	0.789	0.593
Ours	0.878	0.663	0.814	0.557	0.818	0.616

Tab. 5. Comparison results in scenes of illumination variation, low resolution and occlusion.

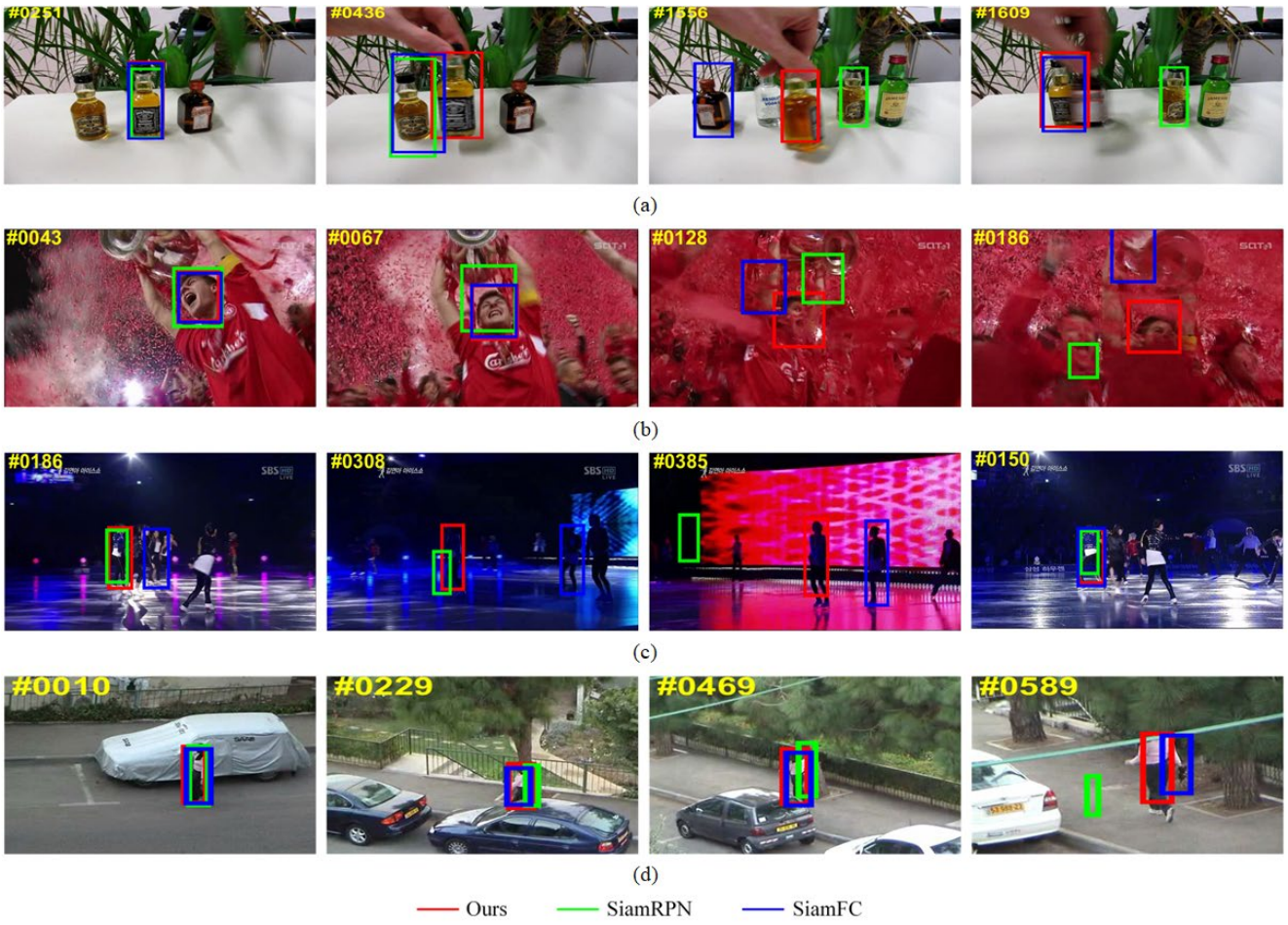


Fig. 8. Visual comparison results of complex scenes on OTB100 Dataset such as (a) liquor, (b) soccer, (c) skating, and (d) woman. The red, green and blue boxes are the tracking results of our approach, SiamRPN and SiamFC, respectively.

the 'liquor' video sequence, objects are disturbed and occluded by similar objects when moving, and the video sequence is long, which is prone to model drift. In the 'soccer' video sequence, the background is mottled and messy. The object and background are almost integrated, which is difficult to distinguish with the naked eye. In the 'skating1' video sequence, the object is not only moving rapidly, but also affected by the illumination change. In the 'woman' video sequence, the scale of the object changes instantaneously.

It can be seen from Fig. 8 that compared with SiamFC and SiamRPN, our approach shows better performance in the face of long-term tracking and complex scenes such as fast movement, mottled background, interference of similar objects, and occlusion.

The excellent results can be obtained thanks to the rich feature expression resulted from the combination of the deep feature extraction network and the attention feature fusion module, as well as the stable template update under the dual-template strategy. The above experimental results verify the ability of our approach to deal with object tracking tasks in complex scenes.

4.3 Experimental Results on VOT2018 Dataset

(1) Objective criteria

The two most important indicators in VOT2018 are accuracy and robustness. Accuracy is used to evaluate the precision of the tracker. The larger the value is, the higher the precision is. The accuracy of the t-th frame of a sequence is defined as

$$\Phi_t = \frac{\mathrm{A}_t^{\mathrm{G}} \bigcap \mathrm{A}_t^{\mathrm{T}}}{\mathrm{A}_t^{\mathrm{G}} \bigcup \mathrm{A}_t^{\mathrm{T}}} \tag{8}$$

where $\mathrm{A}_t^{\mathrm{G}}$ represents the corresponding bounding box of the ground truth in the t-th frame, and $\mathrm{A}_t^{\mathrm{T}}$ represents the bounding box predicted by the tracker in the t-th frame. In more detail, $\Phi_t(i,k)$ is defined as the accuracy of the i-th tracker at the t-th frame in the k-th repetition, and the number of repetitions is set as N_{rep}. So the accuracy on t-th frame is defined as

$$\Phi_t(i) = \frac{1}{N_{\mathrm{rep}}} \sum_{k=1}^{N_{\mathrm{rep}}} \Phi_t(i,k) . \tag{9}$$

N_{valid} represents the number of valid frames, the final accuracy of the i-th tracker is defined as

$$Accuracy = \frac{1}{N_{\text{valid}}} \sum_{t=1}^{N_{\text{valid}}} \Phi_t(i). \tag{10}$$

Robustness is used to evaluate the stability of the tracker. The smaller the value is, the better the stability is. Following the definition of accuracy above, if $F(i,k)$ is defined as the number of times that the i-th tracker failed to track in the k-th repetition, then the final robustness of the i-th tracker is defined as

$$Robustness = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} F(i,k). \tag{11}$$

In fact, the primary index for evaluating tracking performance in the competition is expected accuracy overlap (EAO) [18] which reflects the overall performance of accuracy and robustness to a certain extent. The larger the value is, the better the tracking performance is.

(2) Contrast Experiments

In addition to the selected algorithms on the OTB100 dataset (SRDCF [16], Staple [15], SiamFC [5], SiamRPN [6]), we also selected the state-of-the-art algorithms in VOT2018 competition which ranked in the front, namely ECO [19], SiamRPN++ [7] and SiamFC++[20] to compare with our approach, and the results are shown in Tab. 6 and Tab. 7.

EAO of our method ranks first which is 25.9% higher than the baseline algorithm SiamFC, and 2.1% higher than that of the state-of-the-art algorithm SiamFC++. The accuracy of our approach is slightly lower than SiamRPN++, SiamFC++ and SiamRPN. The reason for this phenomenon may be that the VOT2018 dataset contains multiple small objects, and the proposed approach is less effective in dealing with the boundary regression of small objects. However, with the support of deep feature extraction network and attention feature fusion module, our approach still improves by 5.1% in accuracy compared with baseline algorithm SiamFC. Meanwhile, under the tracking stability provided by the dual-template strategy, the robustness of our approach is 44.3% higher than the baseline algorithm and 4.1% higher than that of the state-of-the-art SiamFC++, achieving the optimal result.

Methods	EAO	Accuracy	Robustness	Speed (fps)
SRDCF	0.119	0.490	0.974	13
Staple	0.169	0.530	0.688	80
SiamFC	0.188	0.503	0.585	86
ECO	0.280	0.484	0.276	20
SiamRPN	0.383	0.586	0.276	160
SiamRPN++	0.414	0.600	0.234	35
SiamFC++	0.426	0.587	0.183	90
Ours	0.447	0.554	0.142	22

Tab. 6. Comparison results of tracking performance on VOT2018 dataset.

Methods	Parameters (M)	FLOPs (G)
SiamFC	2.3	2.7
SiamRPN++	11.2	7.1
SiamFC++	13.1	17.2
Ours	24.9	19.5

Tab. 7. Comparison results of computational complexity on VOT2018 dataset.

It is noteworthy that the tracking speed of the proposed method has decreased significantly. This is largely derived from the increase of computational complexity of the algorithm. As can be seen from Tab. 7, although the FLOPs of our approach is higher than that of SiamFC++ only by 2.3, it is undeniable that the tracking performance brought by the proposed method is the most promising.

In general, the above experimental results further show that the proposed algorithm has excellent tracking performance.

5. Conclusion

In this paper, dual-template Siamese network with attention feature fusion for object tracking is proposed. By using a deeper ResNet-50 and improved network structure, the feature extraction ability of the tracker is improved. Then, the last three residual block features of ResNet-50 are fused under the attention mechanism using the attention feature fusion module. Therefore, effective shallow spatial information and deep semantic information of object can be obtained, and the location ability of the tracker is improved. At the same time, in order to resolve the problem of object template degradation, and adapt to the changes of the appearance and state of the object, a dual-template strategy is introduced to prevent tracking drift by establishing a new template branch. Compared with the state-of-the-art methods on the OTB100 dataset and VOT2018 dataset, our approach obtains the most excellent tracking performance, which verifies that the proposed method can effectively deal with complex scenes such as rapid object movement, mottled background, interference of similar objects, and occlusion etc.

Future research will focus on the template update strategy of object tracking algorithm to further improve tracking performance. In addition, the network structure will be optimized to reduce the computational complexity of the algorithm. We will consider making the code and data public after the confidentiality requirement of the project is relieved.

Acknowledgments

This work is supported by Joint Project of Beijing Natural Science Foundation and Beijing Municipal Education Commission (Grant No. KZ202110011015).

References

[1] LI, X., ZHA, Y. F., ZHANG, T. Z., et al. Survey of visual object tracking algorithms based on deep learning (in Chinese). *Journal of Image and Graphics*, 2019, vol. 24, no. 12, p. 2057–2080.

[2] MENG, L., YANG, X. A survey of object tracking algorithms (in Chinese). *Acta Automatica Sinica*, 2019, vol. 45, no. 7, p. 1244 to 1260.

[3] CHEN, Y. F., WU, Y., ZHANG, W. Survey of target tracking algorithm based on Siamese network structure (in Chinese). *Computer Engineering and Applications*, 2020, vol. 56, no. 6, p. 10–18. DOI: 10.3778/j.issn.1002-8331.1911-0127

[4] TAO, R., GAVVES, E., SMEULDERS, A. W. M. Siamese instance search for tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas (NV, USA), 2016, p. 1420–1429. DOI: 10.1109/CVPR.2016.158

[5] BERTINETTO, L., VALMADRE, J., HENRIQUES, J. F., et al. Fully-convolutional Siamese networks for object tracking. In *Proceedings of European Conference on Computer Vision*. Amsterdam (Netherlands), 2016, p. 850–865. DOI: 10.1007/978-3-319-48881-3_56

[6] LI, B., YAN, J. J., WU, W., et al. High performance visual tracking with Siamese region proposal network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City (UT, USA), 2018, p. 8971–8980. DOI: 10.1109/CVPR.2018.00935

[7] LI, B., WU, W., WANG, Q., et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach (CA, USA), 2019, p. 4277–4286. DOI: 10.1109/CVPR.2019.00441

[8] WANG, Q., ZHANG, L., BERTINETTO, L., et al. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach (CA, USA), 2019, p. 1328–1338. DOI: 10.1109/CVPR.2019.00142

[9] ZHANG, Z. P., PENG, H. W. Deeper and wider Siamese networks for real-time visual tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach (CA, USA), 2019, p. 4586–4595. DOI: 10.1109/CVPR.2019.00472

[10] LONG, J., SHELHAMER, E., DARRELL, T. Fully convolutional networks for semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Boston (MA, USA), 2015, p. 3431–3440. DOI: 10.1109/CVPR.2015.7298965

[11] VALMADRE, J., BERTINETTO, L., HENRIQUES, J., et al. End-to-end representation learning for Correlation Filter based tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu (HI, USA), 2017, p. 5000–5008. DOI: 10.1109/CVPR.2017.531

[12] YANG, Y., GAO, X., WANG. Y., et al. VAMYOLOX: An accurate and efficient object detection algorithm based on visual attention mechanism for UAV optical sensors. *IEEE Sensors Journal*, 2023, vol. 23, no. 11, p. 11139–11155. DOI: 10.1109/JSEN.2022.3219199

[13] HU, J., SHEN, L., SUN, G., et al. Squeeze-and-excitation network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City (UT, USA), 2018, p. 7132 to 7141. DOI: 10.1109/CVPR.2018.00745

[14] WOO, S. H., PARK, J., LEE, J. Y., et al. CBAM: Convolutional Block Attention Module. In *Proceedings of European Conference on Computer Vision*. Munich (Germany), 2018, p. 3–19. DOI: 10.1007/978-3-030-01234-2_1

[15] BERTINETTO, L., VALMADRE, J., GOLODETZ, S., et al. Staple: Complementary learners for real-time tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas (NV, USA), 2016, p. 1401–1409. DOI: 10.1109/CVPR.2016.156

[16] DANELLJAN, M., HAGER, G., KHAN, F. S., et al. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of IEEE International Conference on Computer Vision*. Santiago (Chile), 2015, p. 4310–4318. DOI: 10.1109/ICCV.2015.490

[17] DANELLJAN, M., HAGER, G., KHAN, F. S., et al. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39, no. 8, p. 1561 to 1575. DOI: 10.1109/TPAMI.2016.2609928

[18] KRISTAN, M., PFLUGFELDER, R., LEONARDIS, A., et al. The Visual Object Tracking VOT2013 challenge results. In *Proceedings of IEEE International Conference on Computer Vision.* Sydney (Australia), 2013, p. 98–111. DOI: 10.1109/ICCVW.2013.20

[19] DANELLJAN, M., BHAT, G., KHAN, F. S., et al. ECO: Efficient Convolution Operators for tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.* Honolulu (HI, USA), 2017, p. 6931–6939. DOI: 10.1109/CVPR.2017.733

[20] XU, Y., WANG, Z., LI, Z., et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of AAAI Conference on Artificial Intelligence*. New York (USA), 2020, p. 12549–12556. DOI: 10.1609/aaai.v34i07.6944

About the Authors ...

Minhua LIU was born in 1976. He received his Ph.D. degree from Tsinghua University. Now he is the vice president of Beijing Technology and Business University. His research interests include image processing and the modeling and analysis of complex system.

Jiantong SHI was born in 1996. He received his B.S. degree from Qingdao University of Science and Technology in 2019. He is now a candidate of master degree in the School of Artificial Intelligence, Beijing Technology and Business University, China. His research interests include pattern recognition, image processing and computer vision.

Yu WANG (corresponding author) was born in 1977. She received her Ph.D. degree from the University of Science and Technology Beijing in 2009. She was engaged in scientific research as a post-doctoral in the Beijing Key Laboratory of Multidimensional and Multiscale Computing Photography, Tsinghua University from 2009 to 2011. She is now a Professor and doctoral supervisor of the Beijing Technology and Business University. Her research interests include pattern recognition, image processing and computer vision.