IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2020, accepted June 23, 2020, date of publication June 25, 2020, date of current version July 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004992

SiamFF: Visual Tracking With a Siamese Network Combining Information Fusion With Rectangular Window Filtering

YUAN LUO[1], YUANXIAO CAI[1], BOYU WANG[1], JIE WANG[1], AND YANJIE WANG[2]
[1]Key Laboratory of Optoelectronic Information Sensing and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]College of Data Science, Zhejiang University of Finance and Economics, Hangzhou 310000, China

Corresponding author: Yuanxiao Cai (421860130@qq.com)

This work was supported by the National Nature Science Foundation of China under Grant 61801061.

ABSTRACT Recently, Siamese trackers have shown excellent performance in both accuracy and speed. However, traditional trackers have poor robustness against similar objects due to the use of single deep features and the limitation of cosine windows. In this paper, a novel Siamese network combining information fusion with rectangular window filtering named SiamFF is introduced. First, a multilevel fusion network is proposed. At feature-level, the shallow and deep features of the network are fused through a layer-hopping connection to obtain complementary feature maps. Then, the score maps generated by the complementary feature maps are further fused at the score-level to improve the robustness. In addition, based on the continuity and stationarity of objects movement in reality, a score map filtering strategy is proposed. The relative displacement of the target can be predicted by obtaining the interframe information, and the moving direction is applied to filter the score map to further eliminate the analog interference. Experimental results on OTB2015 and VOT2016 benchmarks indicate that SiamFF performs favorably against many state-of-the-art trackers in terms of accuracy while maintaining real-time tracking speed.

INDEX TERMS Deep learning, visual tracking, Siamese network, multilevel fusion, rectangular window filtering.

I. INTRODUCTION

Target tracking is one of the topical issues in the field of computer vision. After the first frame of the video is initialized, the target is surrounded by a bounding-box generated by the tracker in subsequent frames [1]. Overcoming deformation, occlusion and movement of the target during the tracking process makes visual tracking challenging [2]–[4]. Correlation filters have demonstrated excellent tracking performance, they utilize the characteristics of Fourier transform and cyclic matrices to train the networks, and update the parameters while tracking [5]. Recently, the role of convolutional neural networks (CNN) in image classification has been verified [6]. CNN can be applied to extract deep features to improve tracking accuracy, but online updating greatly reduces the speed of trackers as networks become deeper. Under the CNN framework, Siamese trackers have demonstrated their excellent performance in terms of accuracy and speed for training the network end-to-end without online updating [7]. However, traditional Siamese trackers extract semantic features from only the last layer of the network for similarity matching and ignore the shallow features. Simultaneously, the trackers use cosine windows to suppress the interference points in score maps and have poor robustness against analogs with large influence.

To solve the above problems, a novel Siamese network named SiamFF is proposed in this paper, and the contribution can be divided into two parts:

1). The shallow features of CNN have better robustness to similar interference, and can be fused with deep features to improve tracking performance. We introduced a multilevel fusion network, first, the feature-level fusion is performed where the shallow and deep features are fused to obtain complementary feature maps. Then, the score-level fusion is carried out where the complementary feature maps of two branches are correlated to generate a pair of similarity score

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

maps that are further fused to obtain the final score map. According to experience, fusing layer-by-layer not only generates redundant information but also creates computational complexity [8]; thus, we applied a layer-hopping connection to avoid this issue.

2). Based on continuity and stationarity of object movement between two adjacent frames, there is a mapping relationship between the target's actual motion and the peak point of the score map; therefore, we proposed a score map filtering strategy. By obtaining the motion information between two frames, the displacement direction of the target is predicted, and the score map is filtered along this direction to further eliminate the influence of analogs.

Extensive experimental results show that SiamFF achieves state-of-the-art performance in recent benchmarks. The remaining content of the paper is arranged as follows:

II) Related Work.
III) Our Approach.
IV) Experiments on Benchmarks.
V) Conclusion.

II. RELATED WORK

A. VISUAL TRACKING

Visual tracking can be divided into generative and discriminative models according to the participation of the detection process. The generative model estimates the optimal position of the target by a certain tracking strategy after modeling, and the representative methods include sparse representation and the probability model. The discriminative models regard tracking as a binary classification problem for seeking the decision boundary between the target and the background by incremental learning. Currently, the discriminative model represented by correlation filters and Siamese networks has become the mainstream in visual tracking.

Kernelized correlation filters (KCF) [5] lead the research on correlation filters. Minimum output sum of squared error filter (MOSSE) [9] uses an adaptive training strategy, and pushes tracking speed to a high level. RDCF [10] introduces a penalty factor for filter coefficients to resolve the boundary effect caused by an inaccurate representation of image contents. Multitask correlation particle filter (MCPF) [11] solves the problem of large-scale changes in the target by jointly learning different features.

B. SIAMESE NETWORK BASED TRACKING

The two branches of Siamese networks share weight parameters; their similarity is output after sending two inputs. Siamese networks convert target tracking into a similarity learning problem, which matches the essence of visual tracking well; that is, it finds the similarity between the template and search images. GOTURN [12] uses the Siamese network to extract features and trains a CNN to predict the position of b-boxes relative to the target represented by the previous frame. SiamFC [13], shown in Figure 1, introduces AlexNet into the Siamese network to compare the similarity between

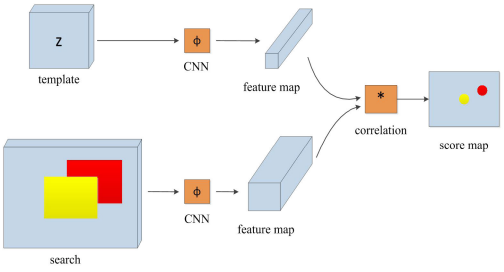


FIGURE 1. The structure of SiamFC.

two frames and predicts the target's position by similarity scores. SiamRPN [14] introduces a region generation network to adapt to the scale change in the target. CA-Siam [15] follows a classification network behind SiamFC to improve tracking performance.

III. OUR APPROACH

A. MULTILEVEL FUSION NETWORK

Compared with traditional trackers, we studied the characteristics of other shallower features while using deep features. As shown in Figure 2, we visualized the feature maps of input samples in each CNN layer. It can be observed that as the number of layers increases and the network deepens, the feature maps not only show a size change, but the resolution gradually decreases. The fifth layer cannot recognize the target appearance, and the shallow features can be easily achieved. This confirms that deep features contain semantic information with low resolution, and they are more robust the target deformation and suitable for classification. However, shallow features with high resolution can capture fine spatial details better, and obtain more background information, and they are suitable for positioning by virtue of the robustness to interference from similar objects. These two types can be fused to complement each other and improve performance. We introduce a modified AlexNet which removes padding layers and modifies the number of network channels. Then, we carry out a multilevel fusion strategy to fuse the shallow and deep features, thereby making full use of the target's spatial and semantic information.

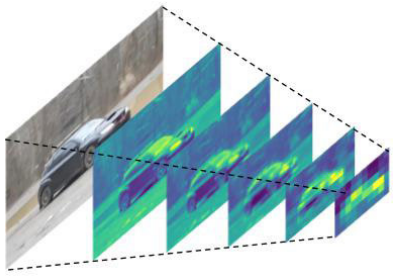


FIGURE 2. The feature maps of CNN shift from delicate appearance features to abstract semantic features, while the scale gradually decreases.

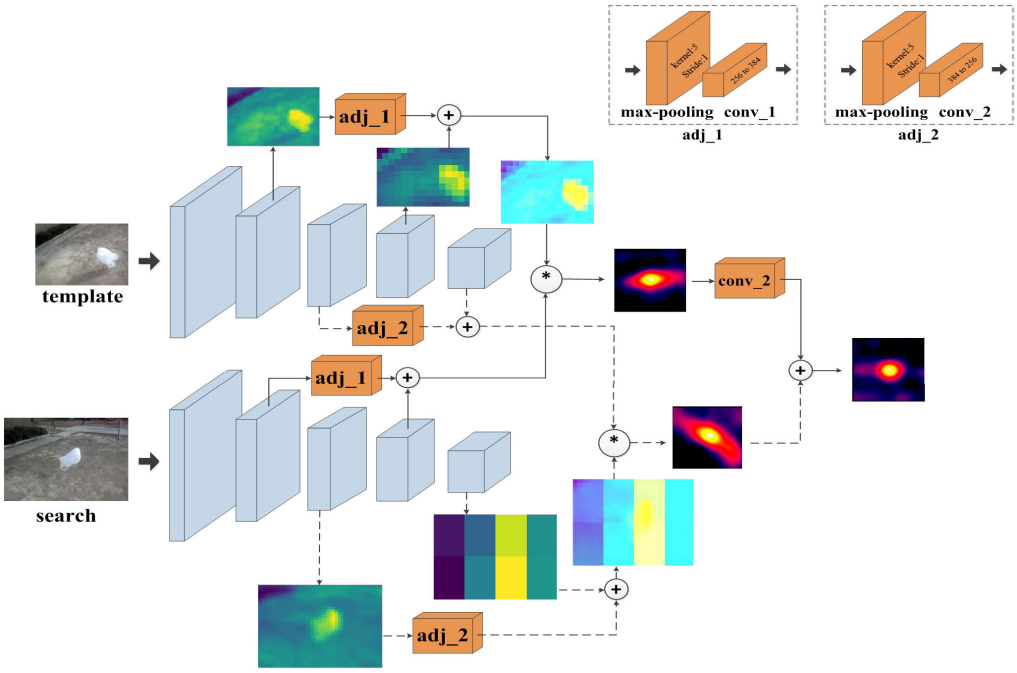


FIGURE 3. The template and search images enter the two branches of the network. The feature maps show the feature-level fusion results, and the score maps show the score-level fusion results. The specific structure of the adjustment modules is shown in the upper-right corner.

A multilevel fusion network is shown in Figure 3. First, feature-level fusion is executed. Considering the layer-by-layer change in the map information indicated in Figure 2 and to avoid information redundancy and computing complexity, we apply a layer-hopping connection in which conv2 is paired with conv4, and conv3 is paired with conv5 to achieve a better complementary effect. The feature maps need to be uniform before fusing due to the pooling layers of the network. Adjustment modules are introduced in this paper to change the maps' sizes and channels; we used adj_1 which consists of max-pooling and a 1×1 convolution to fuse conv2 and conv4 to obtain feature map24. The conv2 feature map downsamples by max-pooling to reduce the image size, then changes the number of channels by a 1×1 convolution. The adjustment modules can not only unify images but also retain original spatial information. Similarly, the adjustment module adj_2 is carried out to fuse conv3 and conv5 to generate feature map35. The above operations are applied in both template and search branches of the network, and the corresponding feature maps are correlated to obtain score map24 and score map35. Then, the second level fusion, which is score-level, is performed to obtain the final score map for target prediction. TABLE 1 shows the algorithmic summary of the framework.

The search size and template images are set to 255×255 and 127×127 respectively, then the score map24 and score map35 are output as $17 \times 17 \times 384$ and $17 \times 17 \times 256$. It can be noticed that the channels of the two score maps are different; we changed the channels of score map24 through conv_2 and obtained the final score map with size $17 \times 17 \times 256$. The multilevel fusion network filters out most analogs with a small impact or long distance, and the peak points are more convergent in score maps without scattered or subtle interference, which improves the accuracy of the tracker. TABLE 2 shows the detailed network structure and parameters.

B. SCORE MAP FILTERING STRATEGY

Information fusion can improve tracking robustness favorably in the case of simple backgrounds, few similarities or low interference. However, we visualized the score maps shown in Figure 4 in training, and indicated that the improvement achieved by simply performing information fusion is limited. It is hard to completely filter out interference for objects that are highly similar to the target. Traditional Siamese trackers utilize a cosine window to filter out similarities far away from the center, but cannot deal with close-distance interference. Simultaneously, the trackers apply the entire area of the score map and directly select the maximum point as the target position. This leads to poor robustness against similarity interference in complex environments, which easily causes tracking drift. Therefore, we propose a score map filtering strategy to help trackers search and locate the target accurately in the final prediction phase.

The motion of objects tends to be continuous and stationary in reality; that is the relative displacement of the target between two adjacent frames in image sequences is small, and there is no situation in which the instantaneous movement is large. Siamese trackers crop the target position of the previous frame as the center to generate the search image for the current frame. Fully convolutional networks eliminate the

TABLE 1. Algorithmic summary of the framework.

Algorithm 1. Multilevel Fusion Network

Input: The template image z, the samples $X=\{x_1, x_2, x_3, ..., x_n\}$ (n indicates the number of video frames).

1. Extract the feature maps conv2-5 of z from the network.
2. Fuse the conv2 and conv4 of z to generate feature map24_z by the adjustment module (adj_1).
3. Fuse the conv3 and conv5 of z to generate feature map35_z by the adjustment module (adj_2).
4. while ($i < n$) {
5. Extract the feature maps conv2-5 of x_i from the network.
6. Fuse the conv2 and conv4 of x_i to generate feature map24_x_i by the adjustment module (adj_1).
7. Fuse the conv3 and conv5 of x_i to generate feature map35_x_i by the adjustment module (adj_2).
8. The feature map24 and feature map35 of the two branches are correlated to obtain score map24 and score map35.
9. Fuse score map24 and score map35 to obtain the final map.
10. Utilize the score map filtering strategy to obtain the tracking result.
11. }

Output: Tracking results with bounding boxes.

TABLE 2. Network structure and parameters.

Layer		kernel	stride	channel	z-size	x-size	channel.image
-		-	-	-	127	255	3
conv1		11	2	96	59	123	96
pool1		3	2	-	29	61	96
conv2		5	1	256	25	57	256
pool2		3	2	-	12	28	256
conv3		3	1	384	10	26	384
conv4		3	1	384	8	24	384
conv5		3	1	256	6	22	256
adj_1	mp	5	1	-	-	-	-
	conv_1	1	-	384	-	-	-
adj_2	mp	5	1	-	-	-	-
	conv_2	1	-	256	-	-	-

need for image pairs to have the same size, and the score map is obtained by a sliding window convolution on a dense grid. Therefore, the video images and the score maps of the network have a mapping relationship. As shown in Figure 5, the relative displacement of the target between two adjacent frames is not only shown on video images but also mapped to the peak points of the score maps. Trackers can select the peak point to obtain the current position of the target through such a mapping relationship.

Based on the above theory, we obtain the motion information of the target between two frames, and predict its relative displacement in the next frame. Then, the displacement direction is utilized to filter the score map, and the target positioning is guided. In experiments, a number axis coordinate system xOy with the same size was introduced to cover the final score map to digitize the position of each point. For score map t-1 of frame $t-1$, the relative displacements dx, dy of the target from the center was measured, and the moving direction D_{t-1} between two frames was obtained

$$D_{t-1} = dy/dx, \tag{1}$$

The image of frame t is sent to the network to generate score map t, and we introduce a rectangular function $rect(n)$ to detect the peak point V_i ($1 < i < m$, m represents the number

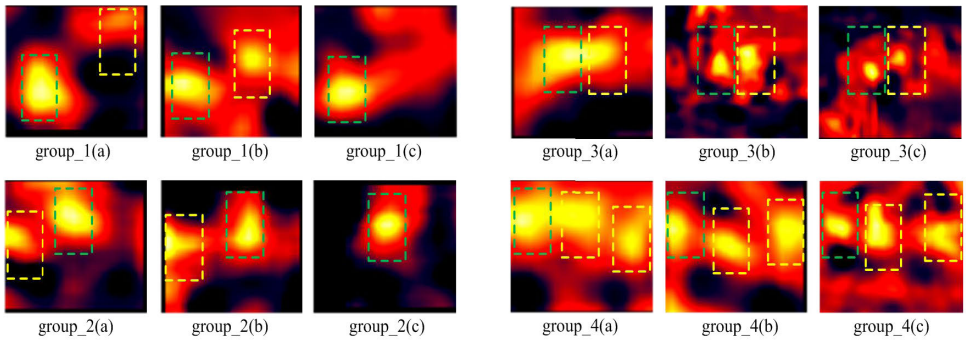


FIGURE 4. The green b-boxes indicate the target, and the yellow indicate similarities. (a) and (b) represent a pair of score maps participating in fusion, (c) represents the fusion result. Group_1 and group_2 show that the multilevel fusion network can obtain obvious improvement when there are few analogs with a small effect, group_3 and group_4 show that the network cannot effectively eliminate interference in complex environments.

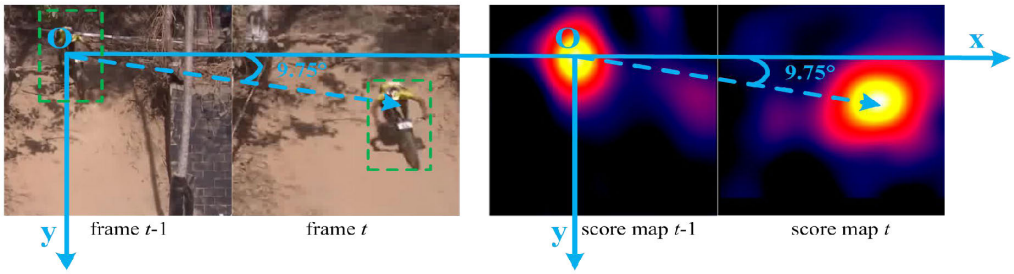


FIGURE 5. The left shows the displacement of the target in video images, and the right indicates the peak points in the corresponding score maps. The motion information of the target in two spaces has a mapping relationship.

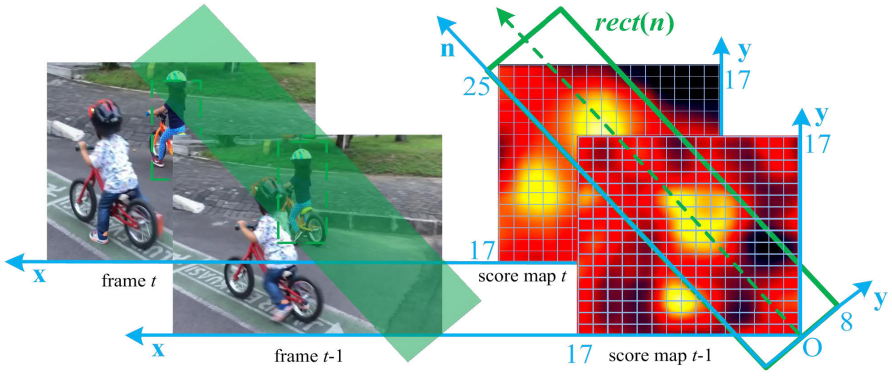


FIGURE 6. Score map t is filtered using the motion information of frame $t-1$. The green dotted line indicates the filtering direction, *rect*(*n*) is defined in a larger coordinate system in the same space as the score map. The left figure shows the corresponding detection range of the tracker in the video.

of peak points) of score map t. As shown in Figure 6, $rect(n)$ obtains the filtered coordinate range according to D_{t-1}, and the peak point V_i not in the range is filtered out. Through the strategy, the search area is limited to a smaller range instead of the entire score map.

$$rect(n) = \begin{cases} 8, & n \leq 25 \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

$$v_i(x_i,\ y_i) = \begin{cases} v_i, & (x_i,\ y_i) \in rect(n) \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

As mentioned earlier, the improvement from information fusion and the cosine window is limited. A score map filtering strategy was utilized to eliminate the influence of high-score analogs, and further improve the tracker's accuracy. We applied sampling statistics, and the results indicated the average size of the target points was 4. To adapt the size of both target points and score maps, the width and length of $rect(n)$ were set to 8 and 25 respectively; a length greater than 24 is acceptable. $rect(n)$ rotates with the fixed size along D_{t-1}, and covers the filtering area in the score map during tracking.

Finally, the situation of the contradiction between the multilevel network and score map filtering strategy was considered. Due to the continuity and stationarity of object motion, the displacement of the target between frames was approximately a continuous curve, which stabilized the score map filtering. Based on the above, the tracker selected the maximum point in the range of $rect(n)$, and experimental results demonstrated the effectiveness.

C. TRAINING DETAILS

We used ILSVRC and GOT-10k to train the network. After template-search image pairs were used to extract feature maps through the CNN, a correlation operation was carried out to generate the score map. The formula can be expressed as

$$S(z, \ x) = f(\phi(z), \ \phi(x)), \tag{4}$$

$\phi(\cdot)$ is the feature representation of an image, $f(\cdot)$ represents the correlation operation, $S(z, x)$ represents the similarity of image pairs, and the goal of the network is to obtain the maximum value of eq.4. The network was trained using logic loss

$$L(y, \ v) = \frac{1}{|D|} \sum_{u \in D} \log(1 + \exp(-yv)), \tag{5}$$

u represents a pixel point in the score map, $v[u]$ represents the similarity score of the point, and $y[u]$ is its ground-truth label. We adopted stochastic gradient descent (SGD) to optimize the loss function to obtain the weight parameters θ. $y[u]$ is defined according to the distance from the target center in the score map (k represents the stride of the network, c represents the target center)

$$y[u] = \begin{cases} +1, & k \|u - c\| \leq R \\ -1, & \text{otherwise}, \end{cases} \tag{6}$$

Image pairs are cropped centered on the target during training. Template and search images were cropped to 127×127 and 255×255 respectively. The range beyond cropping was filled with the mean RGB value of the images.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

The hardware for the experiments in this paper was an Intel Xeon E5 CPU and NVIDIA 2080ti GPU, the system environment was Ubuntu 16.04LTS, and the experiment tool was MATLAB 2018b. Experiments were performed on OTB2015 and VOT2016. Hyper-parameters of the network were set as follows: learning rate = 0.01, batch size = 16, and epoch number = 80.

B. EXPERIMENTS ON OTB2015

OTB2015 uses precision and succession as evaluation indicators and adopts OPE for robust evaluation.

1) PRECISION

Locate the center point of the b-box and calculate the distance between it and the ground-truth, then count the percentage of video frames whose distance is less than a given threshold. A curve can be obtained with different thresholds, and better trackers achieve higher curve values.

2) SUCCESSION

The overlap score (OS) is defined as

$$OS = \frac{|a \cap b|}{|a \cup b|}, \tag{7}$$

"a" represents the b-box generated by the tracker, "b" represents the ground-truth, and $|\cdot|$ represents the number of pixels in an area. The frame whose OS is greater than a set threshold is considered successful, and the percentage of total successful frames in video is succession.

3) ONE PASS EVALUATION (OPE)

OPE indicates that only the first frame of the video is be initialized with the ground-truth, and then running the algorithm to obtain the results.

OTB2015 is an extension of OTB2013 [2]; it includes one hundred videos for testing that cover eleven different scenes, and each video contains a ground-truth. We executed experiments with SiamFF and other state-of-the-art trackers, including KCF [5], DSST [16], SAMF [17], SiamFC [13], and SiamRPN [14]. Among them, KCF [5], DSST [16] and SAMF [17] are trackers based on correlation filters, SiamFC [13] and SiamRPN [14] are based on a Siamese network. Figure 7 shows the experimental results of SiamFF and others. The left figure is precision plot and the right is succession plot. It can be observed that SiamFF outperforms others on both indicators. TABLE 3 indicates the performance differences in detail.

It is clear that SiamFF not only greatly improved over the correlation filtering trackers but also had better performance compared with the same type trackers. The promotion of two indexes were 5.60% and 3.64% over SiamRPN (2nd) [14], 9.52% and 7.91% over SiamFC (3rd) [13]. Due to the multilevel fusion network and the score map filtering strategy, SiamFF had a slower speed but also achieved 53FPS, which ensures real-time tracking. To further test the robustness of the trackers, we implemented experiments in eleven different scenes (including "fast motion", "background clutter", "motion blur", "deformation", "illumination variation", "in-plane-rotation", "out-of-plane rotation", "low resolution", "occlusion", "out-of-view", "scale variation"). The precision plots are shown in Figure 8.

The score map filtering strategy applied by SiamFF can effectively improve the accuracy of target positioning, and greatly reduce the center error with the ground-truth after mapping to video. We can observe from the plots that SiamFF ranks first in nine scenes. It improves most in "fast motion" and "motion blur" with increases of 13.14%

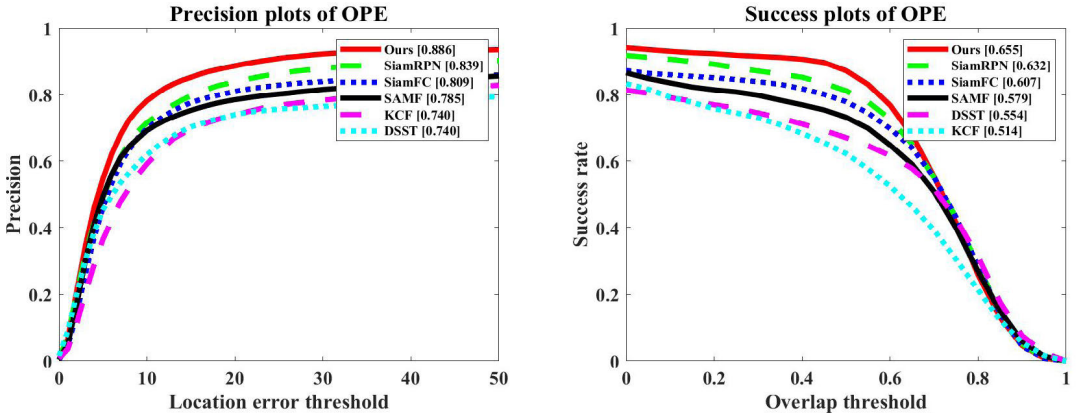


FIGURE 7. The precision and succession plots on OTB2015 with one pass evaluation(OPE).

TABLE 3. Precision and succession scores of trackers, "Improve" indicates our tracker's improvement over the others, and "Speed" indicates the tracking speed.

Tracker	Precision score	Succession score	Improve (pre./succ.) %	Speed (FPS)
SiamRPN	0.839	0.632	5.60/3.64	200
SiamFC	0.809	0.607	9.52/7.91	86
SAMF	0.785	0.579	12.87/13.13	7
KCF	0.740	0.514	19.73/27.43	172
DSST	0.740	0.554	19.73/18.23	24
Ours	0.886	0.655	-	53

TABLE 4. Trackers' scores of five indicators in VOT2016, the best of each indicator is marked in red.

	SiamFC	SiamRPN	SiamAN	ACT	ColorKCF	DSST	KCF	SAMF	TCNN	Ours
EAO	0.2797	0.3411	0.2358	0.1721	0.2262	0.1806	0.1935	0.1851	0.3251	0.3896
Acc.	0.4061	0.4693	0.4003	0.2803	0.3478	0.3255	0.3032	0.3496	0.4872	0.5121
Fail.	19.3910	20.1382	29.8021	42.6031	25.7661	44.8138	38.0820	37.7937	17.9393	15.7913
Overlap	0.5332	0.5804	0.5312	0.4349	0.4944	0.5246	0.4916	0.4965	0.5470	0.5865
FPS	103	200	12	82	111	13	22	5	1	53

and 11.22% over the second-place, respectively. In addition, SiamFF ranks second in "occlusion" and "out-of-view" with a difference of 0.009 from SAMF [17] and 0.013 from SiamFC [13]. In trackers' succession plots shown in Figure 9; SiamFF ranks first in nine scenes except for "low resolution" or "out-of-view". The ascensions are greatest in "fast motion" and "motion blur", which are 10.97% and 8.21% higher than the second-place, respectively. Figure 10 shows the tracking record of six trackers on OTB2015.

C. EXPERIMENTS ON VOT2016

VOT2016 uses expected average overlap (EAO), robustness (failure numbers), overlap, and FPS as evaluation indicators.

1) OVERLAP

Overlap is defined similarly to *OS*, and larger overlap values indicate better tracking performance.

2) ROBUSTNESS

Robustness adopts failure numbers to quantify. The tracking of frame t is considered failed if the overlap is less than a given threshold ($\text{overlap}_t < th$), and the total failed frames are counted. Fewer failed frames indicate better tracking performance.

3) EXPECTED AVERAGE OVERLAP (EAO)

EAO was applied to calculate the overlap and robustness uniformly and obtain the comprehensive performance of the tracker.

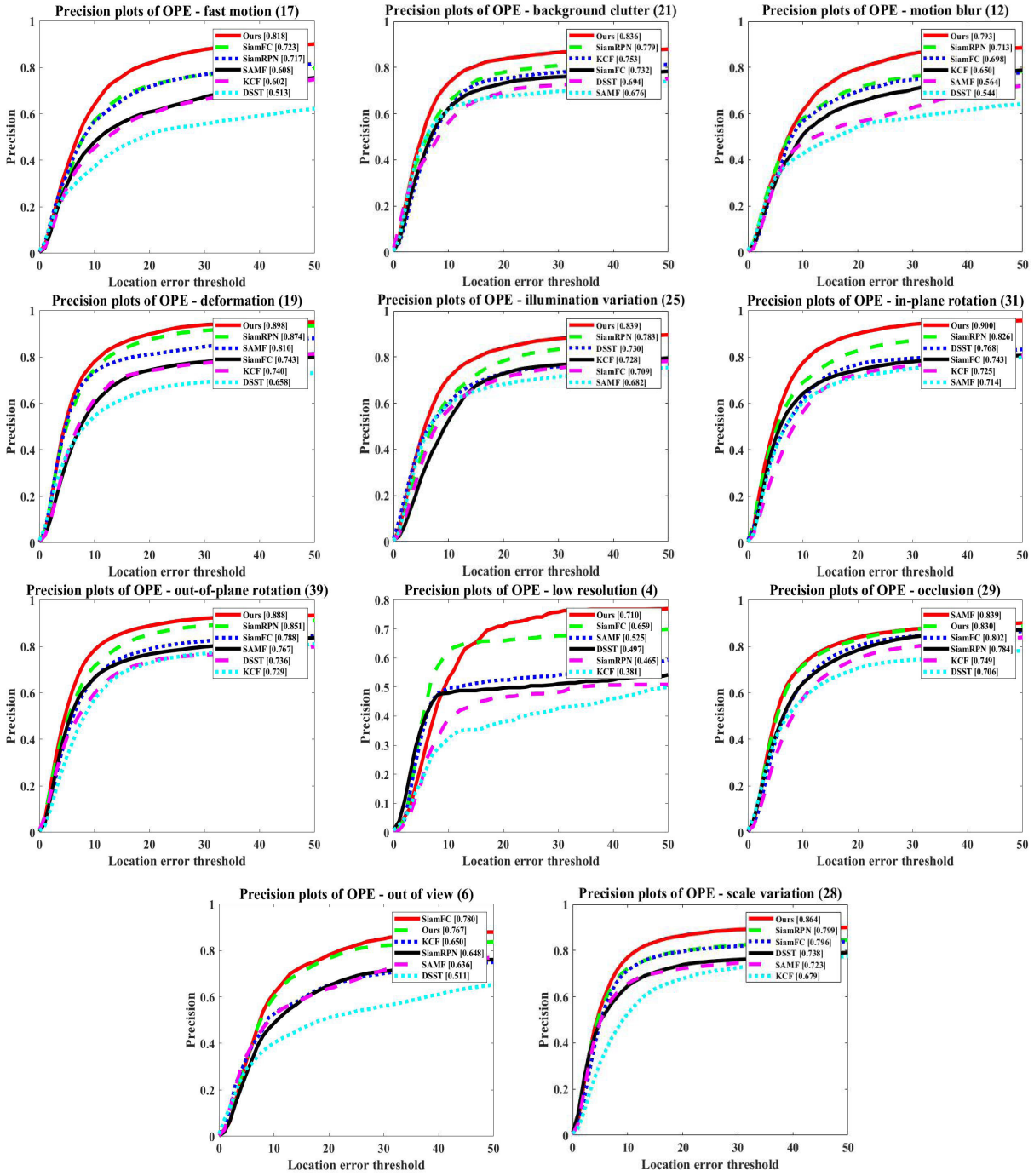


FIGURE 8. Precision plots on OTB2015 over eleven tracking scenes of fast motion, background clutter, motion blur, deformation, illumination variation, in-plane rotation, out-of-plane rotation, low resolution, occlusion, out-of-view, scale variation.

VOT2016 contains sixty test videos with the ground truth. We compared SiamFF with nine trackers including SiamFC [13], SiamRPN [14], SiamAN [4], ACT [4], ColorKCF [18], DSST [16], KCF [5], SAMF [17], TCNN [19]. The results are shown in TABLE 4 with indicators of overlap, robustness (failures), EAO, and FPS. It can be observed that SiamFF performed best on overlap, robustness and EAO. Overlap was 1.05% higher than SiamRPN (2nd) [14], robustness was 11.97% higher than TCNN (2nd) [19], and EAO was 14.22% higher than SiamRPN (2nd) [14]. Figure 11 exhibits the robustness-accuracy ranking of trackers, the abscissa represents robustness and the ordinate represents accuracy. The better tracker is positioned closer to the top-right corner of the figure, and it can be seen that SiamFF has higher accuracy and robustness.

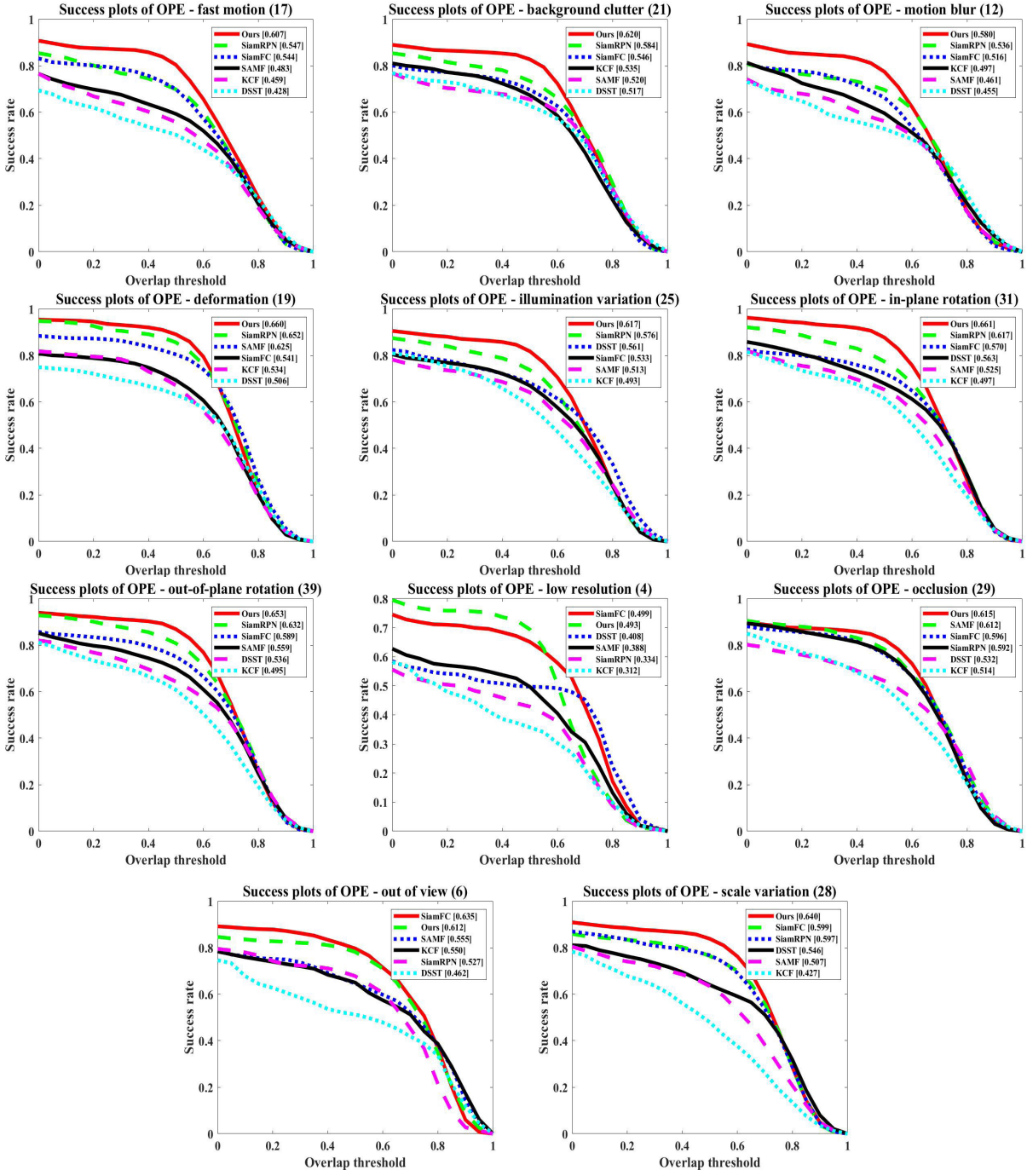


FIGURE 9. Succession plots on OTB2015 over eleven tracking scenes of fast motion, background clutter, motion blur, deformation, illumination variation, in-plane rotation, out-of-plane rotation, low resolution, occlusion, out-of-view, scale variation.

D. ABLATION STUDY

1) DIFFERENT MODULES

To discuss the impact of each proposed module on tracking performance, we listed the experiment results of different modules on benchmarks in TABLE 5. "+" indicates the addition of related modules on the basis of SiamFC. For the module of "feature-level fusion", we only utilized score map35, and for the "score-level fusion" model, we utilized conv2 and conv5 to build two score maps for fusion. From the table, we can observe that the fusion effect at the feature-level was better than that at the score-level because the difference of features in different layers

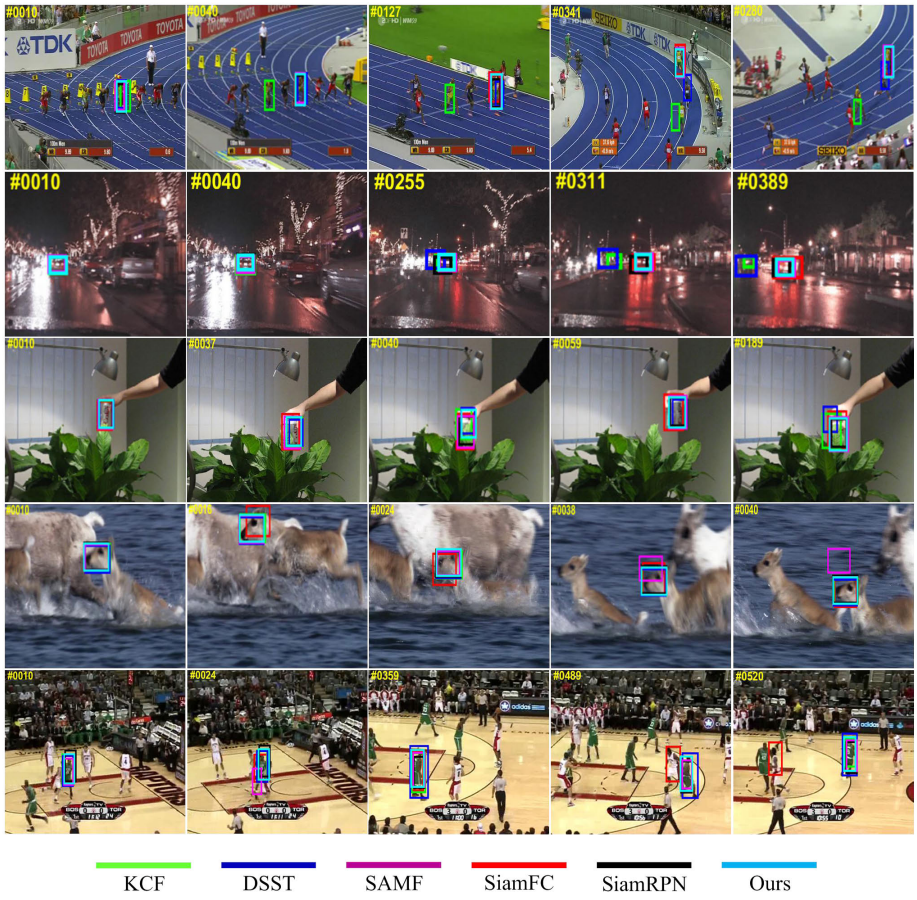


FIGURE 10. The tracking record of SiamFF, KCF, DSST, SAMF, SiamFC and SiamRPN on OTB2015. Five challenging sequences from top to down are bolt, car, coke, deer and basketball.

TABLE 5. Experimental results of different modules on benchmarks.

		SiamFC	+ feature-level fusion	+ score-level fusion	+ multilevel fusion	+ score map filtering	Ours
OTB2015	Prec.	0.809	0.831	0.816	0.839	0.871	0.886
	Succ.	0.607	0.625	0.611	0.633	0.641	0.655
VOT2016	EAO	0.2797	0.3192	0.3023	0.3420	0.3755	0.3896
	Acc.	0.4061	0.4468	0.4233	0.4629	0.4898	0.5121
	Fail.	19.3910	18.0212	19.1131	16.2755	15.9779	15.7913
	Overlap	0.5332	0.5563	0.5366	0.5604	0.5811	0.5865
	FPS	103	70	81	61	87	53

disappears in score maps that position the target by the score value. Furthermore, the table also confirms that the score map filtering strategy improves the tracking performance better than the multilevel fusion network. The object motion attribute enables locating the target in a smaller search range, which is more robust to the interference of analogs than the information fusion. However, the information fusion strategy adds more computations and results in a significant loss in tracker speed. The benchmarks of OTB2015 and VOT2016 show the same trend in indicators for each modules.

2) HISTORICAL FRAMES

The last frame is utilized to predict the target's motion information. To discuss the influence of historical frames on the filtering result, we listed the experimental results of the tracker on benchmarks when using different numbers of frames and shown in TABLE 6. We observe that the tracking performance declines as more historical frames were utilized. In addition, more frames reduced the tracking speed. In the strategy, historical frames record the motion information of the target in the past. Since the target keeps moving, the increase in historical frames cannot accurately predict the

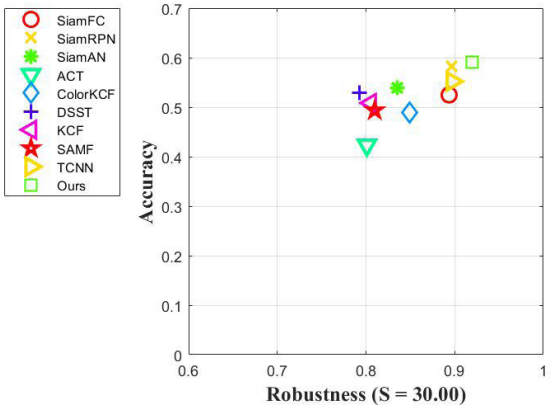


FIGURE 11. The robustness-accuracy ranking of ten trackers; the better tracker is positioned closer to the top-right corner of the figure.

TABLE 6. Experimental results of different historical frames on benchmarks.

	1 frame (last frame)	5 frames	10 frames
Precision scores	0.886	0.863	0.821
Succession scores	0.655	0.637	0.619
EAO	0.3896	0.3702	0.3204
FPS	53	44	28

current motion information of the target, and the accumulation of errors will lead to a decline in tracking performance.

V. CONCLUSION

Considering the problem of poor robustness to similar objects caused by traditional trackers' neglect of shallow features and the limitation of cosine windows, first, a multilevel fusion network was proposed. A layer-hopping connection was utilized to fuse the shallow and deep features at feature-level, and then the similarity information was further fused at the score-level to filter out most analogs. Second, the score map filtering strategy was carried out in the predict stage, which uses the interframe motion information of the target to limit the detection area of the tracker, further filters out similar objects with strong influence and improves tracking performance. In the experiments on OTB2015 and VOT2016 compared with other state-of-the-art trackers, our algorithm ranked at the forefront in accuracy and robustness and showed excellent performance.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for valuable suggestions to improve this paper and the subjects for participation in the experiment.

REFERENCES

[1] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 125–141, May 2008.

[2] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

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

[4] M. Kristan, A. Leonardis, and J. Matas, "The visual object tracking VOT2016 challenge results," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Feb. 2016, pp. 777–823.

[5] J. F. Henriques, R. Caseiro, and P. Martins, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. vol. 37, no. 3, pp. 583–596, Mar. 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

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

[8] Y. Kuai, G. Wen, and D. Li, "Hyper-feature based tracking with the fully-convolutional siamese network," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 472–481.

[9] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[10] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4310–4318.

[11] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4819–4827.

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

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

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

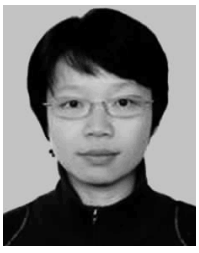
[15] P. Gao, Y. Ma, R. Yuan, L. Xiao, and F. Wang, "Learning cascaded siamese networks for high performance visual tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3078–3082.

[16] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1108–1117.

[17] Z. Liu, Z. Lian, and Y. Li, "A novel adaptive kernel correlation filter tracker with multiple feature integration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 254–265.

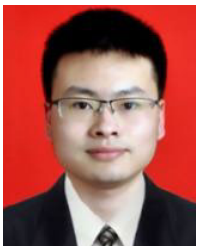
[18] P. Senna, I. N. Drummond, and G. S. Bastos, "Real-time ensemble-based tracker with Kalman filter," in *Proc. 30th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2017, pp. 338–344.

[19] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang, "T-CNN: Tubelets with convolutional neural networks for object detection from videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2896–2907, Oct. 2018.



YUAN LUO received the M.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 1996, and the Ph.D. degree from Chongqing University, Chongqing, in 2003. She was a Visiting Scholar at the Université de Montréal, Canada, in 2006. She is currently a Professor with the CQUPT. Her research interests include computer vision, photoelectric sensing, image processing, and mobile robots.

YUANXIAO CAI received the B.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China, in 2018, where he is currently pursuing the master's degree with the College of Photoelectrics. His current research interests include pattern recognition and visual tracking.



BOYU WANG received the master's degree from the Chongqing University of Posts and Telecommunications, in 2018, where he is currently pursuing the Ph.D. degree. His current research interests include computer vision and pattern recognition.

JIE WANG received the B.S. degree from Hubei Engineering University, Hubei, China, in 2017. He is currently pursuing the master's degree with the College of Photoelectrics, Chongqing University of Posts and Telecommunications (CQUPT). His current research interests include image processing, pattern recognition, and visual tracking.

YANJIE WANG received the B.S. degree from the Jiangxi University of Finance and Economics (JUFE), Nanchang, China, in 2018. He is currently pursuing the master's degree with the College of Data Science, Zhejiang University of Finance and Economics (ZUFE). His current research interests include data mining and processing.