Siamese Network Tracker Based on Dynamic Convolution And Attention Fusion of Shallow And Deep Information

Zhangfang Hu, Hongling Yu, Kehuan Linghu

Abstract—**Given that traditional networks lack the ability to adaptively fine-tune weight parameters within feature layers and solely rely on deep feature data, this paper introduces a pioneering tracking algorithm employing a siamese network architecture. This algorithm is built upon the dynamic convolution and attention fusion of both shallow and deep information. Its aim is to enhance tracking performance by accurately extracting image features. To begin, we adopted ResNeSt as the foundational network architecture. To enable the dynamic adjustment of the network feature layer's weight parameters, we replaced the first three traditional convolutional layers with dynamic convolutional layers, while leaving the last two convolutional layers untouched. Next, we integrated channel and spatial attention mechanisms into each convolutional layer, and fused the third and fifth convolutional layers of the two branch networks to yield a pair of complementary feature mappings. Ultimately, we fused the resulting score map at the fractional level to produce the ultimate score map. This approach effectively mitigates the impact of similarity interference, enhancing the tracker's robustness. The experimental results from OTB2015 and VOT2018 datasets unequivocally demonstrate a notable enhancement in the tracking performance of this algorithm.**

Index Terms—**Dynamic convolution, Attention mechanism, Siamese network, Deep learning, Feature fusion**

I. INTRODUCTION

AS an important branch of machine vision, visual object tracking integrates the related technologies of image detection and image processing, which has important research significance and great challenge. Target tracking technology is widely used in many places, including UAV [1], interaction between individuals and computational systems [2], simulated reality environments [3], and automatic driving [4], playing an important role in modern social life. Currently, the primary challenge in visual target tracking technology is its limited ability to extract distinctive features from the target object when it closely resembles its surrounding environment. Additionally, the technology tends to erroneously capture similar features from the background information, and it struggles when the target object undergoes deformation. The network model cannot extract target features well, and it is easy to lose important feature points and bright environment. Overexposure can readily obscure the key salient attributes within the target image, resulting in the inability of real-time tracking to be effective, ultimately leading to the loss of the tracking target.

As visual object tracking has advanced, deep convolutional neural network (CNN) have emerged as fundamental tools for image processing and classification within the realm of computer vision. Convolution serves as a pivotal constituent within convolutional neural networks. Networks with different convolutional structures have different functions, but in essence they are all used for feature extraction. At present, there are up to 20 convolution methods commonly used in deep learning networks, which are mainly divided into classical convolution and convolution variants. Classical convolutions include convolutional neural networks, group convolutions [5], and depth-separable convolutions [6], etc., while convolution variants include dynamic convolutions [7], asymmetric convolutions [8], and conditional parametric convolutions [9]. Among them, some convolutional blocks will increase parameter count and additional operations in the process of improving target tracking accuracy, but compared with improving target tracking accuracy and feature extraction capability, their disadvantages can be ignored.

To tackle the previously mentioned challenges, this research introduces a siamese network-based tracking a designated target system that leverages dynamic convolution and attention fusion techniques to integrate shallow and deep information. The primary division encompasses two principal segments:

1) In traditional target tracking model networks, CNNs are employed for the extraction of features from target objects, with each convolutional kernel having consistent weights. The backbone network, ResNeSt [10], is utilized, and a dynamic convolution approach is introduced. This substitution involves the replacement of the initial three convolutional kernels within the foundational network architecture with dynamic convolution, while keeping the final two convolutional kernels unaltered. Dynamically

Manuscript received June 19, 2023; revised October 29, 2023. This work was supported in part by the National Natural Science Foundation Youth Fund Project (Grant No.61801061), and the National Natural Science Foundation of China (Grant No.51905065, and 52005070).

Zhangfang Hu is a Professor at the Key Laboratory of Optical Information Sensing and Technology, School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: huzf@cqupt.edu.cn)

Hongling Yu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (corresponding author phone: 182-252-70880; e-mail: 2562863425@qq.com)

Kehuan Linghu is a graduate student of the School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065 China (e-mail: 1353319371@qq.com)

regulating the weighting of individual convolution kernels through the utilization of dynamic convolution.

2) In the context of image feature extraction, our approach involves enhancing the correlation between channel and spatial feature information across convolutional kernels. We achieve this by introducing spatial and channel attention modules following each feature layer in the fundamental network structure. We utilize these attention mechanisms to elevate similarity and accuracy during the feature extraction phase from target objects. The introduction of these mechanisms enables the more effective differentiation between target image details, background details, and analog interference present in the shallow feature maps generated by the network. These enhancements allow for the fusion of in-depth attributes to boost target tracking effectiveness. We employ a fusion network method with jump connections. Initially, the two branches are merged at the feature layer, blending both shallow and deep features to produce complementary feature representations. Following this, the feature representations from individual branches are correlated with the output features from the network's final layer, yielding a pair of similar score maps. Finally, the amalgamation of these score maps results in the production of the ultimate score map.

Extensive experimental results indicate the high performance of the presented algorithm in recent benchmark experiments. The subsequent sections of this paper are organized as follows: Section two discusses related work, Section three delves into our methodology, Section four provides insights into the network's training process and its performance evaluation, and finally, Section five concludes the study.

II. Related Work

J.F. Henriques et al. [11] asserted that KCF method played a pioneering role in the evolution of correlation-based tracking technology, thereby establishing a benchmark for the progress of tracking a designated target methods. D.S. Bleme et al. [12] asserted the minimum mean square error filtering and pioneered the application of correlation-based tracking technology to the domain of target localization and tracking. The MOSEE filter tracker ran at a speed of 669 frames per second, achieving efficient tracking effect.

The Siamese network consists of two identical branches with a unified weight structure. Due to the inherent two-branch nature of the Siamese network, the task of tracking a designated target is effectively converted into a target matching task. This transformation focuses on the similarity mapping between the learning search and the target image, leading to a significant acceleration in the model's tracking speed. SINT [13] pioneered the utilization of the Siamese network for target tracking. Within the course of their tracking procedure, they generated multiple candidate regions and compared them with the initial frame, ultimately selecting the most similar target as the network's output. Bertinetto et al. [14] presented SiamFC, an algorithm for target tracking that is rooted in a fully-convolutional siamese network framework. They integrated AlexNet into the network to estimate the location of targets through feature representations on both branches. The configuration of SiamFC is depicted in Figure. 1. Li Bo et al. [15] put forward SiamRPN (Siamese Region Proposal Network), based on the Fast-RCNN method for object detection. They augmented the original network structure with two branches for classification and regression and introduced the region proposal network to address adaptive bounding box transformations. Qiang Wang et al. [16] introduced a mask segmentation module to propose the SiamMask algorithm, enabling the unified execution of both target tracking and target segmentation.

Saining Xie et al. [17] introduced the ResNeXt network within the ResNet framework, employing group convolution to consolidate the multipath structure into a unified operation. Zhang et al. [18] presented the SiamDW algorithm, which incorporates the internal clipped residual unit (CIR) module to address the issue of network performance degradation as the network's depth increases. Chen et al. asserted dynamic convolution, a technique that dynamically amalgamates several parallel convolution cores kernels based on attention mechanisms without augmenting the network's depth and width. It further allows for the adaptive adjustment of the parameterization of individual convolution kernels in response to the feature points of the target object.

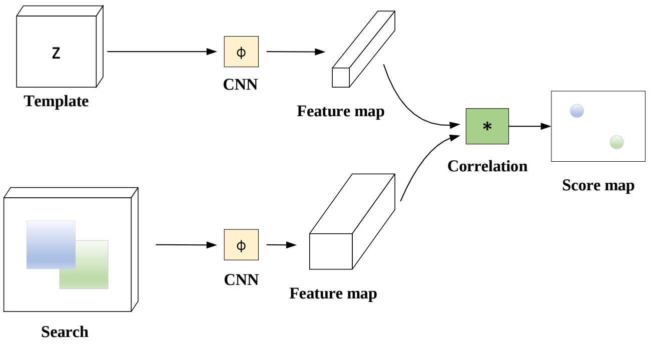


Fig.1. The Structure of SiamFC.

III. Our Approach

A. Dynamic Convolution

Traditional convolutional neural networks classify input information by translation invariants. The feature extraction of input information is completed in the convolution layer, the scaling modification and dimensionality reduction are completed in the pooling layer, and the result classification and output are completed in the fully connected layer. The whole process is transmitted by neurons, as illustrated in Figure. 2. After CNN calculation, feature fusion is carried out on the local region of each feature map. As the convolutional kernel weights remain consistent across each layer, the feature map produced when the surrounding environment closely resembles the target object fails to emphasize the target's distinctive features adequately. Furthermore, traditional deep convolutional neural networks entail relatively low computational demands. For networks with too many convolutional layers and channels, the expression of feature extraction ability in feature layer is limited, which leads to the degradation of target tracking performance. Dynamic convolution can change the structure of the convolutional model without inflating the network's depth and breadth. It dynamically consolidates numerous parallel convolutional cores according to attention, and dynamically adjust the weights of convolutional cores for different feature graphs to generate adaptive convolution, so as to dynamically fine-tune the weight parameters for each individual convolutional kernel.

In contrast to CNN, each of its layers has a group of K concurrent convolution nuclei, represented as $\{\tilde{W}_k, \tilde{b}_k\}$, and these convolution nuclei of each individually input x (such as an image) are dynamically aggregated using the input attention $\{\pi_1(x), \pi_2(x), \cdots \pi_k(x)\}$, Figure. 3. illustrates the attention module. It dynamically modulates the weight parameter $\{\pi_1, \pi_2, \cdots, \pi_k\}$ for the K convolutional kernels based on the input image to achieve the purpose of adaptive dynamic convolution. The configuration of the dynamic convolutional network is depicted in Figure. 4., and the weights and bias linear functions of K convolutional nuclei after aggregation are shown as follows.

$$\tilde{W}(x) = \sum_{k=1}^{K} \pi_k(x) \tilde{W}_k \tag{1}$$

$$\tilde{b}(x) = \sum_{k=1}^{K} \pi_k(x) \tilde{b}_k \tag{2}$$

π_k symbolizes the attention weighting, $\tilde{W}(x)$ denotes the aggregation weight, while $\tilde{b}(x)$ symbolizes the bias after aggregation.

Building on the research discussed above, we present a method to replace traditional convolutional networks with dynamic convolutional networks. Given that the conventional foundational network architecture, AlexNet, has a limited number of network layers, which restricts the tracker's enhancement, we opt for ResNeSt as the foundational network architecture. Additionally, We removed the ultimate full connection layer from the network, as there is no necessity for performing classification operations on the results. In contrast to the conventional foundational network architecture, ResNeSt boasts a greater number of convolutional layers and an increased channel count, enabling the comprehensive utilization of deep-level characteristics of the network in the tracking process. We replace the first three CNN convolution kernels in ResNeSt with dynamic convolution, leaving the last two unchanged. The trajectory of the target subject is consistent. In traditional convolutional kernels, when applied to the feature maps of consecutive frames, the resulting feature maps are unable to swiftly and precisely pinpoint the score locations of the target object. This limitation arises from the minor changes between the two frames and the uniform weights of each convolutional kernel. This leads to the phenomenon of slow target tracking speed and tracking failure. With the integration of dynamic convolution, the attention module within dynamic convolution dynamically adapts the weights of K convolution cores based on the input frames. This dynamic adjustment allows for the rapid determination of the target object's location and facilitates real-time tracking.

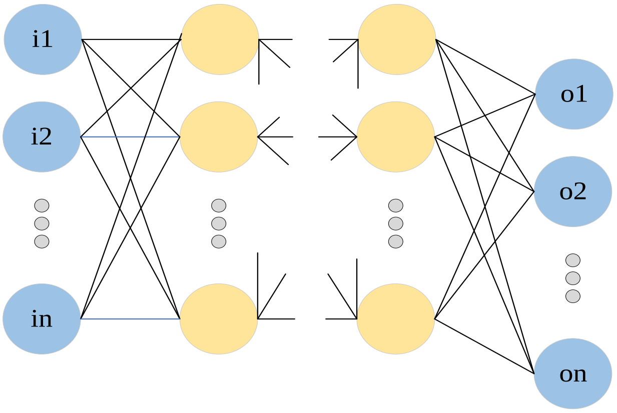


Fig. 2. CNN feature transmission.

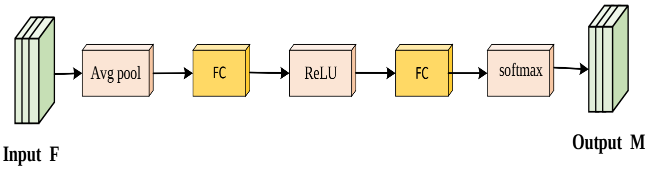


Fig. 3. Attention Module.

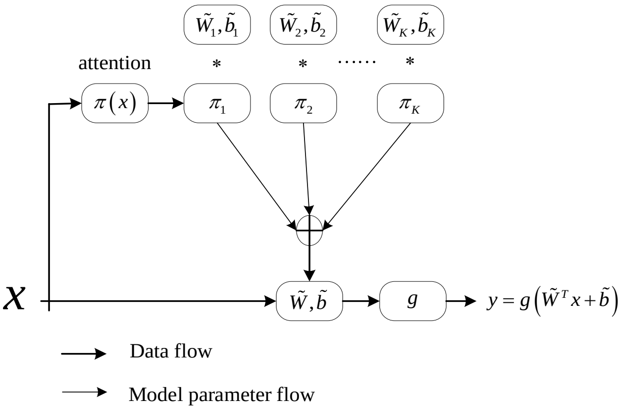

$\tilde{W}_1, \tilde{b}_1$
$\tilde{W}_2, \tilde{b}_2$
$\tilde{W}_K, \tilde{b}_K$
attention
*
*
……
*
$\pi(x)$
π_1
π_2
π_K
x
$\tilde{W}, \tilde{b}$
g
$y = g(\tilde{W}^T x + \tilde{b})$
Data flow
Model parameter flow

Fig. 4. Dynamic Convolutional Networks.

B. Attention Multi-level Information Fusion Siamese Network Tracker

Traditional trackers such as SiamFC and SiamRPN do not perform fractional fusion except for the last layer of feature information after feature extraction. The graph after visualization of each CNN layer is shown in Figure. 5. It is noticeable that as the quantity of network layers rises and the network depth increases, the feature map's resolution diminishes. The first three layers can roughly identify the appearance of the target, and the last two layers can not identify the appearance of the target. Shallow features encompass high-resolution appearance details and can capture fine spatial information within the target image, making them well-suited for target localization. Conversely, deep features carry lower-resolution semantic information, which renders them more resilient in the face of deformable target images and ideal for target classification. To enhance the overall performance of target tracking, we suggest integrating both shallow and deep features to address and complement their respective limitations.

In the process of feature extraction in dynamic convolutional networks, the channel information of feature graphs obtained through the underlying convolution is not perfect enough. To enhance the connectivity of feature information across channels and spatial dimensions among convolutional layers, We propose the incorporation of spatial and channel attention modules following every feature layer within the foundational network architecture. The attention mechanism is employed to enhance the similarity and accuracy of feature extraction. To extract features from input images, we combine two attention mechanisms of CBAM [19] module. This adjustment aims to elevate the importance of feature points relevant to target tracking within the network model while diminishing the significance of irrelevant feature points. Consequently, the tracker focuses more on the image features critical for target tracking, resulting in an enhancement in tracking accuracy. The two attention models are illustrated in Figure. 6. and Figure. 7. respectively. The channel weight obtained by the channel attention mechanism is:

$$M_c = \beta_1 \sigma a_{cMax} \oplus \beta_2 \sigma a_{cAvg} \tag{3}$$

The spatial attention mechanism obtains the following weights:

$$M_s = \sigma(h^{3\times3}(cc([AvgPool(Y)];[MaxPool(Y)]))) \tag{4}$$

σ is the sigmoid activation function, a_{cMax} and a_{cAvg} are channel feature information obtained from maximum pooling and average pooling respectively, β_1 and β_2 are fusion weights, both set to 0.5, M_c is the learned channel attention weight. $h^{3\times3}$ is the 3×3 convolution, cc is fully connected, and M_s represents the weight of spatial attention acquired through learning.

Based on the above research, we propose a Siamese network based on dynamic convolution and shallow and deep information fusion of attention. The network structure is shown in Figure. 8. Set template images to 127×127 and search images to 255×255. Firstly, the input image is fused at the feature level. The method of jump connection is used to fuse Dy_conv3 and conv5 in the two branches after passing attention and spatial attention, so as to achieve the effect of complementation of deep information and shallow information. Given the varying sizes and channel dimensions of feature representations at different layers in the network, we employ maximum pooling and 1×1 convolution modules to standardize the fused feature representations, ensuring their uniform size. This approach not only standardizes the images but also preserves the original spatial information. Then, after the fusion of the image features and the last layer of image features, two feature representations are obtained. Finally, The 1×1 convolution is used to calculate the correlation between the two feature graphs, and the final fraction graph with the size of 25×25 is obtained. The network for merging shallow and deep information, featuring channel attention and spatial attention, effectively filters out minor and long-range interferences. Consequently, the peak points in the score map become more concentrated, eliminating scattered or subtle disturbances and thereby enhancing the tracker's accuracy. The algorithm's procedural steps are delineated in TABLE I.

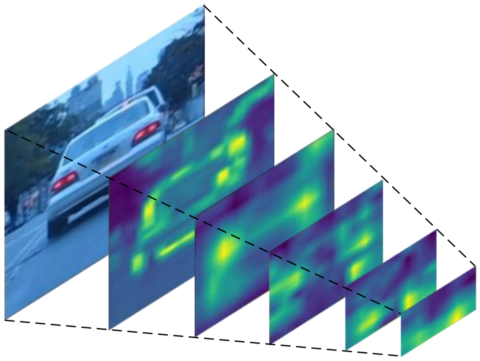


Fig. 5. Visual Graphics.

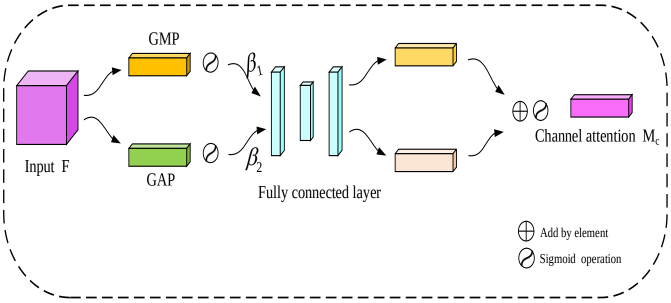


Fig. 6. Channel Attention Mechanism.

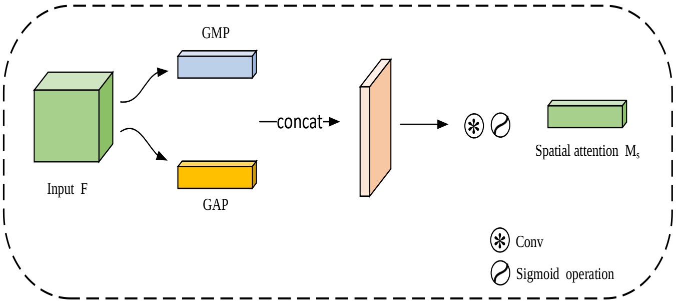


Fig. 7. Spatial Attention Mechanism.

IV. Benchmark Experiments

A. Experimental Details

The system environment used in this experiment is Ubuntu 16.04LTS, the hardware is NVIDIA 2080Ti GPU and an Intel Xeon E5 CPU, while MATLAB 2018b serves as the experimental tool.

The data sets used are GOT-10k [20], ILSVRC [21], OTB2015 [22], and VOT2018 [23]. The network's hyperparameters are configured as follows: learning rate = 0.005, individual sample size = 16, and cycle count = 80.

B. Training Details

Given that this paper replaces the conventional AlexNet

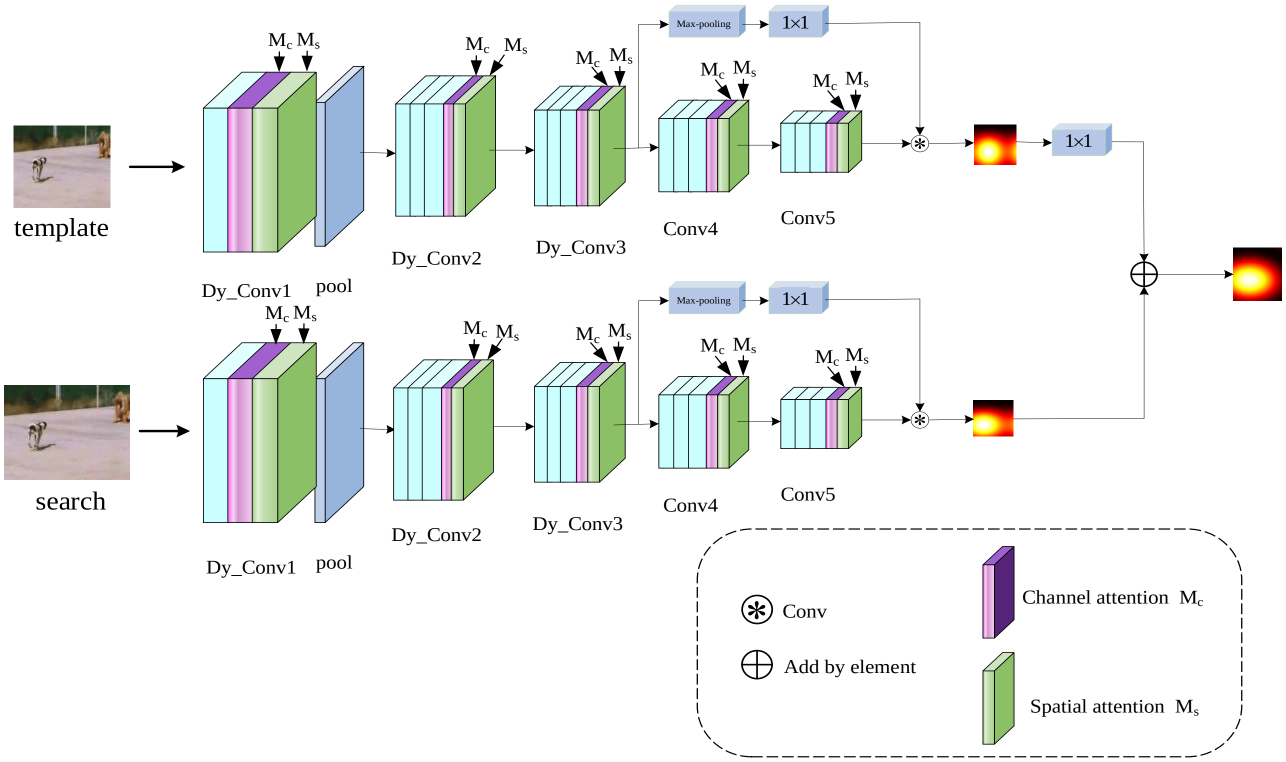


Fig. 8. The Framework of Dynamic Convolutional Attention Shallow and Deep Information Fusion Network (Dy_Conv represents a dynamic convolutional layer).

TABLE I. ALGORITHM SUMMARY FRAMEWORK.

Feature fusion network algorithm
Input: Template image z, $X=\{x_1,x_2,\cdots,x_n\}$ (n represents the number of frames where the target object is moving).
1. Replace the first three layers of CNN of the foundational network architecture with dynamic convolution.
2. Feature extraction is carried out after two attention mechanisms for each layer of the foundational network architecture convolution.
3. Extract the features of z from the network.
4. Extract the features of Dy_Conv1, Dy_Conv2, Dy_Conv3, conv4, and conv5.
5. while (i<n) {
6. Generate feature maps by maximizing pooling and 1 × 1 convolution fusing z's Dy_Conv2 and conv5.
7. Combine a pair of complementary feature representations from two branches to derive the ultimate score map.
8. Filter is used to obtain the ultimate outcome.
9. }
Output: Use the target box to track the target object.

backbone network with ResNeSt, which has already been initialized with image labels from the ImageNet dataset, we proceed to further train the network using ILSVRC and GOT-10k datasets. CNN is utilized for feature extraction from search images and template images, with the subsequent process of generating the final score map post-extraction delineated as follows:

$$S(z,x)=f(\varphi(z),\varphi(x)) \tag{5}$$

$\varphi(z)$ represents template image features, $\varphi(x)$ represents search image features, $f(\bullet)$ represents related operations, $S(z,x)$ denotes the similarity measure between search images and template images, and the primary objective of the network is to maximize the value of $S(z,x)$. The logical loss function is used to train the network, which is defined as:

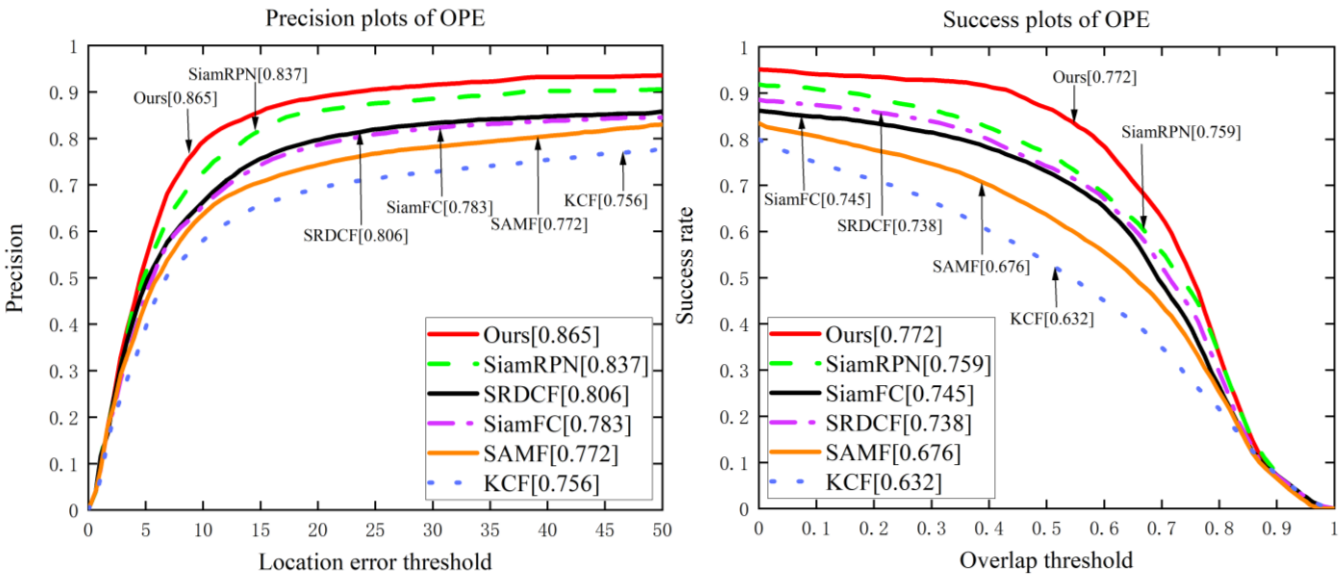
$$L(y,v)=\frac{1}{D}\sum_{u\in D}\log(1+\exp(-y[u]v[u])) \tag{6}$$

u denotes a score point on the score chart, $v[u]$ signifies the similarity score associated with that point, and $y[u]$ denotes the actual label of the score point. The optimization of the loss function is achieved through the application of the stochastic gradient descent (SGD) technique, leading to the update and acquisition of the network's weight parameters. $y[u]$ is defined by the position of any point on the score chart and the target center point:

$$y[u]=\begin{cases}+1 & k\,\|u-c\|\leq R\\ -1 & otherwise\end{cases} \tag{7}$$

k denotes the iteration step of the neural network, and c signifies the centroid of the target image.

In training, the image is cropped with the target's position as the central point, and the search image and template image are resized to dimensions of 127×127 and 255×255, respectively. When the image clipping area is insufficient, the average RGB is used to fill.



(a) Precision. (b) Succession.

Fig. 9. The Precision and Succession Plots on OTB2015 with One Pass Evaluation (OPE).

TABLE II. TRACKER SUCCESSION AND PRECISION SCORES. “IMPROVE" REPRESENTS THE IMPROVEMENT OF OUR TRACKER OVER OTHER TRACKERS, AND "SPEED" REPRESENTS THE TRACKING SPEED.

Tracker	Precision score	Succession score	Improve(%)		Speed(FPS)
			Pre.(%)	Succ.(%)	
SiamRPN	0.837	0.759	3.35	1.71	200
SRDCF	0.806	0.738	7.19	4.61	51
SiamFC	0.783	0.745	10.47	3.62	89
SAMF	0.772	0.676	12.05	14.20	27
KCF	0.756	0.632	14.42	22.15	7
Ours	0.865	0.772	-	-	66

Since the dynamic convolution has K convolution nuclei in each layer, respectively $\{W_1, W_2, \cdots, W_K\}$, which leads to difficult training problems, the attention model is constrained to parameters $\sum_{k=1}^{K} \pi_k(x) = 1$ and $0 \le \pi_k(x) \le 1$ to facilitate the learning process of the model $\pi_k(x)$.

C. OTB2015 Experiment

The evaluation metrics employed in the OTB2015 dataset comprise Success and Precision. The tracker's performance robustness is evaluated using the One Pass Evaluation (OPE) criterion.

1) *Precision:* Compute the proportion of video frames less than a given threshold is calculated based on the Euclidean distance by the following formula:

$$s = \sqrt{(x_u - x_r)^2 + (y_u - y_r)^2} \tag{8}$$

(x_u, y_u) represents the central point of the predicted bounding box, and (x_r, y_r) represents the central point of the ground truth bounding box. A lower value indicates better tracking performance. A curve can be generated using various thresholds, and higher curve values indicate superior tracker performance.

2) *Succession:* Target tracking success can be measured by an overlap score (OS), expressed as follows:

$$OS = \frac{|bounding\ \ box \cap ground\ \ truth\ \ box|}{|bounding\ \ box \cup ground\ \ truth\ \ box|} \tag{9}$$

$|\bullet|$ denotes the pixel count within the specified area. If the OS value of any frame surpasses the predefined threshold, it signifies a successfully calibrated for that frame. Otherwise, the object fails to be calibrated. The typical threshold is established at a value of 0.5.

3) *One Pass Evaluation (OPE):* In the assessment of tracking performance, the testing process exclusively leverages the initial frame within the video sequence to establish the true target position. Subsequently, the algorithm is executed to compute the metrics for succession and precision.

Experimental trials were conducted on the OTB2015 dataset to compare and analyze the proposed algorithm with the current advanced tracking algorithms, including KCF, SAMF [24], SRDCF [25], SiamFC and SiamRPN trackers. Among them, KCF, SAMF and SRDCF are related filter type trackers, and SiamFC and SiamRPN are siamese network type trackers. Figure. 9. depicts a comparative chart showcasing the outcomes of the proposed algorithm in contrast to other tracking algorithms, with "Ours" denoting our proposed algorithm. In Figure. 9(a), the precision chart is presented, with the score in the upper right corner indicating the tracker's performance at a center error of 20 pixels. In Figure. 9(b), the success chart is depicted, and the score in the upper right corner represents the area under the curve. It can be seen from

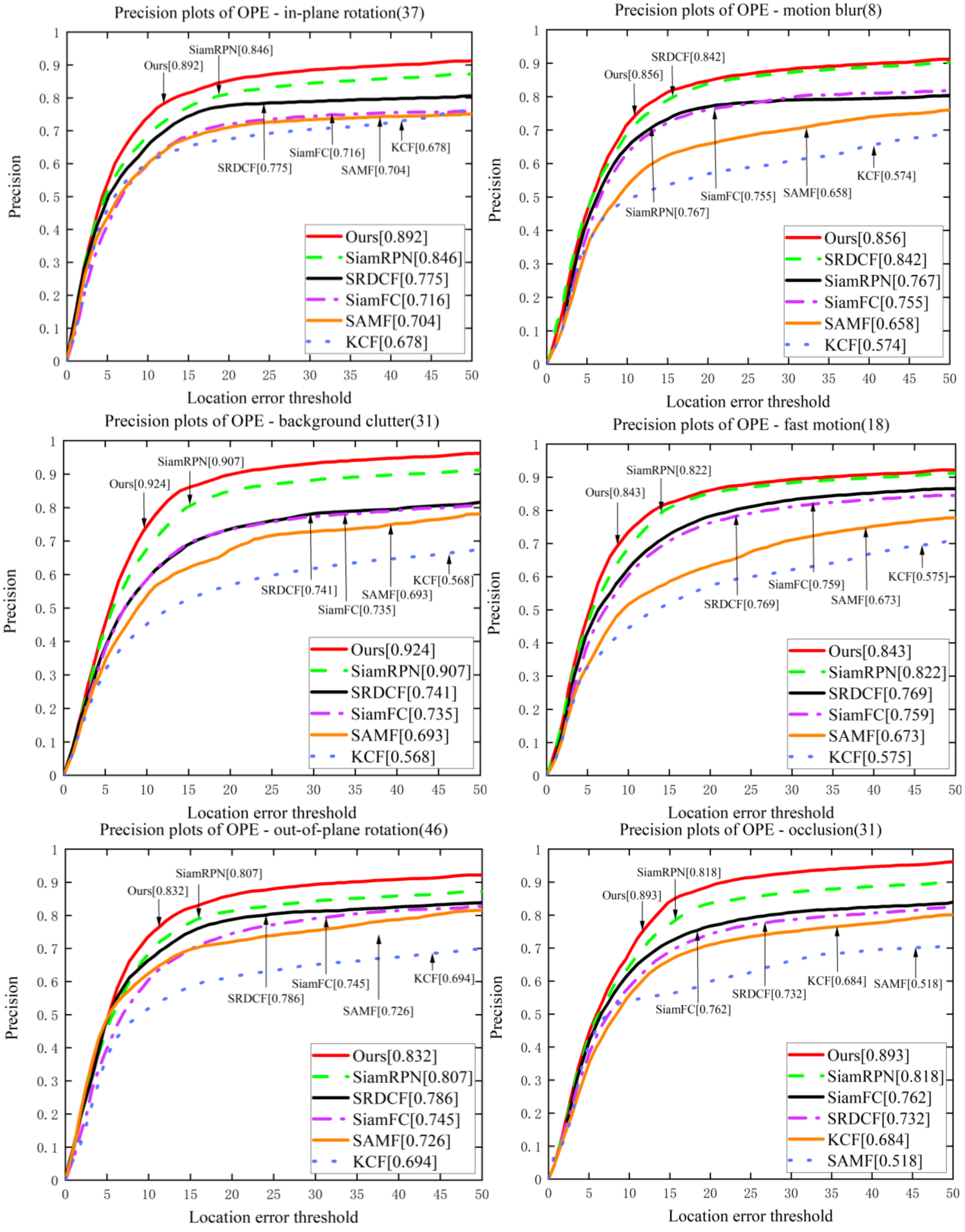


Fig. 10. Precision Plots on OTB2015 Over Six Tracking Scenes of In Plane Rotation, Motion Blur, Background Clutter, Fast Motion, Out of Plane Rotation and Occlusion.

the figure that this algorithm is superior to other tracking algorithms in terms of precision and success rate. For a detailed breakdown of the performance differences across various tracking algorithms on these two metrics, please refer to TABLE II.

In contrast to SiamRPN, the proposed algorithm exhibits enhancements both performance, with improvements of 3.35% and 1.71%, respectively. Additionally, compared to

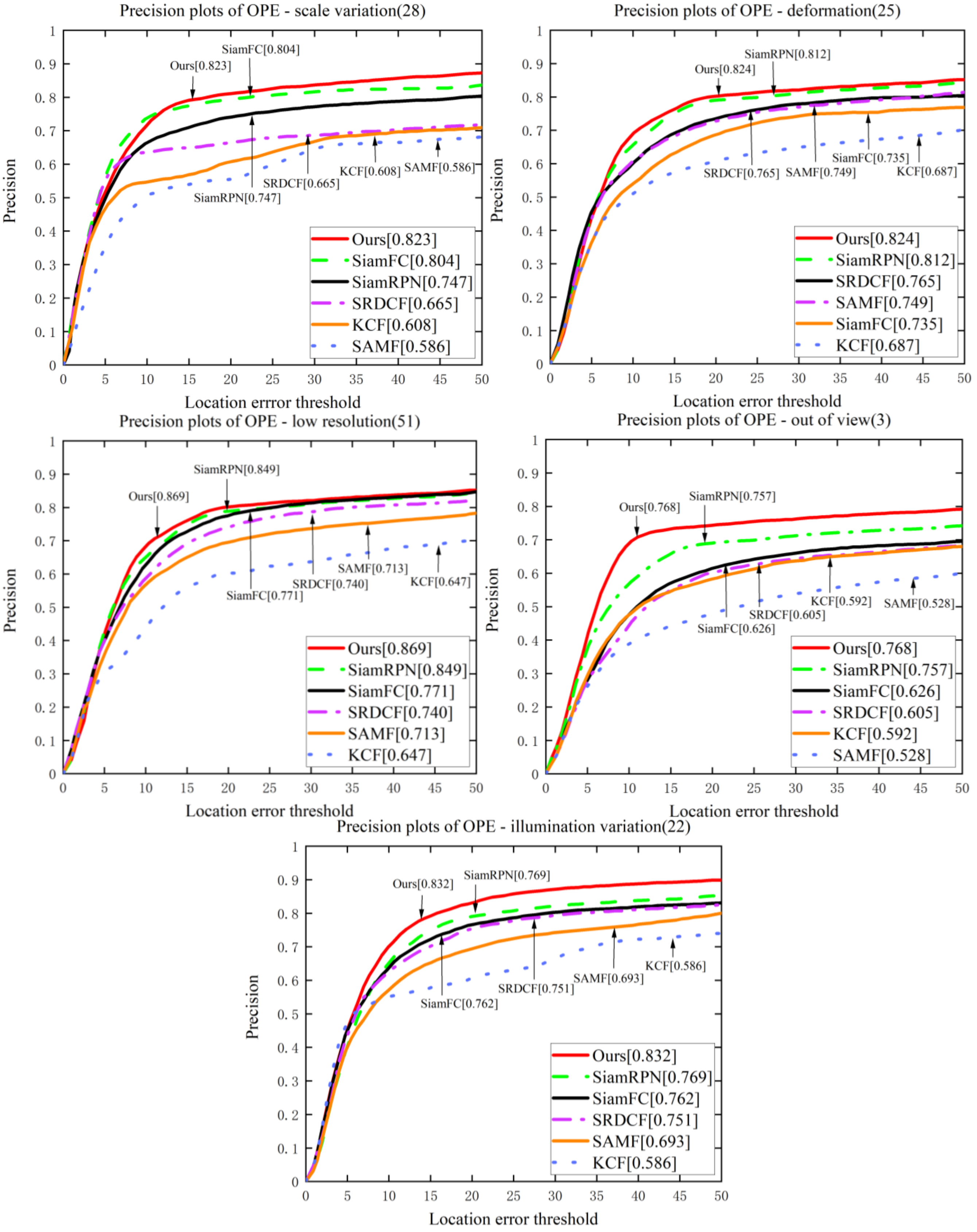


Fig. 11. Precision Plots on OTB2015 Over Five Tracking Scenes of Scale Variation, Deformation, Low Resolution, Out of View and Illumination Variation.

SRDCF, the algorithm achieves remarkable gains, with precision and success rates increasing by 7.19% and 4.61%, respectively. Regarding tracking speed, the algorithm presented in this paper achieves a frame rate of 66 FPS, which is 15FPS, 39FPS and 59FPS faster than that of SRDCF, SAMF and KCF trackers, respectively. Due to the addition of attention mechanisms in the network, and the integration of shallow features and deep features, as the network's computational complexity increases, it leads to a reduction in the tracking speed of the proposed algorithm, but it still meets

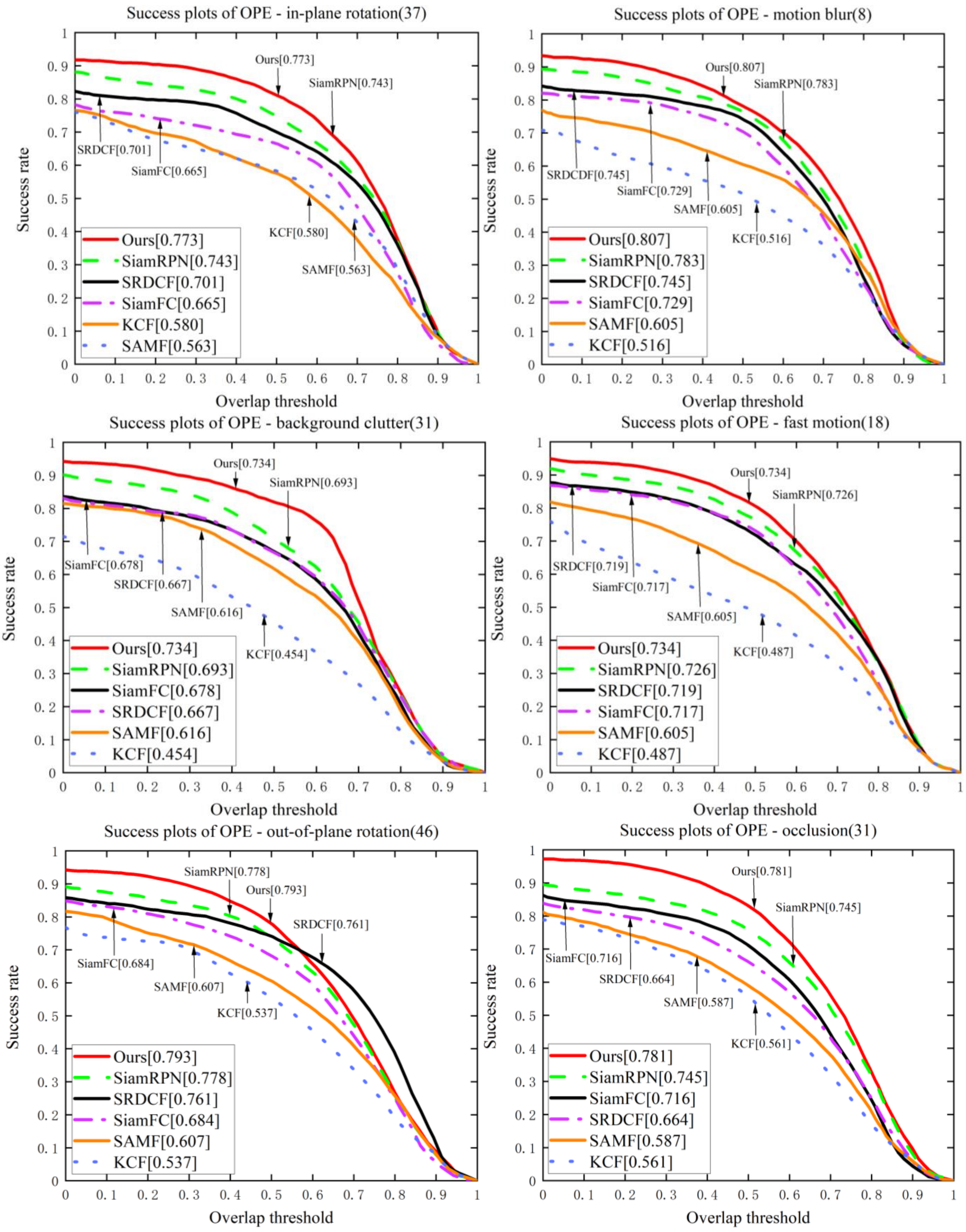


Fig. 12. Succession Pots on OTB2015 Over Six Tracking Scenes of In Plane Rotation, Motion Blur, Background Clutter, Fast Motion, Out of Plane Rotation and Occlusion.

the real-time tracking requirements.

To further investigate the tracking algorithm's performance across varying environmental conditions, we conducted tests in 11 unconstrained environments. The test environment includes "Low Resolution", "Background Clutter", "Out of view", "out-of-plane Rotation", "In-Plane Rotation", "Fast Motion", "Motion Blur", "Deformation", "Occlusion", "Scale variation", "Illumination variation". The precision is illustrated in Figure. 10. and Figure. 11., and the succession is illustrated in Figure. 12. and Figure. 13.

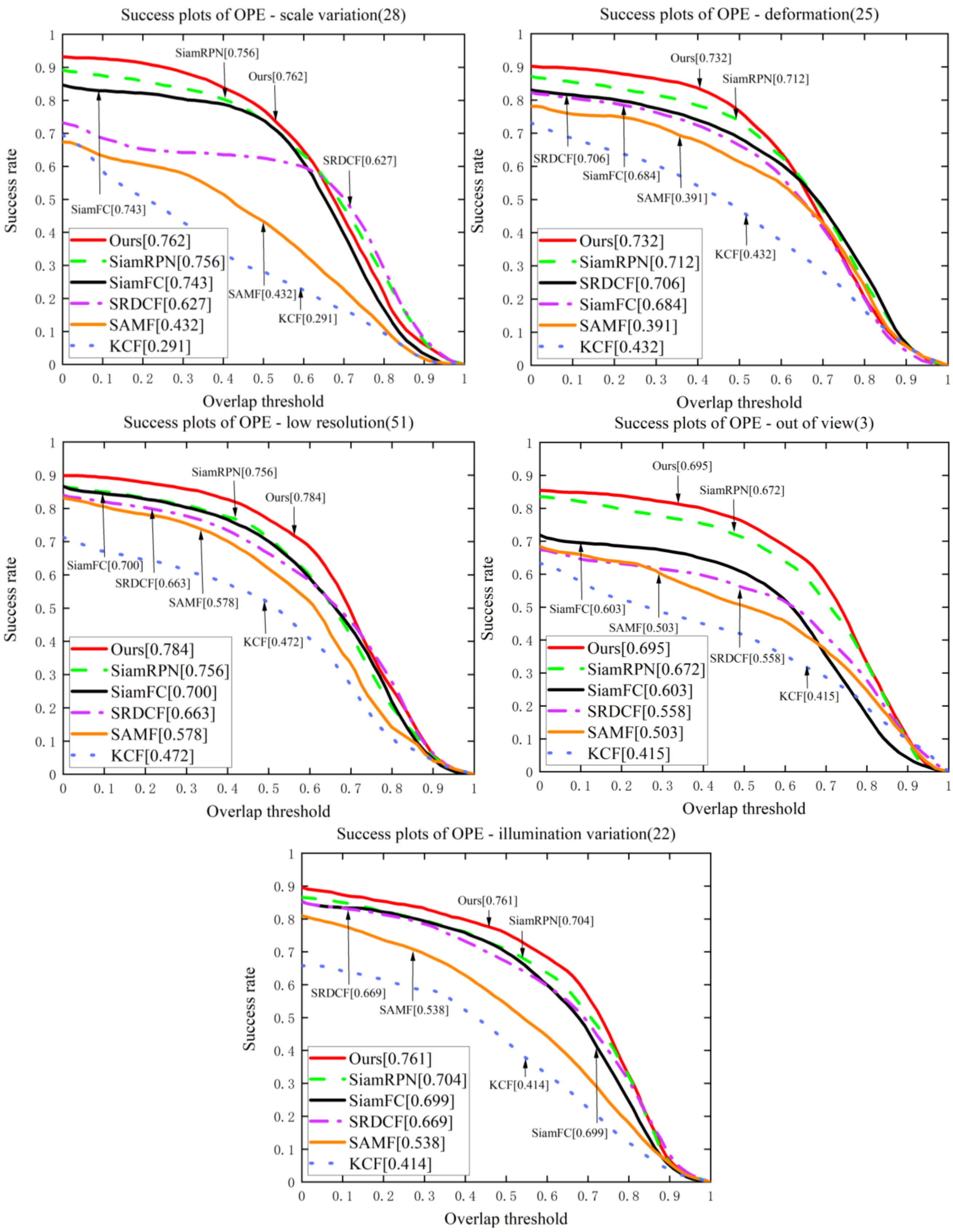


Fig. 13. Succession Pots on OTB2015 Over Five Tracking Scenes of Scale Variation, Deformation, Low Resolution, Out of View and Illumination Variation.

The proposed algorithm integrates attention mechanisms, and increases the weight of the target image and diminishing the impact of the disturbing object during the feature extraction process, thus effectively improving the target positioning precision. The fusion of shallow feature and deep feature leverages the superficial feature's appearance-related information thereby significantly diminishing the center error between the object prediction frame and the genuine object frame. Evident from Figures 10. and 11., our proposed algorithm secures the top position in the ranking across 11

TABLE III. TRACKERS' SCORES OF FIVE INDICATORS IN VOT2018.FOR EACH INDICATOR, RED AND BOLD ARE THE BEST, GREEN AND BOLD ARE THE NEXT BEST.

Tracker	SiamFC	SiamAN	SiamRPN	TCNN	SAMF	KCF	DSST	ColorKCF	ACT	Ours
EAO	0.2597	0.2149	**0.3402**	0.3171	0.1732	0.1732	0.1654	0.2135	0.1521	**0.3894**
Acc.	0.3987	0.3974	0.4601	**0.4856**	0.3387	0.2963	0.3104	0.3312	0.2735	**0.5312**
Fail.	20.4196	31.0820	21.5246	**18.3524**	37.8454	39.0412	45.8868	26.0937	42.8138	**14.8761**
Overlap	0.5312	0.5233	**0.5781**	0.5392	0.1878	0.1878	0.5166	0.4826	0.4244	**0.5912**
FPS	103	12	**200**	1	5	22	13	**111**	82	57

TABLE IV. TEST RESULTS OF DIFFERENT MODULES IN BENCHMARK EXPERIMENTS.

		Without any modules	+Dynamic Convolution	+attention mechanism	+Shallow and deep fusion	Ours
OTB2015	Prec.	0.805	0.816	0.849	0.871	0.885
	Succ.	0.684	0.693	0.698	0.707	0.725
VOT2018	Acc.	0.4425	0.4596	0.4719	0.5064	0.5148
	Fail.	17.5861	17.5374	17.1954	16.6824	15.2489
	Overlap	0.5027	0.5278	0.5634	0.5721	0.5865
	EAO	0.3375	0.3443	0.3515	0.3862	0.3957
	FPS	87	78	69	62	58

TABLE V. EXPERIMENTAL RESULTS OF DIFFERENT HISTORICAL FRAMES ON BENCHMARKS.

	1 frame	2 frames	4 frames	5 frames	7 frames	8 frames	9 frames	10 frames
Prec.	0.892	0.886	0.881	0.873	0.864	0.851	0.847	0.837
Succ.	0.766	0.759	0.752	0.741	0.736	0.733	0.729	0.725
EAO	0.3954	0.3924	0.3885	0.3826	0.3778	0.3714	0.3689	0.3613
FPS	71	68	64	59	52	47	44	42

unconstrained environments. Background Clutter and Occlusion are the two scenarios with the most significant improvement. The precision is 1.87% and 9.17% higher than SiamRPN respectively. It was 24.70% higher than SRDCF and 17.19% higher than SiamFC, which ranked third. As can be seen from Figure. 12. and Figure. 13., our proposed algorithm secures the top position in the ranking across eleven unconstrained environments, and the improvement is most obvious in Scale variation and Out of view scenarios, with a succession 0.79% and 3.42% higher than that of SiamRPN ranked second, respectively. It was 2.56% and 15.26% higher than SiamFC, which ranked third.

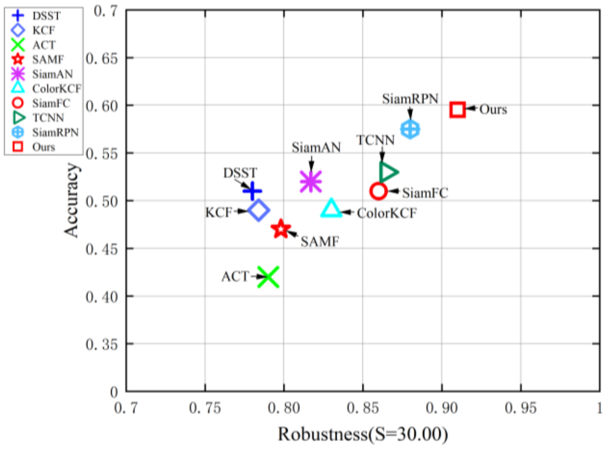


Fig. 14. The Robustness-Accuracy Ranking of Ten Trackers on VOT2018. The Closer The Location is to The Upper Right Corner, The Better The Performance.

D. VOT2018 Experiment

VOT2018 uses robustness, overlap, and Expected Average Overlap (EAO) to evaluate the tracker's performance.

1) Robustness: Robustness is assessed by measuring the count of unsuccessful frames that were tracked, hence, a lower value indicates improved tracker robustness.

2)EAO: The tracker's robustness and accuracy are assessed, with superior performance indicated by higher values.

On VOT2018, the proposed algorithm was tested and compared with KCF, DSST [26], SAMF, ACT [23], ColorKCF [27], TCNN [28], SiamAN [23], SiamFC and SiamRPN. Figure. 14. is a comparison chart of robustness and accuracy in the VOT2018 dataset. The proximity of a tracker's performance to the chart's upper-right corner indicates its superiority in both aspects. The figure suggests that the proposed algorithm exhibits elevated performance in both robustness and accuracy.

TABLE III details the tracking performance of the 10 trackers in terms of EAO, Accuracy, Failures (Robustness), Overlap and FPS. Based on the data presented in the table, it can be deduced that the algorithm introduced in this research attains the highest performance across the four metrics of EAO, Accuracy, Failures (Robustness) and Overlap. Among them, EAO was 14.46% higher than the second-ranked SiamRPN, Accuracy was 9.39% higher than the second-ranked TCNN, Failures was 18.94% higher than the second-ranked TCNN, Overlap was 2.27% higher than the second-ranked SiamRPN.

E. Ablation Study

1) Different Modules: A series of ablation experiments were devised to assess the impact of different functional modules on tracking performance. The results of these experiments are presented in TABLE IV for further analysis. The term "Without any modules" signifies that the sole mechanism employed for target tracking is the ResNeSt backbone network, and "+" means that relevant functional modules are added on the basis of ResNeSt.

TABLE IV Results show that using ResNeSt for target

tracking within the network enhances the tracker's performance. Because of the distinct characteristics exhibited by various layers, replacing traditional deep convolutional networks with dynamic convolutional networks, the performance of the network can be improved by adding attention mechanism on each feature layer of the network and integrating shallow and deep features. Nevertheless, as the network's layer count increases and more modules are incorporated, the amount of network computation increases, resulting in a significant decrease in tracking speed. The module performance variation showed the same trend on both data sets.

2) *Historical Frames:* To further assess the impact of historical frames on tracking outcomes, we selected different history frames to conduct relevant ablation experiments, as shown in TABLE V. Upon reviewing the table, we can discern that, as a greater number of image frames are utilized, the tracking performance of the network gradually diminishes, and the tracker's speed decreases as well. This occurrence can be ascribed to the existence of preceding motion data within the archival frames, wherein the images have been in motion, but the recorded data isn't entirely precise and contains some errors. The continued accumulation of historical frames does not yield improved target tracking performance but, rather, results in a decline in tracking performance.

V. Conclusion

In traditional deep convolutional neural networks, convolution nuclei of the same feature layer have the same weight when generating feature maps. When the number of convolutional layers and convolutional channels is too large, the target tracking performance deteriorates due to the small computation. Furthermore, disregarding the impact of shallow features on tracking performance. This neglect of shallow features contributes to the tracker's diminished robustness when dealing with similar targets. We propose a siamese network based on dynamic convolution and shallow and deep information fusion is proposed. Firstly, the traditional AlexNet backbone network has been substituted with ResNeSt, with the initial three layers of the network being replaced by dynamic convolution, while the convolutional kernels in the final two layers remain unchanged. The weights of feature points within the representations of the initial three network layers are dynamically fine-tuned, facilitating the expeditious determination of the target object's location within the image. Then, channels and spatial attention mechanisms are incorporated into each convolutional layer to enhance the likeness and precision of target feature extraction. Secondly, the convolution results from the third and fifth layers of both network branches are merged following the convolution process. This merger yields a pair of complementary fraction graphs, enhancing the tracker's discriminative ability at a sub-pixel level for target-background differentiation. Finally, the proposed algorithm, along with several other state-of-the-art algorithms, is evaluated using OTB2015 and VOT2018 datasets. Experimental results indicate that the proposed algorithm outperforms in terms of tracking speed, accuracy and robustness, and achieves real-time tracking effect.

References

[1] Y. Zeng and R. Zhang, "Energy-Efficient UAV Communication With Trajectory Optimization," in IEEE Transactions on Wireless Communications, vol. 16, no. 6, pp. 3747-3760, June 2017.

[2] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction." Advances in physiological computing, Springer, London, 2014, pp. 39-65.

[3] D. Schneider, A. Otte, A. S. Kublin, A. Martschenko, P. O. Kristensson and E. Ofek, et al., "Accuracy of Commodity Finger Tracking Systems for Virtual Reality Head-Mounted Displays," 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 2020, pp. 804-805.

[4] M. Kumon, Y. Udo, H. Michihira, M. Nagata, I. Mizumoto and Z. Iwai, "Autopilot System for Kiteplane," in IEEE/ASME Transactions on Mechatronics, vol. 11, no. 5, pp. 615-624, Oct. 2006.

[5] X. Wang, M. Kan, S Shan and X. Chen, "Fully learnable group convolution for acceleration of deep neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9049-9058.

[6] C. S. Vorugunti, V. Pulabaigari, R. K. S. S. Gorthi and P. Mukherjee, "OSVFuseNet: Online Signature Verification by feature fusion and depth-wise separable convolution based deep learning." Neurocomputing, vol. 409, no. 7, pp. 157-172, Oct. 2020.

[7] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan and Z. Liu, "Dynamic convolution: Attention over convolution kernels." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11030-11039.

[8] X. Ding, Y. Guo, G. Ding and J. Han, "Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks." Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1911-1920.

[9] B. Yang, G. Bender, Q. V. Le and J. Ngiam, "Condconv: Conditionally parameterized convolutions for efficient inference." Advances in Neural Information Processing Systems, 2019, pp. 32.

[10] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin and Z. Zhang, et al., “Resnest: Split-attention networks,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Jun. 2022, pp. 2736-2746.

[11] J. F. Henriques, R. Caseiro, P. Martins and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 3, pp. 583-596, 1 March 2015.

[12] D. S. Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, "Visual object tracking using adaptive correlation filters," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, pp. 2544-2550.

[13] R. Tao, E. Gavves and A. W. Smeulders, "Siamese instance search for tracking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1420-1429.

[14] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking." Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14. Springer International Publishing, 2016.

[15] B. Li, J. Yan, W. Wu, Z. Zhu and X. Hu, "High performance visual tracking with siamese region proposal network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8971-8980.

[16] Q. Wang, L. Zhang, L. Bertinetto, W. Hu and P. H. S. Torr, "Fast online object tracking and segmentation: A unifying approach." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1328-1338.

[17] S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1492-1500.

[18] Z. Zhang and H. Peng, "Deeper and wider siamese networks for real-time visual tracking." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4591-4600.

[19] S. Woo, J. Park, J. Lee and I. S. Kweon, "Cbam: Convolutional block attention module." Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3-19.

[20] L. Huang, X. Zhao and K. Huang, "GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 5, pp. 1562-1577, 1 May 2021.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh and S. Ma, et al., "Imagenet large scale visual recognition challenge." International journal of computer vision, vol. 115, pp. 211-252, Apr. 2015.

[22] Y. Wu, J. Lim, M. H. Yang. “Online object tracking: A benchmark,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2013, pp. 2411-2418

[23] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder and L. C. Zajc, et al., "The sixth visual object tracking vot2018 challenge results." Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0-0.

[24] Y. Li and J. Zhu. "A scale adaptive kernel correlation filter tracker with feature integration." Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13. Springer International Publishing, 2015.

[25] D. Martin, H. Gustav, S. K. Fahad and F. Michael, "Learning spatially regularized correlation filters for visual tracking." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4310-4318.

[26] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, ‘‘Accurate scale estimation for robust visual tracking,’’ British Machine Vision Conference, 2014, pp. 1108–1117.

[27] P. Senna, I. N. Drummond and G. S. Bastos, "Real-Time Ensemble-Based Tracker with Kalman Filter," 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Niteroi, Brazil, Oct. 2017, pp. 338-344.

[28] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang and T. Xiao, et al., "T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896-2907, Oct. 2018.