Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 6731717, 19 pages
https://doi.org/10.1155/2023/6731717

Research Article

Robust Visual Object Tracking Based on Feature Channel Weighting and Game Theory

Sugang Ma,[1] **Bo Zhao**,[1] **Zhiqiang Hou**,[1] **Wangsheng Yu**,[2] **Lei Pu**,[3] **and Lei Zhang**[4]

[1]*School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China*
[2]*School of Information and Navigation, Air Force Engineering University, Xi'an 710077, China*
[3]*School of Operational Support, Rocket Force Engineering University, Xi'an 710025, China*
[4]*School of Automation, Northwestern Polytechnical University, Xi'an 710129, China*

Correspondence should be addressed to Bo Zhao; jw133zz@163.com

Received 17 November 2022; Revised 17 March 2023; Accepted 20 July 2023; Published 31 July 2023

Academic Editor: Paolo Gastaldo

Although the discriminative correlation filter- (DCF)-based tracker improves tracking performance, some object representation issues can still be further optimized. On the one hand, the DCF tracker's deep convolutional features contain many noisy channels, and assigning the same weights to multiple channels cannot distinguish the importance of different channels. On the other hand, a simple weighted fusion approach cannot fully utilize the benefits of different feature types. We propose a visual object tracking algorithm based on adaptive channel weighting and feature game fusion to solve these problems. In this study, an adaptive channel weighting strategy is designed to assign suitable weights to each channel based on the average energy ratio of the target and background regions in the feature channels and prune the channels with low weights to improve feature robustness and reduce computational complexity. Simultaneously, the game theory concept is introduced in the multifeature fusion. The handcrafted features are combined with shallow and deep convolutional features according to feature complementarity. Then, the two combined features are seen as two sides of the game, continuously gamed during the tracking process to generate a feature model with a higher representation capacity. Extensive experiments are conducted on four mainstream visual tracking benchmark datasets, including OTB2015, VOT2018, LaSOT, and UAV123. The experimental results show that the proposed algorithm performs outstandingly compared to the state-of-the-art trackers.

1. Introduction

Visual object tracking is a fundamental topic in computer vision [1–4], with various real-world scenario applications such as intelligent surveillance [5], human-computer interaction [6], robotics [7], and intelligent transportation systems [8]. Due to illumination variations, target occlusion, motion blur, and other factors in realistic settings, fast and accurate robust tracking remains a difficult challenge.

DCF-based object tracking algorithms have been one of the most popular object tracking frameworks in recent years due to their good balance of tracking performance and computational efficiency. Although the performance is not as good as the current deep learning-based trackers, it may be better suited for resource-constrained areas regarding computational efficiency and tracking speed, such as UAVs [9–11]. The typical DCF algorithm formulates the tracking task as a ridge regression problem. It converts spatial domain matrix operations to frequency domain element multiplication, which decreases computing complexity and considerably improves tracking speed. Most early DCF trackers used handcrafted features such as grayscale features, histogram of oriented gradient (HOG), and color name (CN) to build the target appearance model [12–14]. For instance, Henriques et al. [12] replaced single-channel grayscale features with multichannel HOG features. Danelljan et al. [15] suggested CN features by associating the RGB color space with linguistic color labels. The scale adaptive kernel correlation filter (SAMF) [13] tracker combines grayscale, HOG, and CN features to create new features for object tracking.

With the development of deep learning, most of the current advanced DCF trackers use deep features [16–18]

although deep DCF trackers still have the following two issues. First, the multichannel deep features include a high number of noisy channels that are unrelated to the target information. Because current CNN models for visual tracking are pretrained with a large set of image samples, each feature channel in the object appearance model reflects different graphical class information. Still, visual tracking tasks require more attention to feature channels in a specific target region. Second, the high-dimensional deep features bring much computational complexity when involved in filter learning, which affects the tracking speed. The majority of current trackers for improving deep feature dimensions use a basic channel pruning strategy; for example, Che et al. [19] suggested a feature channel pruning algorithm to evaluate the validity of channels by calculating the feature average energy ratio between the target region and the search region in the initial frame, pruning the invalid channels to achieve better tracking performances. Ma et al. [17] utilized adaptive feature channel selection to obtain more robust tracking by calculating the energy relationship between the background and foreground in the feature channels. Nevertheless, these algorithms assign the same weight to the selected channels, which do not reflect the importance of effective channels.

It is noticed that various types of features focus on different visual attributes; for example, CN features reflect the color attributes of the image, HOG features describe the oriented gradient density distribution, and likewise shallow features in convolutional features focus on the edge contour information, while deep features contain better semantic information [20–22]. Therefore, a better fusion strategy can obtain features with greater representation capacity. In order to increase tracking accuracy, advanced correlation filter trackers such as ECO [18], ASRCF [21], and CFWCR [23] use multifeature fixed-weighted fusion. However, different features do not exhibit the same robustness in complex scenarios, and the simple weighted fusion mechanism cannot fully utilize the complementing characteristics of various features. Jin et al. [24] proposed a game theory-based object tracking algorithm that incorporates the game theory concept into the feature fusion mechanism within the mean shift tracking framework and treats color and texture features as two sides of the game that are adaptively fused to maximize gain and accomplish quick and effective tracking. However, this algorithm solely uses handcrafted features and ignores the complementarity property between multiple features. Liu et al. [25] achieved adaptive feature fusion by assigning appropriate weights to handcrafted and deep features using response map peak-to-sidelobe ratio (PSR) and smooth constraint. Nevertheless, only one fusion could not fully utilize each feature's benefits. Xia et al. [26] employed an adaptive joint weight method to combine color histograms effectively and HOG features to better cope with distortion and occlusion. However, the target could not be adequately modeled using simply handcrafted features.

To address the abovementioned issues, we propose a correlation filter tracking algorithm based on adaptive feature channel weighting and game theory feature fusion. On the one hand, to better reflect the importance of different channels, we consider the mapping of the target's bounding box on the feature map as the foreground region and the rest as the background region and we use the average energy ratio of the foreground and background regions as the weight to measure the importance of the feature channels. Meanwhile, we introduce the idea of channel pruning, define a weight threshold, remove feature channels below the threshold, and weight the effective channels. On the other hand, a feature fusion strategy based on game theory is proposed to achieve adaptive fusion between different types of features. First, two kinds of feature combinations are constructed according to the characteristics of different features: one is HOG, CN, and shallow features and the other is HOG, CN, and deep features, and the response maps generated by these two multichannel feature combinations are treated as game objects through continuous game iteration to achieve adaptive fusion and finally obtain more feature representation.

The main contributions of this paper are as follows:

(i) To better reflect the importance of different channels and reduce the influence of noisy channels on tracking performance, an adaptive feature channel weighting method is proposed in this study by combining the ideas of channel weighting and channel pruning. We calculated each channel's weight score by calculating the average energy ratio of its object and background regions, defining a weight score threshold, pruning the channels below the threshold, and performing weighted fusion on the remaining channels.

(ii) To give full play to the complementary advantages of different features and improve the fusion effect between multiple feature fusions, this study proposes a feature fusion method based on game theory. Based on the complementarity of two feature combinations, they are treated as two sides of the game and iterated continuously to obtain the optimal effect of fusion, which effectively increases the algorithm's tracking performance. By evaluating the response maps of various features, a new response map evaluation indicator, the deep-handcraft peak ratio (DHPR), is proposed, effectively expressing the differences between multiple features. This indicator is used to construct the gain function of both sides of the game.

(iii) Evaluated on four popular datasets, OTB2015, LaSOT, VOT2018, and UAV123, the extensive experimental results show that our tracker can better handle complex tracking environments such as illumination change, target rotation, fast motion, out-of-view, and deformation than recent state-of-the-art algorithms, which verified the accuracy and robustness of the proposed algorithm.

The rest of this paper is structured as follows: we give the previous works related to our work in Section 2, describe the proposed tracker in detail, including the classical DCF framework, the general framework of the proposed tracker, channel differentiation, adaptive channel weighting, and

feature game fusion in Section 3, and demonstrate the implementation details and experimental results in Section 4. Lastly, the conclusion and future works are drawn in Section 5.

2. Related Work

In this section, we briefly describe the related work of the proposed algorithm. The related work includes two aspects: tracking by DCFs and tracking by CNNs.

2.1. Tracking by DCFs. In recent years, discriminative correlation filter- (DCF)-based object tracking algorithms have demonstrated superior performance and speed advantages on many objects tracking benchmark datasets. The predecessor of the DCF-based tracker is the Minimum Output Sum of Squared Error (MOSSE) tracker proposed by Bolme et al. in 2010 [27], which uses grayscale images to train the filter and then correlates with the original image to obtain the target region. To address the issue of insufficient samples, Henriques et al. [28] proposed a circulant structure kernel algorithm named CSK, which employs the kernel cycle concept to train samples and the cycle matrix to solve the filter more effectively in the frequency domain. Based on the CSK tracker, KCF [12] substitutes the single grayscale features in CSK with multichannel HOG features to increase tracking accuracy while detecting the target location quickly. The sampling density achieved by utilizing the cyclic matrix, on the other hand, creates the boundary effect problem. To address these issues, the SRDCF [29] tracker uses the spatial regularization term to limit the response of the background region. The BACF [14] tracker effectively increases the number and quality of samples by cropping each sample. Based on SRDCF, the STRCF [30] tracker introduces temporal regularization to prevent model corruption and make the tracker more robust in the face of target occlusion. Aiming at the object scale variation problem, DSST [31] separates the scale estimation and position estimation, trains the scale filter separately, and uses the feature pyramid concept to find the optimal scale for the object scale change problem. SAMF [13] achieves object-adaptive scale adjustment by introducing the concept of a scale pool. IBCCF [32] separates the scale filters for the left, right, top, and bottom boundaries, allowing trackers to cope flexibly with the aspect ratio variation problem.

Recently, DCF trackers have shown significant advantages in the field of UAV tracking, and current research focuses on enhancing tracking performance by integrating new regularization terms in filter training. For example, ARCF [10] utilizes the previous frame's response to incorporate temporal cues into the tracking framework, which improves tracking performance even further. Based on this, IBRI [9] extends the historical time information into three frames and penalizes the interference region around the object, significantly improving the algorithm's accuracy and robustness. DRCF [33] uses saliency detection algorithms for spatial-dynamic regularization and a dual regularization strategy to achieve accurate tracking. AutoTrack [34] utilizes an online learning approach to achieve adaptive adjustment of spatial-temporal regularization hyperparameters.

2.2. Tracking by CNNs. With the development of deep learning, convolutional neural networks have demonstrated significant advantages in visual tracking, inspiring many studies. On the one hand, DCF trackers started using CNNs to extract objects' deep features. Ma et al. [20] substituted HOG features in KCF with shallow, middle, and deep features extracted from the VGG-19 network, considerably increasing tracking accuracy. To achieve reliable and quick tracking, STCCF [35] uses a channel distillation method to choose channels with high significance scores. ACSDCF [36] employs adaptive group elastic networks and introduces independent sparsity and temporal smoothness to the DCF framework, successfully optimizing the filter model and considerably reducing noisy channel interference. SCSTCF [37] proposes a spatial-channel selection and temporal regularization tracker that combines background information, spatial-channel constraint, and temporal consistency to obtain a more robust appearance model. Zhang et al. [16] alleviated the tracker shifting problem by introducing a distractor-aware map to reduce the weight of interference regions in multilevel features. On the other hand, the end-to-end deep learning tracking framework uses a well-designed deep network structure. From this aspect, SiamFC [38] first utilizes a Siamese network for object tracking, extracting object template features and searching area features for cross-correlation operations, leading to excellent real-time offline tracking. CFNet [39] extends the SiamFC network structure with CF layers, allowing it to be trained with fewer network layers while maintaining accuracy. Zhang et al. [40] developed the spatial attention extraction (SAE) block, which incorporates the template and search region features to generate multiscale spatial attention, efficiently separating the foreground and background. To accomplish precise localization, SiamOA [41] proposes an offset-aware tracking framework that accurately predicts the offset of the target's bounding box in the interval. Li et al. [42] designed a regression network to evaluate each channel's importance, effectively improving feature representation through a weighted fusion strategy.

3. The Proposed Tracker

In this section, we will first review the basic principles of traditional DCF methods and then describe the proposed feature adaptive channel weighting and feature game fusion visual object tracking algorithm.

3.1. Revisit of DCF Framework. The DCF tracking algorithm's central concept is to train the appropriate multichannel filter f in the objective sample set $\{(x_k, y_k)\}_{k=1}^{t}$, where each training sample $x_k = [x_k^1, x_k^2, \ldots, x_k^d]$ consists of a D-dimensional feature map of size $M \times N$, $f = [f^1, f^2, \ldots, f^d]$ represents the corresponding D-dimensional multichannel filter, and given a desired label function y_k

with the peak at the center of the target, the filter f is implemented by minimizing the following objective function:

$$\arg\min_{f} \left\| \sum_{d=1}^{D} x_k^d \otimes f^d - y_k \right\|^2 + \lambda \sum_{d=1}^{D} \left\| f^d \right\|^2, \tag{1}$$

where $\otimes$ represents the cyclic correlation operator and $\lambda\,(\lambda > 0)$ represents the weight of the regularization term. To improve the calculation efficiency, by utilizing the convolution property of the discrete Fourier transform (DFT), the closed solution of the filter on the d-th channel is given by the following equation:

$$\widehat{f}^d = \frac{\left(\widehat{x}_k^d\right)^* \odot \widehat{y}_k}{\sum_{d=1}^{D}\left(\widehat{x}_k^d\right)^* \odot \widehat{x}_k^d + \lambda}, \tag{2}$$

where $\odot$ denotes the element product, $\widehat{(\cdot)}$ denotes the discrete Fourier transform (e.g., $\widehat{f}^d = \mathscr{F}(f^d)$), and $(\widehat{x})^*$ is the conjugate of $\widehat{x}$. When the $(k+1)$-th frame arrives, the current frame's candidate region z (which is the same size as the sample x) is extracted, and the response R of region z is calculated using the trained filter model, the maximum response position serving as the target position:

$$R = \mathscr{F}^{-1}\left(\sum_{d=1}^{D} \widehat{f}^d \odot \widehat{z}_{k+1}^* \right), \tag{3}$$

where $\mathscr{F}^{-1}$ denotes the inverse Fourier transform.

When the model is updated, using the online update rule, the numerator $\widehat{n}_d$ and denominator $\widehat{m}_d$ of the filter f are updated as follows:

$$\begin{aligned} \widehat{n}_d^t &= (1-\eta)\widehat{n}_d^{t-1} + \eta \widehat{y} \odot \widehat{x}_d^{*t}, \\ \widehat{m}_d^t &= (1-\eta)\widehat{m}_d^{t-1} + \eta \sum_{i=1}^{D} \widehat{x}_i^{*t} \odot \widehat{x}_i^t, \end{aligned} \tag{4}$$

where t represents the index of the current frame and η represents the learning rate.

3.2. General Framework of the Algorithm. The general framework of the proposed algorithm is shown in Figure 1. First, for each new input image frame, the region of interest is clipped at the predicted target center location p_0 from the previous frame. In addition, handcrafted features are extracted, as well as the shallow and deep features extracted from the pretrained VGG-16 network model. A feature channel weighting strategy adaptively weights and fuses CNN features to produce new multichannel features. Next, the different multichannel features correlate with their associated filters to generate the response maps R_{HS} and R_{HD}. Then, using game theory concepts, R_{HS} and R_{HD} are regarded as two sides of the game, with continuous game iteration to fully integrate them, obtaining their expected central positions p_1 and p_2, respectively, and choosing whether to finish the game by judging the Euclidean distance between p_1 and p_2. Suppose their distance is less than the set threshold φ; they are regarded to have accomplished the optimal fusion, and the final predicted target center position is obtained directly by the fused response map. In contrast, they are deemed to have failed to attain the optimal equilibrium point, and the game is repeated with $(p_1 + p_2)/2$ as the new central position until the end condition is satisfied.

3.3. Channel Differentiation. Most leading DCF trackers use multichannel features to train correlation filters [12–14, 29]. With the advancement of deep learning, more and more researchers have begun to use deep features with a greater number of channels as feature extraction to increase tracking performance [16, 18, 21, 43]. However, computational complexity issues arise when directly using features taken from pretrained CNN network models improves tracking performance. Specifically, different deep feature channels do not exhibit the same tracking robustness during the tracking process. Some channels create larger activation values around the target region, generating a tremendous amount of interference information unrelated to the target while participating in filter training, causing a tracking shift. Focusing on channels with higher activation values in a given target region is more significant for improving feature representation.

To better show channel differences, we choose specific frames from the *MotorRolling*, *Soccer*, and *Human*4 sequences, extract features by using the conv5-2 layer of the VGG-16 net, and visualize them. As shown in Figure 2, most of the channels (e.g., channel 452 in *Human*4, channel 34 in *MotorRolling*, and channel 110 in *Soccer*) have higher activation values in the region around the target, and these feature channels generate much redundant information during filter training, causing a shift in the prediction center. Therefore, more emphasis should be placed on the channels with the highest energy in the target region (e.g., channel 228 in *Human*4, channel 506 in *MotorRolling*, and channel 26 in *Soccer*), and these channels should be utilized to improve tracking accuracy and robustness.

In this study, an adaptive channel weighting method is suggested to give appropriate weights to each channel based on its respective scores and discard channels with too small weights to improve feature quality.

3.4. Adaptive Channel Weighting. This paper proposes an adaptive channel weighting strategy that improves effective channels while suppressing interference channels. We use a metric to assess the impact of different channels on tracking performance and assign appropriate weights to them and then introduce the concept of channel pruning, which improves feature representation capability while significantly reducing the computational complexity problem caused by multichannel features.

To better demonstrate the variable importance of each channel in the tracking process, we calculate channel weights based on the feature maps' background-foreground average energy ratio (BFAER) [17]. Specifically, for the D-dimensional deep feature x with size $M \times N$ extracted from a layer of the CNN pretrained model, we define the feature energy value of position (m, n) in the feature map

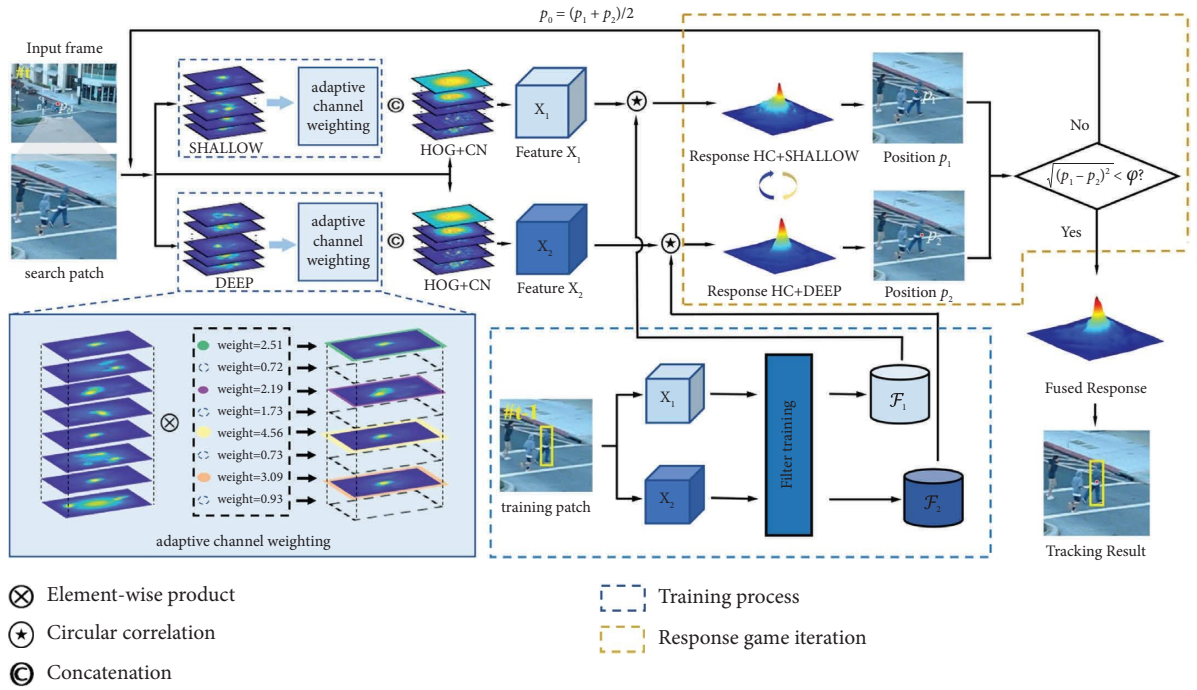


FIGURE 1: The overall framework of the proposed algorithm. Our adaptive channel weighting strategy (solid blue line) improves the effective channels in CNN features and combines them with handcrafted features to form different feature combinations. Their respective response maps are optimally fused by the feature game fusion strategy (yellow dashed line) to obtain the predicted object center.

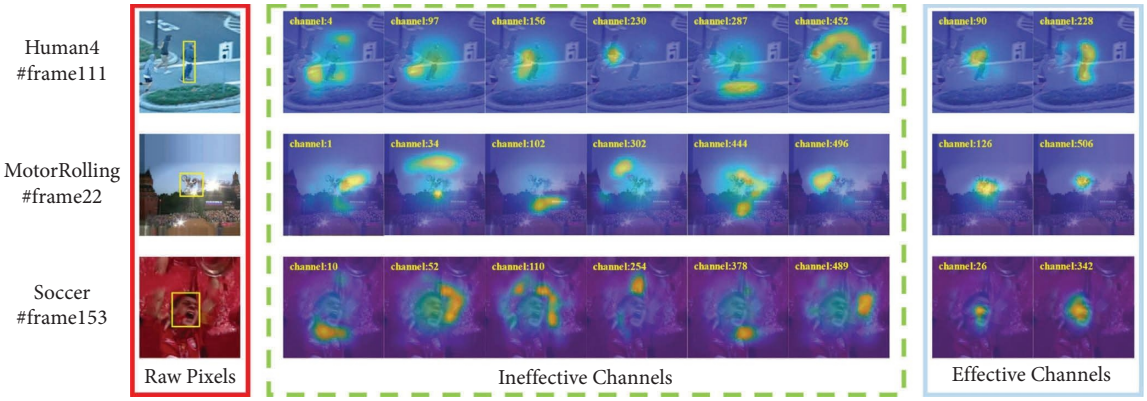


FIGURE 2: Visualization of different feature channels corresponding to selected frames of *Human*4, *MotorRolling*, and *Soccer*, where the red rectangle represents the original target search area, and the green dashed line and blue rectangle represent the activated pixels of the invalid and valid channels, respectively.

x^d ($d \in \{1, \ldots, D\}$) as $x^d(m, n)$. The following equation gives the overall average energy value on this feature x^d:

$$G\left(x^d\right) = \frac{\sum_m^M \sum_n^N x^d(m, n)}{M \times N}, \tag{5}$$

where $m = 1, 2, \ldots, M$ and $n = 1, 2, \ldots, N$. The average pixel value of a region is defined as its overall energy value, and a simple calculation can reflect the importance of different feature channels.

In order to effectively separate the foreground and background regions of the feature map, we first resize the feature map following bilinear interpolation, which equals the size of the original image patch. Following that, the foreground region is defined as the mapping of the target's bounding box on the feature map, and the rest is the background region, as illustrated in Figure 3. Denote the foreground region's center (i.e., the center of the target's bounding box) as (a, b) and the size of the foreground region as $I \times J$, where $I < M$ and $J < N$. The average energy value $G_f(x^d)$ of the foreground region on feature x^d is, therefore, defined as follows:

$$G_f\left(x^d\right) = \frac{\sum_w^W \sum_h^H x^d(w, h)}{I \times J}, \tag{6}$$

where $w = a - I/2$, $W = a + I/2$, $h = b - J/2$, and $H = b + J/2$. As a result, the BFAER score for the feature x^d on the d-th channel is defined as follows:

$$\mathrm{BFAER}\left(x^d\right) = \begin{cases} \dfrac{G_f\left(x^d\right)}{G\left(x^d\right) - G_f\left(x^d\right)}, & \text{if } \mathrm{BFAER}\left(x^d\right) > \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $G_f(x^d)$ and $G(x^d) - G_f(x^d)$ represent the feature energy values of the foreground and background regions, respectively, and τ is a predefined weight threshold. After that, we prune the channels with low BFAER scores according to the weight threshold τ, and the remaining channels are weighted and fused to form new features.

We find that the BFAER score is higher when the channel is more concerned with information about the target region and lower when there is much disturbing information. As a result, BFAER may visually represent a channel's tracking robustness. When the channel weight is less than the threshold, the channel is deemed invalid and contains much disturbing information, and a pruning operation is performed on it. When the channel weight exceeds the threshold, appropriate weights are adaptively assigned to the channels based on the relevant BFAER scores. Notably, the BFAER weight score is only calculated in the initial frame and the result is followed in subsequent frames. Following the above analysis, the proposed adaptive channel weighting strategy can better suppress invalid channel interference, enhance the robustness of effective channels, and reduce computational complexity while improving tracking accuracy.

3.5. Feature Fusion Based on Game Theory. Feature fusion improves feature representation by integrating features with different properties. In this paper, we implement an adaptive fusion of features while taking into account the complementary characteristics of multiple features, preweighted handcrafted (HOG + CN) feature responses with shallow (conv4-3 of VGG-16 net) and deep (conv5-2 of VGG-16 net) feature responses to form different combinations of feature responses. Then, they are adaptively fused at the decision level to obtain the best fusion via game theory concepts.

The spatial resolution of the VGG-16 network varies per layer, and deeper levels have a low-spatial resolution. Therefore, we provide deeper features with a higher weight to ensure we get all the information when features are preweighted. The combination of response maps for handcrafted features and conv4-3 layer deep features is given by the following equation:

$$R_1(x) = \mathscr{F}^{-1}\left(W_1 \times \sum_{a=1}^{D_h} \widehat{f}^a \odot \left(\widehat{x}^a\right)^* + W_2 \times \sum_{b=1}^{D_s} \widehat{f}^b \odot \left(\widehat{x}^b\right)^*\right), \tag{8}$$

where different features' convolutional responses are used, and D_h and D_s stand for the dimensionalities of the handcrafted features and conv4-3 layer deep features, respectively. The weights W_1 and W_2 represent the importance of the handcrafted features and conv4-3 layer deep features filters' responses. Similarly, the combination of handcrafted feature responses with conv5-2 layer deep features can be expressed as

$$R_2(x) = \mathscr{F}^{-1}\left(W_1 \times \sum_{a=1}^{D_h} \widehat{f}^a \odot \left(\widehat{x}^a\right)^* + W_3 \times \sum_{c=1}^{D_p} \widehat{f}^c \odot \left(\widehat{x}^c\right)^*\right), \tag{9}$$

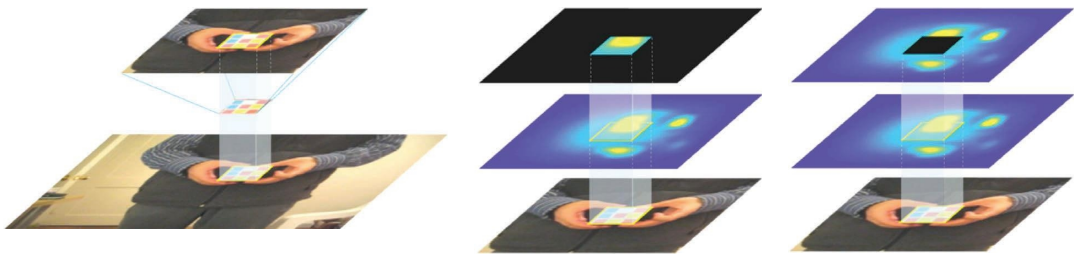
where D_p and W_3 denote the dimensionality of the conv5-2 layer deep features and the weights of their response maps.

We employ a game theory-based feature fusion strategy to achieve an optimal fusion of the two feature response combinations mentioned above. Game theory is a mathematical theory that studies the optimization strategies of various individuals with competitive characteristics by considering their predicted and actual behaviors. This work employs a classic game theory strategy known as Nash equilibrium. In Nash equilibrium, each participant's equilibrium strategy is designed to maximize their expected interests and does not change easily.

The above two different feature response combinations are treated as two participants in the game, and the set of participants is denoted as $P = \{1, 2\}$, where x_1 and x_2 are the handcrafted features combined with conv4-2 and conv5-3 layer deep features, respectively. $R_1(x_1)$ and $R_2(x_2)$ are the filter convolution responses. Each participant's revenue function is defined as follows:

$$\begin{cases} \mathscr{G}_1(x_1) = R_1(x_1), \\ \mathscr{G}_2(x_2) = R_2(x_2), \end{cases} \tag{10}$$

where $\mathscr{G}_1$ and $\mathscr{G}_2$ are the two participants' revenue functions. Considering the Nash equilibrium, we consider the game participants' self-interest and the opponent's interest to be common interests, giving in a win-win situation for both participants, and the revenue function can be extended as follows:



(a) (b) (c)

Figure 3: The calculation principle of the proposed weight formula BFAER: (a) The foreground and background regions obtained by the target's bounding box in the input frame, (b, c) The foreground and background region mappings on the feature map. The weights of the corresponding channels are calculated based on the average energy ratio of the foreground and background regions.

$$\begin{cases} \mathcal{G}_1(x_1) = R_1(x_1) + \omega_2 R_2(x_2), \\ \mathcal{G}_2(x_2) = R_2(x_2) + \omega_1 R_1(x_1), \end{cases} \tag{11}$$

where ω_1 and ω_2 denote the fusion factor of two response maps, and the fusion factor indicates the response map's quality.

Currently, the primary evaluation indicators for determining the quality of response maps in object tracking are the peak-to-sidelobe ratio (PSR) and the average peak-to-correlation energy (APCE). The main idea is to regard the response map with a higher peak and minor oscillation as a good-quality response map. However, the response maps of handcrafted features have many perturbations, which might result in excessive standard deviations after combining the response maps, resulting in low PSR and APCE scores and affecting the fusion results. Therefore, inspired by the above two indicators, we propose a new response map evaluation indicator, the deep-handcraft peak ratio (DHPR). It can balance the response map peaks and surrounding perturbations while effectively reflecting the differences between the convolutional features of conv5-2 and conv4-3 layers in the feature combinations, and DHPR is defined as follows:

$$\mathrm{DHPR} = \frac{\max(R_{\mathrm{total}}) - \mathrm{mean}(R_{\mathrm{HC}})}{\mathrm{mean}(R_{\mathrm{total}}) + \mathrm{mean}(R_{\mathrm{HC}})}, \tag{12}$$

where $\max(\cdot)$ is the peak value of the response map, $\max(\cdot)$ is the mean value of the response map, and R_{HC} and R_{total} represent, respectively, the responses of handcrafted features alone and handcrafted features combined with various deep features. Therefore, the fusion factors ω_1 and ω_2 can be calculated by the following equation:

$$\begin{cases} \omega_1 = \dfrac{\mathrm{DHPR}(R_1)}{\mathrm{DHPR}(R_1) + \mathrm{DHPR}(R_2)}, \\ \omega_2 = \dfrac{\mathrm{DHPR}(R_2)}{\mathrm{DHPR}(R_1) + \mathrm{DHPR}(R_2)}. \end{cases} \tag{13}$$

To verify the indicator's effectiveness, we compared it to PSR and APCE on sequences from OTB2015 [44] with different challenge attributes as shown in Figure 4. The overall test results on the OTB2015 dataset are given in Table 1, which shows that the proposed DHPR can outstandingly cope with tracking challenges such as illumination variations, fast motion, and object deformation and significantly outperforms the other two indicators in terms of overlap rate.

4. Experiments

4.1. Implementation Details and Evaluation Metrics. The proposed algorithm is implemented on the MATLAB 2018a platform, which runs on a *PC with an Intel Xeon Silver* 4216 CPU 2.10 GHz, 128 GB RAM, and an NVIDIA GTX 1080Ti GPU. The MatConvNet [45] toolbox is used to extract deep features from the pretrained CNN model.

The experimental parameters are as follows: the conv4-3 and conv5-2 layers in imagenet-vgg-verydeep-16 are used to extract the target's convolutional features, respectively, and the HOG and CN features are extracted in the same way as the tracker SRDCF [29]. In Section 3.4, we set the BFAER threshold for deep features to $\mu_1 = 1.75$ and the BFAER threshold for shallow features to $\mu_2 = 1.55$. We set the ratio of the fusion weights of equations (8) and (9) to $W_2/W_1 = 1.5$ and $W_3/W_1 = 2$ for the feature preweighted section. The feature game threshold $\varphi = 1$ was set for the feature game fusion section. The rest parameters are consistent with the tracker STRCF [30].

We evaluated the effectiveness of the proposed algorithm on four public tracking benchmarks, including OTB2015 [44], VOT2018 [46], UAV123 [47], and LaSOT [48]. For the OTB2015, LaSOT, and UAV123 benchmarks, we apply the one-pass evaluation (OPE) protocol and use the success rate and precision under this evaluation to quantify the algorithm's tracking performance. A frame is regarded as successful if the overlap between the algorithm's predicted bounding box and the ground-truth bounding box is larger than a given threshold. The area under the curve (AUC) is used in the success rate plot to demonstrate the algorithm's robustness. The overlap precision (OP) is the success rate score corresponding to an overlap rate threshold of 0.5. The precision plot represents the percentage of frames in which the center distance between the predicted bounding box and the ground truth is smaller than a given threshold, plotted from 0 to 50 pixels. Distance precision (DP) is the value corresponding to a distance pixel threshold of 20 and is used

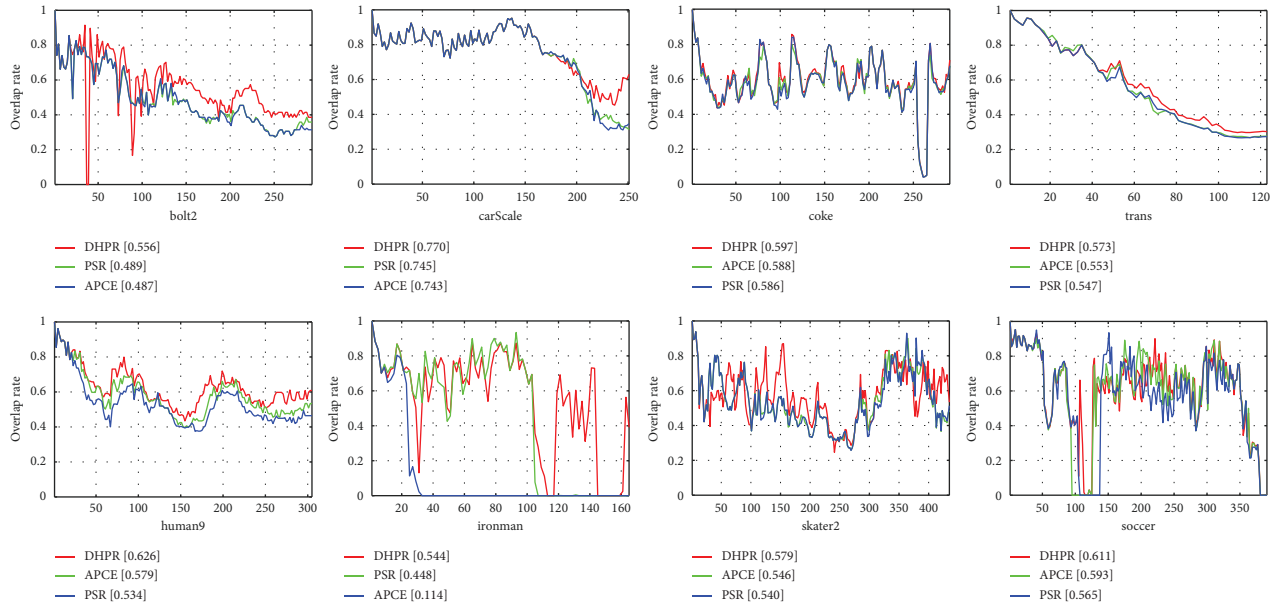


Figure 4: Comparison of overlap rate curves of three different evaluation indicators on partial sequences of the OTB2015 dataset. The horizontal coordinate represents the number of video frames, and the vertical coordinate represents the overlap rate. The video sequences are *Bolt*2, *CarScale*, *Coke*, *Trans*, *Human*9, *Ironman*, *Skater*2, and *Soccer*.

Table 1: The effect of three different evaluation indicators on tracking performance.

	DHPR	PSR	APCE
Success rate	**0.691**	0.686	0.681
Precision	**0.916**	0.911	0.909

The best results are shown in bold.

Input: The target center position p_{t-1} of the $(t-1)$-th frame; Filter model f_{t-1}.
Output: The predicted target center position p_t of the t-th frame; Filter model f_t.
Target tracking:
(1) Crop the search region according to the position p_{t-1}.
(2) Extract handcrafted features $[x_1^a]_{a=1}^{D_h}$ (HOG and CN features), shallow features $[x_2^b]_{b=1}^{D_m}$ (conv4-3 layer of VGG-16), and deep features $[x_3^c]_{c=1}^{D_p}$ (conv5-2 layer of VGG-16) of the search region.
(3) (7) is used to adaptively channel weight the convolutional features x_2 and x_3, with x_1, form new features combination X_1 and X_2.
(4) Using equations (8) and (9), the response maps R_1 and R_2 of feature combinations X_1 and X_2 are calculated, and the respective predicted central positions p_1 and p_2 are obtained.
(5) **if** $\sqrt{(p_1 - p_2)^2} > \varphi$ **then**
(6) Take $p_{t-1} = (p_1 + p_2)/2$ and return to 1.
(7) **else**
(8) The game ends and the final position p_t is obtained.
(9) **end if**
Filter learning:
(10) Get the image patch at position p_t and extract handcrafted features x_1, shallow features x_2, deep features x_3.
(11) The optimized deep features $[x_2^b]_{b=1}^{N}$ and $[x_3^c]_{c=1}^{M}$ (where $N < D_m$, $M < D_p$) are obtained using the adaptive channel weighting method in Section 3.4.
(12) The filter model f_t is obtained using equation (2)

Algorithm 1: The proposed tracking algorithm

for performance evaluation of precision plot. The average center distance between the algorithms predicted bounding box and the ground truth is defined as the center location error (CLE). For the VOT2018 benchmark, we used expected average overlap (EAO), accuracy, and robustness as evaluation metrics. Where accuracy is the average overlap

between the predicted bounding box and ground truth, robustness counts the number of tracking failures, and EAO considers both accuracy and robustness to reflect the tracker's overall performance.

4.2. Qualitative Analysis. As shown in Figure 5, we present the qualitative analysis results of our algorithm and various advanced algorithms (i.e., DaHCF [16], ECO [18], C-COT [49], SRECF [50], BACF [14], PrDiMP-18 [51], and SiamFC [38]) on video sequences with varying challenge properties. When there is severe occlusion or the object is out-of-view (e.g., *Bird*1 and *Box*), the proposed algorithm can track it accurately, whereas none of the other algorithms can. For sequences with severe background interference (e.g., *Soccer* and *Ironman*), the algorithm can enhance the channels with effective expression capability and suppress the channels with interference information through an adaptive channel weighting strategy to achieve accurate target localization. For video sequences with illumination variations and motion blur in the scene (e.g., *MotorRolling* and *Skiing*), our algorithm fuses different features of various properties by gaming, maximizing the complementarity between different features and efficiently mitigating obstacles such as illumination variations. Furthermore, for video sequences with object rotation and scale variation (e.g., *MotorRolling*, *Diving*, and *Clifbar*), our algorithm can perform stable tracking and accurate scale estimate at the same time. Overall, our algorithm can maintain excellent robustness and tracking accuracy even in the face of these interference difficulties.

4.3. Quantitative Analysis

4.3.1. OTB2015 Dataset. The OTB2015 [44] dataset is one of the most authoritative benchmarks in the field of object tracking, containing 100 completely labeled video sequences with 11 different challenge attributes, including illumination variation (IV), scale variation (SV), occlusion (OCC), deformation (DEF), motion blur (MB), fast motion (FM), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), background clutters (BC), and low resolution (LR). We compared the proposed algorithms to 16 top-performing deep feature-based DCF trackers (DaHCF [16], AFCSCF [17], MEGTCF [52], DeepSTRCF [30], ECO [18], CFWCR [39], and C-COT [49]), handcrafted feature-based DCF trackers (SRECF [50], DRCF [33], AutoTrack [34], BACF [14], and SAMF [13]), and deep learning-based trackers (PrDiMP-18 [51], GradNet [53], DaSiamRPN [54], and SiamFC [38]). Figure 6 shows the proposed algorithm's tracking results compared to 16 other algorithms on the OTB2015 dataset. As illustrated in the figure, the proposed algorithm has an AUC score of 69.1% and a DP score of 91.6%, which is optimal in terms of success rate and precision and has a relative gain of 0.6% in terms of precision over the second-place tracker (ECO). In addition, our approach beats the trackers AFCSCF, MEGTCF, and DeepSTRCF, which also use deep features, by 1.0%/0.1%, 1.3%/0.2%, and 1.6%/3.6%, respectively, thanks to our proposed adaptive channel weighting strategy. Furthermore, we consider the complementary qualities of features and apply the game theory approach to accomplish feature adaptive fusion, which more effectively exploits the advantages of deep features and handcrafted features. As a consequence, the suggested algorithm outperforms handcrafted feature-based algorithms BACF and SRECF by 7.0%/9.4% and 9.3%/10.9%, respectively.

We also compare the proposed algorithm with other algorithms in further detail, using CLE, OP, and speed (fps) as metrics. As demonstrated in Table 2, the proposed algorithm obtains an 85.8% mean OP, the most outstanding performance among other sophisticated trackers. It also achieves 10.8 pixels in the mean CLE, 0.3/4 pixels less than the CLE of the second and third places (MEGTCF and ECO), attaining the least faults. As shown in Table 2, the algorithm in this article has an average computing speed of 3.0 fps, which is inferior to deep learning-based trackers (SiamRPN [55] and GradNet) and handcrafted feature-based DCF trackers (SRECF and ECO-HC [18]). Because our tracker constructs feature combinations using two deep features and iterates through a continuous game to find the optimal fusion between different types of features. It is worth noticing that our tracker performs well among deep feature-based DCF algorithms, outperforming DaHCF, MCPF [56], and ECO. In general, our tracker performs better than the other 13 advanced trackers.

To demonstrate the tracking performance of each tracker under varied challenge attributes, we display the success rate and precision plots of each tracker under 11 different attributes, as shown in Figures 7 and 8. To more intuitively express the differences between trackers, we compared the ten trackers with the best success rates on the OTB2015 dataset. Figure 9 shows that the suggested algorithm performs optimally in both fast motion and motion blur video sequences because the proposed adaptive channel weighting strategy significantly suppresses the effect of noisy channels and achieves steady tracking in the face of rapidly moving objects. At the same time, our tracker still performs well in the face of object rotation, scale changes, and object out-of-view challenges. Notably, our tracker outperforms in low-resolution attributes, thanks to the aforementioned feature game fusion strategy, which fully advantages handcrafted features while learning high-level semantic information during filter training. Overall, our tracker performs excellently in a complex tracking environment.

4.3.2. LaSOT Dataset. On the LaSOT [48] dataset, we compared the proposed tracker to 35 trackers to better illustrate its performance. As a large-scale, long-term tracking dataset, the LaSOT dataset has 1400 video sequences. Each video has an average of 2512 frames, is classified into 70 categories, and contains different challenge attributes. We only analyze the proposed tracker on a test set of 280 video sequences using the OPE strategy, again using AUC and DP scores to reflect the algorithm's performance. Figure 10 shows the test results of several trackers on the LaSOT dataset, and as shown, our tracker's AUC and DP scores of

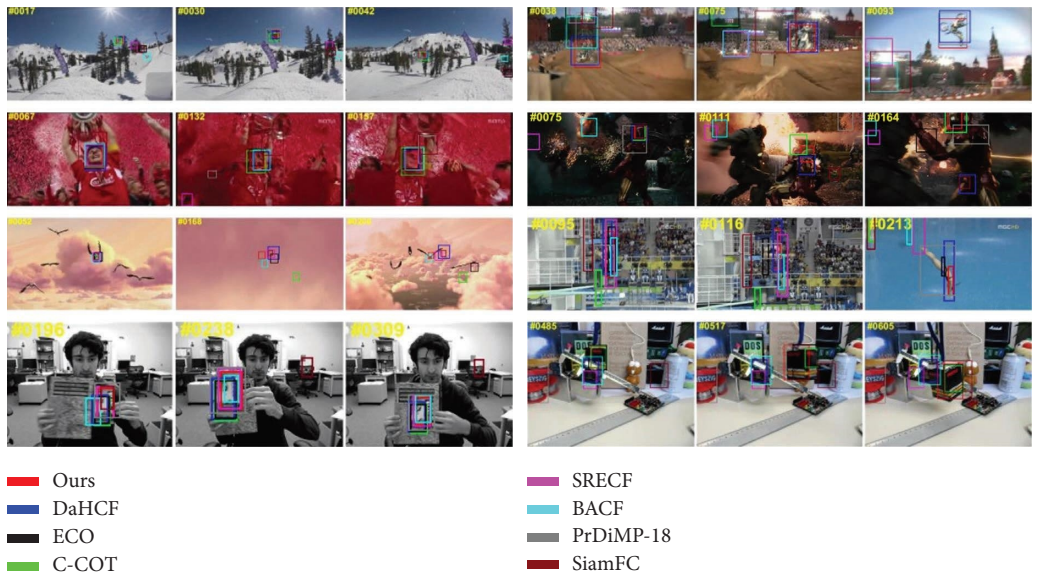


Figure 5: Qualitative analysis of the proposed algorithm with DaHCF [16], ECO [18], C-COT [49], SRECF [50], BACF [14], PrDiMP-18 [51], and SiamFC [38] on ten challenge sequences from the OTB2015 dataset. From top to bottom and left to right, these sequences are *Skiing*, *Soccer*, *Bird*1, *Clifbar*, *MotorRolling*, *Ironman*, *Diving*, and *Box*.

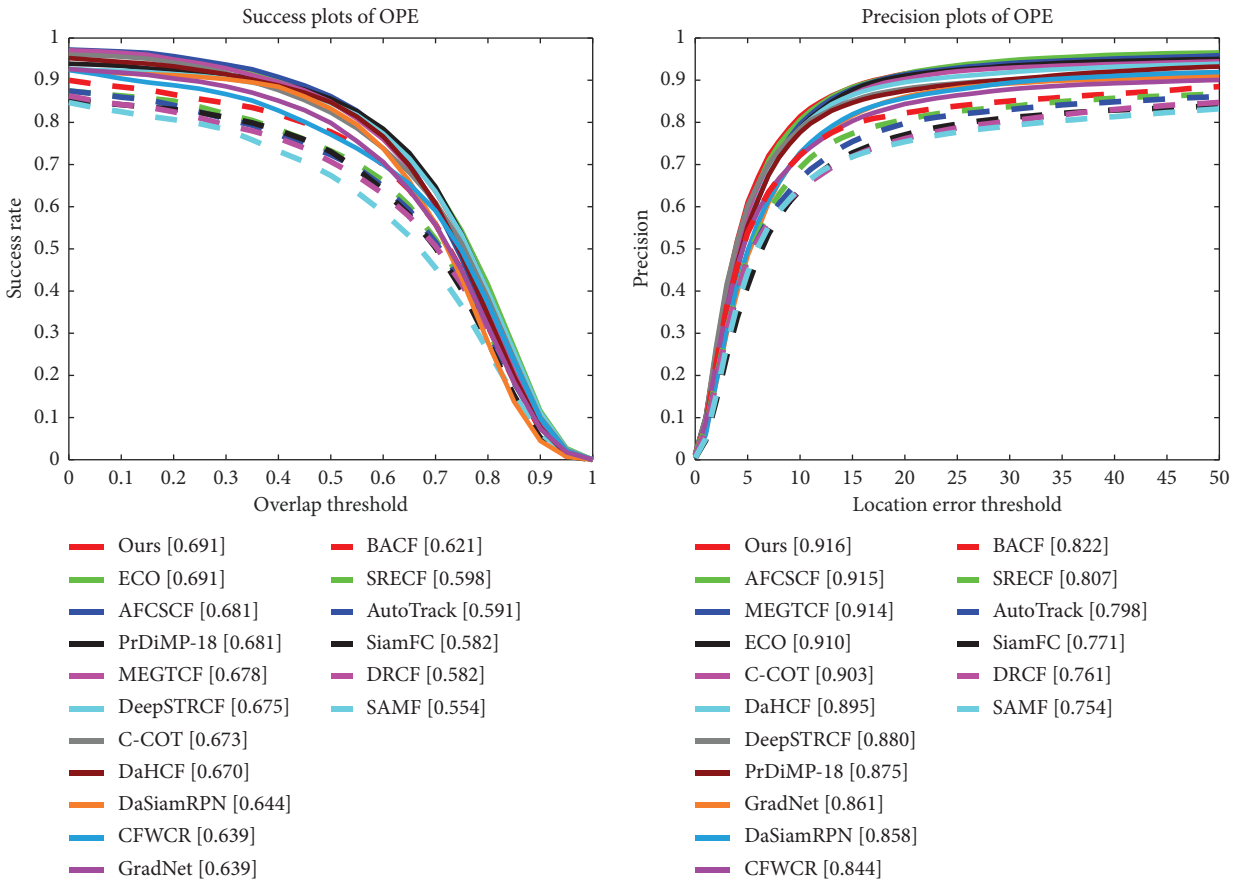


Figure 6: Comparison of success rate and precision on the OTB2015 dataset.

36.0%/35.6%, ranking seventh among all algorithms and exceeding most state-of-the-art trackers. Remarkably, the top six algorithms (LTMU [57], GlobalTrack [58], SiamRPN++ [59], SPLT [60], MDNet [61], and VITAL [62]) are all deep learning-based trackers. The main reason is that the end-to-end tracking framework employs additional

Table 2: Performance of the proposed algorithm compared with other algorithms on the OTB2015 dataset.

	SiamRPN	GradNet	ATOM	PrDiMP-18	ECO-HC	STRCF	AutoTrack
Mean OP (%)	81.6	79.9	82.2	**85.6**	78.5	80.0	72.3
Mean CLE	19.6	18.7	16.4	16.8	22.7	20.0	32.0
Speed (fps)	**34.2**	32.8	11.5	16.2	24.6	15.8	22.0
	SRECF	ECO	MCPF	DeepSTRCF	DaHCF	MEGTCF	Ours
Mean OP (%)	73.2	84.9	78.1	84.6	84.8	84.9	*85.8*
Mean CLE	32.8	14.8	20.9	17.8	14.9	**11.1**	*10.8*
Speed (fps)	*37.9*	1.0	0.5	4.0	0.02	10.8	3.0

The best three results are highlighted in italic, bold, and underline.

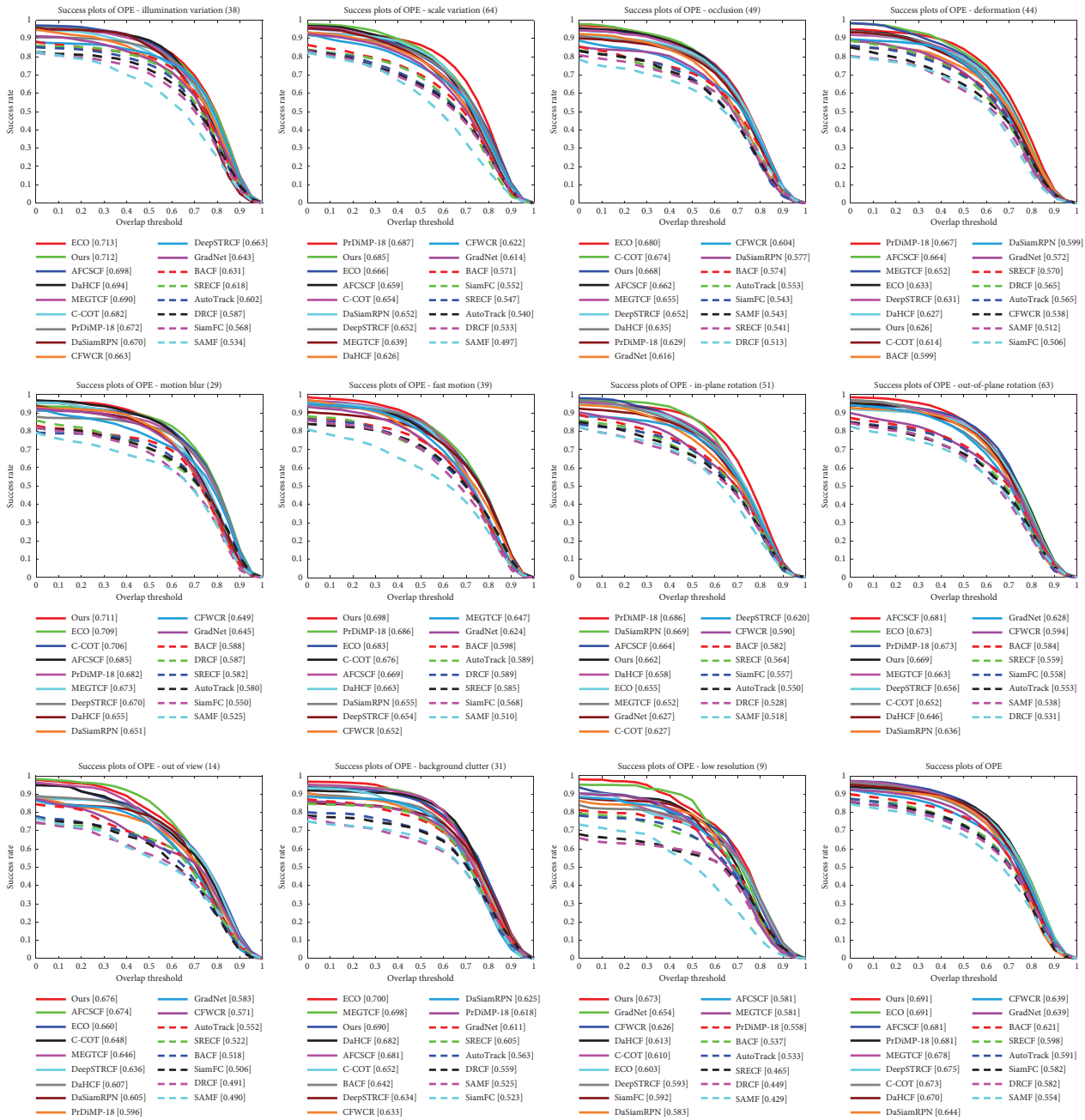


Figure 7: Success rate plot of each tracker under 11 challenge attributes on the OTB2015 dataset.

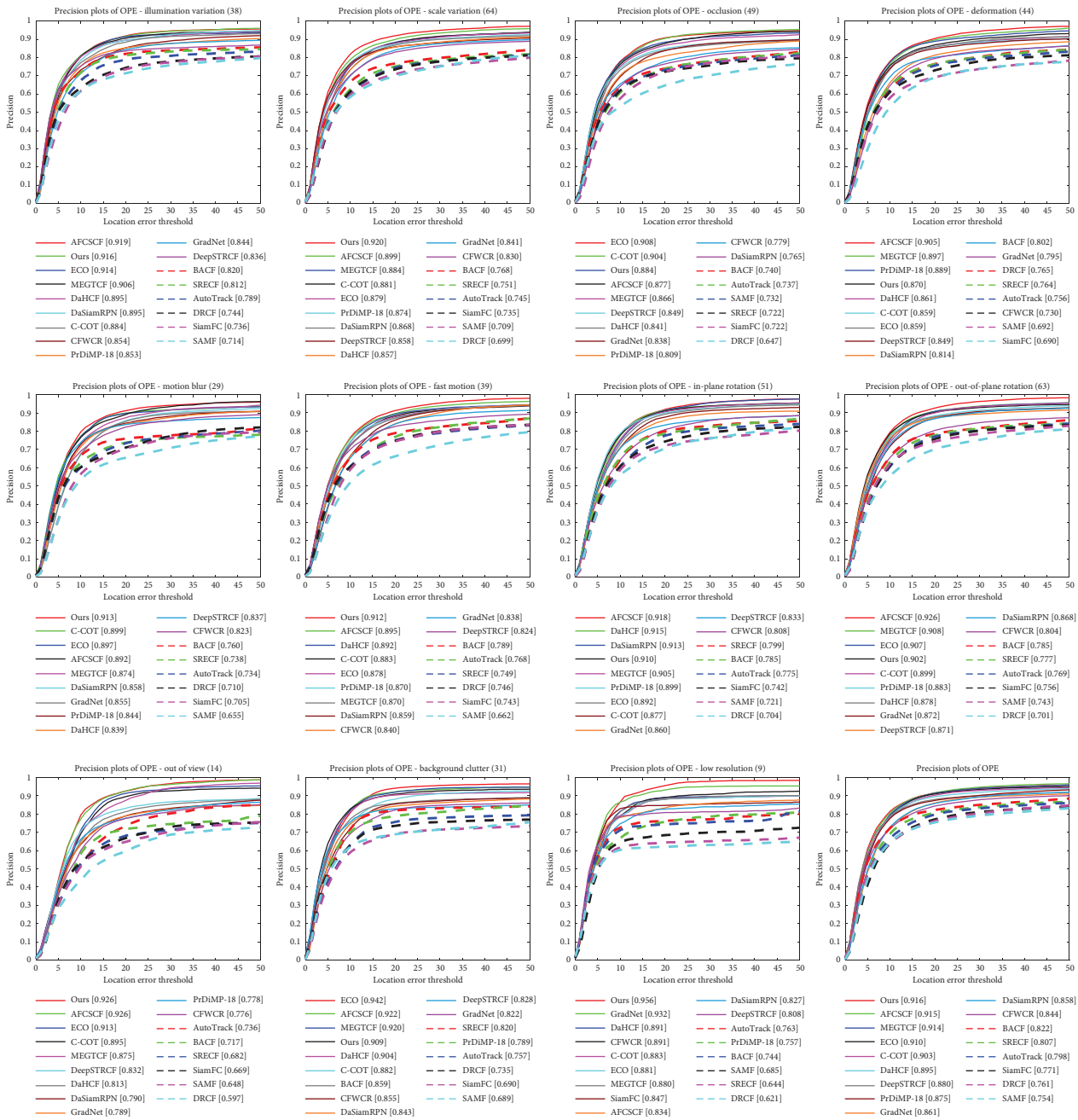


Figure 8: Precision plot of each tracker under 11 challenge attributes on the OTB2015 dataset.

strategies to cope with target scale and aspect ratio variations. In contrast, the DCF framework uses a constant aspect ratio. In addition, among many DCF-based trackers, our tracker performs the best, with a relative gain of 1.6%/2.5% over the second-ranked ASRCF [21], verifying the effectiveness of the proposed algorithm. In conclusion, the proposed algorithm has a better competitive advantage over many other advanced tracking algorithms.

4.3.3. VOT2018 Dataset. We also evaluated our tracker on the VOT2018 [46] dataset, which includes 60 video sequences with five different challenge attributes: camera motion, illumination change, object size change, object motion change, and occlusion. To better use the dataset, VOT adds a reinitializing attribute that allows the tracker to reinitialize five frames after detecting a failure track. This section compares the proposed tracker with eight state-of-the-art trackers on the VOT2018 dataset, including KCF [12], SRDCF [29], Staple [63], SiamFC [38], UpdateNet [64], DCFNet [65], CSRDCF [66], and C-COT [49]. To reflect the tracker's performance, we utilize three evaluation metrics: EAO, accuracy, and robustness, and the experimental results are displayed in Table 3 and Figure 11. Our tracker takes first place in both EAO and robustness, with a relative increase of 0.7%/1.4% over the winner of VOT2016 and C-COT. For

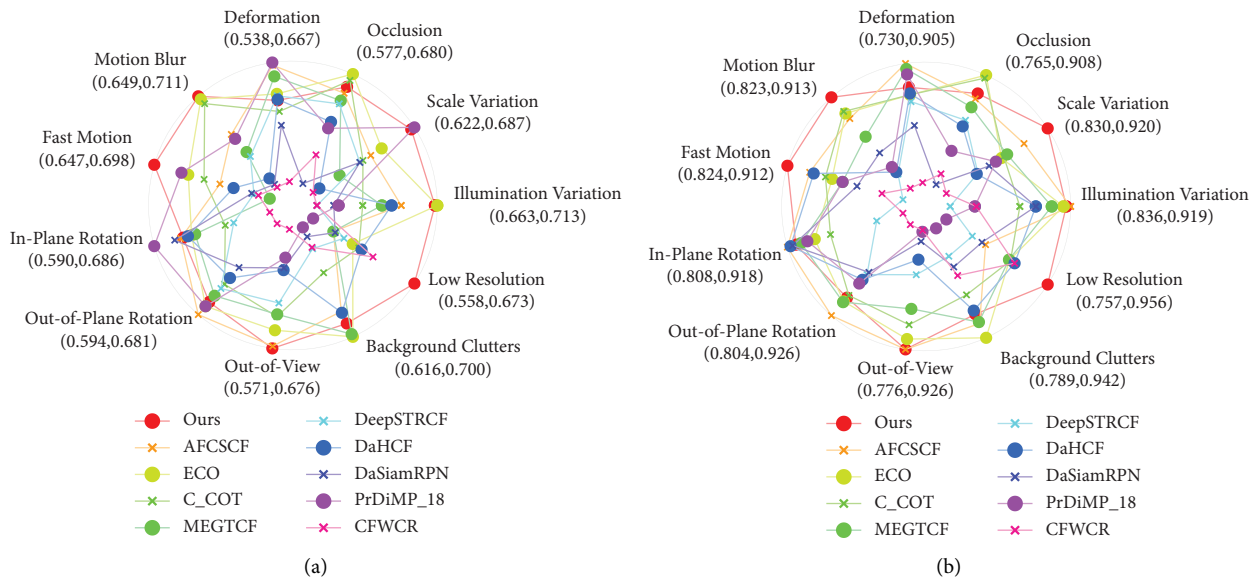


Figure 9: AUC (a) and DP (b) scores of different trackers under 11 attributes on the OTB2015 dataset.

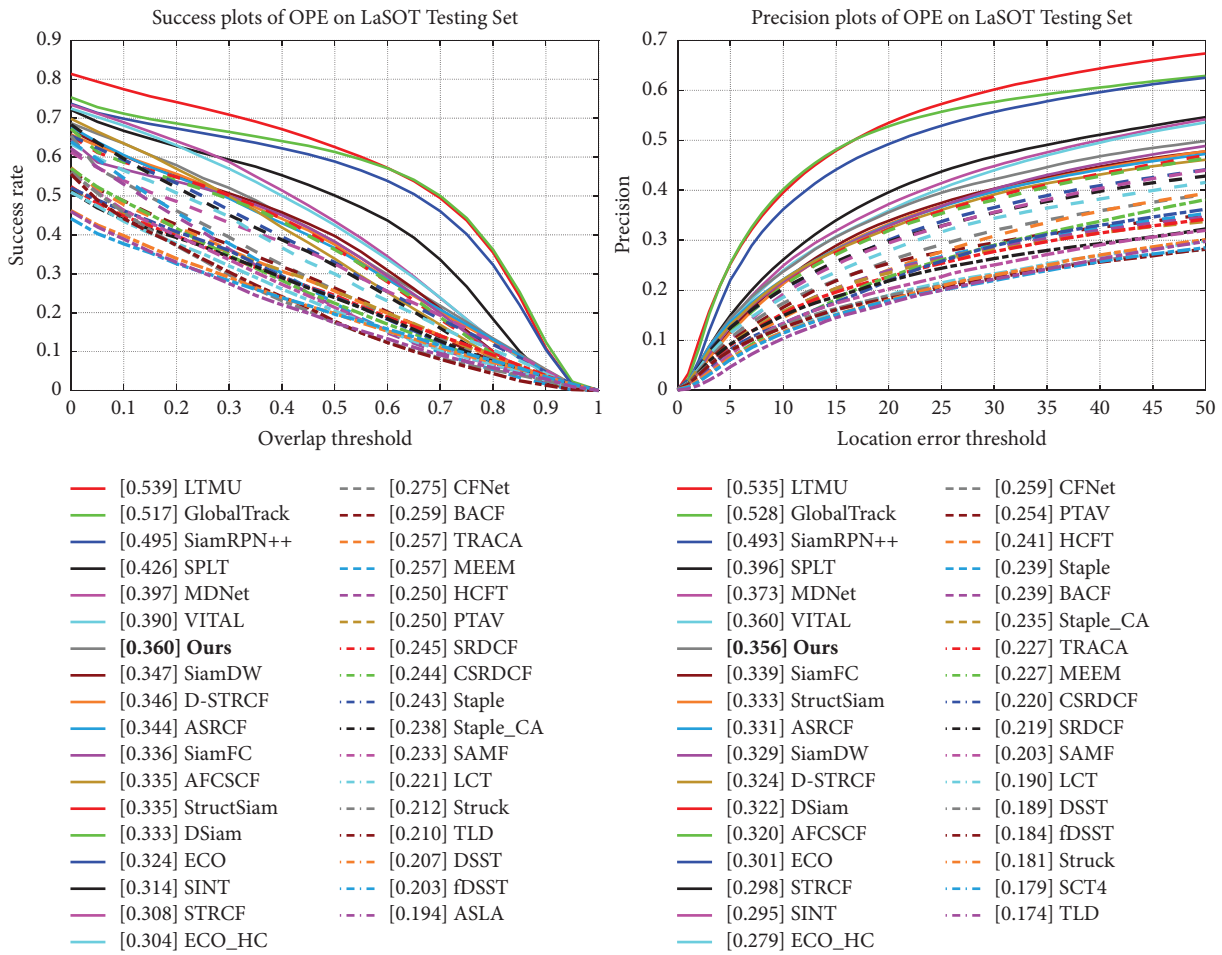


Figure 10: Comparison of the success rate and precision on the LaSOT dataset.

Table 3: Comparison with state-of-the-art trackers in terms of EAO, accuracy, and robustness.

	KCF	SRDCF	Staple	SiamFC	UpdateNet	DCFNet	CSRDCF	C-COT	Ours
EAO	0.135	0.118	0.169	0.188	0.244	0.183	0.256	**0.267**	*0.274*
Accuracy	0.448	0.489	*0.528*	0.501	**0.520**	0.470	0.492	0.494	0.495
Robustness	0.773	0.974	0.688	0.585	0.454	0.543	0.356	**0.318**	*0.304*

The best three results are highlighted in italic, bold, and underline.

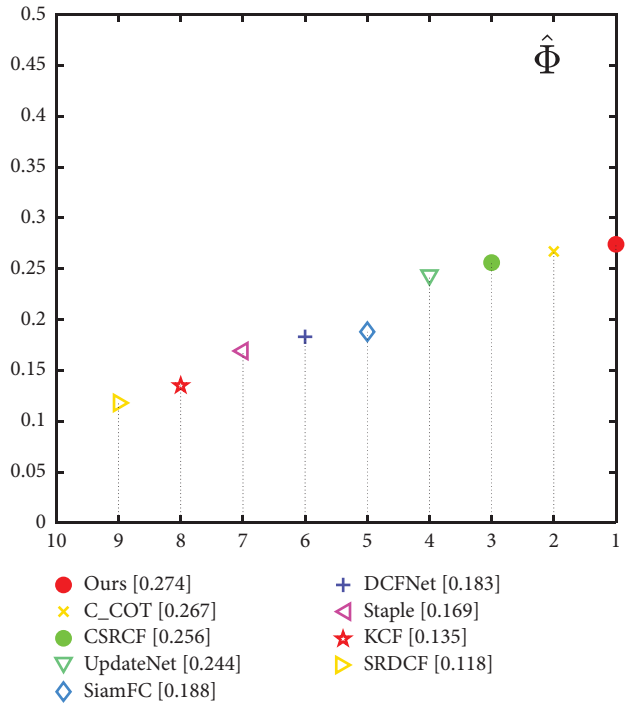


Figure 11: EAO ranking plot on the VOT2018 dataset.

accuracy, our tracker ranks fourth (0.495) behind UpdateNet, SiamFC, and Staple. Notably, our tracker's EAO and robustness values are significantly greater than that of these trackers. Overall, our trackers demonstrate good accuracy and resilience, confirming the efficacy of our proposed method.

4.3.4. UAV123 Dataset. The UAV12350 dataset is a high-resolution dataset of 123 completely annotated aerial video sequences captured by an unmanned aerial vehicle (UAV) and annotated with various challenge attributes (e.g., aspect ratio change, camera motion, and viewpoint change). The UAV also employs the OPE method to evaluate the tracker, considering both success rate and precision. We compared our tracker with 10 leading-edge trackers, including DaHCF [16], MEGTCF [52], ASRCF [21], MCCT [22], ECO [18], MRCF [11], DRCF [33], STRCF [30], BACF [14], and SAMF [13], and the evaluation results are shown in Figure 12. Compared to the most recent state-of-the-art trackers (MRCF, MEGTCF, and DaHCF), our tracker surpasses the majority of its competitors, with the best DP score (75.3%) and the second highest AUC score (51.4%). Overall, the experimental results on the UAV123 dataset fully demonstrate the efficacy and robustness of our tracker.

4.4. Ablation Experiments. To verify the effectiveness of different components of the proposed algorithm, including adaptive channel weighting (ACW) and feature game fusion (FGF), we conducted ablation experiments of our tracker on the OTB2015 dataset, and the experimental results are shown in Figure 13. The different variants in the experiments are explained as follows: "Baseline" refers to introducing spatial-temporal regularization [30] in the standard DCF formulation to optimize the boundary effects and computational complexity. "Baseline + D" is to add the deep features from VGG-16 on the basis of the handcrafted features to enhance the expression of the features. "Baseline + D + ACW" means to introduce adaptive channel weighting into the "Baseline + D" tracker. "Baseline + D + ACW*" introduces a weight threshold to optimize the deep features further. "Baseline + D + ACW* + FGF" represents our final version, which uses adaptive channel weighting and feature game fusion.

Compared to the baseline, "Baseline + D + ACW" improves the AUC and DP scores by 2.4% and 3.0%, respectively. By pruning the channels with low weights, "Baseline + D + ACW*" improves 2.7%/3.7% over the baseline. In addition, our final tracker achieves the best performance by combining all components and improves

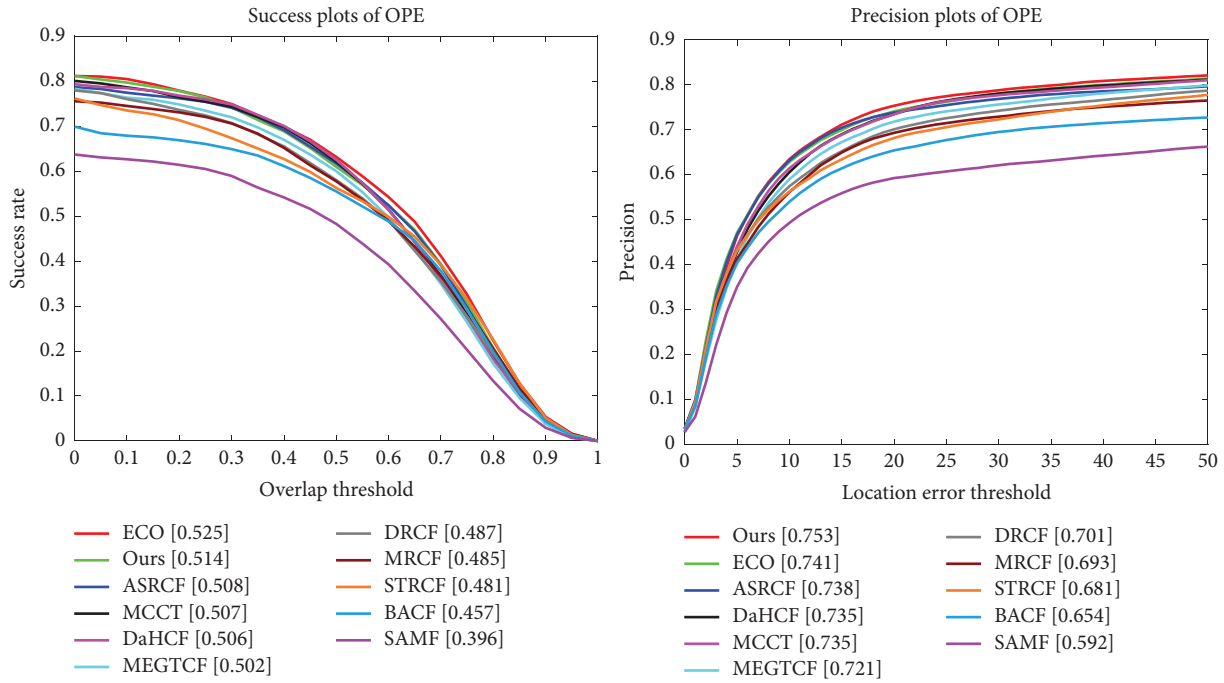


Figure 12: Comparison of success rate and precision on the UAV123 dataset.

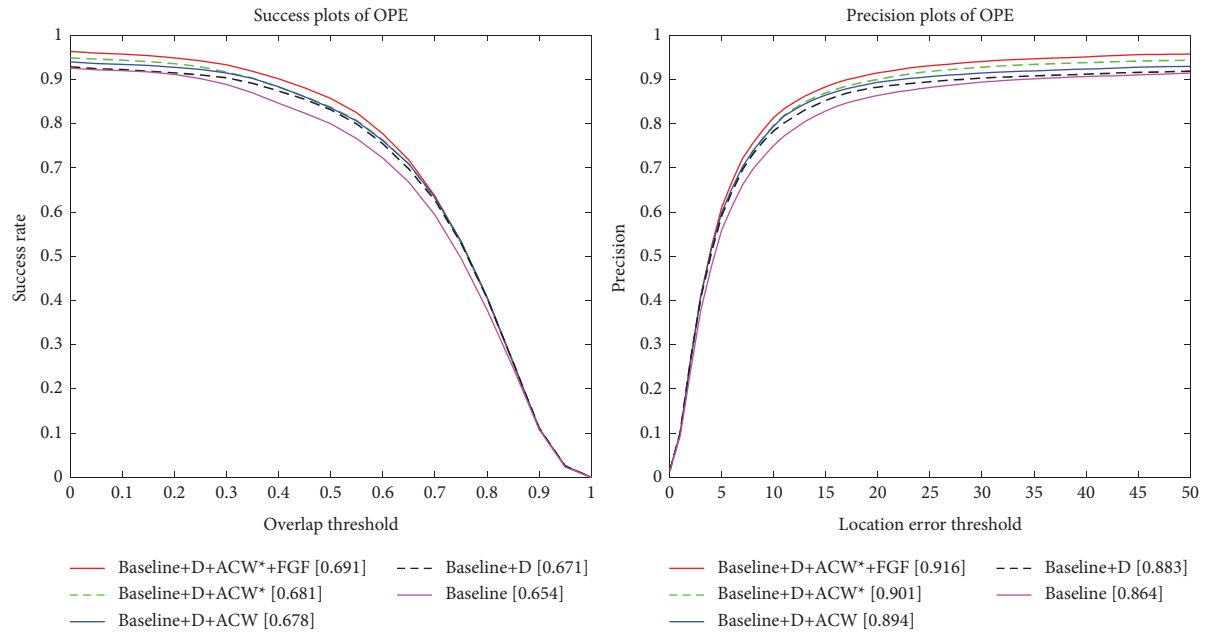


Figure 13: Performance evaluation of different variants of the proposed algorithm on the OTB2015 dataset.

the AUC and DP by 3.7% and 5.2% from the baseline, further demonstrating the effectiveness of the proposed adaptive channel weighting approach and feature game fusion strategy.

In the feature preweighted part, for the responses generated by different convolutional layers, the spatial resolution and dimensionality of their convolutional features should be considered; for instance, the deeper the features tend to have a smaller spatial resolution. To avoid missing information, we assign the responses generated by the deeper features more weight when fusing them. Set $\sigma_1 = W_2/W_1$ and $\sigma_2 = W_3/W_1$, where W_1, W_2, and W_3 represent the response map weights corresponding to handcrafted, shallow, and deep features. The effect of varied weight ratios σ_1 and σ_2 for tracking performance on the OTB2015 dataset is shown in Figure 14. When σ_1 and σ_2 are both 1, which means that all feature response maps have the same importance, it will result in the missing information. If the

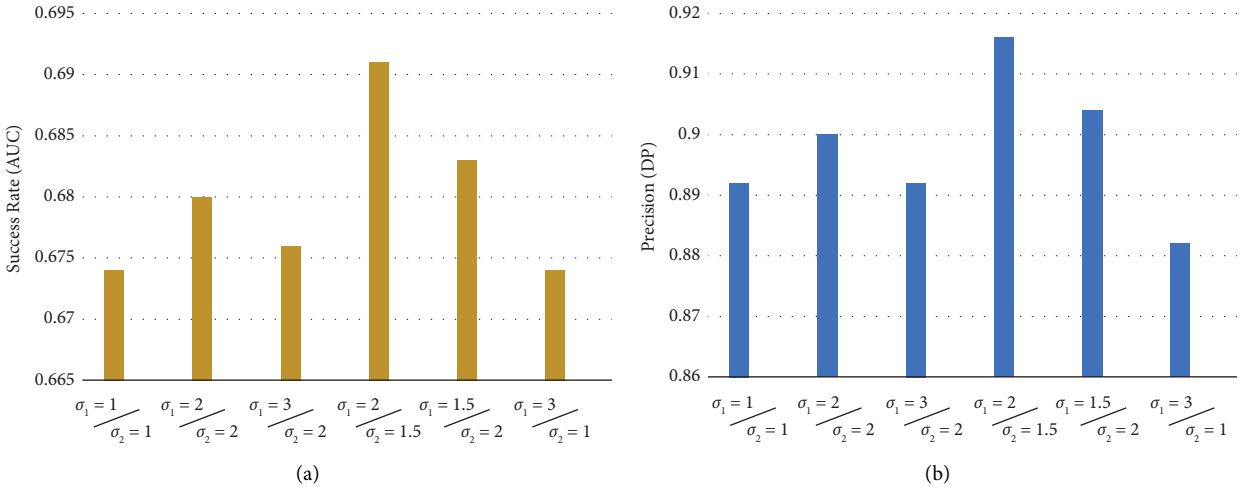


Figure 14: Analysis for success rate (left) and precision (right) of different weights on the OTB2015 dataset.

weights are not selected correctly, it causes performance degradation. The experimental results show that when considering the relevance of different convolutional layers, the success rate and precision on the OTB2015 dataset improve significantly.

5. Conclusion

This research proposes a tracking algorithm based on adaptive channel weighting and feature game fusion. Initially, we utilize channel weight scores to evaluate the importance of the feature channels to the tracker, and then, we prune the noisy channels with low scores and enhance the effective channels through an adaptive channel weighting strategy. On this basis, a game theory-based feature fusion strategy is proposed. We use complementary characteristics between features to construct different feature combinations and utilize game theory to achieve feature-adaptive fusion. Extensive experiments on the OTB2015, LaSOT, VOT2018, and UAV123 datasets show that the proposed algorithm has excellent tracking performance. The suggested method can eliminate the effects of noise channels on tracking performance while achieving optimal fusion between multiple complementary features, effectively improving feature representation. The adaptive channel weighting method suggested in this study is only calculated in the initial frame, making it ineffective for dealing with changes in feature channel relevance induced by changes in target appearance during the tracking process. The next step will be to train the neural network to automatically learn the feature channel weights for each image frame to improve the algorithm's performance and robustness.

Data Availability

The four datasets used in this paper are all well-known public datasets in the field of object tracking, and they are introduced and cited in the paper's experimental part. The raw data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 62072370), the Natural Science Foundation of Shaanxi Province (Grant no. 2023-JC-YB-598), and the Science and Technology Project of Xi'an City (Grant no. 22GXFW0125).

References

[1] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: a comprehensive survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 3943–3968, 2022.

[2] Y. F. Huang, L. P. Shih, C. H. Tsai, and G. T. Shen, "Describing video scenarios using deep learning techniques," *International Journal of Intelligent Systems*, vol. 36, no. 6, pp. 2465–2490, 2021.

[3] J. Wang, B. Tao, Z. Gong, S. Yu, and Z. Yin, "A mobile robotic measurement system for large-scale complex components based on optical scanning and visual tracking," *Robotics and Computer-Integrated Manufacturing*, vol. 67, Article ID 102010, 2021.

[4] S. Liu, D. Liu, G. Srivastava, D. Połap, and M. Woźniak, "Overview and methods of correlation filter algorithms in object tracking," *Complex and Intelligent Systems*, vol. 7, no. 4, pp. 1895–1917, 2021.

[5] A. A. Ahmed and M. Echi, "Hawk-eye: an AI-powered threat detector for intelligent surveillance cameras," *IEEE Access*, vol. 9, pp. 63283–63293, 2021.

[6] Y. Ming, Y. Yang, R. P. Fu et al., "IPMC sensor integrated smart glove for pulse diagnosis, braille recognition, and human-computer interaction," *Advanced Materials Technologies*, vol. 3, no. 12, Article ID 1800257, 2018.

[7] M. C. dos Santos, R. H. C. Palácios, M. Mendonca, J. A. Fabri, and W. F. Godoy, "A neural autonomous robotic manipulator with three degrees of freedom," *International Journal of Intelligent Systems*, vol. 37, no. 9, pp. 5597–5616, 2022.

[8] S. E. Bekhouche, Y. Ruichek, and F. Dornaika, "Driver drowsiness detection in video sequences using hybrid selection of deep features," *Knowledge-Based Systems*, vol. 252, Article ID 109436, 2022.

[9] C. Fu, J. Ye, J. Xu, Y. He, and F. Lin, "Disruptor-aware interval-based response inconsistency for correlation filters in real-time aerial tracking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6301–6313, 2021.

[10] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time UAV tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2891–2900, Seoul, Korea (South), June 2019.

[11] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-Regularized correlation filter for UAV tracking and self-localization," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6004–6014, 2022.

[12] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[13] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *Proceedings of the Computer Vision- ECCV 2014 Workshops*, pp. 254–265, Zurich, Switzerland, September 2015.

[14] H. Kiani Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1135–1143, Venice, Italy, October 2017.

[15] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and V. D. J. Weijer, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1090–1097, Columbus, OH, USA, June 2014.

[16] J. Zhang, Y. Liu, H. Liu, J. Wang, and Y. Zhang, "Distractor-aware visual tracking using hierarchical correlation filters adaptive selection," *Applied Intelligence*, vol. 52, no. 6, pp. 6129–6147, 2022.

[17] S. Ma, L. Zhang, Z. Hou, X. Yang, L. Pu, and X. Zhao, "Robust visual tracking via adaptive feature channel selection," *International Journal of Intelligent Systems*, vol. 37, no. 10, pp. 6951–6977, 2022.

[18] M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg, "ECO: efficient convolution operators for tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6638–6646, Honolulu, HI, USA, June 2017.

[19] M. Che, R. Wang, Y. Lu, Y. Li, H. Zhi, and C. Xiong, "Channel pruning for visual tracking," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 70–82, Tel Aviv, Israel, October 2018.

[20] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3074–3082, Santiago, Chile, October 2015.

[21] K. Dai, D. Wang, H. Lu, C. Sun, and J. Li, "Visual tracking via adaptive spatially-regularized correlation filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4670–4679, Long Beach, CA, USA, June 2019.

[22] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4844–4853, Salt Lake City, UT, USA, June 2018.

[23] Z. He, Y. Fan, J. Zhuang, Y. Dong, and H. Bai, "Correlation filters with weighted convolution responses," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 1992–2000, Venice, Italy, June 2017.

[24] Z. Jin, Z. Hou, W. Yu, C. Chen, and X. Wang, "Game theory-based visual tracking approach focusing on color and texture features," *Applied Optics*, vol. 56, no. 21, pp. 5982–5989, 2017.

[25] B. Liu, X. Chang, D. Yuan, and Y. Yang, "HCDC-SRCF tracker: learning an adaptively multi-feature fuse tracker in spatial regularized correlation filters framework," *Knowledge-Based Systems*, vol. 238, Article ID 107913, 2022.

[26] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter," *Journal of King Saud University- Computer and Information Sciences*, vol. 34, no. 8, pp. 6008–6018, 2022.

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

[28] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proceedings of the Computer Vision – ECCV 2012*, pp. 702–715, Florence, Italy, October 2012.

[29] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4310–4318, Santiago, Chile, December 2015.

[30] F. Li, C. Tian, W. Zuo, L. Zhang, and M. H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4904–4913, Salt Lake City, UT, USA, June 2018.

[31] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference*, pp. 1–11, Nottingham, UK, September 2014.

[32] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M. H. Yang, "Integrating boundary and center correlation filters for visual tracking with aspect ratio variation," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2001–2009, Venice, Italy, June 2017.

[33] C. Fu, J. Xu, F. Lin, F. Guo, T. Liu, and Z. Zhang, "Object saliency-aware dual regularized correlation filter for real-time aerial tracking," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8940–8951, 2020.

[34] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "AutoTrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11923–11932, Seattle, WA, USA, October 2020.

[35] Y. Liang, Y. Liu, Y. Yan, L. Zhang, and H. Wang, "Robust visual tracking via spatio-temporal adaptive and channel selective correlation filters," *Pattern Recognition*, vol. 112, Article ID 107738, 2021.

[36] T. Xu, Z. Feng, X. J. Wu, and J. Kittler, "Adaptive channel selection for robust visual object tracking with discriminative correlation filters," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1359–1375, 2021.

[37] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiah, "SCSTCF: spatial-Channel Selection and temporal regularized correlation filters for visual tracking," *Applied Soft Computing*, vol. 118, Article ID 108485, 2022.

[38] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European conference on computer vision*, pp. 850–865, Glasgow, UK, September 2016.

[39] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2805–2813, Honolulu, HI, USA, June 2017.

[40] J. Zhang, B. Huang, Z. Ye, L. D. Kuang, and X. Ning, "Siamese anchor-free object tracking with multiscale spatial attentions," *Scientific Reports*, vol. 11, no. 1, Article ID 22908, 2021.

[41] J. Zhang, X. Xie, Z. Zheng, L. D. Kuang, and Y. Zhang, "SiamOA: siamese offset-aware object tracking," *Neural Computing and Applications*, vol. 34, no. 24, pp. 22223–22239, 2022.

[42] Z. Li, J. Zhang, Y. Li et al., "Learning feature channel weighting for real-time visual tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 2190–2200, 2022.

[43] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg, "Convolutional features for correlation filter based visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 58–66, Santiago, Chile, June 2015.

[44] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: a benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2411–2418, Portland, OR, USA, June 2013.

[45] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, Las Vegas, NV, USA, June 2016.

[46] M. Kristan, A. Leonardis, and J. Matas, "The sixth Visual Object Tracking VOT2018 challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 3–53, Munich, Germany, September 2018.

[47] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proceedings of the European conference on computer vision*, pp. 445–461, Amsterdam, The Netherlands, October 2016.

[48] H. Fan, L. Lin, and F. Yang, "LaSOT: a high-quality benchmark for large-scale single object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5374–5383, Long Beach, CA, USA, June 2019.

[49] M. Danelljan, A. Robinson, F. Shahbaz Khan, and M. Felsberg, "Beyond correlation filters: learning continuous convolution operators for visual tracking," in *Proceedings of the European conference on computer vision*, pp. 472–488, Amsterdam, The Netherlands, October 2016.

[50] C. Fu, J. Jin, F. Ding, Y. Li, and G. Lu, "Spatial reliability enhanced correlation filter: an efficient approach for real-time UAV tracking," *IEEE Transactions on Multimedia*, pp. 1–15, 2021.

[51] M. Danelljan, L. V. Gool, and R. Timofte, "Probabilistic regression for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7183–7192, Seattle, WA, USA, June 2020.

[52] S. Ma, Z. Zhao, Z. Hou, L. Zhang, X. Yang, and L. Pu, "Correlation filters based on multi-expert and game theory for visual object tracking," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[53] P. Li, B. Chen, W. Ouyang, D. Wang, X. Yang, and H. Lu, "Gradnet: gradient-guided network for visual object tracking," in *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 6162–6171, Seoul, Korea (South), June 2019.

[54] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 101–117, Munich, Germany, September 2018.

[55] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8971–8980, Salt Lake City, UT, USA, June 2018.

[56] T. Zhang, C. Xu, and M. H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4335–4343, Honolulu, HI, USA, June 2017.

[57] K. Dai, Y. Zhang, D. Wang, J. Li, H. Lu, and X. Yang, "High-performance long-term tracking with meta-updater," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6298–6307, Seattle, WA, USA, June 2020.

[58] L. Huang, X. Zhao, and K. Huang, "Globaltrack: a simple and strong baseline for long-term tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11037–11044, Washington DC, USA, March 2020.

[59] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4282–4291, Long Beach, CA, USA, June 2019.

[60] B. Yan, H. Zhao, D. Wang, H. Lu, and X. Yang, "'skimming-perusal' tracking: a framework for real-time and robust long-term tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2385–2393, Seoul, Korea (South), June 2019.

[61] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293–4302, Las Vegas, NV, USA, June 2016.

[62] Y. Song, C. Ma, and X. Wu, "VITAL: VIsual tracking via adversarial learning," in *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 8990–8999, Salt Lake City, UT, USA, June 2018.

[63] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: complementary learners for real-time tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1401–1409, Las Vegas, NV, USA, June 2016.

[64] L. Zhang, A. Gonzalez-Garcia, W. Jvd, M. Danelljan, and F. S. Khan, "Learning the model update for siamese trackers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4010–4019, Seoul, Korea (South), October 2019.

[65] Q. Wang, J. Gao, J. Xing, M. Zhang, and W. Hu, "Dcfnet: discriminant correlation filters network for visual tracking," 2017, https://arxiv.org/abs/1704.04057.

[66] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative correlation filter with channel and spatial reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6309–6318, Honolulu, HI, USA, June 2017.