Neural Networks 186 (2025) 107210

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Learning temporal regularized spatial-aware deep correlation filter tracking via adaptive channel selection

Sathiyamoorthi Arthanari, Dinesh Elayaperumal, Young Hoon Joo *

School of IT Information and Control Engineering, Kunsan National University, 558 Daehak-ro, Gunsan-si, Jeonbuk 54150, Republic of Korea

ARTICLE INFO

Keywords:
Deep correlation filter
Adaptive channel selection
Temporal regularization
Spatial-aware
Statistical color model
Visual object tracking

ABSTRACT

In recent years, deep correlation filters have demonstrated outstanding performance in robust object tracking. Nevertheless, the correlation filters encounter challenges in managing huge occlusion, target deviation, and background clutter due to the lack of effective utilization of previous target information. To overcome these issues, we propose a novel temporal regularized spatial-aware deep correlation filter tracking via adaptive channel selection. To do this, we first presented the adaptive channel selection approach, which efficiently handles target deviation by adaptively selecting suitable channels during the learning stage. In addition, the adaptive channel selection method allows for dynamic adjustments to the filter based on the unique characteristics of the target object. This adaptability enhances the tracker's flexibility, making it well-suited for diverse tracking scenarios. Second, we propose the spatial-aware correlation filter with dynamic spatial constraints, which effectively reduces the filter response in the complex background region by distinguishing between the foreground and background regions in the response map. Hence, the target can be easily identified within the foreground region. Third, we designed a temporal regularization approach that improves the target accuracy when the case of large appearance variations. Additionally, this temporal regularization method considers the present and previous frames of the target region, which significantly enhances the tracking ability by utilizing historical information. Finally, we present a comprehensive experiments analysis of the OTB-2013, OTB-2015, TempleColor-128, UAV-123, UAVDT, and DTB-70 benchmark datasets to demonstrate the effectiveness of the proposed approach against the state-of-the-trackers.

1. Introduction

Visual object tracking is one of the principal challenges in computer vision, which has been widely applied in numerous applications, such as autonomous driving, surveillance, robotics, medical image processing, and so on (Dinesh & Joo, 2021; Fan, Li, Zhou, Liu, & He, 2021; Han, Lekamalage, & Huang, 2022; Sathishkumar & Joo, 2021). However, the main difficulties in object tracking are dealing with unexpected changes in target, such as the presence of occlusions, background clutter, variations in illumination, object deformation, camera motion, and many others. In recent years, several tracking methods, such as correlation filters (CF) and deep learning-based trackers implemented to address these challenges.

Recently, the CF-based trackers (Li et al., 2020; Wen, Chu, Lai, Xu, & Shen, 2023) have gained much popularity among researchers because of their impressive performance on benchmark datasets. Despite substantial advancements achieved by CF-based trackers, challenges remain in developing computationally efficient and accurate real-time trackers. To overcome these challenges, the authors have presented CF-based trackers such as KCF (Henriques, Caseiro, Martins, & Batista, 2014), and CSK (Henriques, Caseiro, Martins, & Batista, 2012). Moreover, the CF-based trackers deal with the frequency domain transformed from the time domain for getting an impressive tracking speed. Although the frequency domain improves the computational performance of CF-based trackers, the negative samples have been formed through the circulant shift in the positive samples, which did not adequately represent the target appearance. In addition, the negative samples get boundary effects due to circulant shifts, which significantly reduces the tracking performance. In recent years, several approaches have been suggested to handle this boundary effect issue (Danelljan, Hager, Khan, & Felsberg, 2015b; Galoogahi, Hamed, & Lucey, 2017). The authors in Galoogahi et al. (2017) have presented a BACF tracker, which utilizes the foreground and background region information of the image samples to enhance the tracker efficiency. Similarly, the authors in Danelljan et al. (2015b) have proposed the SRDCF method that utilizes the spatial-regularization technique, which helps to reduce the impact of the boundary effect by capturing the object's appearance

* Corresponding author.
E-mail address: yhjoo@kunsan.ac.kr (Y.H. Joo).

https://doi.org/10.1016/j.neunet.2025.107210
Received 18 February 2023; Received in revised form 11 June 2024; Accepted 22 January 2025
Available online 15 February 2025

more accurately. Despite the CF-based trackers improving tracking efficiency, the tracker may lose its precise target location when the tracking approach is more complex. To address these problems, spatio-temporal information has been integrated into CF-based methods to boost tracker performance across diverse analyses (Li, Tian, Zuo, Zhang & Yang, 2018; Teng et al., 2017). The authors in Teng et al. (2017) have introduced a TSN deep network architecture that incorporates spatio-temporal information to accurately predict the target location. Moreover, the authors in Li, Tian et al. (2018) have presented the STRCF tracking approach, which is designed to handle fast-moving objects by incorporating spatial and temporal information. Although the CF-based tracking methods obtained better tracking performance using hand-crafted features, there are still unavoidable problems due to object deformation and occlusion. It is known that recent deep convolution features can solve this problem by providing a detailed analysis of the video sequence to the tracker model.

On the other hand, deep learning approaches have been recently used to develop new strategies for object tracking tasks (Elayaperumal & Joo, 2023; Moorthy & Joo, 2023; Nam & Han, 2016; Wu, Lan, Zhang, & Xiang, 2023) that perform better than conventional CF due to their robust feature extraction ability and strong learning capability. Hence, the convolutional neural network (CNN) is known as one of the most often utilized deep learning techniques for object tracking. In addition, the CF-based trackers utilized the pre-trained CNN methods (Fan & Ling, 2017; Lu et al., 2017) instead of conventional hand-crafted features to represent the target object and achieve better performance. Specifically, the MDNet (Nam & Han, 2016) tracking technique employed the pre-trained CNN method on a large-scale video tracking sequence. These pre-trained models were adopted in CF-based tracking methods to obtain high-level representations of the target. In addition, the SANet (Fan & Ling, 2017) tracking approach utilized the same technique as the MDNet tracker and introduced a recurrent neural network (RNN)-based framework to enhance the object representation. Moreover, the authors in Sun et al. (2022) have presented the Siamese network architecture to improve the robustness and video tracking performance. Meanwhile, the deep learning-based offline training approach is time-consuming, and it may be unable to track specific object because the learned target object representations lack distinctiveness. To address this issue, the authors in Lu et al. (2017) have presented a feed-forward CNN to produce an adequate object representation for visual tracking. Recently, the authors in Nai, Li, and Wang (2022) have proposed a residual network (ResNet-50) model, which is used to extract features more quickly than traditional methods and obtain adequate target representation.

Inspired by the above analysis, we present a novel CF-based tracking approach using a temporal regularized spatial-aware deep correlation filter (TRSADCF) via adaptive channel selection. The major contribution of the proposed approach is described as follows:

1. Initially, a novel adaptive channel selection (ACS) method for multi-channel feature representation is proposed, which effectively deals with target deviations by adaptively selecting a suitable channel during the learning stage.
2. A spatial-aware correlation filter with dynamic spatial constraints is presented to effectively track the target region when the spatial distribution deviates from the target. Afterward, the statistical color model (SCM) utilizes a dynamic spatial constraint to distinguish the foreground and background regions in the response map, which effectively improves the filter response in the foreground region.
3. A temporal regularization approach is designed to improve the target accuracy when the case of large appearance variations and cluttered backgrounds. Specifically, the temporal regularization method considers the present and previous frames of the target region, which significantly enhances the tracking ability by utilizing historical information of the previous frame.
4. A multi-feature fusion strategy is presented that combines histogram-oriented gradient (HOG), ColorName (CN), Intensity (IC), and ResNet-50 features to improve the tracking efficiency and robustness.
5. Finally, we conducted experimental analysis on the benchmark datasets OTB-2013, OTB-2015, TempleColor-128, UAV-123, UAVDT, and DTB-70 to demonstrate that the proposed TRSADCF tracker has consistently outperformed state-of-the-art trackers.

2. Related works

In this section, we explore various object-tracking approaches. In Section 2.1, we analyze the correlation filter-based methods. Sections 2.2 and 2.3, we examined deep learning and transformer-based trackers, respectively.

2.1. Correlation filter-based tracking approach

Over the past few years, the discriminative correlation filter (DCF)-based tracking approaches such as KCF (Henriques et al., 2014), CSK (Henriques et al., 2012), BACF (Galoogahi et al., 2017), SRDCF (Danelljan et al., 2015b), SASR (Fu, Xiong, Lin, & Yue, 2020), TLD (Kalal, Mikolajczyk, & Matas, 2011), and MRCT (Hu, Ma, Shen, & Shao, 2017) have gained much attention due to their remarkable tracking performance as well as high tracking speed. The authors in Henriques et al. (2014) have introduced the kernelized correlation filter (KCF) tracking approach, which obtains a dense sampling effect on the training samples by utilizing characteristics of the circulant matrix in the frequency domain. Therefore, this KCF tracker significantly increased the number of training samples as well as improving the learning and detection speed. Specifically, the authors in Henriques et al. (2012) have proposed a circulant structure of tracking-by-detection with the kernel (CSK) to improve robustness and tracking performance. Specifically, the DCF-based trackers benefited from this method's use of circulant matrices and kernel space to probe the potential of dense sampling. In addition, the BACF (Galoogahi et al., 2017) trackers enhance tracking by considering background information to reduce drift, which employs spatial regularization to focus the correlation filter on the target object while suppressing background influence. Additionally, the authors in Danelljan et al. (2015b), have introduced the SRDCF tracking method, which adapts to scale changes and can handle complex scenarios such as occlusions and abrupt motion. Although more computationally demanding than simpler correlation filters, SRDCF provides a significant improvement in robustness and accuracy, making it effective for diverse tracking challenges. Also, the SASR (Fu et al., 2020) tracker leverages surrounding patches to refine the localization of the target object. By incorporating information from adjacent patches, it can better distinguish the target from the background, leading to more accurate and stable tracking. Moreover, the authors in Kalal et al. (2011) have proposed a tracking-learning-detection (TLD) method, which exploits an ensemble of weak classifiers to solve the problem of global object re-detection. However, this approach fails to classify the target object due to the huge number of scanning windows. The authors in Hu et al. (2017) have examined the manifold regularized correlation tracking (MRCT) approach for online tracking using a blockwise fast detection technique. Despite the tracker obtaining great performance, the tracking approach remains complex because of boundary effects issues. In this work, we propose a temporal regularized spatial-aware approach to CF-tracker, which is used to avoid boundary effects and provide a more robust appearance model. In addition, the proposed tracking method proves that the tracking performance is greatly improved compared to the other trackers.

2.2. Deep learning-based tracking approach

In the last few years, the deep CNN structure has attained remarkable performance, which has become increasingly popular in the tracking community due to its superior tracking performance in object tracking (Danelljan, Hager, Khan, & Felsberg, 2015a; Qi et al., 2016). Further, the trackers are classified into two categories depending on their integration of deep features and CF trackers: (1) Pre-trained CNN, and (2) End-to-end trained CNN. On the one hand, the CF employs a pre-trained CNN approach to extract deep features and improve tracking efficiency. Recently, various deep feature extraction techniques have been used in object tracking, such as CNN, RNN, ResNet, VGGNet, LSTM, and so on. More specifically, the extracted deep features is employed in the CF framework to enhance the tracker efficiency. The hedged deep tracking (HDT) (Qi et al., 2016) are used in the CF trackers to learn the multi-layer convolutional features and infer the target position by combining response maps of multiple layers. Further, the authors have proposed CF-based deepSRDCF in Danelljan et al. (2015a), which can be used to enhance the tracking performance through robust feature extraction.

On the other hand, the CF-based trackers employ end-to-end training with well-designed deep network structures such as MDNet (Nam & Han, 2016), Siamese network (Sun et al., 2022), SiamRPN (Li, Yan, Wu, Zhu, & Hu, 2018), CFNet (Valmadre, Bertinetto, Henriques, Vedaldi, & Torr, 2017), and ResNet (Nai et al., 2022). In this aspect, the authors in Nam and Han (2016) have introduced the multi-domain CNN (MDNet) based end-to-end object tracking framework, which significantly increases the tracking efficiency and at the same time, decreases the speed. A short time ago, the authors in Sun et al. (2022) have proposed the Siamese network architecture, which enables the end-to-end training of a Siamese network for object tracking. In addition, the authors in Li, Yan et al. (2018) have proposed the Siamese region proposal network (SiameseRPN), which integrates Siamese-FC and RPN structures to achieve excellent tracking efficiency. Also, the authors have introduced CFNet (Valmadre et al., 2017) as an end-to-end object tracking strategy in correlation filter, which achieves outstanding tracking performance using a trained CNN model. In Nai et al. (2022), the ResNet-50 model has been used in object tracking for feature extraction and it helps to achieve better performance of the CF tracker. Therefore, we adopt the ResNet-50 model to demonstrate the superior performance of the proposed tracker through comparison with existing methods.

2.3. Tracking based on transformer approaches

Recently, transformers-based architecture introduced for natural language processing tasks has demonstrated remarkable success in various computer vision domains, prompting researchers to explore their potential for object tracking. The transformer's ability to capture long-range dependencies and contextual information makes it well-suited for addressing the challenges posed by object appearance variations, occlusions, and scale changes. In the context of object tracking, transformer-based methods (Gu, Lu, & Cai, 2022; Wu, Wen, Yuan, Xu, & Liu, 2024; Zhang, Piao, & Qi, 2024) leverage the self-attention mechanism to efficiently model the relationships between different regions of an image. The authors in Gu et al. (2022) introduced the RPformer architecture, specifically designed to effectively incorporate the feature relationship between the template and the search region. This design facilitates to capture the rich contextual information, thereby minimizing information loss and enhancing the utilization of global feature information. Following that, SIFT (Zhang et al., 2024) proposes a transformer-based tracker that effectively combines spatial and temporal features for robust object tracking, which utilizes a spatial attention mechanism to capture the target object's appearance in the current frame and a temporal attention mechanism to learn the object's motion patterns across frames. Moreover, the authors in Wu et al. (2024) propose a novel transformer-based tracker that effectively utilizes a deformable transformer module and spatio-temporal information to handle object deformations and long-range appearance dependencies, which helps to enhance the tracker's efficiency and robustness.

3. Proposed method and implementation

In the present section, we mainly discuss the proposed temporal regularized spatial-aware deep correlation filter via the adaptive channel selection approach. First, we briefly revisit the BACF approach in Section 3.1. Then, in Section 3.2, we investigate the adaptive channel selection approach via elastic net regularization. Furthermore, we describe the spatial-aware approach to reduce the filter response in the background region in Section 3.3. Moreover, we have introduced a temporal regularization approach to enhance long-term tracking performance in Section 3.4. Finally, the overall diagram of the proposed approach is illustrated in Fig. 1.

3.1. Background-aware correlation filter (Baseline tracker)

In this work, we consider the BACF (Galoogahi et al., 2017) method as our baseline tracker that crops the true negative sample by using the binary matrix, which effectively handles the boundary effect issues. Then, the filter trains a classifier on the image patch using dense sampling to increase tracker performance. Hence, the objective function for the BACF tracker can be expressed as:

$$E(h_t) = \frac{1}{2}\sum_{j=1}^{T}\left\| y(j) - \sum_{k=1}^{C}(h_t^k)^{\mathsf{T}} P X_t[\Delta\tau_j]\right\|_{\mathcal{F}}^2 + \frac{\lambda}{2}S(\mathcal{H}), \tag{1}$$

where $X_t \in \mathbb{R}^T$ denotes the training sample, $y \in \mathbb{R}^T$ represents the correlation output and $h_t \in \mathbb{R}^D$ indicates filter. Further, the P denotes the $D \times T$ binary matrix and $[\Delta\tau_j]$ is the circulant shift operator. The C indicates the number of the feature channels, T denotes the transpose matrix, λ is a regularization parameter, $\mathcal{F}$ represents the Frobenius norm, and $S(\mathcal{H})$ denotes the regularization term.

3.2. Adaptive channel selection via elastic net

The multi-channel feature-based adaptive channel selection approach is presented, which is designed to efficiently handle target deviation by adaptively selecting suitable channels during the learning stage. Subsequently, the adaptive channel selection feature enables the filter to make real-time adjustments according to the unique attributes of the target object. This adaptability enhances the tracker's flexibility, making it well-suited for diverse tracking scenarios. Specifically, the multi-channel feature-based ACS approach is illustrated in Fig. 1. Moreover, the ACS employs an adaptive group elastic net in the learning framework to combine the l_1-norm and l_2-norm regularization. Specifically, the adaptive group elastic net approach is presented along with the group variable selection to improve the grouping effect during the variable selection. In particular, the authors in Zou and Hastie (2005) proposed the regularization and variable selection via the elastic net approach. Further, the authors in Nie, Huang, Cai, and Ding (2010) introduced an efficient and robust feature selection approach. Based on their third dimension channel, the elements in $\mathcal{H}$ are classified into specific groups. The grouping operator for the high-dimensional filtering system constructs the groups of variables. The different group of information is fused and conveyed to each group by introducing a balancing function (BF) $\Psi()$. Therefore, we define the BF $\Psi()$, as the Frobenius norm for each $N \times N$ matrix h_t^k, i.e., $\Psi_t^k = \Psi(h_t^k) = \left\|h_t^k\right\|_{\mathcal{F}}$. In Eq. (1), we enforce the BF in the regularization term $S(\mathcal{H})$. Finally, the regularization term is described in the following structure:

$$S(\Psi_t) = \beta\left\|\Psi_t\right\|_1 + (1-\beta)\left\|\Psi_t - \Psi_{t-1}\right\|_2^2. \tag{2}$$

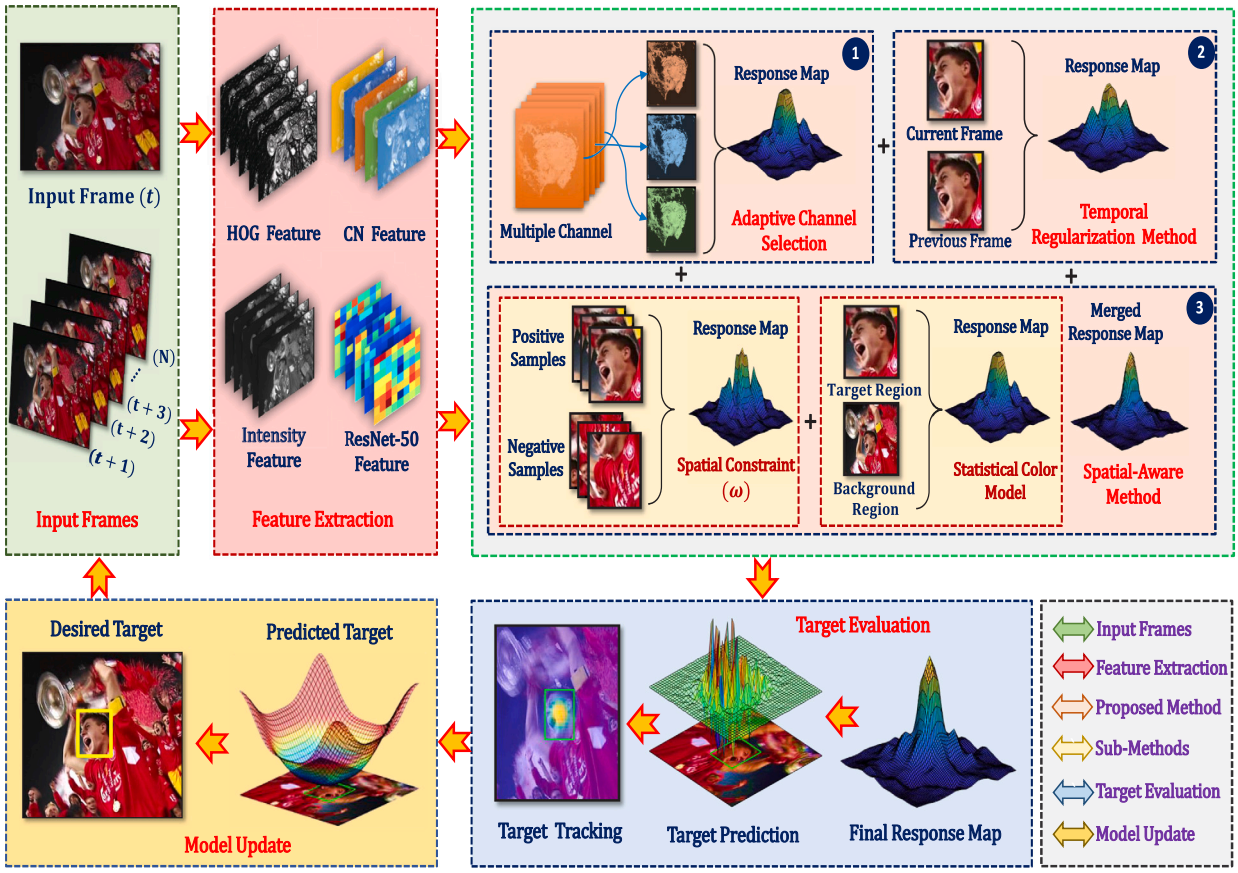


Fig. 1. The diagram illustrates the overall framework of our proposed approach. Initially, we extract input frames using a combination of hand-crafted and deep features. Following this, the extracted features are directed to three proposed approaches: (1) Adaptive channel selection, (2) Temporal regularization method, and (3) Spatial-aware module. Subsequently, the proposed approach independently generates a response map, and these individual response maps are integrated into a final response. Afterward, the target prediction is determined by analyzing the information contained in this final response map. Finally, the predicted target is updated to the upcoming frames.

Now we expand the last term in Eq. (2), and we can be described as follows:

$$S(\Psi_t) = \beta \left\|\Psi_t\right\|_1 + (1-\beta)\left\|\Psi_t\right\|_2^2 + (1-\beta)(\Psi_{t-1} - 2\Psi_t)^{\mathrm{T}}\Psi_{t-1}. \tag{3}$$

In Eq. (3), the term $(1-\beta)\left\|\Psi_t\right\|_2^2$ is represented as elastic net regularization and $(1-\beta)(\Psi_{t-1}-2\Psi_t)^{\mathrm{T}}\Psi_{t-1}$ describes as adaptive regularization. To integrate the adaptive elastic net method into our baseline tracker (1), the balancing function $\Psi_t^k = \Psi(h_t^k) = \left\|h_t^k\right\|_F$ is enforced into (2). Therefore, we can obtain the regularization function as follows:

$$\begin{aligned} S(\mathcal{H}) &= \beta \left\|\Psi(\mathcal{H})_t\right\|_1 + (1-\beta)\left\|\Psi(\mathcal{H})_t - \Psi(\mathcal{H})_{t-1}\right\|_2^2 \\ &= \beta \sum_{k=1}^{C} \left\|h_t^k\right\|_F + (1-\beta)\sum_{k=1}^{C}\left(\left\|h_t^k\right\|_F - \left\|h_{t-1}^k\right\|_F\right)^2 \\ &\le \beta \sum_{k=1}^{C} \left\|h_t^k\right\|_F + (1-\beta)\sum_{k=1}^{C}\left\|h_t^k - h_{t-1}^k\right\|_F^2. \end{aligned} \tag{4}$$

Therefore, the objective function is defined as follows:

$$\begin{aligned} E(h_t) = \frac{1}{2}\sum_{j=1}^{T}\left\| y(j) - \sum_{k=1}^{C}(h_t^k)^{\mathrm{T}} P X_t[\Delta\tau_j]\right\|_F^2 \\ + \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|h_t^k\right\|_F + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|h_t^k - h_{t-1}^k\right\|_F^2, \end{aligned} \tag{5}$$

where, we consider the $\beta\lambda$ term as λ_1 and $(1-\beta)\lambda$ term as λ_2.

3.3. Spatial-aware correlation filter

The spatial distribution of the target representation is recognized as a crucial feature in object tracking. However, the spatial features response value significantly reduces when the target object disappears due to the object's deformation, rotation, and scale variation. To address these problems, we proposed a spatial-aware correlation filter with dynamic spatial constraints (ω) that effectively differentiates the positive and negative samples. By leveraging the positive samples with the help of the spatial constraint, the spatial-aware approach accurately tracks the target region. Following this, the statistical color model utilizes a dynamic spatial constraint to distinguish the foreground and background regions in the response map and effectively improve the filter response in the foreground region. Specifically, the comprehensive examination of the spatial constraint and SCM is presented in Fig. 1 and Section 3.8. The objective function is written as follows:

$$\begin{aligned} E(h_t) = \frac{1}{2}\sum_{j=1}^{T}\left\| y(j) - \sum_{k=1}^{C}(h_t^k)^{\mathrm{T}} P X_t[\Delta\tau_j]\right\|_F^2 + \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|\omega h_t^k\right\|_F \\ + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|h_t^k - h_{t-1}^k\right\|_F^2. \end{aligned} \tag{6}$$

3.4. Temporal regularized correlation filter

The correlation filter-based trackers encounter challenges in the presence of target object occlusion, background clutter, and deformation. In cases where occlusion occurs within video sequences, the filter may lose track of the target, leading to the tracker's inability to redetect the target's position until the end of the video sequences. To overcome these issues, we designed a temporal regularization approach that improves the target accuracy when the case of large appearance variations and cluttered backgrounds. In addition, this temporal regularization

method considers the present and previous frames of the target region, which significantly enhances the tracking ability by utilizing historical information. The objective function can be described as follows:

$$E(h_t) = \frac{1}{2}\sum_{j=1}^{T}\left\| y(j) - \sum_{k=1}^{C}(h_t^k)^{\mathsf{T}} P X_t[\Delta\tau_j]\right\|_F^2 + \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|\omega h_t^k\right\|_F + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|h_t^k - h_{t-1}^k\right\|_F^2 + \frac{\eta}{2}\sum_{k=1}^{C}\left\|y - (h_t^k)^{\mathsf{T}} X_{t-1}\right\|_F^2, \tag{7}$$

where λ_1, λ_2 and η denote the regularization parameters ($\lambda_1, \lambda_2, \eta \geq 0$), respectively.

3.5. Frequency domain transformation

The Fourier transform is typically used to learn the correlation filter in the frequency domain, which improves the computation performance. In this way, the objective function Eq. (7) is described in the frequency domain as follows:

$$E(h_t, \hat{g}_t) = \frac{1}{2}\left\|\hat{y} - \sum_{k=1}^{C}\hat{X}_t \odot \hat{g}_t^k\right\|_F^2 + \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|\omega h_t^k\right\|_F + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \hat{g}_{t-1}^k\right\|_F^2 + \frac{\eta}{2}\sum_{k=1}^{C}\left\|\hat{y} - (h_t^k)^{\mathsf{T}}\hat{X}_{t-1}\right\|_F^2.$$
$$s.t.\ \ \hat{g}_t^k = \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k, \tag{8}$$

where $\hat{g}_t^k$ denotes the auxiliary variable, $\wedge$ represents the Fourier transform operation, $\odot$ denotes the element-wise multiplication, $\otimes$ is the Kronecker product, F denotes the $P * P$ Fourier matrix, I^k represents the identity matrix.

3.6. Augmented Lagrangian method

To solve the Eq. (8), we utilize the ALM technique (Boyd, Parikh, Chu, Peleato, & Eckstein, 2011) in the frequency domain.

$$\mathcal{L}(h_t, \hat{g}_t, \hat{\zeta}) = \frac{1}{2}\left\|\hat{y} - \sum_{k=1}^{C}\hat{X}_t \odot \hat{g}_t^k\right\|_F^2 + \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|\omega h_t^k\right\|_F + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \hat{g}_{t-1}^k\right\|_F^2 + \frac{\eta}{2}\sum_{k=1}^{C}\left\|\hat{y} - (h_t^k)^{\mathsf{T}}\hat{X}_{t-1}\right\|_F^2 + \sum_{k=1}^{C}(\hat{\zeta}^k)^{\mathsf{T}}(\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k) + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k\right\|_F^2, \tag{9}$$

where μ denotes the regularization term ($\mu \geq 0$) and ζ represents the Lagrangian term. $\mathcal{L}(h_t, \hat{g}_t, \hat{\zeta})$ can be fixed via ADMM (Boyd et al., 2011) approach, in which the sub-problems $\hat{g}_{t+1}^*$ and h_{t+1}^* have a closed-form solution.

3.6.1. Sub-problem h_{t+1}^*

The sub-problem h_{t+1}^* can be solved as follows :

$$h_{t+1}^* = \underset{h_t}{\operatorname{argmin}}\Big\{ \frac{\lambda_1}{2}\sum_{k=1}^{C}\left\|\omega h_t^k\right\|_F + \frac{\eta}{2}\sum_{k=1}^{C}\left\|\hat{y} - (h_t^k)^{\mathsf{T}}\hat{X}_{t-1}\right\|_F^2 + \sum_{k=1}^{C}(\hat{\zeta}^k)^{\mathsf{T}}(\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k) + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k\right\|_F^2\Big\}. \tag{10}$$

An exact solution to h_{t+1}^* can be expressed as:

$$h_{t+1}^* = T(\mu T + \eta\hat{s}_x(t))^{-1}(\mu g + \zeta + \eta\hat{s}_y(t) - \frac{\lambda_1}{2}\omega), \tag{11}$$

where $g = \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)\hat{g}_t$, $\hat{s}_x(t) = \hat{x}^{\mathsf{T}}(t)\hat{x}(t)$, $\hat{s}_y(t) = \hat{x}^{\mathsf{T}}(t)\hat{y}(t)$.

3.6.2. Sub-problem $\hat{g}_{t+1}^*$

The sub-problem g_{t+1}^* is determined as follows:

$$\hat{g}_{t+1}^* = \underset{\hat{g}_t}{\operatorname{argmin}}\Big\{ \frac{1}{2}\left\|\hat{y} - \sum_{k=1}^{C}\hat{X}_t \odot \hat{g}_t^k\right\|_F^2 + \frac{\lambda_2}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \hat{g}_{t-1}^k\right\|_F^2 + \sum_{k=1}^{C}(\hat{\zeta}^k)^{\mathsf{T}}(\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k) + \frac{\mu}{2}\sum_{k=1}^{C}\left\|\hat{g}_t^k - \sqrt{T}(FP^{\mathsf{T}} \otimes I^k)h_t^k\right\|_F^2\Big\}. \tag{12}$$

After solving Eq. (12), we can get as follows:

$$\hat{g}_{t+1}^* = \left(\hat{x}(t)\hat{x}^{\mathsf{T}}(t) + \lambda_2 T + \mu T\right)^{-1}\left(\hat{x}(t)\hat{y}(t) + T\lambda_2\hat{g}_{t-1}(t) - T\hat{\zeta}(t) + T\mu\hat{h}(t)\right). \tag{13}$$

In order to improve the computational efficiency and avoid the inverse operation, we employ the Sherman–Morrison method (Sherman & Morrison, 1950). Therefore, Eq. (13) is redefined as follows:

$$\hat{g}_{t+1}^* = \frac{1}{(\lambda_2 + \mu)}\left(T\hat{x}(t)\hat{y}(t) + \lambda_2\hat{g}_{t-1}(t) - \hat{\zeta}(t) + \mu\hat{h}(t)\right) - \frac{\hat{x}(t)}{(\lambda_2 + \mu)b}\left(T\hat{s}_x(t)\hat{y}(t) + T\lambda_2\hat{s}_g(t) - \hat{s}\zeta(t) + \mu\hat{s}_h(t)\right), \tag{14}$$

where, $\hat{s}_x(t) = \hat{x}^{\mathsf{T}}(t)\hat{x}(t)$, $\hat{s}_g(t) = \hat{x}^{\mathsf{T}}(t)\hat{g}_{t-1}(t)$, $\hat{s}\zeta(t) = \hat{x}^{\mathsf{T}}(t)\hat{\zeta}(t)$, $\hat{s}_h(t) = \hat{x}^{\mathsf{T}}(t)\hat{h}(t)$ and $b = \hat{s}_x(t) + T(\lambda_2 + \mu)$ are scalar.

3.6.3. Lagrangian multiplier update

We updated the Lagrangian multiplier as follows:

$$\hat{\zeta}_{t+1} \leftarrow \hat{\zeta}_t + \mu(\hat{g}_{t+1}^* - \hat{h}_{t+1}^*),$$
$$\mu_{t+1} = min(\mu_{max}, \beta\mu_t), \tag{15}$$

where $\hat{\zeta}_t$ denotes the Lagrangian term, β indicates the scale factor, and μ_{max} is the predefined maximum value of μ. Besides, the t and $t+1$ represent the iteration index.

3.7. Model update

Online model updating is an essential component of object tracking. The appearance of an object frequently varies during the tracking process due to the background clutter, scale variations, occlusion, rotation, object deformation, and so on. It improves the efficiency and robustness of the proposed approach by training the filter through an online update technique similar to existing CF-based trackers such as Galoogahi et al. (2017) and Henriques et al. (2014). The target appearance model is defined as follows:

$$\hat{x}_t^{model} = (1 - \gamma)\hat{x}_{t-1}^{model} + \gamma\hat{x}_t, \tag{16}$$

where γ denotes the learning rate. In the training stage, we adopt $\hat{x}_t^{model}$ instead of $\hat{x}_t$ in Eq. (14) to learn the CF. Further, t represents the current frame, $t-1$ represents the learned model of the previous frame and $\hat{x}_t$ is the current learned filter.

3.8. Statistical color model

The target representation in the spatial distribution is considered an essential feature during object tracking. Initially, the spatial response value is significantly reduced when the spatial features undergo object deformation, rotation, and scale variation. As a result, the filter struggles to effectively capture the target region. To solve these problems, we integrate the SCM with the spatial-aware method that utilizes the spatial constraint (ω) to track the exact target location when the spatial distribution deviates from the target region. Specifically, spatial constraint effectively reduces the filter response in the complex background region and improves the response value in the foreground region to boost the tracker performance. Finally, the objective function

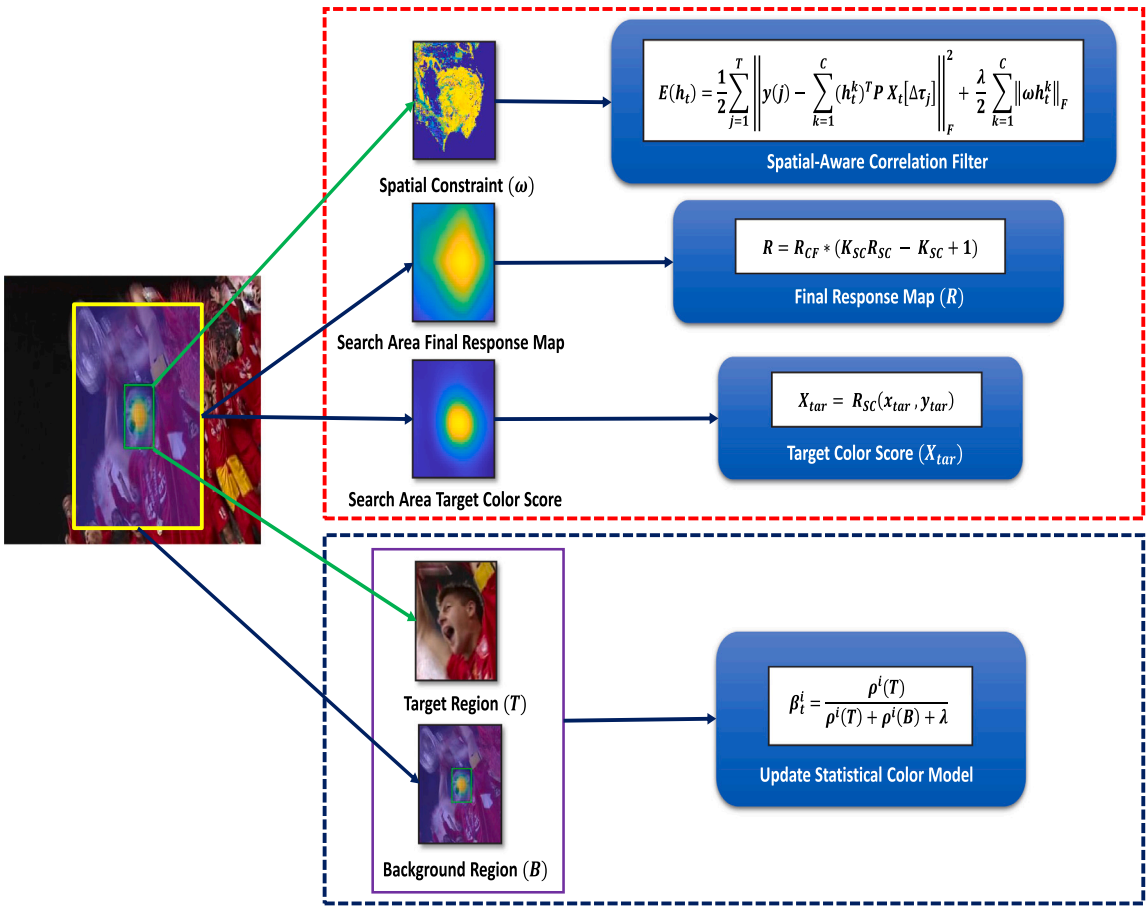


Fig. 2. The overall framework of the statistical color model consists of two main blocks. The first block presents the proposed tracker functions and the second block illustrates the updating rule of the statistical color model.

$E_{SC}(\beta)$ of the SCM (Bertinetto, Valmadre, Golodetz, Miksik, & Torr, 2016) is expressed as follows:

$$E_{SC}(\beta) = \sum_{i=1}^{M}\left[\frac{N^i(\mathcal{T})}{|\mathcal{T}|}.(\beta^i-1)^2 + \frac{N^i(\mathcal{B})}{|\mathcal{B}|}.(\beta^i)^2\right], \tag{17}$$

where M represents the number of channels, $\mathcal{T}$ represents the target area, $\mathcal{B}$ represents the background area, β indicates the histogram weight vector, $N^i(\mathcal{T})$ denotes the number of pixels in target region, and $N^i(\mathcal{B})$ represents the number of pixels in background region. Finally, the updated SCM (17) solution is defined as follows:

$$\beta_t^i = \frac{\rho^i(\mathcal{T})}{\rho^i(\mathcal{T}) + \rho^i(\mathcal{B}) + \lambda}. \tag{18}$$

The SCM is trained by computing the $\rho^i(\mathcal{T}) = N^i(\mathcal{T})/|\mathcal{T}|$ and $\rho^i(\mathcal{B}) = N^i(\mathcal{B})/|\mathcal{B}|$ for each and every channel in the image. Further, the updated parameters can be described as the following equation:

$$\rho_t(\mathcal{T}) = (1-k_l)\rho_{t-1}(\mathcal{T}) + k_l\rho_t'(\mathcal{T}), \tag{19}$$

$$\rho_t(\mathcal{B}) = (1-k_l)\rho_{t-1}(\mathcal{B}) + k_l\rho_t'(\mathcal{B}), \tag{20}$$

where t and $t-1$ represent the present instants and previous instants, respectively.

The SCM computes the response value in the target location. Moreover, the SCM is employed to compute the dynamic spatial constraint (ω), target color score ($\mathcal{X}_{tar}$), and final response map ($\mathcal{R}$), as illustrated in Fig. 2.

3.8.1. Dynamic spatial constraint (ω)

In the present section, we introduce a dynamic spatial constraint (ω), which utilizes the presented model to reduce the response values of the background region. Moreover, conventional CF trackers such as SRDCF (Danelljan et al., 2015b) and STRCF (Li, Tian et al., 2018) adopt a stable Gaussian diffusion control in the search region to enhance the filter ability. However, it is essential to utilize fixed controls in the search region rather than the sample area to design more accurate filter constraints. Alternatively, the proposed tracker employs the SCM, which produces the spatial constraint in the image samples. Moreover, the tracker initially locates the target bounding box in the current frame. Then, the SCM computes the bounding box area pixel score. Finally, the proposed tracker is trained using the new spatial constraint based on the pixel score. The spatial constraint differentiates the foreground and background regions, effectively decreasing the filter response to the background region.

3.8.2. Final response $\mathcal{R}$

The proposed tracker uses the SCM to compute the target region color response in the search region. Specifically, the SCM enhances distorted targets from the filter to enhance the tracker's robustness. Moreover, we obtain the final response $\mathcal{R}$ by integrating the SCM into our filter. Therefore, the final response $\mathcal{R}$ is described as follows:

$$\mathcal{R} = \mathcal{R}_{CF} * (\mathcal{K}_{SC}\mathcal{R}_{SC} - \mathcal{K}_{SC} + 1), \tag{21}$$

where $\mathcal{R}_{CF}$ is the response of the search region, $\mathcal{K}_{SC}$ is the constant weight, $\mathcal{R}_{SC}$ is the response map achieved by the SCM, and $\mathcal{R}$ is the maximum value in the target location.

3.8.3. Target color score $\mathcal{X}_{tar}$

The SCM computes the target color score that describes the color statistics of the bounding box regions. Further, when the target reappears after background clutter, the target's spatial distribution is

significantly changed. In particular, the target color scores $\mathcal{X}_{tar}$ generated by the SCM, which is employed as a significant indicator for the re-detection of the lost target. Finally, the target color score $\mathcal{X}_{tar}$ is obtained from $\mathcal{R}_{SC}$ as follows:

$$\mathcal{X}_{tar} = \mathcal{R}_{SC}(\mathrm{X}_{tar}, \mathrm{Y}_{tar}), \tag{22}$$

where X_{tar} and Y_{tar} represent the target position and $\mathcal{X}_{tar}$ denotes the target color score.

4. Experimental results and analysis

In this section, we first introduce the comprehensive overview of the experiment to maintain impartiality. Subsequently, we carry out an ablation study to validate the effectiveness of the proposed contributions. Afterward, we conducted the feature analysis section to illustrate the performance of our method across different features. Finally, we compare our proposed method with other state-of-the-art trackers.

4.1. Experimental setup

In the present section, we discuss our experimental setup and features representation details. The proposed TRSADCF tracking method is conducted in MATLAB. In addition, the experiments are carried out on a PC with an Intel(R) Core(TM) i5-12400 CPU at 2.50 GHz and 16 GB RAM. The proposed work employs hand-crafted and deep features such as HOG, IC, CN, and ResNet-50 that are used to extract the feature map in the CF tracker. Specifically, our study conducted a feature analysis to showcase the effectiveness of our proposed approach across various features, encompassing both hand-crafted and deep features. Moreover, these features demonstrate robust feature representations that play a key role in achieving better efficiency in object tracking.

4.2. Evaluation setting

In this study, we present the experimental results on six standard challenging datasets, such as OTB-2013 (Wu, Lim, & Yang, 2013), OTB-2015 (Wu et al., 2013), TempleColor-128 (Liang, Blasch, & Ling, 2015), UAV-123 (Mueller, Smith, & Ghanem, 2016), UAVDT (Du et al., 2018), and DTB-70 (Li & Yeung, 2017) datasets. Moreover, the performance of the proposed approach is calculated through one-pass evaluation (OPE) and center location error (CLE). Finally, the video sequences in all datasets are classified with 11 different challenging attributes as follows: low resolution (LR), out-of-view (OV), illumination variation (IV), occlusion (OCC), fast motion (FM), in-plane rotation (IPR), deformation (DEF), scale variation (SV), out-of-plane rotation (OPR), motion blur (MB), and background clutter (BC).

4.3. Parameter analysis

In our experiments, we evaluate the regularization parameters λ_1 and λ_2 in Eqs. (11) and (14). The parameters λ_1 and λ_2 are fixed values, such that, $\lambda_1 = 1.0\times10^{-3}$ and $\lambda_2 = 0.5$. We fixed the learning rate for all experiments $\gamma = 1.3 \times 10^{-2}$. Further, we analyze the η and μ parameter in Eq. (11). The experimental results are exhibited in Fig. 3 as well as the DP scores of the η and μ parameters have presented in Tables 1 and 2. Specifically, when the η parameter differs from 0.000001 to 0.00009, the maximum DP score is achieved at $\eta = 0.00003$. However, when the η value exceeds 0.00003, the DP score gradually decreases as shown in Fig. 3. Similarly, when the μ parameter varies from 0.1 to 1, the maximum DP score is obtained at $\mu = 0.5$. Nevertheless, when the μ value exceeds 0.5, the DP score gradually reduces as shown in Fig. 3. Finally, as shown in Tables 1 and 2, we obtained better tracking performance than other methods when we set $\eta = 1.3 \times 10^{-5}$ and $\mu = 0.5$.

Besides, the parameter analysis section is a crucial component of our work, enabling us to fine-tune and optimize the performance of our tracking method. We evaluate the regularization parameters, λ_1 and λ_2, along with the learning rate, γ, to ensure optimal model training. Additionally, we analyze the η and μ parameters to determine their impact on the tracking performance. Through comprehensive experimentation and analysis, we identify optimal parameter values that maximize the DP score, as demonstrated in Fig. 3. Overall, our parameter analysis serves as a backbone for refining our tracking method and achieving state-of-the-art performance compared to existing approaches.

4.4. Ablation studies

We perform an ablation analysis on the OTB-2015 dataset to evaluate the proposed tracker performance variations of each component. Initially, we compared our proposed approach with the baseline tracker and different variants of the proposed approach, such as spatial-aware (SA), temporal regularization (TR), and adaptive channel selection (ACS). The ablation analysis results are shown in Table 3 and . From these, we can demonstrate that the baseline tracker obtained the distance precision (DP) score of (80.2%) and area under curve (AUC) score of (61.0%). Further, we noticed that the tracking performance is slightly increased after including the SA term with our baseline tracker ($Baseline + SA$). Also, by integrating TR into the baseline tracker, we showed that the tracking performance significantly ($Baseline + TR$) improved in DP/AUC scores (0.5%/0.6%), respectively.

We can see that the proposed tracker efficiency is significantly enhanced by integrating the SA+TR term with the baseline tracker ($Baseline + SA + TR$). Moreover, we have examined the influence of integrating deep features into the tracking environment. By integrating the deep features with SA + TR into the BACF tracker, we confirm that the tracking performance is further improved compared to the baseline tracker ($Baseline + Deep\ Features + SA + TR$). Finally, we have proven that the proposed tracker ($Baseline + Deep\ Features + SA + TR + ACS$) achieves better performance in DP/AUC scores (91.1%)/(68.8%) compared to the baseline.

4.5. Feature analysis

The feature analysis for our proposed method involves a thorough examination of different feature components, each playing a distinct role in enhancing the overall performance. Initially, we evaluated the performance of the proposed TRSADCF method by employing different combinations of four features such as HOG, HOG+CN, HOG+CN+IC, and HOG+CN+IC+ResNet-50. Furthermore, Table 4 illustrates the DP and AUC scores on the OTB-2015 dataset. When integrating the HOG feature with the TRSADCF method (TRSADCF + HOG), we obtained the DP and AUC scores of (81.5%/62.9%). Afterward, we confirmed that the DP score of 85.2% and AUC score of 64.8% enhanced after adding the CN feature with the TRSADCF method (TRSADCF + HOG + CN). By combining the IC feature with our proposed approach (TRSADCF + HOG + CN + IC), we can see that the DP and AUC scores (86.5%/66.4%) significantly improved. Specifically, the DP and AUC scores (91.1%/68.8%) greatly increased after integrating the ResNet-50 feature with the TRSADCF method (TRSADCF + HOG + CN + IC + ResNet-50). Notably, we demonstrate that our proposed method achieved better tracking performance by incorporating hand-crafted and deep features.

4.6. Experimental evaluation

In the present section, we analyze the performance of the proposed approach on standard benchmark datasets, such as OTB-2013 (Wu et al., 2013), OTB-2015 (Wu et al., 2013), TempleColor-128 (Liang et al., 2015), UAV-123 (Mueller et al., 2016), UAVDT (Du et al., 2018) and DTB-70 (Li & Yeung, 2017). Moreover, we compare the proposed TRSADCF tracker with various other trackers such as MEVT (Sathishkumar & Joo, 2021), KCF (Henriques et al., 2014), CSK (Henriques et al.,

Table 1
Learning parameter η setting.

Parameter(η)	0.000001	0.00001	0.00002	0.00003	0.00004	0.00005	0.00006	0.00007	0.00008	0.00009
OTB-2015-DP	83.2	86.9	84.7	**91.1**	83.4	85.1	86.8	78.1	85.0	83.5

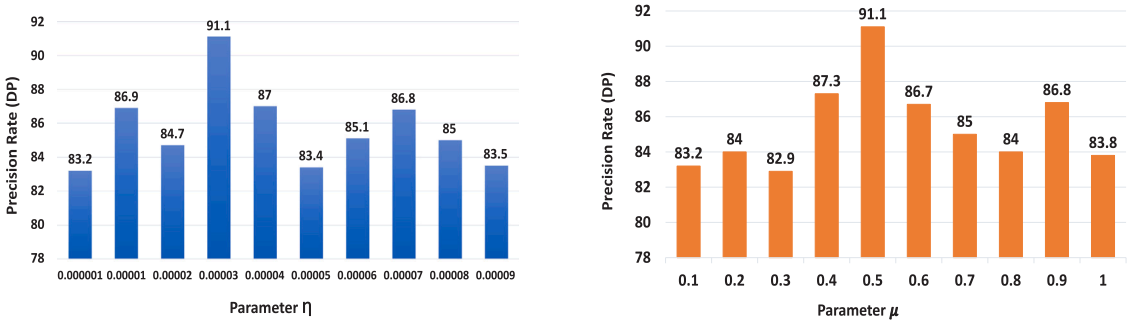
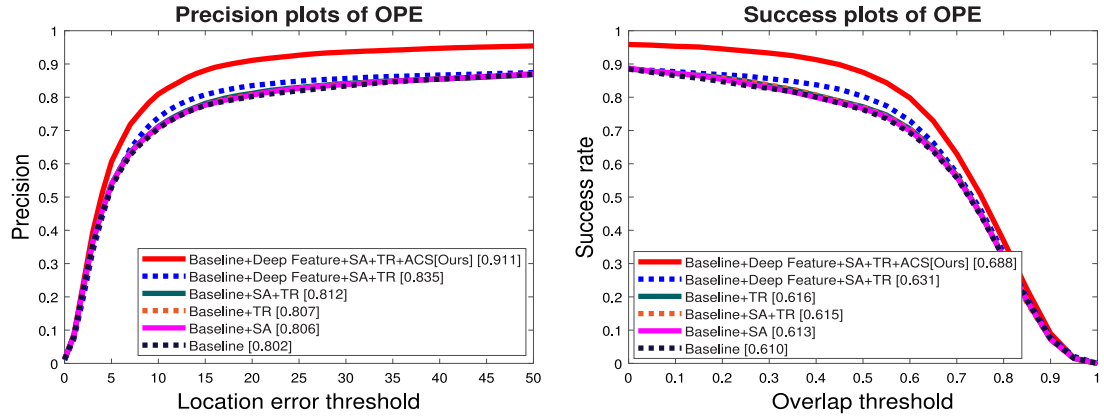


Fig. 3. Parameter analysis.



(a) (b)

Fig. 4. The DP and AUC results are illustrated in Fig. 4(a) and 4(b), respectively, through the ablation analysis of the proposed tracking method on the OTB-2015 dataset.

Table 2
Learning parameter μ setting.

Parameter(μ)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
OTB-2015-DP	83.2	84	82.9	87.3	**91.1**	86.7	85	84	86.8	83.8

Table 3
Ablation study on OTB-2015 dataset.

Methods	DP (%)	AUC (%)
Baseline + Deep Feature + SA + TR + ACS [Ours]	91.1	68.8
Baseline + Deep Feature + SA + TR	83.5	63.1
Baseline + SA + TR	81.2	61.6
Baseline + TR	80.7	61.5
Baseline + SA	80.6	61.3
Baseline	80.2	61.0

Table 4
Feature analysis on OTB-2015 dataset.

Methods	DP (%)	AUC (%)
HOG+CN+IC+ResNet-50	91.1	68.8
HOG + CN + IC	86.5	66.4
HOG + CN	85.2	64.8
HOG	81.5	62.9

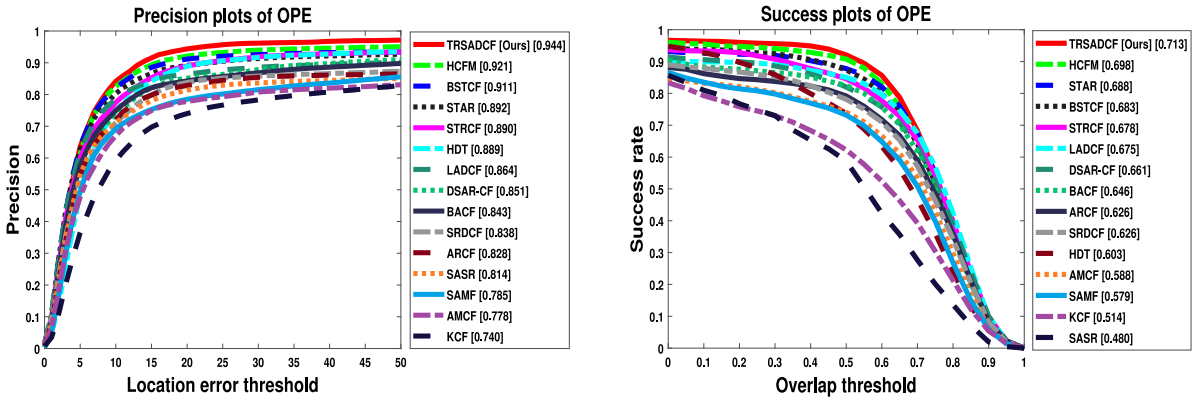
2012), BACF (Galoogahi et al., 2017), SRDCF (Danelljan et al., 2015b), STRCF (Li, Tian et al., 2018), MDNet (Nam & Han, 2016), HDT (Qi et al., 2016), SASR (Fu et al., 2020), ARCF (Huang, Fu, Li, Lin, & Lu, 2019), HCFM (Zhang, Liu, Liu, Wang, & Zhang, 2022), BSTCF (Zhang et al., 2022), LADCF (Xu, Feng, Wu, & Kittler, 2019), DSAR-CF (Feng, Han, Guo, Zhu, & Wang, 2019), STAR (Xu et al., 2022), AMCF (Li, Fu, Ding, Huang, & Pan, 2020), MCCT (Wang et al., 2018), SAMF (Li & Zhu, 2015).

4.6.1. *Evaluation on OTB-2013 dataset*

We evaluate the proposed approach with different conventional methods on the OTB-2013 dataset. The experimental results are exhibited in Fig. 5. Table 5. As shown in Fig. 5, we can see that our proposed approach obtained excellent performance with a DP score (94.4%) and an AUC score (71.3%) when compared to conventional trackers. Furthermore, when compared to the baseline method, our proposed technique attains a great improvement in DP/AUC scores (10.1%)/(6.7%). Specifically, when compared with hand-crafted feature-based tracking methods such as STRCF (Li, Tian et al., 2018), SRDCF (Danelljan et al., 2015b), and ARCF (Huang et al., 2019), the presented method achieves superior results in DP/AUC scores (5.4%/3.5%), (10.6%/8.7%), and (11.6%/8.7%), respectively. Moreover, when compared with deep convolutional feature-based approaches such as HCFM (Zhang et al., 2022), BSTCF (Zhang et al., 2022), STAR (Xu et al., 2022), and SASR (Fu et al., 2020), we demonstrated that our suggested tracker secured the best improvement of (2.3%/1.5%), (3.3%/3.0%), (5.2%/2.5%), and (13.0%/23.3%) in DP and AUC scores on the OTB-2013 benchmark dataset.

Table 5
Comparison results on the OTB-2013 dataset with 50 sequences.

Methods	Metrics	Ours	HCFM	BSTCF	STAR	STRCF	HDT	DSAR-CF	BACF	LADCF	SASR	SRDCF	ARCF	SAMF	AMCF	KCF
OTB-2013	DP	94.4	92.1	91.1	89.2	89.0	88.9	85.1	84.3	86.4	81.4	83.8	82.8	78.5	77.8	74.0
	AUC	71.3	69.8	68.3	68.8	67.8	60.3	66.1	64.6	67.5	48.0	62.6	62.6	57.9	58.8	51.4
FPS		11.5	2.7	19	5.55	5.8	10	16	25.4	1.5	–	6.2	26.2	28.5	42.4	246



(a) (b)

Fig. 5. The distance precision and success plots on the OTB-2013 dataset are exhibited in Fig. 5(a) and 5(b), respectively.

4.6.2. Statistical analysis

In this section, we performed the statistical analysis of the proposed method with 5 cutting-edge approaches on the OTB-2013 dataset, which is illustrated in Table 6. The first column represents the precision score for each sequence, while the second column shows the tracker ranking position. As shown in Table 6, our proposed approach achieved the best result in the precision score and racking position compared to the cutting-edge methods such as ARCF (Huang et al., 2019), LADCF (Xu et al., 2019), BACF (Galoogahi et al., 2017), SASR (Fu et al., 2020), and STRCF (Li, Tian et al., 2018), which is highlighted in the red color font. In addition, the LADCF and STRCF trackers achieved the second-best results in terms of precision score and ranking position, which is indicated in the blue color font. Moreover, the average precision score and ranking position of each tracker are displayed in the last row of the table. Based on the above discussions, our proposed method demonstrates superior performance in terms of the precision score and ranking position.

4.6.3. Evaluation on OTB-2015 dataset

To estimate the proposed TRSADCF tracking performance, we conduct the experimental analysis on the OTB-2015 dataset. Also, the experimental outcomes are shown in Fig. 6 and Table 7. As shown in Fig. 6, we can see that our proposed approach has obtained excellent performance with a DP score (91.1%) and AUC score (68.8%), which is better than the other conventional trackers. Moreover, the performance of the proposed approach is compared with 11 attributes as shown in Fig. 7 and comparisons of the DP and AUC scores are also shown in Tables 8 and 9. Further, the comparative performance of the TRSADCF tracker with the other five modern trackers is exhibited in Fig. 13. From these results, we noticed that the suggested approach obtained superior performance by achieving DP and AUC scores (10.9%/7.8%) compared to the baseline tracker. Moreover, when compared to hand-crafted feature-based methods such as STRCF (Li, Tian et al., 2018), ARCF (Huang et al., 2019), and SRDCF (Danelljan et al., 2015b), the proposed tracker improves great outcomes in precision and success scores (4.5%/3.2%), (10.5%/8.2%), and (12.3%/9.1%), respectively. Furthermore, compared with the deep feature-based trackers like HCFM (Zhang et al., 2022), BSTCF (Zhang et al., 2022), STAR (Xu et al., 2022), and SASR (Fu et al., 2020), we noticed that the proposed method attains the significant development in precision and success scores (0.2%/1.3%), (0.5%/1.3%), (3.6%/1.6%), and (10.4%/21.2%), on the OTB-2015 dataset, respectively. Finally, we conclude that from these deep analyses, our proposed tracker achieves the best tracking performance when compared to handcrafted and deep feature-based trackers.

4.6.4. Evaluation on TempleColor-128 dataset

We conduct the experimental evaluation on the TempleColor-128 dataset, which contains 128 sequences. The experimental outcomes are illustrated in Fig. 8 and Table 10. From these results, we noticed that the proposed approach achieved a better performance (81.7%/58.3%) in terms of precision and success scores. In addition, we confirm that the proposed TRSADCF method outperforms the baseline tracker in terms of precision and success scores (16.7%/9.3%). In particular, compared with STRCF (Li, Tian et al., 2018), DSAR-CF (Feng et al., 2019), and SRDCF (Danelljan et al., 2015b) trackers, our proposed tracker obtains excellent improvement in precision and success scores (7.3%/3.5%), (13.7%/7.8%), and (15.4%/9.8%), respectively. Finally, compared with HCFM (Zhang et al., 2022), MEVT (Sathishkumar & Joo, 2021), BSTCF (Zhang et al., 2022), STAR (Xu et al., 2022), and SASR (Fu et al., 2020) deep feature trackers, this method outperforms by (2.7%/0.3%), (2.9%/0.6%), (3.6%/0.2%), (4.3%/1.5%), and (7.8%/7.0%) in DP/AUC scores, respectively.

4.6.5. Evaluation on UAV-123 dataset

To evaluate the tracker performance, we perform the experimental analysis on the UAV-123 dataset. Moreover, the comparative performance is shown in Fig. 9 and Table 11. We observed from Table 11 that our proposed approach has obtained better performance in DP and AUC scores compared to other modern trackers. From these analyses, the presented approach attains better outcomes in precision and success scores of (75.2%/53.1%). Notably, when compared with the baseline approach, the presented TRSADCF approach increased the tracking ability in terms of precision and success scores (9.4%/19.1%). Moreover, compared with STRCF (Li, Tian et al., 2018), SRDCF (Danelljan et al., 2015b), and DSAR-CF (Feng et al., 2019) tracker, the proposed method achieves (7.1%/17.9%), (7.6%/6.7%), and (10.4%/7.0%) best improvements in terms of precision and success scores, respectively. In particular, compared with deep convolutional feature-based methods like

Table 6
The left column under each tracker name contains the precision value for each video sequence on the OTB-2013 dataset, while the right column shows the position in the ranking with respect to the rest of the trackers. The top and second best precision score and ranking position are highlighted in red and blue color fonts.

Dataset	Ours	Rank	ARCF	Rank	LADCF	Rank	BACF	Rank	SASR	Rank	STRCF	Rank
Basketball	0.98	2	1	1	1	1	0.98	2	0.26	3	1	1
Bolt	1	1	1	1	1	1	1	1	1	1	1	1
Boy	1	1	1	1	1	1	1	1	1	1	1	1
Car4	1	1	0.97	3	1	1	0.98	2	1	1	1	1
CarDark	1	1	1	1	1	1	1	1	1	1	1	1
CarScale	0.71	2	0.73	1	0.68	3	0.71	2	0.67	4	0.68	3
Coke	0.96	2	0.88	5	0.97	1	0.84	6	0.95	3	0.93	4
Couple	1	1	1	1	1	1	1	1	0.97	2	1	1
Crossing	1	1	1	1	1	1	1	1	1	1	1	1
David	1	1	1	1	1	1	1	1	1	1	1	1
David2	1	1	1	1	1	1	1	1	1	1	1	1
David3	1	1	1	1	1	1	1	1	1	1	1	1
Deer	1	1	1	1	1	1	0.97	2	1	1	1	1
Dog1	0.90	4	0.93	3	1	1	1	1	0.85	5	0.98	2
Doll	0.99	1	0.99	1	0.99	1	0.99	1	0.87	2	0.99	1
Dudek	0.95	1	0.61	5	0.86	3	0.80	4	0.58	6	0.88	2
FaceOcc1	0.34	4	0.40	3	0.62	1	0.30	5	0.34	4	0.48	2
FaceOcc2	0.90	1	0.68	6	0.76	3	0.71	5	0.72	4	0.77	2
Fish	1	1	1	1	1	1	1	1	1	1	1	1
FleetFace	0.76	1	0.64	4	0.69	2	0.53	6	0.57	5	0.68	3
Football	1	1	1	1	1	1	1	1	1	1	1	1
Football1	1	1	1	1	1	1	1	1	1	1	1	1
Freeman1	1	1	1	1	1	1	0.37	2	1	1	1	1
Freeman3	1	1	0.80	4	1	1	0.95	2	0.90	3	1	1
Freeman4	0.93	3	0.99	1	0.98	2	0.18	5	0.17	6	0.80	4
Girl	0.99	2	0.98	3	1	1	0.95	4	0.89	5	1	1
Ironman	0.80	1	0.18	4	0.18	4	0.16	5	0.61	2	0.21	3
Jogging1	0.97	2	0.98	1	0.98	1	0.98	1	0.98	1	0.98	1
Jogging2	0.99	2	0.17	5	0.95	3	0.78	4	0.95	3	1	1
Jumping	1	1	0.99	2	1	1	0.98	3	0.11	4	1	1
Lemming	1	1	0.28	5	0.92	3	0.78	4	0.28	5	0.98	2
Liquor	0.63	4	0.69	3	0.96	1	0.85	2	0.25	5	0.96	1
Matrix	0.96	1	0.01	5	0.30	4	0.31	3	0.77	2	0.29	5
Mhyang	1	1	1	1	1	1	0.93	3	0.97	2	1	1
MotorRolling	0.96	1	0.09	5	0.30	4	0.43	2	0.37	3	0.30	4
MountainBike	1	1	1	1	1	1	1	1	1	1	1	1
Shaking	1	1	0.03	4	0.98	3	0.98	3	0.99	2	0.99	2
Singer1	1	1	1	1	1	1	1	1	0.99	2	1	1
Singer2	0.99	2	1	1	0.33	4	1	1	0.36	3	0.36	3
Skating1	0.69	2	0.73	1	0.60	3	0.21	6	0.52	5	0.57	4
Skiing	1	1	0.07	5	0.12	4	0.74	2	0.14	3	0.14	3
Soccer	0.90	1	0.27	2	0.22	5	0.24	4	0.24	4	0.25	3
Subway	1	1	1	1	1	1	1	1	1	1	1	1
Suv	0.89	3	0.98	1	0.98	1	0.98	1	0.97	2	0.97	2
Sylvester	0.98	3	1	1	0.99	2	0.89	4	0.84	5	0.98	3
Tiger1	0.96	1	0.89	4	0.95	2	0.61	6	0.94	3	0.84	5
Tiger2	0.95	1	0.90	3	0.45	6	0.55	5	0.91	2	0.72	4
Trellis	1	1	1	1	0.99	2	1	1	1	1	1	1
Walking	1	1	1	1	1	1	1	1	1	1	1	1
Walking2	1	1	1	1	1	1	1	1	0.39	2	1	1
Woman	1	1	0.94	4	0.94	4	0.99	2	0.93	5	0.98	3
Average	**0.94**	**1.45**	0.79	2.27	**0.85**	**1.90**	0.81	2.49	0.76	2.62	**0.85**	**1.90**

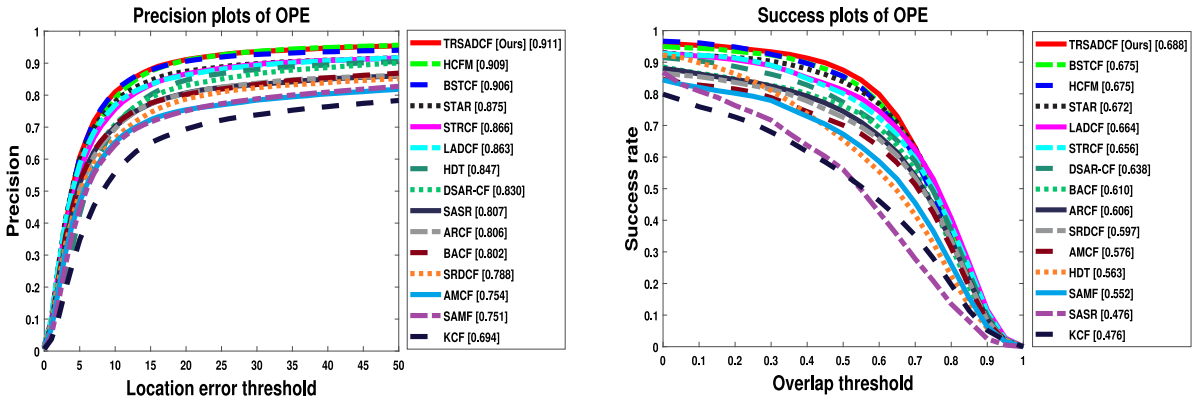
Table 7
Comparison results of existing trackers with the proposed tracker on the OTB-2015 dataset with 100 sequences.

Methods	Metrics	Ours	HCFM	BSTCF	STAR	STRCF	LADCF	HDT	DSAR-CF	BACF	SASR	SRDCF	ARCF	AMCF	SAMF	KCF
OTB-2015	DP	**91.1**	90.9	90.6	87.5	86.6	86.3	84.7	83.0	80.2	80.7	78.8	80.6	75.4	75.1	69.4
	AUC	**68.8**	67.5	67.5	67.2	65.6	66.4	56.3	63.8	61.0	47.6	59.7	60.6	57.6	55.2	47.6
FPS		**11.2**	**2.3**	**15.6**	**2.5**	**24.3**	**0.9**	**2.7**	**6.1**	**26.7**	–	**5.2**	**10.2**	**34.4**	**23.2**	**171.8**

BSTCF (Zhang et al., 2022), HCFM (Zhang et al., 2022), STAR (Xu et al., 2022), MEVT (Sathishkumar & Joo, 2021), and SASR (Fu et al., 2020), we ensure that the proposed method enhances the better outcomes in precision and success scores (0.1%/1.5%), (2.4%/2.1%), (2.9%/1.5%), (3.2%/2.8%), and (3.6%/6.3%), respectively.

4.6.6. Evaluation on UAVDT dataset

To estimate the presented TRSADCF tracking performance, we conduct the experimental analysis on the UAVDT dataset. The experimental outcomes are exhibited in Fig. 10 and Table 12. As shown in Table 12, we ensure that the presented approach obtains better results in precision score (74.1%) and success score (48.2%). In addition, when compared to the baseline approach, we noticed that the presented method increases the best outcomes by achieving (5.5%) in precision scores and (5.0%) in success scores. In addition, compared with DSAR-CF (Feng et al., 2019), SRDCF (Danelljan et al., 2015b), and STRCF (Li, Tian et al., 2018) trackers, we observe that the proposed approach gained good performance in DP and AUC scores (6.2%/5.2%),



(a) (b)

Fig. 6. The distance precision and success plots on the OTB-2015 dataset are exhibited in Fig. 6(a) and 6(b), respectively.

Table 8
Evaluation of the proposed method's efficacy on the 11-attributes OTB-2015 dataset (DP scores). Red, blue, and pink fonts indicate the top three values for each attribute, respectively.

Methods	Ours	HCFM	BSTCF	STAR	STRCF	DSAR-CF	BACF	HDT	LADCF	SASR	SRDCF	ARCF	AMCF	KCF	SAMF
IV	94.6	89.1	88.4	86.0	83.8	81.3	80.9	81.7	80.5	79.1	78.8	71.5	75.1	71.5	71.3
SV	88.8	87.6	88.5	0.84	84.3	82.4	74.3	80.6	83.5	77.5	74.4	76.8	69.6	63.3	70.5
OCC	88.6	86.4	88.3	83.6	82.6	77.1	69.7	76.6	82.4	73.2	72.5	73.1	68.4	61.7	71.9
DEF	90.0	89.5	87.5	83.8	84.1	78.9	76.3	81.9	81.3	76.7	72.9	76.5	68.5	61.1	68.4
MB	92.2	88.0	86.1	82.3	84.1	81.3	73.1	78.9	81.4	76.3	77.6	76.8	72.6	61.0	67.4
FM	90.7	87.5	84.1	81.3	79.5	78.4	75.8	79.8	77.8	77.9	75.7	75.9	70.0	61.4	65.4
OPR	90.9	89.9	89.8	87.3	85.5	81.0	77.1	80.6	83.4	79.3	74.1	77.1	71.3	67.5	73.7
IPR	92.4	91.8	87.2	83.8	80.9	77.9	75.7	84.0	80.1	78.8	73.5	77.8	74.9	69.2	71.4
OV	87.1	87.8	87.5	77.0	75.7	72.7	70.9	68.6	82.8	72.2	62.4	69.6	67.3	53.4	65.2
BC	91.6	91.4	88.0	85.9	87.3	81.5	77.8	84.4	84.4	79.1	77.5	76.0	73.8	71.3	68.9
LR	89.3	89.4	81.3	81.2	75.6	75.8	69.0	76.6	75.4	70.6	63.1	71.4	56.6	54.6	56.6

Table 9
Evaluation of the proposed method's efficacy on the 11-attributes OTB-2015 dataset (AUC scores). Red, blue, and pink fonts indicate the top three values for each attribute, respectively.

Methods	Ours	HCFM	BSTCF	STAR	STRCF	DSAR-CF	BACF	HDT	LADCF	SASR	SRDCF	ARCF	AMCF	KCF	SAMF
IV	71.8	69.0	67.6	68.2	65.2	64.8	63.0	53.2	64.6	48.6	61.0	59.6	58.6	47.5	53.1
SV	66.5	63.6	65.2	63.9	63.6	62.7	55.9	48.6	63.8	42.5	56.2	56.0	53.6	39.3	49.7
OCC	65.5	64.0	66.6	64.5	62.3	60.3	54.8	51.7	63.9	38.7	54.9	55.1	52.4	42.9	53.0
DEF	64.2	63.8	63.7	62.8	60.5	59.0	57.7	54.0	59.7	44.1	54.0	58.0	52.2	43.1	50.6
MB	72.2	80.0	67.2	66.8	67.0	64.4	57.6	56.0	65.9	45.6	60.7	61.5	57.6	45.0	53.0
FM	69.8	64.9	64.6	65.1	62.9	61.4	58.7	54.7	62.0	46.8	59.2	58.8	57.0	44.4	50.1
OPR	67.4	65.6	65.5	65.3	62.7	61.1	57.7	52.9	63.0	43.7	54.6	55.7	52.8	44.9	53.2
IPR	67.1	65.3	63.1	62.1	59.8	57.8	56.3	54.7	59.6	44.8	53.5	55.4	55.0	46.0	51.0
OV	65.9	65.6	65.4	60.0	57.8	57.5	54.0	50.0	63.0	35.4	48.9	52.9	53.5	42.9	50.9
BC	68.9	68.7	65.7	66.0	64.8	63.7	60.0	57.8	64.9	48.1	58.3	58.9	57.5	49.8	82.5
LR	64.5	62.6	63.3	60.3	55.8	57.1	50.6	0.42	56.5	38.5	48.0	49.4	42.6	30.7	45.9

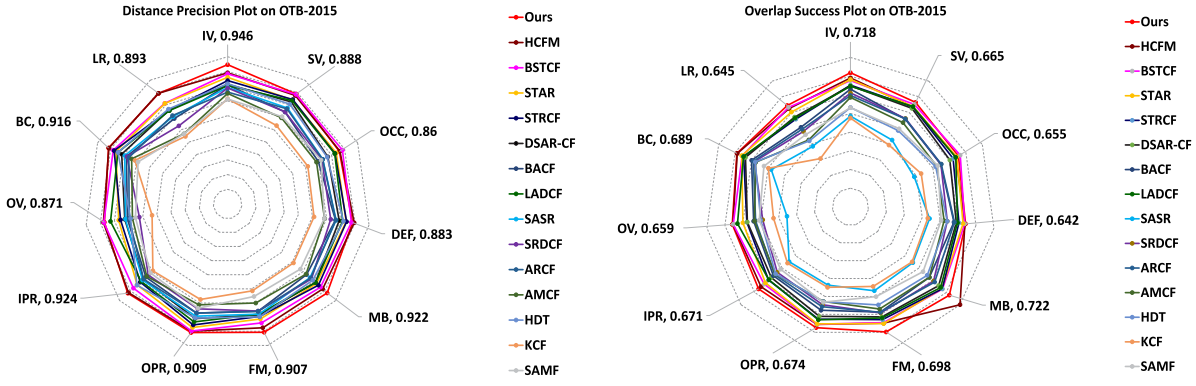
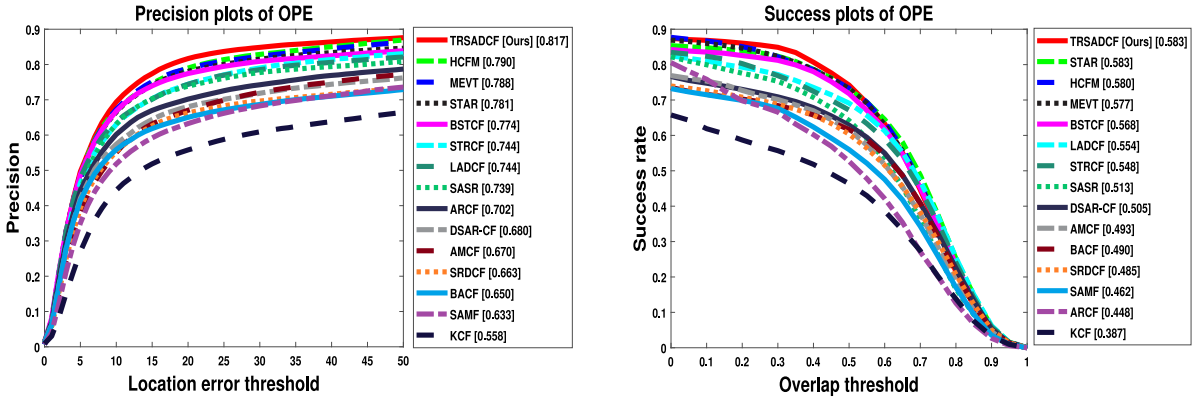


Fig. 7. Attribute-based distance precision and overlap success score plots on OTB-2015 dataset.

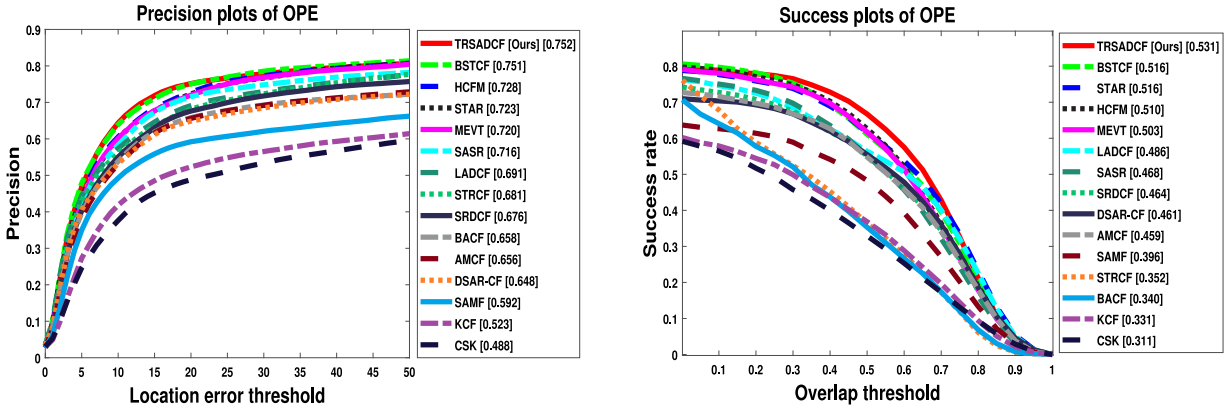
Table 10
Comparison results on the TempleColor-128 dataset with 128 sequences.

Methods	Metrics	Ours	HCFM	MEVT	STAR	BSTCF	STRCF	LADCF	SASR	ARCF	DSAR-CF	AMCF	SRDCF	BACF	SAMF	KCF
TempleColor-128	DP	**81.7**	79.0	78.8	78.1	77.4	74.4	74.4	73.9	70.2	68.0	67.0	66.3	65.0	63.3	55.8
	AUC	**58.3**	58.0	57.7	58.3	56.8	54.8	55.4	51.3	44.8	50.5	49.3	48.5	49.0	46.2	38.7



(a) (b)

Fig. 8. The distance precision and success plots on the TempleColor-128 dataset are exhibited in Fig. 8(a) and 8(b), respectively.



(a) (b)

Fig. 9. The distance precision and success plots on the UAV-123 dataset are exhibited in Fig. 9(a) and 9(b), respectively.

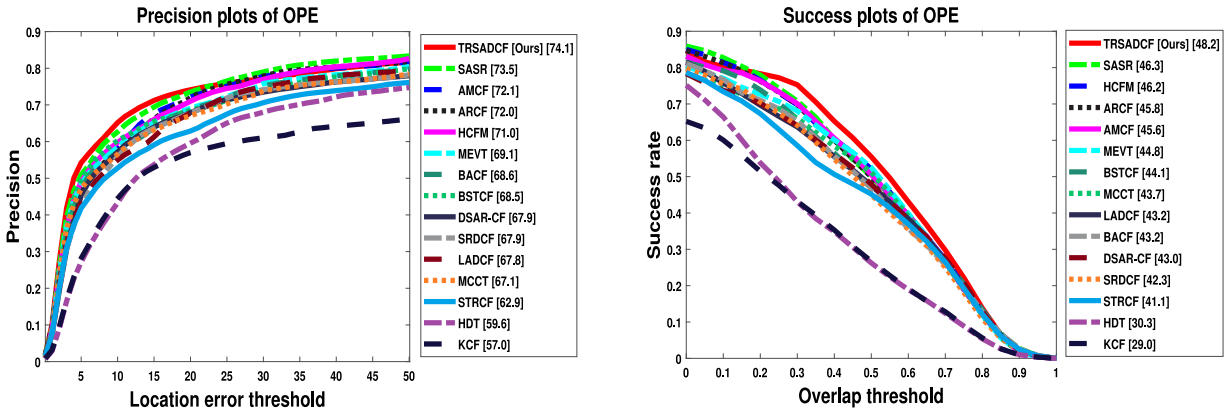
Table 11
Comparison results on the UAV-123 dataset with 123 sequences.

Methods	Metrics	Ours	BSTCF	HCFM	STAR	MEVT	SASR	LADCF	STRCF	SRDCF	BACF	AMCF	DSAR-CF	SAMF	KCF	CSK
UAV-123	DP	**75.2**	75.1	72.8	72.3	72.0	71.6	69.1	68.1	67.6	65.8	65.6	64.8	59.2	52.3	48.8
	AUC	**53.1**	51.6	51.0	51.6	50.3	46.8	48.6	35.2	46.4	34.0	45.9	46.1	39.6	33.1	31.1

Table 12
Comparison results on the UAVDT dataset with 50 sequences.

Methods	Metrics	Ours	SASR	AMCF	ARCF	HCFM	MEVT	BACF	BSTCF	DSAR-CF	SRDCF	LADCF	MCCT	STRCF	HDT	KCF
UAVDT	DP	**74.1**	73.5	72.1	72.0	71.0	69.1	68.6	68.5	67.9	67.9	67.8	67.1	62.9	59.6	57.0
	AUC	**48.2**	46.3	45.6	45.8	46.2	44.8	43.2	44.1	43.0	42.3	43.2	43.7	41.1	30.3	29.0

(6.2%/5.9%), and (11.2%/7.1%), respectively. Moreover, we observe that when compared to SASR (Fu et al., 2020), HCFM (Zhang et al., 2022), MEVT (Sathishkumar & Joo, 2021), and BSTCF (Zhang et al., 2022) deep feature trackers, the proposed approach achieves better tracking

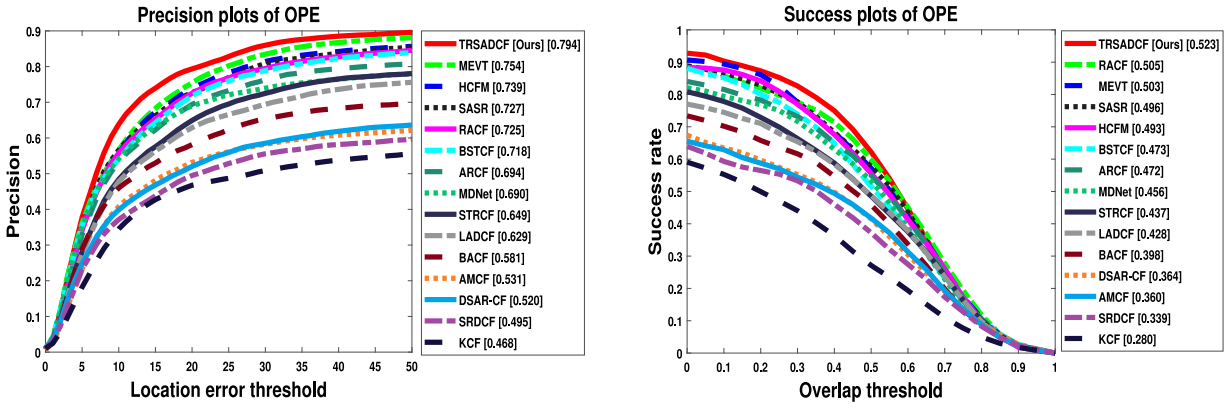


(a) (b)

Fig. 10. The distance precision and success plots on the UAVDT dataset are exhibited in Fig. 10(a) and 10(b), respectively.

Table 13
Comparison results on the DTB-70 dataset with 70 sequences.

Methods	Metrics	Ours	MEVT	HCFM	SASR	RACF	BSTCF	ARCF	MDNet	STRCF	LADCF	BACF	AMCF	DSAR-CF	SRDCF	KCF
DTB-70	DP	**79.4**	75.4	73.9	72.7	72.5	71.8	69.4	69.0	64.9	62.9	58.1	53.1	52.0	49.5	46.8
	AUC	**52.3**	50.3	49.3	49.6	50.5	47.3	47.2	45.6	43.7	42.8	39.8	36.0	36.4	33.9	28.0



(a) (b)

Fig. 11. The distance precision and success plots on the DTB-70 dataset are exhibited in Fig. 11(a) and 11(b), respectively.

performance in terms of DP and AUC scores (0.6%/1.9%), (3.1%/2.0%), (5.0%/3.4%), and (5.6%/4.1%), respectively.

4.6.7. Experiments on DTB-70 dataset

We perform the experimental analysis on the DTB-70 dataset with 70 sequences. The overall experimental evaluation results are illustrated in Fig. 11 and Table 13. As shown in Table 13, we ensure that our presented method improves the best results in terms of precision and success scores (79.4%/52.3%). Moreover, we can see that compared with the baseline tracker, the proposed TRSADCF approach improves the DP score (21.3%) and AUC score (12.5%), respectively.

Further, compared with STRCF (Li, Tian et al., 2018), DSAR-CF (Feng et al., 2019), and SRDCF (Danelljan et al., 2015b) trackers, we noticed that our method improves the better results in terms of DP/AUC scores of (14.5%/8.6%), (27.4%/15.9%), and (29.9%/18.4%), respectively. Finally, compared with the deep feature-based trackers such as MEVT (Sathishkumar & Joo, 2021), HCFM (Zhang et al., 2022), SASR (Fu et al., 2020), and BSTCF (Zhang et al., 2022), the proposed approach acquires best tracking outcomes in terms of precision and success scores (4.0%/2.0%), (5.5%/3.0%), (6.7%/2.7%), and (7.6%/5.0%), respectively.

4.7. Theoretical analysis

This study employed six standard benchmark datasets to evaluate the proposed TRSADCF approach. The proposed tracker adopts the three main approaches with hand-crafted and deep features to obtain better results across the various scenarios. Initially, we employ the multi-channel feature-based ACS approach, which is designed to effectively handle the target deviation by adaptively selecting the suitable channel. In addition, we utilize the spatial-aware term that leverages the spatial constraint and SCM methods to obtain the crucial features in the target region. Following this, the temporal regularization method is employed that considers the present and previous frames of the target region, which significantly enhances the tracking ability by utilizing historical information.

Moreover, we conducted the component analysis with different proposed approaches (Baseline, DF, SA, TR, and ACS) on the OTB-2015 dataset. To analyze the various components of the proposed approach,

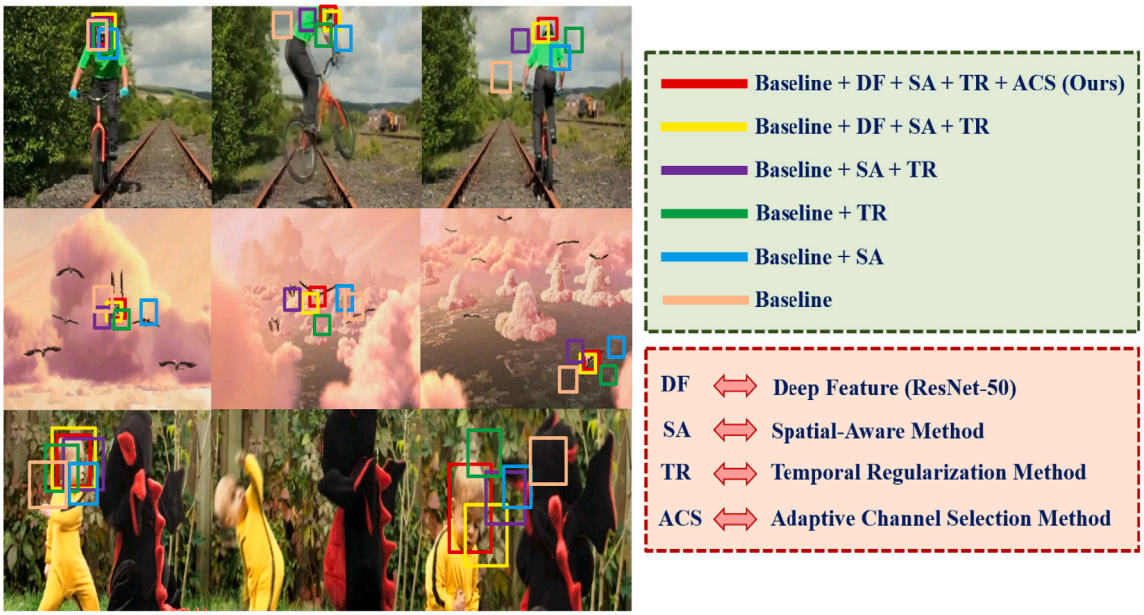


Fig. 12. Component analysis of the six different combinations of the proposed method with three challenging sequences on the OTB-2015 dataset.

we employed the three challenging sequences such as Biker, Bird1, and DragonBaby. The analysis of the proposed different components is showcased in Fig. 12. Initially, when the fast motion and object rotation occur in the Biker and DragonBaby sequences, the different combinations of proposed methods deviate from the target region. Specifically, the proposed method successfully predicts the target location despite the presence of fast motion and object rotation in the Biker and DragonBaby sequences as shown in Fig. 12. Furthermore, when object rotation occurs in the Bird1 sequence, the proposed TRSADCF approach predicts the target location accurately. As a result, we confirm that the proposed approach obtained superior performance when compared to the different combinations of proposed approaches.

4.8. Qualitative analysis

We perform the qualitative analysis of our proposed method with other state-of-the-art-trackers such as BACF (Galoogahi et al., 2017), KCF (Henriques et al., 2014), SRDCF (Danelljan et al., 2015b), HDT (Qi et al., 2016), and SASR (Fu et al., 2020). Also, the qualitative comparison results are exhibited in Fig. 13. As shown in Fig. 13, we compared the proposed tracker on six challenging video sequences. In addition, the BACF, KCF, SRDCF, HDT, and SASR trackers fail to capture the target location accurately due to the fast motion, object rotation, and occlusion. Specifically, when the target undergoes fast motion and object rotation in the Biker, Bird1, and DragonBaby sequences, the BACF, KCF, SRDCF, HDT, and SASR trackers failed to monitor the target region. Furthermore, when integrating the SCM and spatial constraint terms into the spatial-aware method, the proposed tracker tracks the target region accurately despite the presence of fast motion and object rotation in the Biker, Bird1, and DragonBaby sequences. Afterward, when encountering fast motion in the Liquor sequence, the BACF, HDT, and SASR trackers exhibit deviations from the target region. However, the proposed TRSADCF tracker performs effectively despite the challenges posed by fast motion in the Liquor sequence. Moreover, when occurring light variations and occlusion in the Shaking and Soccer sequences, the proposed tracker performs well until the end of the video sequence compared to the other trackers. Overall, we can confirm that the proposed tracker performs well in all these scenarios compared to the other conventional trackers.

4.9. Discussion

The proposed TRSADCF approach demonstrated superior performance across six standard benchmark datasets, incorporating a combination of hand-crafted and deep features to handle diverse tracking scenarios effectively. Our tracker leverages three main strategies: the multi-channel feature-based ACS approach, spatial-aware correlation filter with dynamic spatial constraints, and temporal regularization method. Initially, the adaptive channel selection approach in our method addresses target deviation by adaptively selecting the most suitable channel. This is an enhancement over conventional correlation filter methods such as BACF (Galoogahi et al., 2017) and KCF (Henriques et al., 2014), which use fixed channels, thereby limiting their adaptability in dynamic scenarios. In particular, the incorporation of spatial-aware terms and spatial constraint methods in our approach ensures crucial features in the target region are effectively captured. This builds on the work of SRDCF (Danelljan et al., 2015b), which uses spatial regularization but lacks the dynamic adaptability offered by our approach. Additionally, the temporal regularization leverages historical information from previous frames to enhance tracking robustness. This strategy extends the capabilities of existing trackers like HDT (Qi et al., 2016) and SASR (Fu et al., 2020), which primarily rely on frame-by-frame analysis without effectively utilizing temporal continuity. In conclusion, the TRSADCF approach enhances object tracking by integrating adaptive channel selection, spatial constraints, and temporal regularization, outperforming state-of-the-art trackers. Combining hand-crafted and deep features improves robustness in dynamic scenarios. Overall, the proposed TRSADCF tracker outcomes demonstrate its potential for practical applications in dynamic environments.

5. Conclusion

In this study, a novel temporal-regularized spatial-aware deep correlation filter via an adaptive channel selection approach has been presented. Notably, the proposed channel selection method has proven effective in addressing target deviations by adaptively selecting appropriate channels. This adaptive selection process plays a crucial role in accurately determining the target's position throughout the tracking scenario. Meanwhile, a spatial-aware correlation filter with dynamic spatial constraints has been proposed to effectively track the

Fig. 13. Comparison of TRSADCF tracker with other five modern trackers such as BACF (Galoogahi et al., 2017), KCF (Henriques et al., 2014), SRDCF (Danelljan et al., 2015b), HDT (Qi et al., 2016), and SASR (Fu et al., 2020) in six different challenging sequences on the OTB-2015 dataset. The sequential order is arranged from top to bottom as follows: Biker, Bird1, DragonBaby, Liquor, Shaking, and Soccer.

target when the spatial distribution varies from the target region. Following this, the SCM utilized a dynamic spatial constraint to distinguish between foreground and background regions in the response map, which effectively improved the filter response in the foreground region. By employing a spatial-aware correlation filter with dynamic spatial constraints, we mitigate the challenges of light variation and fast motion, which emphasizes stable spatial features, adapts to object appearance changes, and adjusts to fast motion in real-time scenarios. Subsequently, we designed a temporal regularization approach that improved target accuracy in cases of large appearance variations and cluttered backgrounds. This temporal regularization method considered the present and previous frames of the target region, which significantly enhanced tracking ability by leveraging historical information. In the end, the experimental results on the benchmark datasets OTB-2013, OTB-2015, TempleColor-128, UAV-123, UAVDT, and DTB-70 have demonstrated the superiority of our proposed method compared to other trackers.

Despite the proposed TRSADCF tracker achieving better tracking performance, it may encounter challenges with motion blur and low resolution. In future work, we will address the challenges posed by motion blur and low resolution in the current TRSADCF tracker by incorporating advanced deep feature extraction techniques and transformer-based methods. These enhancements aim to improve the robustness and accuracy of the tracker under adverse conditions. Specifically, we will leverage convolutional neural networks (CNNs) and vision transformers (ViTs) to capture more discriminative features and better model complex motion patterns. Additionally, we plan to explore multi-object tracking (MOT) to extend the tracker's capabilities beyond single object scenarios. This will involve developing algorithms that can simultaneously track multiple objects while maintaining high performance.

CRediT authorship contribution statement

Sathiyamoorthi Arthanari: Writing – original draft, Methodology, Conceptualization. **Dinesh Elayaperumal:** Writing – review & editing, Validation, Formal analysis. **Young Hoon Joo:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the Basic Science Research Program through the National Research Foundation, Republic of Korea (NRF) funded by the Ministry of Education of South Korea (NRF-2016R1A6A1A03013567, NRF-2021R1A2B5B01001484).

Data availability

Data will be made available on request.

References

Bertinetto, Luca, Valmadre, Jack, Golodetz, Stuart, Miksik, Ondrej, & Torr, Philip H. S. (2016). Staple: Complementary learners for real-time tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1401–1409).

Boyd, Stephen, Parikh, Neal, Chu, Eric, Peleato, Borja, & Eckstein, Jonathan (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Danelljan, Martin, Hager, Gustav, Khan, Fahad Shahbaz, & Felsberg, Michael (2015a). Convolutional features for correlation filter based visual tracking. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 58–66).

Danelljan, Martin, Hager, Gustav, Khan, Fahad Shahbaz, & Felsberg, Michael (2015b). Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 4310–4318).

Dinesh, Elayaperumal, & Joo, Young Hoon (2021). Aberrance suppressed spatio-temporal correlation filters for visual object tracking. *Pattern Recognition*, *115*, Article 107922.

Du, Dawei, Qi, Yuankai, Yu, Hongyang, Yang, Yifan, Duan, Kaiwen, Li, Guorong, et al. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision* (pp. 370–386).

Elayaperumal, Dinesh, & Joo, Young Hoon (2023). Learning spatial variance-key surrounding-aware tracking via multi-expert deep feature fusion. *Information Sciences*, *629*, 502–519.

Fan, Nana, Li, Xin, Zhou, Zikun, Liu, Qiao, & He, Zhenyu (2021). Learning dual-margin model for visual tracking. *Neural Networks*, *140*, 344–354.

Fan, Heng, & Ling, Haibin (2017). Sanet: Structure-aware network for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 42–49).

Feng, Wei, Han, Ruize, Guo, Qing, Zhu, Jianke, & Wang, Song (2019). Dynamic saliency-aware regularization for correlation filter-based object tracking. *IEEE Transactions on Image Processing*, *28*(7), 3232–3245.

Fu, Changhong, Xiong, Weijiang, Lin, Fuling, & Yue, Yufeng (2020). Surrounding-aware correlation filter for UAV tracking with selective spatial regularization. *Signal Processing*, *167*, Article 107324.

Galoogahi, Kiani, Hamed, Ashton Fagg, & Lucey, Simon (2017). Learning background-aware correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision* (pp. 1135–1143).

Gu, Fengwei, Lu, Jun, & Cai, Chengtao (2022). RPformer: A robust parallel transformer for visual tracking in complex scenes. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–14.

Han, Wei, Lekamalage, Chamara Kasun Liyanaarachchi, & Huang, Guang-Bin (2022). Efficient joint model learning, segmentation and model updating for visual tracking. *Neural Networks*, *147*, 175–185.

Henriques, Joao F., Caseiro, Rui, Martins, Pedro, & Batista, Jorge (2012). Exploiting the circulant structure of tracking-by-detection with kernels. In *Computer vision–ECCV 2012: 12th European conference on computer vision, florence, Italy, Octobe 7–13, 2012, proceedings, part IV 12* (pp. 702–715). Springer Berlin Heidelberg.

Henriques, João F., Caseiro, Rui, Martins, Pedro, & Batista, Jorge (2014). High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(3), 583–596.

Hu, Hongwei, Ma, Bo, Shen, Jianbing, & Shao, Ling (2017). Manifold regularized correlation object tracking. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1786–1795.

Huang, Ziyuan, Fu, Changhong, Li, Yiming, Lin, Fuling, & Lu, Peng (2019). Learning aberrance repressed correlation filters for real-time UAV tracking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2891–2900).

Kalal, Zdenek, Mikolajczyk, Krystian, & Matas, Jiri (2011). Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(7), 1409–1422.

Li, Yiming, Fu, Changhong, Ding, Fangqiang, Huang, Ziyuan, & Pan, Jia (2020). Augmented memory for correlation filters in real-time UAV tracking. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1559–1566). IEEE.

Li, Xin, Liu, Qiao, Fan, Nana, Zhou, Zikun, He, Zhenyu, & Jing, Xiao-yuan (2020). Dual-regression model for visual tracking. *Neural Networks*, *132*, 364–374.

Li, Feng, Tian, Cheng, Zuo, Wangmeng, Zhang, Lei, & Yang, Ming-Hsuan (2018). Learning spatial–temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4904–4913).

Li, Bo, Yan, Junjie, Wu, Wei, Zhu, Zheng, & Hu, Xiaolin (2018). High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8971–8980).

Li, Siyi, & Yeung, Dit-Yan (2017). Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1)*.

Li, Yang, & Zhu, Jianke (2015). A scale adaptive kernel correlation filter tracker with feature integration. In *Computer vision-ECCV 2014 workshops: zurich, Switzerland, September 6–7 and 12, 2014, proceedings, part II 13* (pp. 254–265). Springer International Publishing.

Liang, Pengpeng, Blasch, Erik, & Ling, Haibin (2015). Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing*, *24*(12), 5630–5644.

Lu, Allen, Zontak, Maria, Parajuli, Nripesh, Stendahl, John C., Boutagy, Nabil, Eberle, Melissa, et al. (2017). Dictionary learning-based spatiotemporal regularization for 3D dense speckle tracking. In *Medical imaging 2017: ultrasonic imaging and tomography (Vol. 10139)* (pp. 17–24). SPIE.

Moorthy, Sathishkumar, & Joo, Young Hoon (2023). Learning dynamic spatial–temporal regularized correlation filter tracking with response deviation suppression via multi-feature fusion. *Neural Networks*, *167*, 360–379.

Mueller, Matthias, Smith, Neil, & Ghanem, Bernard (2016). A benchmark and simulator for uav tracking. In computer vision–ECCV 2016. In *14th European conference, amsterdam, the Netherlands, October 11–14, 2016, proceedings, part I 14* (pp. 445–461). Springer International Publishing.

Nai, Ke, Li, Zhiyong, & Wang, Haidong (2022). Dynamic feature fusion with spatial–temporal context for robust object tracking. *Pattern Recognition*, *130*, Article 108775.

Nam, Hyeonseob, & Han, Bohyung (2016). Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4293–4302).

Nie, Feiping, Huang, Heng, Cai, Xiao, & Ding, Chris (2010). Efficient and robust feature selection via joint l2, 1-norms minimization. *Advances in Neural Information Processing Systems*, *23*.

Qi, Yuankai, Zhang, Shengping, Qin, Lei, Yao, Hongxun, Huang, Qingming, Lim, Jongwoo, et al. (2016). Hedged deep tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4303–4311).

Sathishkumar, Moorthy, & Joo, Young Hoon (2021). Multi-expert visual tracking using hierarchical convolutional feature fusion via contextual information. *Information Sciences*, *546*, 996–1013.

Sherman, Jack, & Morrison, Winifred J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, *21*(1), 124–127.

Sun, Xinglong, Han, Guangliang, Guo, Lihong, Yang, Hang, Wu, Xiaotian, & Li, Qingqing (2022). Two-stage aware attentional siamese network for visual tracking. *Pattern Recognition*, *124*, Article 108502.

Teng, Zhu, Xing, Junliang, Wang, Qiang, Lang, Congyan, Feng, Songhe, & Jin, Yi (2017). Robust object tracking based on temporal and spatial deep networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1144–1153).

Valmadre, Jack, Bertinetto, Luca, Henriques, Joao, Vedaldi, Andrea, & Torr, Philip H. S. (2017). End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2805–2813).

Wang, Ning, Zhou, Wengang, Tian, Qi, Hong, Richang, Wang, Meng, & Li, Houqiang (2018). Multi-cue correlation filters for robust visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4844–4853).

Wen, Jiajun, Chu, Honglin, Lai, Zhihui, Xu, Tianyang, & Shen, Linlin (2023). Enhanced robust spatial feature selection and correlation filter learning for UAV tracking. *Neural Networks*, *161*, 39–54.

Wu, Yue, Lan, Yuan, Zhang, Luchan, & Xiang, Yang (2023). Feature flow regularization: Improving structured sparsity in deep neural networks. *Neural Networks*, *161*, 598–613.

Wu, Yi, Lim, Jongwoo, & Yang, Ming-Hsuan (2013). Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2411–2418).

Wu, Ruixu, Wen, Xianbin, Yuan, Liming, Xu, Haixia, & Liu, Yanli (2024). Visual tracking based on deformable transformer and spatiotemporal information. *Engineering Applications of Artificial Intelligence*, *127*, Article 107269.

Xu, Tianyang, Feng, Zhen-Hua, Wu, Xiao-Jun, & Kittler, Josef (2019). Learning adaptive discriminative correlation filters via temporal consistency preserving spatial feature selection for robust visual object tracking. *IEEE Transactions on Image Processing*, *28*(11), 5596–5609.

Xu, Libin, Kim, Pyoungwon, Wang, Mengjie, Pan, Jinfeng, Yang, Xiaomin, & Gao, Mingliang (2022). Spatio-temporal joint aberrance suppressed correlation filter for visual tracking. *Complex & Intelligent Systems*, 1–13.

Zhang, Jianming, Liu, Yang, Liu, Hehua, Wang, Jin, & Zhang, Yudong (2022). Distractor-aware visual tracking using hierarchical correlation filters adaptive selection. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*(6), 6129–6147.

Zhang, Hao, Piao, Yan, & Qi, Nan (2024). STFT: Spatial and temporal feature fusion for transformer tracker. *IET Computer Vision*, *18*(1), 165–176.

Zou, Hui, & Hastie, Trevor (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.