mathematics

MDPI

Article

Learning a Context-Aware Environmental Residual Correlation Filter via Deep Convolution Features for Visual Object Tracking

Sachin Sakthi Kuppusami Sakthivel , Sathishkumar Moorthy, Sathiyamoorthi Arthanari, Jae Hoon Jeong and Young Hoon Joo *

School of IT Information and Control Engineering, Kunsan National University, 588 Daehak-ro, Gunsan-si 54150, Republic of Korea; sachin@kunsan.ac.kr (S.S.K.S.); sathishkumar@kunsan.ac.kr (S.M.); sathya@kunsan.ac.kr (S.A.); jh7129@kunsan.ac.kr (J.H.J.)

* Correspondence: yhjoo@kunsan.ac.kr

Abstract: Visual tracking has become widespread in swarm robots for intelligent video surveillance, navigation, and autonomous vehicles due to the development of machine learning algorithms. Discriminative correlation filter (DCF)-based trackers have gained increasing attention owing to their efficiency. This study proposes "context-aware environmental residual correlation filter tracking via deep convolution features (CAERDCF)" to enhance the performance of the tracker under ambiguous environmental changes. The objective is to address the challenges posed by intensive environment variations that confound DCF-based trackers, resulting in undesirable tracking drift. We present a selective spatial regularizer in the DCF to suppress boundary effects and use the target's context information to improve tracking performance. Specifically, a regularization term comprehends the environmental residual among video sequences, enhancing the filter's discrimination and robustness in unpredictable tracking conditions. Additionally, we propose an efficient method for acquiring environmental data using the current observation without additional computation. A multi-feature integration method is also introduced to enhance the target's presence by combining multiple metrics. We demonstrate the efficiency and feasibility of our proposed CAERDCF approach by comparing it with existing methods using the OTB2015, TempleColor128, UAV123, LASOT, and GOT10K benchmark datasets. Specifically, our method increased the precision score by 12.9% in OTB2015 and 16.1% in TempleColor128 compared to BACF.

Keywords: selective spatial regularizer; context-aware; object tracking; correlation filters; multi-feature fusion

MSC: 68U10; 68T45

check for updates

Citation: Kuppusami Sakthivel, S.S.; Moorthy, S.; Arthanari, S.; Jeong, J.H.; Joo, Y.H. Learning a Context-Aware Environmental Residual Correlation Filter via Deep Convolution Features for Visual Object Tracking. *Mathematics* **2024**, *12*, 2279. https://doi.org/10.3390/math12142279

Academic Editors: Laura Cornejo-Bueno and Jorge Pérez-Aracil

Received: 12 June 2024
Revised: 9 July 2024
Accepted: 18 July 2024
Published: 21 July 2024

1. Introduction

Developing continuous object-tracking methods is a hot and demanding area in the field of video surveillance systems. The aim of object tracking is to determine the target's location in subsequent video frames. Visual object tracking (VOT) has been widely used in every aspect of practical situations, from autonomous vehicles to robotic-assisted surgery. Furthermore, fields beyond object tracking, such as classification, segmentation, and detection, also play crucial roles in advancing computer vision applications [1,2]. Technically, it has drawn widespread concerns in several applications ranging from automatic driving cars, homeland security, and traffic monitoring to robotics [3–6]. For the development of good VOT methods, numerous datasets, including OTB50, OTB2015, TempleColor128, UAV123, LASOT, and GOT10K, were introduced as benchmarks for the evaluation of each method. Moreover, several object-tracking challenges have intrigued the curiosity of many scholars and aided in the advancement of the field. Despite significant advances, VOT remains a difficult task in computer vision because of a variety of challenges, including variation illumination, cluttered background, scale estimation, deformation, and partial occlusion. Therefore, it is important to improve the response of the tracker and its adaptability to target variations in order to continuously and smoothly track a moving object without interruption.

DCF-based trackers have extensive popularity due to their remarkable tracking speed and sensible adaptability to photometric and geometric differences. More specifically, the authors of [7–11] proposed some DCF-based trackers that can train filters using ridge regression to learn target positions. According to these methods, it can be seen that the environmental information of the tracking target has a significant impact on the tracking performance. Bolme et al. [7] presented a DCF for object tracking in the MOSSE tracker that learns a filter through the minimum output sum of squared error. Since then, numerous advancements have been made to enhance the efficacy and precision of DCF-based trackers, which achieved the best results in current benchmarks. Inspired by [7], the authors in [12] proposed a CSK tracker which replaces the sum of squared error with a circulant structure of kernels. Furthermore, Henriques et al. [8] have since reinterpreted the existing CSK by utilizing the kernel function and HOG features. In other words, due to the circular assumption, DCF has a boundary effect, which is one of its most significant drawbacks. As a result, it is known that CF trackers typically contain limited ambient information that leads to some tracking failures, such as motion blur, deformation, or a cluttered background. To overcome this constraint, the authors in [13] presented an approach which considers contextual information and integrates it into the filter. By introducing this framework, most CA–CF trackers have achieved better tracking results [13–15]. In addition, the environmental variations significantly hinder trackers from ignoring temporal environmental changes between successive frames, leading to undesirable tracking failure. To solve these issues, the authors of [16–18] considered more reasonable environmental samples during CF learning. Siamese-based trackers, renowned for their robustness and accuracy in visual object tracking tasks, have garnered significant attention due to their ability to learn feature representations and perform effective similarity matching between target and search regions [19–21]. Nevertheless, the problems previously discussed remained unsolved. To address this problem, the authors of [22] presented a tracking method to deal with environmental changes.

Feature representation plays an important role in VOT, which can extract useful information to describe the target appearance. Typically, the best feature representation should model the appearance of the object effectively and efficiently. To do this, various hand-crafted (HC) features were utilized in DCF tracking, which includes color histograms, Haar-like histograms, and histograms of oriented gradients. Nonetheless, these features still performed poorly in distinguishing targets by the appearance of dynamic objects. To address the aforementioned constraints, considerable research effort has been devoted to exploring deep learning-based feature representation, resulting in significant advancements. Since the advent of deep learning technologies, convolutional neural networks (CNNs) have consistently demonstrated their exceptional ability to represent data in many different computer vision applications. For instance, the authors of [23] brought significant performance improvement using the convolution function in the CF tracking framework. In certain difficult circumstances, a single feature is unable to store combined global and local data, which degrades performance. This is where the concept of fusing several features to improve tracking efficiency originated. Consequently, the appearance model is exemplified by the transition from a single-channel to a multi-channel illustration of features.

Inspired by the above analysis, in this study, a context information-based environmental residual CF with a multi-feature fusion strategy is presented. A visual representation of the devised methodology is depicted in Figure 1.

The key contributions of our study can be outlined as follows:

1. In this study, an environmental residual term is introduced with the purpose of maintaining the tracker's smooth operation despite fluctuations in environmental con-

ditions. This addition serves to stabilize the tracker, enabling it to adapt effectively to changes in its surroundings, thereby augmenting its overall reliability and robustness.

2. The proposed correlation filter framework effectively reduces tracking drift caused by external distractions through the strategic inclusion of context patches around the target during the filter learning phase, enhancing overall tracking accuracy and reliability.
3. In this approach, a selective spatial regularizer is incorporated to safeguard essential object information while concurrently mitigating boundary effects. This regularization technique selectively preserves crucial object details, thereby enhancing tracking accuracy and robustness across varying environmental conditions and object orientations.
4. A multi-feature fusion approach, incorporating handcrafted and deep features, is introduced to enhance the tracker's performance, leveraging the complementary strengths of these feature types across various tracking scenarios.
5. In conclusion, experimental evaluations conducted on benchmark datasets such as OTB2015, TempleColor128, UAV123, LASOT, and GOT10K affirm the favorable performance of the proposed tracker when compared to other state-of-the-art tracking methods.

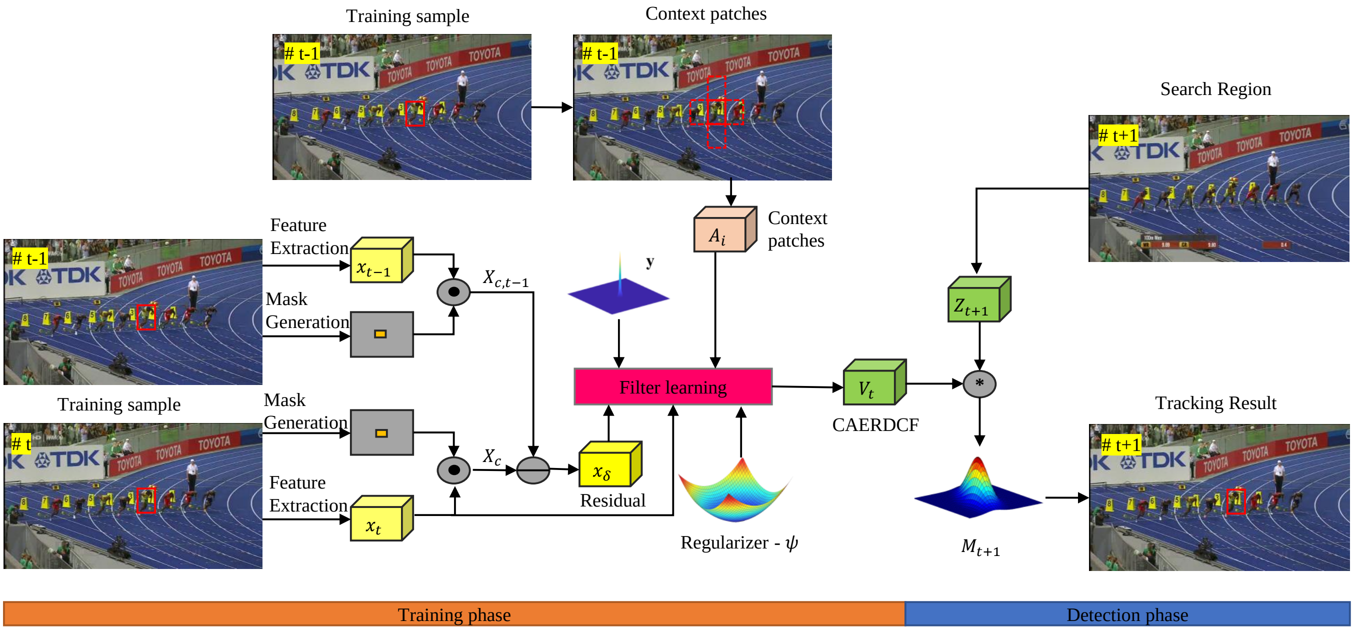


Figure 1. The overall framework of the CAERDCF method. During the training stage, CAERDCF utilizes both environment residual and context information to produce an efficient filter v_t. Finally, our proposed approach tracks the target in the detection stage.

2. Related Works

Several tracking algorithms for object tracking have been proposed through many studies. Recently, the VOT tracking community has carried out extensive research and has achieved impressive results in tracking. The sections that follow examine some well-known tracking approaches that are primarily linked to CAERDCF. Each section is interconnected to an important component of CAERDCF.

2.1. Correlation Filter-Based Tracking

In general, tracking algorithms are characterized as either generative or discriminative. Generative tracking techniques create a model of the target object's presence in the keyframe. After that, the tracking problem is modeled to search the target region whose appearance is most similar to the current frame. Numerous strategies have been proposed for understanding an appearance model, such as the Kalman filter, particle filter, mean-shift, subspace learning, and sparse representations. Instead, the discriminative model learns a binary classifier for distinguishing the target object from its environment. Machine learning concepts are used to train the classifier in discriminative methods, and then the optimal region in the next frame is found using the learned classifier, which reflects higher tracking performance than generative tracking methods. On a variety of difficult benchmark datasets, the trackers based on DCFs exhibit excellent precision and success rates. The authors in [24] presented a DCF tracking method using the color names feature (CN) [25], which is complementary to histograms of oriented gradients (HOG) and represents the outline of the object.

On the other hand, subsequent trackers, which exhibit cutting-edge performance while preserving significant computational performance frequently use the feature fusion of handcrafted features like HOG and CN as in [26,27]. To cope with the varying size of the target, the authors in [28] proposed a tracking algorithm that learns a separate filter for scale estimation. The authors in [27] merged color features and HOG features to provide real-time tracking. The boundary effect is a major problem in DCF-based trackers owing to the cyclic assumption, which deteriorates the performance. Since then, numerous solutions have emerged to solve the boundary effect problem. Furthermore, Martin et al. [29] proposed a regularization term which lessens the impact from the background. Similarly, the authors in [30] used the differences between the foreground and background to create a discriminative classifier. The tracker is trained using negative samples by immediately employing a circulant shift to the full frame and clipping major elements from each shifted sample. Inspired by [29], Li et al. [31] employed a temporal regularization in the STRCF tracker, resulting in spatial–temporal regularized DCFs, which results in the best accuracy and speed. Even though CF-based methods have gained tremendous progress, the trackers are still hindered by the boundary effect, which will be an inherent fault resulting from a circulant shift. In addition, the majority of such trackers hardly evaluate the gray feature or the HOG feature, excluding more significant features. However, real-world experiments have proven that these methods still have less characteristic information because of the simplicity of HC features.

2.2. CNN Feature-Based Tracking

Deep learning-based tracking techniques are becoming more popular due to their enhanced speed and efficacy. Recently, convolutional neural networks (CNN) have been applied in VOT for deep feature extraction thanks to their outstanding representation abilities [32]. For the first time, the authors of [23] presented convolutional features that are extracted from pre-trained VGG-Net in the CF tracker and achieved remarkable performance improvement. The authors of [33] introduced feature maps of various resolutions and proposed a method for mapping learning problems to continuous spatial domain. As a feature fusion strategy, the authors of [34] developed a multi-expert tracker that combines various features to form an expert pool. In contrast, the HDT tracker, described in [35], makes use of multi-layer features alongside the Hedge algorithm to enhance the effectiveness of tracking. However, the aforementioned trackers use limited layers for feature extraction, which are less robust for targeting appearance variations. In this study, a novel tracker using the functions of multiple convolutional layers that effectively improves the tracking performance compared to the conventional CNN-based tracker is proposed. Figure 2 illustrates the feature extraction and training stages of the CAERDCF tracker.

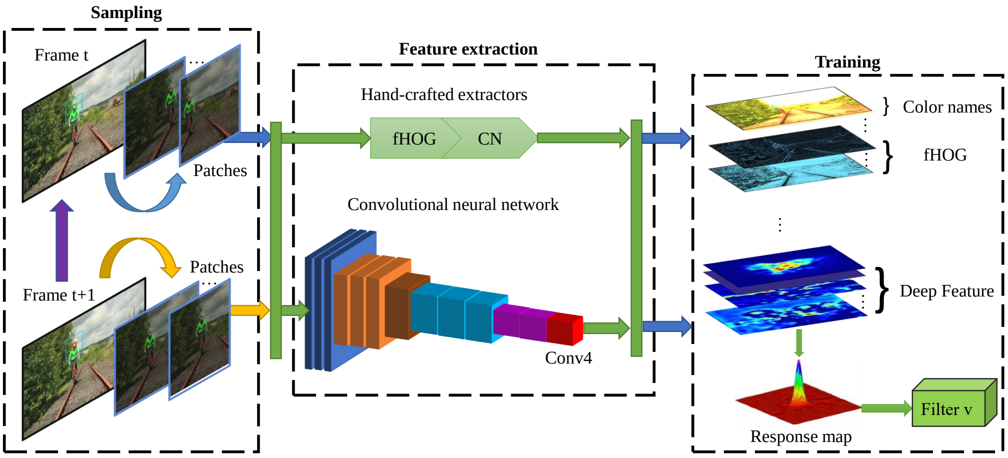


Figure 2. CNN feature-based tracking: In the sampling stage we obtain context patches from the input image which are then used to extract multi features in the feature extraction process. Finally, in the training stage, we use a multi-feature fusion strategy to combine hand-crafted features as well as deep features to formulate a filter which tracks the object efficiently.

3. Proposed Methodology

3.1. Background-Aware Correlation Filter

For the proposed CAERDCF method, we consider the BACF algorithm [30] as our baseline tracker. A typical filter $h \in R^N$ of the framework can be obtained as follows:

$$E(h) = \frac{1}{2}\|\sum_{d=1}^{D} Bx^d \odot h^d - y\|_2^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\|h^d\|_2^2, \tag{1}$$

where $x^d \in R^N$ is the sample vector extracted from a given image sample. The desired response of the object patch is denoted as $y \in R^N$. The objective of the CF is to find the target patch in the consecutive frames using filter $h \in R^N$. The correlation output y is obtained from filter h and input sample x. Moreover, the filter h is generated by minimizing the loss function.

3.2. Selective Spatial Regularized CF

Existing methods employed in correlation filter tracking commonly encounter boundary effects, wherein uniform regularization across spatial regions may result in diminished performance, particularly in scenarios characterized by dynamism and clutter. To circumvent these challenges, we propose the utilization of a selective spatial regularization technique within our manuscript. Diverging from conventional methodologies, which uniformly impose penalties across all spatial regions, our approach dynamically adjusts regularization parameters contingent upon local image attributes. Leveraging deep convolutional features, our method discerns and prioritizes salient spatial regions, facilitating customized regularization that effectively suppresses noise and augments the tracker's discriminative capabilities. By mitigating boundary effects and affording adaptability to heterogeneous environmental conditions, our selective spatial regularization method signifies a notable advancement in environmental tracking systems. Hence, our objective function becomes

$$E(h) = \frac{1}{2}\|\sum_{d=1}^{D} Bx^d \odot h^d - y\|_2^2 + \frac{1}{2}\sum_{d=1}^{D}\|\varphi h^d\|_2^2. \tag{2}$$

In this context, the symbol $\odot$ denotes element-wise multiplication, while φ represents the selective spatial regularizer:

$$\varphi(p,q) = \begin{cases} \varphi\left(\frac{w}{2}, \frac{h}{2}\right), & \text{if}(p,q) \in \kappa, \\ \varphi_0 + \eta\left(\frac{p^2}{w^2}, \frac{q^2}{h^2}\right), & \text{otherwise} \end{cases} . \tag{3}$$

In this context, w refers to width, and h signifies height. φ_0 represents the minimum value of the regularizer. More precisely, κ denotes the area enclosed within the circumellipse of the groundtruth. This helps the tracker to hold the precise information within the bounding box. without it, the tracker would be unable to effectively assess the scale variation of the object.

3.3. Context-Aware Selective Spatial Regularized CF

Due to the boundary effect, CF-based methods must multiply the cosine window formerly implementing the filter computation to repress the response of the surrounding information. It results in a near-zero response of background pixels and a smaller search area for the tracker. Building upon the methodology outlined in [13], we integrate a context-aware framework into our proposed method, effectively harnessing background information from the surrounding object. The tracker's effectiveness is enhanced by the incorporation of the context data via the correlation filter. A new equation is derived as follows:

$$\begin{aligned} E(h) = & \frac{1}{2}\left\| \sum_{d=1}^{D} Bx_0^d \odot h^d - y \right\|_2^2 + \frac{1}{2}\left\| \sum_{d=1}^{D} \varphi h^d \right\| \\ & + \frac{\lambda}{2}\left\| \sum_{k=1}^{K}\sum_{d=1}^{D} Bx_k^d \odot h^d \right\|_2^2 . \end{aligned} \tag{4}$$

In this context, k refers to the context patch in x_k, where x_0 denotes target, and λ represents the regularization parameter.

3.4. Context-Aware Environmental Residual CF

The overall objective function can be obtained as follows:

$$\begin{aligned} E(h) = & \frac{1}{2}\left\| \sum_{d=1}^{D} Bx_0^d \odot h^d - y \right\|_2^2 + \frac{1}{2}\left\| \sum_{d=1}^{D} \varphi h^d \right\| + \frac{\lambda}{2}\left\| \sum_{k=1}^{K}\sum_{d=1}^{D} Bx_k^d \odot h^d \right\|_2^2 \\ & + \frac{\gamma}{2}\left\| \sum_{d=1}^{D} B(x_c^d - x_{c,t-1}^d) \odot h^d \right\|_2^2 , \end{aligned} \tag{5}$$

where x_c^d, γ, and D represent the environment patch, sensitivity, and total number of channels, respectively.

By training the environment term $x_c - x_{c,t-1}$, our CAERDCF maintians durability in the following frames.

To enhance the efficiency of computation, CFs are typically converted to the frequency domain. In Equation (5), we introduce $v^d = B^T h^d \in R^N$ and $X_\delta = x_c - x_{c,t-1,}$ as the auxiliary variable and environment residual:

$$E(h, \hat{v}) = \frac{1}{2N}\left\| \hat{X}^d \hat{v}^d - \hat{Y} \right\|_2^2 + \frac{1}{2}\left\| \varphi h^d \right\| + \frac{\gamma}{2}\left\| \hat{X}_\delta \odot \hat{v}^d \right\|_2^2, \tag{6}$$

where $\hat{X}^d = [(\hat{x}_0^d), \sqrt{\lambda}(\hat{x}_1^d), ..., \sqrt{\lambda}(\hat{x}_k^d)]^T, \hat{Y} = [\hat{y}, 0, ..., 0]^T$, $\odot$, and $\wedge$ denotes the element-wise multiplication and the DFT.

3.5. ADMM

The Lagrangian function of Equation (6) can be expressed as follows:

$$\begin{aligned} E(h,\hat{v},\hat{\zeta}) =& \frac{1}{2N}\| \sum_{d=1}^{D} \hat{X}^d \hat{v}^d - \hat{Y}\|_2^2 + \frac{1}{2}\sum_{d=1}^{D} \|\varphi h^d\| + \frac{\gamma}{2N}\sum_{d=1}^{D} \|\hat{X}_\delta^d \odot \hat{v}^d\|_2^2 \\ &+ \frac{\mu}{2}\sum_{d=1}^{D} \|\hat{v}^d - \sqrt{N}FB^T h^d\|_2^2 + \sum_{d=1}^{D} \hat{\zeta}^d(\hat{v}^d - \sqrt{N}FB^T h^d). \end{aligned} \tag{7}$$

In this equation, μ stands for the penalty factor, and $\zeta = [\zeta_1^T, ..., \zeta_K^T]^T$ represents the Langrangian vector in the Fourier domain.

$$\begin{cases} h_{t+1}^* = & \arg\min_x \{ \frac{1}{2}\|\varphi h^d\| + \frac{\mu}{2}\|\hat{v}^d - \sqrt{N}FB^T h^d\|_2^2 \\ & +\hat{\zeta}^d(\hat{v}^d - \sqrt{N}FB^T h^d) \}, \\ \hat{v}_{i+1} = & \arg\min_v \{ \frac{1}{2N}\|\hat{X}^d \hat{v}^d - \hat{Y}\|_2^2 + \frac{1}{2}\|\varphi h^d\| \\ & +\frac{\gamma}{2}\|\hat{X}_\delta \odot \hat{v}^d\|_2^2 + \frac{\mu}{2}\|\hat{v}^d - \sqrt{N}FB^T h^d\|_2^2 \\ & +\hat{\zeta}^d(\hat{v}^d - \sqrt{N}FB^T h^d), \\ \hat{\zeta}_{i+1} = & \hat{\zeta}_i + \mu(\hat{v}_{i+1} - \hat{h}_{i+1}), \end{cases} \tag{8}$$

where i denotes the number of iterations.

3.6. Solution to Subproblem h^*

Given v, ζ, the h is optimized:

$$h^d = \frac{\zeta^d + \mu v^d}{\frac{\varphi^2}{N} + \mu}. \tag{9}$$

3.7. Solution to Subproblem $\hat{v}^*$

By splitting Equation (8), $\hat{v}$ can be calculated as follows:

$$\begin{aligned} \hat{v}(n)^* =& \frac{1}{2N}\|\hat{X}(n)^T \hat{v}(n) - \hat{Y}(n)\|_2^2 + \frac{\gamma}{2N}\|\hat{X}_\delta(n)^T \hat{v}(n)\|_2^2 \\ &+ \frac{\mu}{2}\|\hat{v}(n) - \hat{h}(n)\|_2^2 + (\hat{v}(n) - \hat{h}(n))^T \hat{\zeta}(n). \end{aligned} \tag{10}$$

For better understanding, we describe X and X_δ as X_0 and X_1, respectively. Finally, we obtain the solution with the help of the Sherman–Morrison formula:

$$\begin{aligned} \hat{v}(n)^* =& \frac{1}{\mu N}\Big(\hat{X}(n)\hat{y}(n) - N\hat{\zeta}_f + \mu N\hat{h}(n)\Big) - \frac{\sum_{k=0}^{1} p_k \hat{X}_k(n)}{\mu\theta} \\ & \left(\frac{1}{N}\eta\hat{y}(n) - \sum_{k=0}^{1} \hat{X}_k(n)^T \hat{\zeta}(n) + \mu\sum_{k=0}^{1} \hat{X}_k(n)^T \hat{h}(n)\right), \end{aligned} \tag{11}$$

where $p_0 = 1$, $p_1 = \gamma$, $\theta = \mu N + \sum_{k=0}^{1} p_k \hat{X}_k(n)^T \hat{X}_k(n)$, and $\eta = \sum_{k=0}^{1} p_k \hat{X}_k(n)\hat{X}(n)$.

3.8. Langrangian Update

We can update parameter $\hat{\zeta}$ as follows:

$$\hat{\zeta}_{i+1} = \hat{\zeta}_i + \mu(\hat{v}_{i+1} - \hat{h}_{i+1}), \tag{12}$$

where $\hat{\zeta}$ represents the Langrangian parameter, t and $t+1$ represent t^{th}, and $(t+1)^{th}$ iteration, respectively.

3.9. Fast Environment Awareness

In the proposed CAERDCF tracker, an efficient strategy is used for obtaining environmental information by multiplying environmental mask m_e with the input sample x:

$$x_c = x \odot m_e. \tag{13}$$

Similar to training sample size, a matrix is generated and the target area values are set to zero. Then, the matrix is rescaled to acquire a mask m_e. With the help of the above strategy, our CAERDCF avoids extra feature extraction along with insignificant background data.

3.10. Online Update

The online updating strategy for the proposed approach is as follows:

$$\hat{x}^M = (1-\beta)\hat{x}^M_{t-1} + \beta\hat{x}, \tag{14}$$

where $\hat{x}^M$ and $\hat{x}^M_{t-1}$ represent the current and last frame models, $\hat{x}$ denotes the observation model of the current frame, and β denotes the learning rate.

3.11. Tracking Model

To procure the search region z in the new frame, we used a motion-aware search strategy similar to [36]. Finally, we obtained the response map M_{t+1} from the following equation:

$$M_{t+1} = F^{-1}\Big(\sum_{d=1}^{D} \hat{z}^d_{t+1} \odot \hat{v}^d_t\Big), \tag{15}$$

where F^{-1} is the inverse DFT.

3.12. Experimental Setup

In this section, we discuss our experimental setup and feature representation details. The proposed approach is implemented in MATLAB2021a, and the hardware platform is an Intel Core(TM) I5-6500 CPU @3.20 GHz, with 16 GB of memory. Our method employs deep and hand-crafted features such as fHOG, CN, and VGG-m-2048 [32] that are used to extract the feature map in the CF tracker. The fHOG feature is an improved version of HOG that delivers greater speed by preserving all information. Moreover, these features demonstrate robust feature representations that play a key role in achieving better efficiency in object tracking. The following parameters, learning rate $\beta = 0.0199$ and sensitivity factor $\gamma = 0.2$, have been determined based on a number of experimental findings and prior research. In our experiments, we set the number of ADMM iterations to 2. This value was chosen based on empirical testing to achieve a balance between convergence speed and accuracy. Additionally, we set the minimum and maximum values of the region window parameter φ to 1×10^{-3} and 1×10^{5}, respectively. These values were selected to provide a flexible range for the window size and are similar to those used in related works such as [16,22] for consistency.

4. Experiments

First, we present the comparison outcomes of the proposed tracker with the standard benchmark datasets. Next, we conduct an ablation study on the OTB2015 benchmark to evaluate the impact of each component. Following the ablation study, a parameter analysis is performed to further understand the influence of different design choices. Finally, an attribute analysis and qualitative comparison are carried out on the OTB2015 dataset to assess the tracker's performance visually.

4.1. Comparison

We conduct the experiments on five well-known datasets including OTB2015 [37], TempleColor128 [38], UAV123 [39], LASOT [40], and GOT10K [41] on different trackers including ECO [42], SiamFC++ [20], EFSCF [43], STRCF [31], CSACF11 [44], SiamRPN [45], BACF [30], FRATCF [46], RBSCF [47], SRDCF [29], AutoTrack [48], HDT [35], SASR [16], KCF [8], SiamRAAN [49], SiamDMU [19], Staple [27], DaSiamRPN [50], SAMF [26], and CSK [12].

4.1.1. Experiments on the OTB2015 Dataset

The OTB2015 dataset is a widely recognized benchmark in visual tracking, offering annotated video sequences to evaluate algorithm performance under diverse real-world conditions. We assess the proposed CAERDCF tracker on the OTB2015 dataset, as depicted in Table 1. Specifically, the CAERDCF tracker achieves the highest DP score of 93.1% and an AUC score of 69.3%, as shown in Figure 3. Furthermore, our tracker surpasses siamese-based trackers such as SiamFC++ and SiamRPN by (1.7% and 7.9%) on DP, respectively. Additionally, our method outperforms ECO, EFSCF, and STRCF by (2.1%/0.3%), (5.7%/1.9%), and (6.5%/3.7%) in DP/AUC scores, respectively. Finally, we confirm that our proposed method enhances DP from 80.2% to 93.1% and AUC from 61.0% to 69.3%, respectively, compared to the baseline BACF tracker. In summary, our method demonstrates robust performance across various metrics, showcasing its effectiveness in tracking compared to state-of-the-art approaches.

Table 1. Performance evaluation on the OTB2015 dataset with 100 sequences.

Methods	Metrics	CAERDCF	SiamFC++	ECO	EFSCF	STRCF	SiamRPN	HDT	CSACF11	SASR	BACF
OTB2015	DP	93.1	91.7	91.0	87.4	86.6	85.3	84.7	83.7	80.7	80.2
	AUC	69.3	68.7	69	67.4	65.6	63.7	56.3	64.4	47.6	61

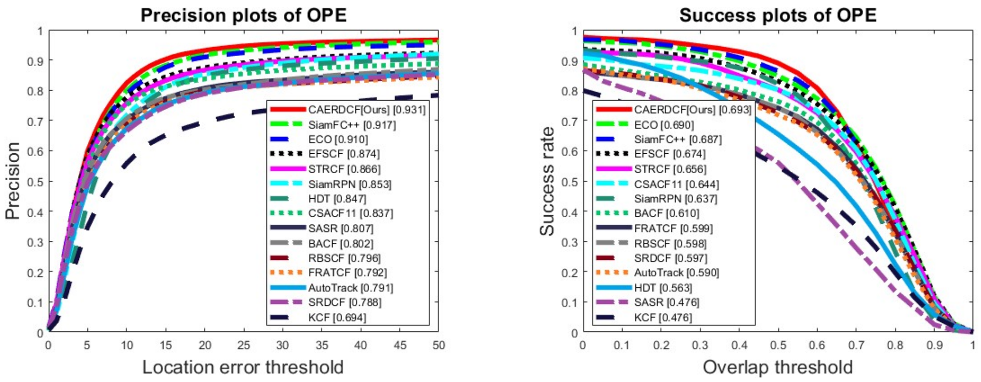


Figure 3. Precision and success plots on the OTB2015 benchmark.

4.1.2. Experiments on the TempleColor128 Dataset

The TempleColor128 dataset consists of 128 high-resolution video sequences, each annotated with ground truth bounding boxes, widely utilized for evaluating visual tracking algorithm performance across diverse environmental conditions. Experiments were conducted on the TempleColor benchmark, with the results shown in Table 2. Out of the trackers assessed, our approach stands out as the leading performer, achieving a precision score of 81.1% and a success score of 59.2%, as depicted in Figure 4. Thanks to the benefits of multi-feature fusion, our method outperforms the STRCF, SASR, and RBSCF methods, improving the DP score by 6.7%, 7.2%, and 8.8%, respectively. Compared to the BACF tracker, the CAERDCF tracker shows a significant improvement of 16.1% in precision. Furthermore, our CAERDCF algorithm surpasses conventional trackers like KCF and CSK, with gains of 25.3% and 41.3% in distance precision scores, respectively. Overall, as observed in Figure 4, our technique consistently outperforms modern trackers.

Table 2. Performance evaluation on the TempleColor128 dataset with 128 sequences.

Methods	Metrics	CAERDCF	STRCF	SASR	RBSCF	FRSTCF	AutoTrack	HDT	SRDCF	BACF	KCF
TempleColor128	DP	81.1	74.4	73.9	72.3	71	70.5	68.7	66.3	65	55.8
	AUC	59.2	54.8	51.3	53.5	52.7	51.6	48	48.5	49	38.7

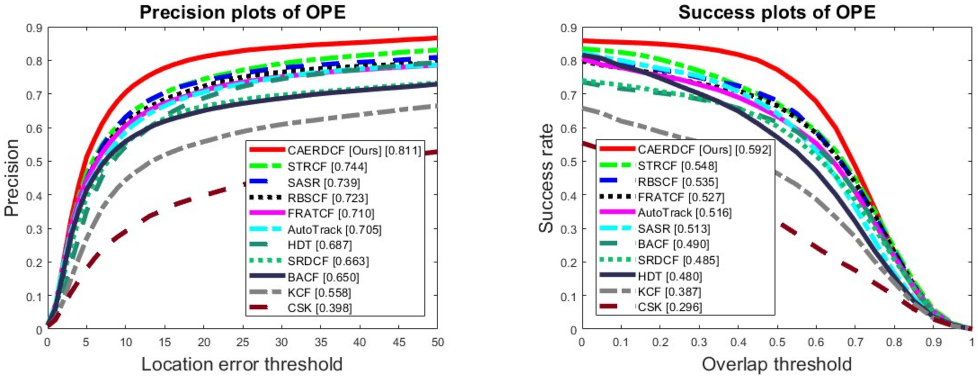


Figure 4. Precision and success plots on the TC128 benchmark.

4.1.3. Experiments on the UAV123 Dataset

The UAV123 dataset comprises 123 video sequences captured by unmanned aerial vehicles (UAVs), annotated with ground truth bounding boxes, providing a standardized benchmark for evaluating visual tracking algorithms in aerial surveillance contexts. To assess the performance of the CAERDCF approach, we conducted comparisons using the UAV123 dataset, as shown in Table 3. As illustrated in Figure 5, our CAERDCF achieves a DP score of 79.1%, marking a 1.0% improvement over DaSiamRPN's performance. In the precision plot, our tracker surpasses CF-based trackers CSACF, EFSCF, and RBSCF by 7.3%, 7.9%, and 9.2%, respectively. Furthermore, compared to conventional methods like SRDCF (67.6%) and KCF (52.3%), our tracker delivers higher scores. Additionally, relative to the baseline tracker, our CAERDCF achieves comparable performance with a 13.3% improvement in precision. Overall, the evaluation results affirm the efficacy of the CAERDCF approach in addressing diverse tracking conditions.

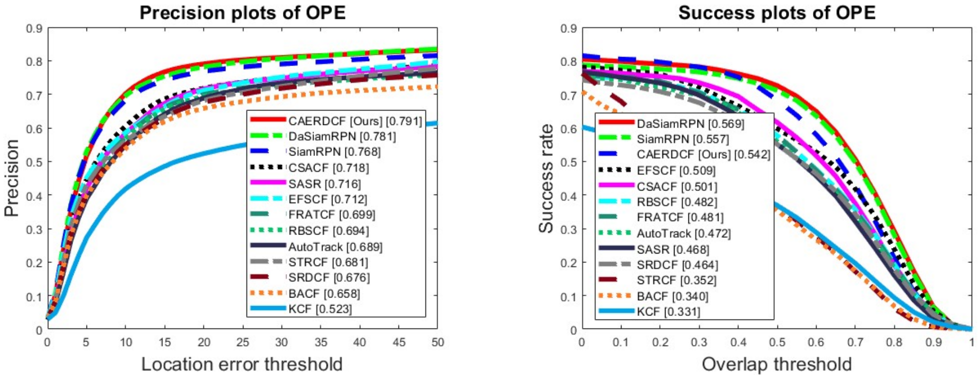


Figure 5. Precision and success plots on the UAV123 dataset.

Table 3. Performance evaluation on the UAV123 dataset with 123 sequences.

Methods	Metrics	CAERDCF	DaSiamRPN	SiamRPN	CSACF	SASR	EFSCF	FRATCF	RBSCF	AutoTrack	STRCF	SRDCF	BACF
UAV123	DP	79.1	78.1	76.8	71.8	71.6	71.2	69.9	69.4	68.9	68.1	67.6	65.8
	AUC	54.2	56.9	55.7	50.1	46.8	50.9	48.1	48.2	47.2	35.2	46.4	34

4.1.4. Experiments on LASOT Dataset

The LaSOT (large-scale single object tracking) dataset consists of a large number of video sequences with diverse challenges, making it a comprehensive benchmark for evaluating visual object tracking algorithms in complex real-world scenarios. Additionally, our investigation encompasses an assessment of the effectiveness of the CAERDCF methodology utilizing the LASOT dataset in conjunction with contemporary tracking algorithms. The LASOT dataset stands as a widely recognized benchmark, notable for its comprehensive representation of diverse real-world scenarios and challenges encountered in object tracking. As illustrated in Figure 6, our CAERDCF algorithm achieves a noteworthy precision score of 46.7% and 40.1% in success scores. Importantly, our tracker emerges as the third-best performer among the evaluated methodologies, underscoring its competitive standing in the field of object tracking. Noteworthy among the examined methodologies are the precision scores of 62.2%, 57.7%, and 34% achieved by the SiamRAAN, SiamDMU, and STRCF techniques, respectively. Notably, our CAERDCF tracker surpasses the foundational BACF by a considerable difference of 18.4% in precision performance. Furthermore, our findings validate the superior precision tracking capabilities of the proposed method compared to traditional trackers such as Staple (27.8%), SRDCF (24.8%), KCF (19%), and CSK (14.9%).

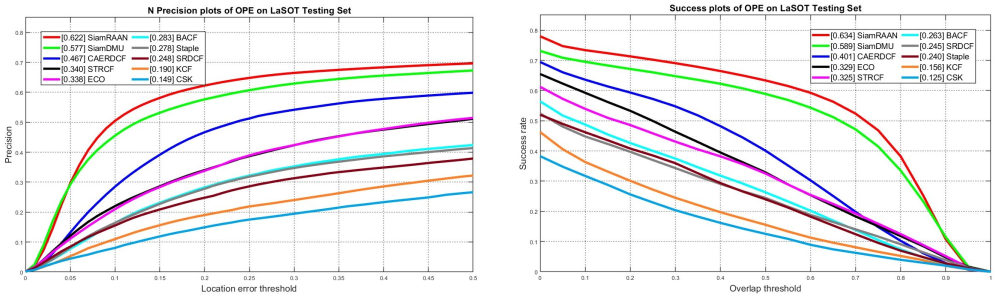


Figure 6. Precision and success plots on the LASOT benchmark.

4.1.5. Experimetns on the GOT-10K Dataset

The GOT-10k dataset is a significant benchmark in visual tracking, featuring a diverse collection of 10,000 video sequences annotated with ground truth bounding boxes. It is widely used to evaluate and compare the performance of visual tracking algorithms across various challenging scenarios and conditions. Our investigation also includes an evaluation of the CAERDCF methodology using the GOT-10k dataset alongside modern tracking techniques. Renowned for its wide array of real-world scenarios, the GOT-10k dataset serves as a valuable benchmark for assessing object tracking systems. As shown in our analysis (Figure 7), our CAERDCF algorithm achieves a significant success score of 44.7%. Among the methodologies scrutinized, the SiamRPN, ECO, and CF2 techniques achieved success scores of 36.7%, 31.6%, and 31.5%, respectively. Additionally, our proposed method demonstrated superior performance compared to the baseline, achieving a notable increase of 18.7% in success performance. These findings underscore the superior tracking capabilities of our approach compared to traditional trackers like SAMF (24.6%), Staple (24.6%), SRDCF (23.6%), and KCF (20.3%).

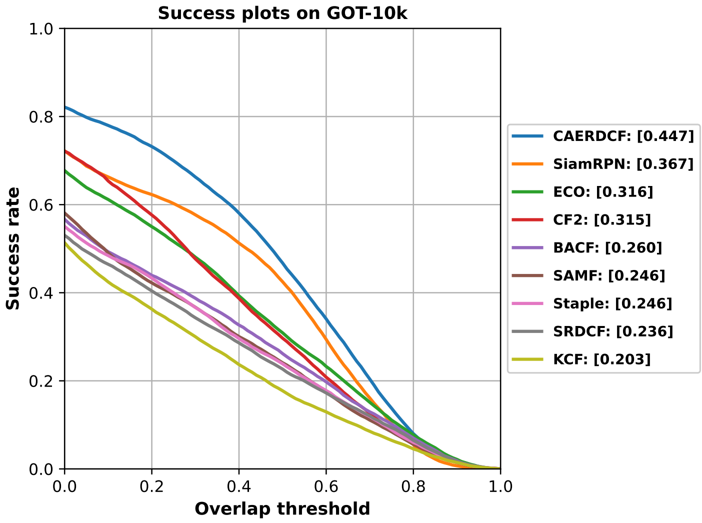


Figure 7. Success plots of the GOT-10K dataset.

4.2. Ablation Studies

In this section, we undertake ablation studies on the OTB2015 dataset to evaluate the impact of each element within our approach. In Table 4, we show the characteristics and tracking results of OTB2015. We compared the experimental results of the proposed approach with different groups, including selective spatial regularization (SSR), context-information (CA), and environmental residual (ER). In our experiments, we take BACF as the baseline [30]. In the analysis, it is observed that the baseline tracker achieves precision and success values of 80.2% and 61.0%, respectively. Following the incorporation of selective spatial regularization (SSR) into the baseline tracker (*Baseline* + *SSR*), notable improvements are observed in precision and success scores, reaching 83.6% and 63.3%, respectively, as delineated in the provided data. The variant *Baseline* + *SSR* + *CA* represents the context-aware selective spatial regularization tracker, which utilizes surrounding information around the target that achieves (86.1%/66.0%) in DP/AUC scores. Furthermore, it is evident that the tracker performance experiences enhancement upon integrating the environmental residual term with the Baseline + SSR + CA + ER term $(Baseline + SSR + CA + ER)$. Subsequently, we delve into analyzing the impact of utilizing deep features within the tracking environment. Ultimately, the culmination of all components within the proposed technique yields significant results, with precision and success scores reaching 93.1% and 69.3%, respectively.

Table 4. Ablation study of the proposed CAERDCF on the OTB2015 dataset. The best results are shown in red font.

Methods	Precision (%)	Success (%)
CAERDCF [Ours]	93.1	69.3
Baseline + SSR + CA + ER	89.5	67.1
Baseline + SSR + CA	86.1	66.0
Baseline + SSR	83.6	63.3
Baseline	80.2	61.0

4.3. Sensitivity Analysis of Key Parameters

We conducted an examination of our proposed approach's performance across various parameter selections. Initially, we scrutinized the impact of the learning rate (β) outlined in Equation (14), as illustrated in Figure 8a. We observed a continuous increase in the AUC score as β ranged from 0.0001 to 1, peaking at β = 0.0199. Following this, we proceeded with a sensitivity analysis concerning the regularization parameter (γ) as delineated in Equation (10). While keeping β fixed at 0.0199, γ was varied from 0.002 to 0.5. Notably, excessively large values of γ led to diminished performance. Consequently, the AUC score reached its pinnacle at 0.634 when γ was set to 0.2. These findings, depicted in Figure 8b, underscore that the optimal performance characteristics were achieved with β set to 0.0199 and γ to 0.2.

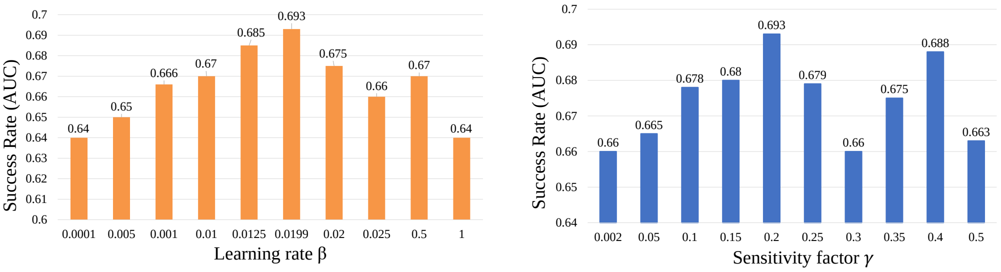
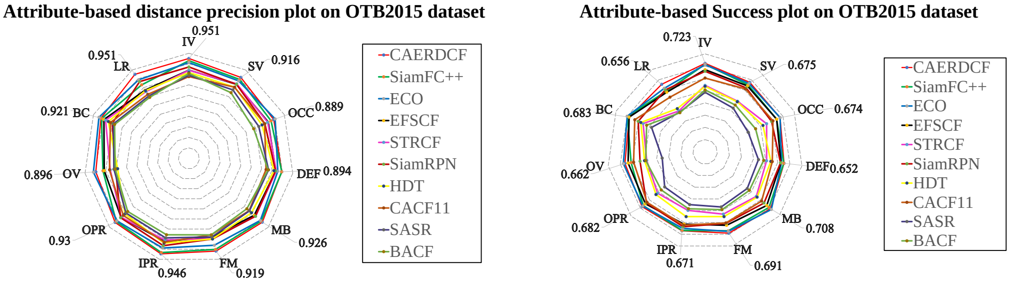


Figure 8. Impact of the β and γ Parameters on CAERDCF's success Scores on OTB2015.

4.4. Analysis Based on Different Attributes on OTB2015 Benchmark

We assess the proposed CAERDCF tracker against various trackers across multiple criteria, utilizing precision (DP) and success (AUC) scores as shown in Figure 9. Our analysis reveals that our tracker excels in handling sequences with diverse challenges, consistently outperforming all other trackers across most attributes. While many trackers exhibit high accuracy in the absence of occlusion, their performance declines significantly under severe occlusion. The superior robustness of our tracker stems from its comprehensive utilization of environmental information pertaining to the tracked target. Notably, the CAERDCF tracker achieves precision scores of 81.8% and 82.8%, respectively, in cases of occlusion and scale variation, outperforming the second-best tracker in these scenarios. Particularly, the CAERDCF method demonstrates optimal performance in both DP and AUC metrics when facing illumination variation (95.1%/72.3%), fast motion (91.9%/69.1%), out-of-view (89.6%/66.2%), and low resolution (95.1%/65.6%). In summary, our proposed method exhibits substantial performance enhancements across most attributes in the OTB2015 benchmark. The radar plot results underscore the outstanding performance achieved by our tracker.


Attribute-based distance precision plot on OTB2015 dataset

Attribute-based Success plot on OTB2015 dataset

Figure 9. Attribute-based comparison of CAERDCF on the precision and success scores of the OTB2015 dataset.

4.5. Qualitative Comparisons

Our proposed approach is evaluated alongside various methods (SRDCF [29], BACF [30], KCF [8], and CSK [12]) on five challenging sequences. Within the Biker sequence, showcased in Figure 10, the object undergoes rapid motion and encounters occlusion. While other methods struggle to maintain tracking, our proposed tracker successfully tracks the target under these challenging conditions. The DragonBaby sequence poses various challenges, such as motion blur, scale variation, fast motion, and rotation. Our tracking algorithm demonstrates consistent performance across most frames due to its selective spatial regularizer. Similarly, in the KiteSurf sequence, which involves rotation, illumination variation, and occlusion, our

CAERDCF method effectively leverages contextual information with environmental residuals to excel across various challenging scenarios. The Shaking sequence has dramatic lighting shifts, posing a challenge for object tracking. In contrast, our approach shows favorable performance compared to other algorithms, indicating robustness to illumination variation. In the soccer sequence, background clutter and occlusion hinder target visibility. While DSAR-CF and CSK lose track of the target, our tracker quickly adapts and maintains tracking even under challenging circumstances. Additionally, Table 5 highlights the pros and cons of our proposed method relative to traditional methods.

Table 5. Pros and cons of the proposed method compared to conventional techniques.

Method	Pros	Cons
Proposed Method (CAERDCF)	- High precision and robustness under ambiguous environmental changes - Efficient acquisition of environment data without additional computation - Enhanced target presence through multi-feature integration - Significant improvement in benchmark performance (e.g., 12.9% increase in OTB2015 precision score)	- Higher computational complexity due to advanced algorithms - Potential need for more computational resources compared to simpler methods
SRDCF	- Effective handling of scale variations - Good performance in various challenging conditions	- Higher computational cost compared to simpler methods - May still struggle with significant occlusions or rapid movements
BACF	- Balanced performance with efficient tracking - Effective in handling boundary effects through context-aware filters	- May not perform as well in highly cluttered environments - Limited by the feature representation used
KCF	- High-speed tracking - Efficient in terms of computational resources	- Less effective in handling scale variations and occlusions - Performance drops in cluttered scenes
CSK	- Very fast and efficient due to simple kernelized correlation filters - Good baseline performance in standard scenarios	- Struggles with scale variation and occlusions - Limited robustness in dynamic and cluttered environments

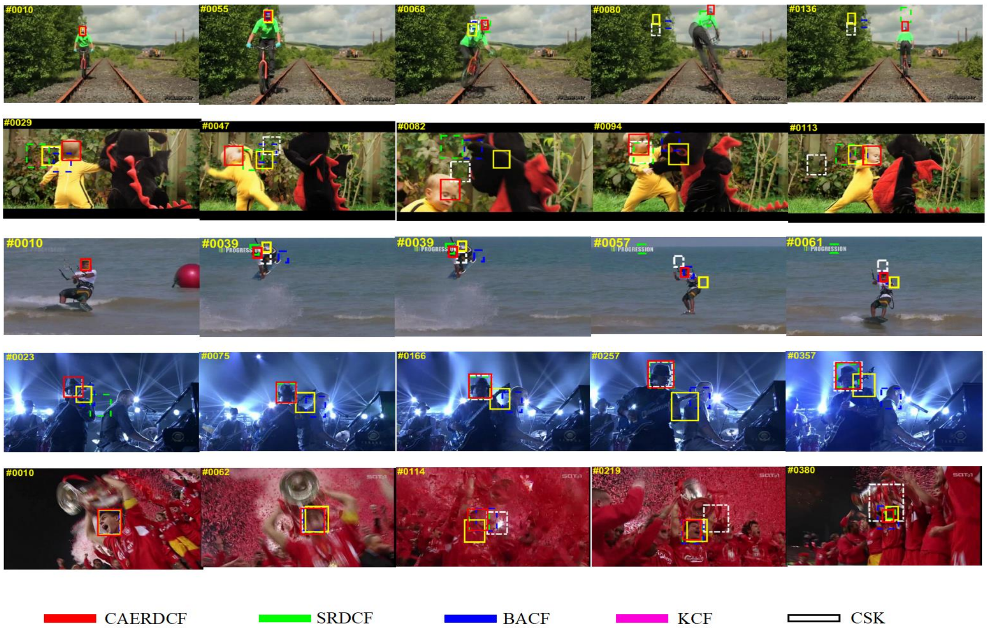


Figure 10. Qualitative performance evaluation of CAERDCF on Biker, DragonBaby, KiteSurf, Shaking, and Soccer from the OTB2015 benchmark.

5. Future Study and Implications

In future research, we aim to enhance the robustness and versatility of the proposed CAERDCF method by addressing its performance under an even broader spectrum of challenging conditions. This includes delving into advanced techniques for real-time processing to ensure the method's viability in practical applications such as autonomous navigation and intelligent surveillance. Additionally, we plan to integrate our approach with other cutting-edge tracking systems to evaluate potential synergies and further performance improvements. These future studies will not only solidify the practical applicability of CAERDCF but also potentially set new benchmarks in the field of visual tracking. The implications of this work are far-reaching, promising substantial advancements in the accuracy and reliability of tracking algorithms in dynamic and complex environments, thereby contributing significantly to the fields of computer vision and machine learning.

6. Conclusions

In summary, this study introduces a novel approach to context-aware environmental residual CF tracking through deep convolution features. The method addresses the challenge of maintaining tracking robustness amidst significant environmental variations by leveraging environmental residual terms and effectively incorporating contextual and background information to enhance tracking accuracy. Additionally, a selective spatial regularization technique is proposed to mitigate boundary effects and preserve object information. Moreover, a method of fusing multiple features is introduced to enhance the appearance of the target under varied and uncontrolled environments. The integration of both handcrafted and deep features contributes to achieving high precision and accuracy in the tracking system. Looking ahead, an avenue for future research lies in the exploration of Siamese and Transformer-based architectures, which hold the potential for further enhancing tracking capabilities. Their incorporation could lead to significant advancements in tracking performance and robustness. In conclusion, the experimental results on benchmark datasets, including OTB2015, TempleColor128, UAV123, LASOT, and GOT10K, validate the superiority of the proposed method over existing approaches, and the outlined directions for future development offer exciting prospects for advancing tracking methodologies.

Author Contributions: Conceptualization, S.S.K.S.; Methodology, S.S.K.S. and S.M.; Validation, S.S.K.S. and S.M.; Investigation, S.S.K.S., S.M., S.A. and J.H.J.; Writing—original draft, S.S.K.S.; Writing—review & editing, S.S.K.S., S.M., S.A. and Y.H.J.; Visualization, S.S.K.S. and S.A.; Supervision, J.H.J. and Y.H.J.; Funding acquisition, Y.H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Basic Science Research Program under Grant NRF -2016R1A6A1A03013567 and Grant NRF-2021R1A2B5B01001484 and by the framework of the International Cooperation Program under Grant NRF-2022K2A9A2A06045121 through the National Research Foundation of Korea (NRF) funded by the Ministry of Education.

Data Availability Statement: The data used in the study are available from the authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rao, R.S.; Kalabarige, L.R.; Alankar, B.; Sahu, A.K. Multimodal imputation-based stacked ensemble for prediction and classification of air quality index in Indian cities. *Comput. Electr. Eng.* **2024**, *114*, 109098. [CrossRef]
2. Patro, P.; Kumar, K.; Kumar, G.S.; Sahu, A.K. Intelligent data classification using optimized fuzzy neural network and improved cuckoo search optimization. *Iran. J. Fuzzy Syst.* **2023**, *20*, 155–169.
3. Wang, W.; Zhang, K.; Lv, M.; Wang, J. Discriminative visual tracking via spatially smooth and steep correlation filters. *Inf. Sci.* **2021**, *578*, 147–165. [CrossRef]
4. Moorthy, S.; Choi, J.Y.; Joo, Y.H. Gaussian-response correlation filter for robust visual object tracking. *Neurocomputing* **2020**, *411*, 78–90. [CrossRef]
5. Elayaperumal, D.; Joo, Y.H. Aberrance suppressed spatio-temporal correlation filters for visual object tracking. *Pattern Recognit.* **2021**, *115*, 107922. [CrossRef]
6. He, X.; Chen, C.Y.C. Learning object-uncertainty policy for visual tracking. *Inf. Sci.* **2022**, *582*, 60–72. [CrossRef]
7. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2544–2550.
8. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 583–596. [CrossRef]
9. Yang, K.; Zhang, H.; Zhou, D.; Dong, L. PaaRPN: Probabilistic anchor assignment with region proposal network for visual tracking. *Inf. Sci.* **2022**, *598*, 19–36. [CrossRef]
10. Lee, H.; Kim, S. SSPNet: Learning spatiotemporal saliency prediction networks for visual tracking. *Inf. Sci.* **2021**, *575*, 399–416. [CrossRef]
11. Chen, K.; Tao, W.; Han, S. Visual object tracking via enhanced structural correlation filter. *Inf. Sci.* **2017**, *394*, 232–245. [CrossRef]
12. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. Exploiting the circulant structure of tracking-by-detection with kernels. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 702–715.
13. Mueller, M.; Smith, N.; Ghanem, B. Context-aware correlation filter tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1396–1404.
14. Moorthy, S.; Joo, Y.H. Multi-expert visual tracking using hierarchical convolutional feature fusion via contextual information. *Inf. Sci.* **2021**, *546*, 996–1013. [CrossRef]
15. Elayaperumal, D.; Joo, Y.H. Robust visual object tracking using context-based spatial variation via multi-feature fusion. *Inf. Sci.* **2021**, *577*, 467–482. [CrossRef]
16. Fu, C.; Xiong, W.; Lin, F.; Yue, Y. Surrounding-aware correlation filter for UAV tracking with selective spatial regularization. *Signal Process.* **2020**, *167*, 107324. [CrossRef]
17. Li, Y.; Fu, C.; Huang, Z.; Zhang, Y.; Pan, J. Intermittent contextual learning for keyfilter-aware UAV object tracking using deep convolutional feature. *IEEE Trans. Multimed.* **2020**, *23*, 810–822. [CrossRef]
18. Yan, Y.; Guo, X.; Tang, J.; Li, C.; Wang, X. Learning spatio-temporal correlation filter for visual tracking. *Neurocomputing* **2021**, *436*, 273–282. [CrossRef]
19. Liu, J.; Wang, H.; Ma, C.; Su, Y.; Yang, X. SiamDMU: Siamese Dual Mask Update Network for Visual Object Tracking. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 1656–1669. [CrossRef]
20. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12549–12556.
21. Kuppusami Sakthivel, S.S.; Jeong, J.H.; Joo, Y.H. A multi-level hybrid siamese network using box adaptive and classification approach for robust tracking. *Multimed. Tools Appl.* **2024**, 67, 1–26.

22. Zhang, F.; Ma, S.; Zhang, Y.; Qiu, Z. Perceiving Temporal Environment for Correlation Filters in Real-Time UAV Tracking. *IEEE Signal Process. Lett.* **2021**, *29*, 6–10. [CrossRef]
23. Ma, C.; Huang, J.B.; Yang, X.; Yang, M.H. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3074–3082.
24. Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; Van de Weijer, J. Adaptive color attributes for real-time visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1090–1097.
25. Van De Weijer, J.; Schmid, C.; Verbeek, J.; Larlus, D. Learning color names for real-world applications. *IEEE Trans. Image Process.* **2009**, *18*, 1512–1523. [CrossRef]
26. Li, Y.; Zhu, J. A scale adaptive kernel correlation filter tracker with feature integration. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 254–265.
27. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, P.H. Staple: Complementary learners for real-time tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1401–1409.
28. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative scale space tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1561–1575. [CrossRef]
29. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
30. Kiani Galoogahi, H.; Fagg, A.; Lucey, S. Learning background-aware correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1135–1143.
31. Li, F.; Tian, C.; Zuo, W.; Zhang, L.; Yang, M.H. Learning spatial-temporal regularized correlation filters for visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4904–4913.
32. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
33. Danelljan, M.; Robinson, A.; Khan, F.S.; Felsberg, M. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 472–488.
34. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Multi-cue correlation filters for robust visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4844–4853.
35. Qi, Y.; Zhang, S.; Qin, L.; Yao, H.; Huang, Q.; Lim, J.; Yang, M.H. Hedged deep tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4303–4311.
36. Fu, C.; Ding, F.; Li, Y.; Jin, J.; Feng, C. Learning dynamic regression with automatic distractor repression for real-time UAV tracking. *Eng. Appl. Artif. Intell.* **2021**, *98*, 104116. [CrossRef]
37. Wu, Y.; Lim, J.; Yang, M.H. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1834–1848. [CrossRef]
38. Liang, P.; Blasch, E.; Ling, H. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5630–5644. [CrossRef]
39. Mueller, M.; Smith, N.; Ghanem, B. A benchmark and simulator for uav tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 445–461.
40. Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Harshit; Huang, M.; Liu, J.; et al. Lasot: A high-quality large-scale single object tracking benchmark. *Int. J. Comput. Vis.* **2021**, *129*, 439–461. [CrossRef]
41. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef]
42. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
43. Wen, J.; Chu, H.; Lai, Z.; Xu, T.; Shen, L. Enhanced robust spatial feature selection and correlation filter learning for UAV tracking. *Neural Netw.* **2023**, *161*, 39–54. [CrossRef] [PubMed]
44. Ma, J.; Lv, Q.; Yan, H.; Ye, T.; Shen, Y.; Sun, H. Color-saliency-aware correlation filters with approximate affine transform for visual tracking. *Vis. Comput.* **2023**, *39*, 4065–4086. [CrossRef]
45. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
46. Yuan, Y.; Chen, Y.; Jing, Y.; Zhou, P.; Zhang, Y. FRATCF: Feature-Residue Real-Time UAV Tracking Based on Automatic Spatio-Temporal Regularization Correlation Filter. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 11–15 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
47. Lin, J.; Peng, J.; Chai, J. Real-time UAV Correlation Filter Based on Response-Weighted Background Residual and Spatio-Temporal Regularization. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6005405. [CrossRef]

48. Li, Y.; Fu, C.; Ding, F.; Huang, Z.; Lu, G. 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, Seattle, WA, USA, 13–19 June 2020; pp. 11923–11932.
49. Xin, Z.; Yu, J.; He, X.; Song, Y.; Li, H. SiamRAAN: Siamese Residual Attentional Aggregation Network for Visual Object Tracking. *Neural Process. Lett.* **2024**, *56*, 98. [CrossRef]
50. Zhu, Z.; Wang, Q.; Li, B.; Wu, W.; Yan, J.; Hu, W. Distractor-aware siamese networks for visual object tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 101–117.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.