2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Joint Response and Background Learning for UAV Visual Tracking

Biao Wang[1], Wenling Li[1,*], Bin Zhang[2] and Yang Liu[1]

***Abstract*— Correlation filter (CF)-based approaches have gained widespread attention in the field of unmanned aerial vehicle (UAV) visual tracking due to their light-weight characteristics. However, CFs are prone to generating low-quality response in challenging UAV scenarios, *e.g.*, fast motion and background clutter. In this paper, in order to model the tracker more robustly, we first conduct an effective regularization analysis from the perspectives of response- and background-learning. Specifically, to address response degradation, we propose a module for learning temporal consistency and reversibility of response, supplemented by a novel background-aware module to enhance the ability to learn from negative samples. In addition, we propose a fast coarse-to-fine scale search strategy, which alleviates the challenges in estimating bounding boxes under non-uniform aspect ratios. We have developed two tracker versions, namely RBLT and DeepRBLT, based on the depth of the features. Comprehensive experiments on four UAV benchmarks and one generic benchmark have indicated the superiority of our trackers compared to other state-of-the-art trackers, with enough speed for real-time applications.**

I. INTRODUCTION

Thanks to the maneuverability and autonomy of unmanned aerial vehicles (UAVs), visual object tracking (VOT) has found widespread practical application in the field of UAVs, *e.g.*, path planning [1], [2], autonomous landing [3], [4], pose estimation [5] and flying vehicle tracking [6]. However, UAV visual tracking remains a challenging task due to the complex aerial scenarios, including background clutter, fast motion and object deformation. Besides, the intrinsic defects of UAVs such as the limitation of computing resource and battery life also impede the development of UAV tracking.

Recently, to solve the above problems, correlation filter (CF)-based methods [7], [8] have been introduced due to its balance between accuracy and computational efficiency. These methods learn a filter by using ridge regression to minimize the square error between expected and actual correlation response maps. And the learned filters can separate the specified target from the background in new frames. The high computational efficiency in CFs originates from transforming spatial domain correlation operations into element-wise multiplication operations in the frequency domain through discrete Fourier transform (DFT).

Although existing CFs have gained excellent performance, there are still some problems. Firstly, the discriminative ability of the CFs is entirely dependent on the quality of the response. However, due to the inherent scarcity of training samples in CFs, the learned filters often generate low-quality response when facing complex scenes, resulting in tracking drift or even failure. Secondly, the periodic assumption of the DFT leads to the problem of spectral leakage, commonly known as the boundary effect in tracking research. By expanding the search area, the boundary effects can be alleviated, but at the cost of introducing significant background noise. Most trackers address this issue by introducing spatial regularization [9] or performing filter pruning [10], but in complex aerial scenes, they can not learn from negative samples sufficiently. Thirdly, existing scale estimation methods [11]–[13] typically assume that the aspect ratio of the target bounding box remains unaltered. However, these methods encounter challenges in accurately estimating the bounding box when there are changes in the perspective of the UAV.

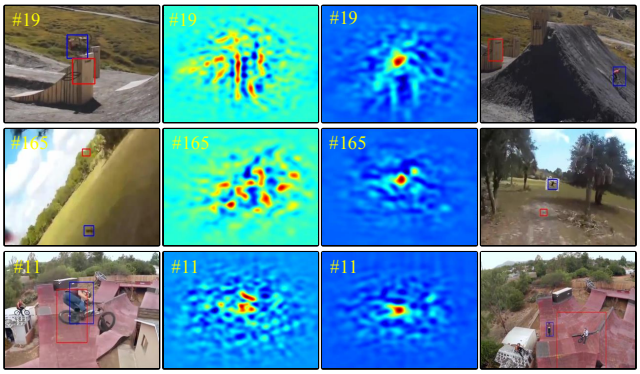


Fig. 1. Comparison between our DeepRBLT and the advanced ECO tracker. DeepRBLT successfully tracks all three challenging sequences and produces higher-quality responses compared to ECO.

In this work, starting from generating high-quality response, we propose the joint **r**esponse and **b**ackground **l**earning **t**racker, i.e., **RBLT**. And the main contributions of our work can be summarized as follows:

- We propose a comprehensive response learning module, which enables the filter to learn the temporal consistency and reversibility of the response, supplemented by a novel background-aware module to enhance the ability to learn from negative samples. In addition, we incorporate historical and current information into our modeling process to avoid over-fitting the filter.
- We implement a real-time CPU-based tracker (RBLT) and a high-performance GPU-based tracker (DeepRBLT). And we propose a coarse-to-fine scale search strategy, which alleviates the challenges in estimating bounding boxes under non-uniform aspect ratios.

*Corresponding Author
[1]Biao Wang, Wenling Li and Yang Liu are with School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. lwlmath@buaa.edu.cn
[2]Bin Zhang is with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China.

- Comprehensive experiments on four UAV benchmarks (DTB70 [14], UAV123@10fps [15], VisDrone2019 [16] and UAVDT [17]) and one generic benchmark (GOT10K [18]) have indicated the superiority of our trackers compared to other state-of-the-art CPU- and GPU-based trackers.

II. RELATED WORKS

A. CF for UAV Visual Tracking

D. Bolme *et al.* [7] were the first to introduce CF into visual tracking, and subsequently, Henriques *et al.* [8] contributed a fundamental framework for CF-based tracking methods. Recently, the performance of CF-based methods is further enhanced by kernel trick [19], scale estimation [11]–[13], multi-channel features [20], [21], boundary effects suppression [9], [10], [22], integration strategy [24], [25], and deep features from convolutional neural networks [26]–[29]. Nevertheless, standard CFs typically overlook interventions on the response. Additionally, they solely depend on spatial regularization to learn about the background. Therefore, these trackers are still weak to handle challenging scenarios, *e.g.*, fast motion, viewpoint changes, and background clutter.

B. Response Learning Methods for CF

Huang *et al.* [30] proposed to repress the response aberrance by constraining the response variation in two adjacent frames. Ye *et al.* [31] further smoothed the variations of the response by regularizing the second-order difference. Li *et al.* [32] proposed the RISTrack, which aimed to keep the target area response consistent in adjacent frames. However, the methods mentioned above share three common weaknesses: (1) They only utilize non-real response from the detection phase to train the filters and required some complex shifting operations. (2) Excessive suppression of the response can lead to the inability of the filter to adapt to the current environment. (3) These methods fail to accommodate features with multiple resolutions.

C. Background Learning Methods for CF

M. Danelljan *et al.* [9] proposed to expand the search area for alleviating boundary effects and introduced spatial regularization to suppress background noise. BACF [10] achieved the same objective by pruning the target region of the filter. Furthermore, some methods [22], [23] aimed to learn background information more accurately by sparse spatial masks. However, these methods only focus on the background information contained in the filter, neglecting further learning of the current background features. In addition, some studies [33], [34] proposed to learn the background by repressing the response of context patches, but extracting features from surrounding context is extremely time-consuming.

To address these issues, we propose a joint response and background learning framework. And we incorporate historical and current information into the CF learning process.

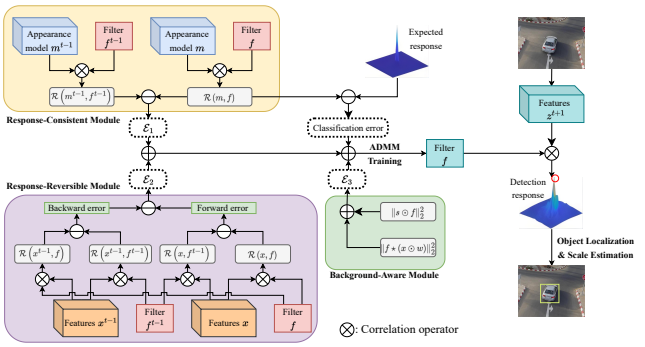


Fig. 2. A flowchart of the proposed RBLT.

III. CF-BASED TRACKING PARADIGM

The objective function of standard CFs is as follows:

$$\mathcal{E}(f) = \sum_{d}^{D} \|y - f_d \star m_d\|_2^2 + \lambda \sum_{d}^{D} \|f_d\|_2^2, \tag{1}$$

where $m_d \in \mathbb{R}^T$ and $f_d \in \mathbb{R}^T$ refer to the d-th channel of the appearance model and filter respectively, and D is the number of feature channels. $y \in \mathbb{R}^T$ is the desired correlation response, λ is a regularization parameter, and $\star$ is the spatial correlation operator. Filters and appearance models across all channels are concatenated and grouped as $f = [f_1, f_2, ..., f_D]$ and $m = [m_1, m_2, ..., m_d]$ for clarity.

The correlation operation can be simplified by transforming Eq. (1) from the spatial domain into the Fourier domain. During the detection phase, we perform correlation operations between the learned filters f and the feature maps $z^{t+1} \in \mathbb{R}^{T \times D}$ of the new (i.e. $(t+1)$-th) frame's search area. This process generates a response map $\mathcal{R}$:

$$\mathcal{R} = \mathcal{R}\left(f \star z^{t+1}\right) = \sum_{d}^{D} f_d \star z_d^{t+1}. \tag{2}$$

By searching for the peak of $\mathcal{R}$, the predicted location of the target can be obtained. Moreover, the appearance model can be adaptively updated at each frame:

$$m^{t+1} = (1-\eta)m^t + \eta x^{t+1}, \tag{3}$$

where η denotes the learning rate and $x \in \mathbb{R}^{T \times D}$ denotes the feature maps of the template area.

IV. PROPOSED METHODS

A. Response-Consistent Module

The tracked target often encounters various challenging scenarios in UAV tracking. As shown in Fig. 1, the response, which is generated by the filter in the detection stage, usually contains multiple peaks that represent the possibility of becoming a target in the corresponding position. When the interfering peak is higher than the target peak, the target will fail to be located. To efficiently address the low-quality response, the response-consistent module is introduced, which aims to constrain the response variations to prevent the occurrence of distractor peaks. The formulation is as follows:

$$\mathcal{E}_1 = \gamma_1 \sum_{d=1}^{D} \left\| f_d \star m_d - f_d^{t-1} \star m_d^{t-1} \right\|_2^2, \tag{4}$$

where γ_1 is a regularization parameter. f_d^{t-1} and m_d^{t-1} represent the filter and appearance model of the previous frame, respectively. In this way, the filter f can inherit the discriminative ability of the f^{t-1}.

B. Response-Reversible Module

The consistency regularization imposes a unidirectional strong constraint on the filter and overly focuses on historical information. When the viewpoint of the UAV changes or the target deforms, filters often fail to make timely adjustments, making it difficult for them to adapt to long-term tracking tasks. Therefore, we introduce a relaxed reversibility regularization term to dynamically adjust the filters by minimizing the response error between forward tracking and backward relocation. Unlike BiCF [35], we incorporate the current information into the modeling process. This change alleviates over-fitting of the filter and is suitable for long-term tracking. The formulation is as follows:

$$\mathcal{E}_2 = \gamma_2 \sum_{d=1}^{D} \left\| \Delta_F \mathcal{R}_d - \Delta_B \mathcal{R}_d \right\|_2^2, \tag{5}$$

where γ_2 denotes the regularization parameter. $\Delta_F\mathcal{R}$ and $\Delta_B\mathcal{R}$ denote the forward tracking error and historical backward relocation error, respectively. And $\Delta_F\mathcal{R}_d$ and $\Delta_B\mathcal{R}_d$ can be computed by:

$$\begin{cases} \Delta_F \mathcal{R}_d = x_d \star f_d - x_d \star f_d^{t-1} \\ \\ \Delta_B \mathcal{R}_d = x_d^{t-1} \star f_d^{t-1} - x_d^{t-1} \star f_d, \end{cases} \tag{6}$$

where x_d denotes the d-th channel of the current feature maps rather than the appearance model. We further substitute Eq. (6) into Eq. (5) to simplify the regularization term:

$$\mathcal{E}_2 = \gamma_2 \sum_{d=1}^{D} \left\| (x_d + x_d^{t-1}) \star (f_d - f_d^{t-1}) \right\|_2^2. \tag{7}$$

C. Background-Aware Module

To address the boundary effect, we extend the search area, similar to existing methods, but this also introduced significant background noise. Therefore, we introduce two similar attention mask matrices to enhance the filter's learning ability for the background, i.e., the negative samples. The formulation for the background-aware module is as follows:

$$\mathcal{E}_3 = \gamma_3 \sum_{d=1}^{D} \left\| s \odot f_d \right\|_2^2 + \gamma_4 \sum_{d=1}^{D} \left\| f_d \star (w \odot x_d) \right\|_2^2, \tag{8}$$

where $\odot$ denotes the Hadamard product. $s \in \mathbb{R}^T$ is borrowed from SRDCF [9], and $w \in \mathbb{R}^T$ shares the same outline as s. In s and w, the values belonging to the target region are significantly lower than the background region. Note that s and w remain constant during the tracking process.

D. Modeling and Optimization of the RBLT

As shown in Fig. 2, the overall optimization problem of our RBLT is to minimize the training error $\mathcal{E}$:

$$\mathcal{E} = \sum_{d=1}^{D} \left\| y - f_d \star m_d \right\|_2^2 + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3. \tag{9}$$

Note that $\mathcal{E}$ can be decomposed into D error terms $\mathcal{E}_d (d = 1, 2, ..., D)$ for optimization, since the filter is trained independently on each channel. For simplified presentation, the subscript $()_d$ is omitted in the following derivation. The d-th sub-problem is expressed as follows:

$$\begin{aligned} \mathcal{E}_d(f) = & \left\| y - f \star m \right\|_2^2 + \gamma_1 \left\| f \star m - f^{t-1} \star m^{t-1} \right\|_2^2 \\ & + \gamma_2 \left\| \left(x + x^{t-1}\right) \star \left(f - f^{t-1}\right) \right\|_2^2 \\ & + \gamma_3 \left\| s \odot f \right\| + \gamma_4 \left\| f \star x_w \right\|, \end{aligned} \tag{10}$$

where x_w is a simplified representation of $w \odot x$.

For optimization, we introduce an auxiliary variable g in order to keep the forth term in the spatial domain by ordering $\hat{g} = \sqrt{T}\mathbf{F}f$ where $\mathbf{F} \in \mathbb{C}^{T \times T}$ denotes the orthonormal matrix. Then Eq. (10) is converted into the frequency domainm (Note that $f_{t-1} = f^{t-1}$.) :

$$\begin{aligned} \mathcal{E}_d(\hat{g}, f) = & \left\| \hat{y} - \hat{g}^* \odot \hat{m} \right\|_2^2 + \gamma_1 \left\| \hat{g}^* \odot \hat{m} - \hat{g}_{t-1}^* \odot \hat{m}_{t-1} \right\|_2^2 \\ & + \gamma_2 \left\| (\hat{x} + \hat{x}_{t-1}) \odot \left(\hat{g}^* - \hat{g}_{t-1}^*\right) \right\|_2^2 \\ & + \gamma_3 \left\| s \odot f \right\|_2^2 + \gamma_4 \left\| \hat{g}^* \odot \hat{x}_w \right\|, \end{aligned} \tag{11}$$

where the superscript $\hat{}$ and * are the DFT of a signal and the conjugate of a complex vector, respectively. Due to the convexity of the proposed formulation, we apply the augmented Lagrange method to optimize Eq. (11):

$$\mathcal{L}_d(\hat{g}, f, \hat{v}) = \mathcal{E}_d(\hat{g}, f) + \theta \left\| \hat{g} - \sqrt{T}\mathbf{F}f + \frac{1}{\theta}\hat{v} \right\|_2^2, \tag{12}$$

where $\hat{v} \in \mathbb{C}^T$ denotes the Fourier transform of the Lagrange multiplier and θ denotes the step size regularization parameter. Then the Alternating Direction Method of Multipliers (ADMM) [36] is employed to perform iterative optimization with guaranteed convergence [37] as follows:

$$\hat{g}^{(i+1)} = \frac{\hat{m} \odot \hat{y}^* + \left(\gamma_1 \hat{S}_m + \gamma_2 \hat{S}_x\right) \odot \hat{g}_{t-1} + \theta^{(i)} \hat{f}^{(i)} - \hat{v}^{(i)}}{(1 + \gamma_1)\, \hat{m} \odot \hat{m}^* + \gamma_2 \hat{S}_x + \gamma_4 \left(\hat{x}_w \odot \hat{x}_w^*\right) + \theta^{(i)}}, \tag{13a}$$

$$f^{(i+1)} = \frac{\mathcal{F}^{-1}\left(\theta^{(i)} \hat{g}^{(i+1)} + \hat{v}^{(i)}\right)}{\frac{\gamma_3}{T}\left(s \odot s^*\right) + \theta^{(i)}}, \tag{13b}$$

$$\hat{v}^{(i+1)} = \hat{v}^{(i)} + \theta^{(i)} \left(\hat{g}^{(i+1)} - \hat{f}^{(i+1)}\right), \tag{13c}$$

where $\hat{S}_m = \hat{m} \odot \hat{m}_{t-1}^*$ and $\hat{S}_x = (\hat{x} + \hat{x}_{t-1}) \odot (\hat{x}^* + \hat{x}_{t-1}^*)$. The superscripts $^{(i)}$ indicates the i-th iteration. Moreover, the complexity of RBLT is bounded by $O(N_{ADMM} DT \log T)$, where N_{ADMM} denotes the number of ADMM iterations.

E. Object Localization and Scale Estimation

By searching the maximum value in the response calculated by Eq. (14), the tracker can estimate the optimal location of the target.

$$\mathcal{R}^{t+1} = \mathcal{F}^{-1}\left(\hat{z}^{t+1} \odot \hat{g}^*\right). \tag{14}$$

For scale estimation, some CFs [11] apply the learned filter on multiple resolutions of the searching area to estimate scale changes, and then select the optimal scale with the maximum response, while others employ an additional discriminative

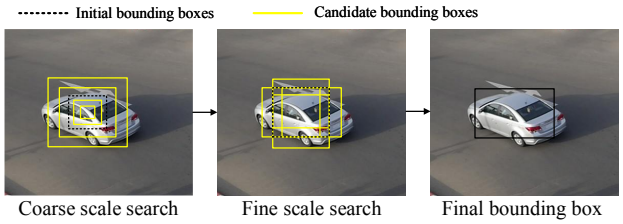


Fig. 3. Visualization of the coarse-to-fine scale search strategy.

scale filter [12] to estimate scale changes for faster speed. However, both of these methods assume that the aspect ratio of the target is constant. When the aspect ratio changes, they may mistakenly identify part of the background as the target.

In this work, we propose a fast coarse-to-fine scale search strategy. As shown in Fig. 3, in the first stage, a multi-resolution coarse scale search is conducted while assuming a constant aspect ratio. In the second stage, a fine search with varying aspect ratios is carried out near the optimal scale obtained from the coarse search.

F. Design of the DeepRBLT

The fundamental difference between RBLT and DeepRBLT lies in the different features used for object representation. The RBLT uses a combination of histogram of oriented gradients (HOG) [38] and color names (CN) [21] features. And the DeepRBLT's feature representation is a combination of the first, second and last convolutional layer in the VGG-m [39] network, along with HOG features. On this basis, we make the following adjustments:

Firstly, we initialize a projection matrix $P \in \mathbb{R}^{D\times C}$ by principal component analysis (PCA) to perform dimensionality reduction on the features. The new features $x_p \in \mathbb{R}^{T\times C}$ can be computed by $x_p = xP$.

Secondly, we assign values to each parameter based on the depth distinction of features. Generally, features with deeper layers have stronger invariance. Therefore, in order to enhance the robustness, the learning rate η for deeper features should be smaller, along with smaller values for each regularization parameter.

V. EXPERIMENTS

A. Experimental Setups

1) Implementation details: The proposed RBLT and DeepRBLT are implemented in MATLAB R2021a on a computer with an i5-13600K CPU and an NVIDIA GeForce RTX 4060 GPU. The specific parameter settings can be found in our code, which is available here : https://github.com/Wangbiao2/RBLT-tracker.

2) Evaluation metrics: In our work, two metrics of precision and success rate (SR) are employed to evaluation all trackers by the one-pass evaluation [40]. The precision is used to measure the center location error (CLE) between the estimated bounding box and the ground-truth (GT) bounding box. The precision plot represents the curve formed by the percentage of bounding boxes whose CLE is within a given threshold. SR is applied to measure the intersection over union (IoU) of the estimated and the GT bounding boxes.

TABLE I

AVERAGE AUC, DP, AND FPS OF THE TOP-6 CPU-BASED TRACKERS ARE REPORTED BY AVERAGING THE RESULTS OF THE FOUR UAV BENCHMARKS. RED, GREEN AND BLUE FONTS INDICATE THE TOP THREE RESULTS, RESPECTIVELY.

	STRCF	ECO-HC	BiCF	ARCF	AutoTrack	**RBLT**
Avg. AUC ⇑	0.468	0.477	0.491	**0.497**	**0.495**	**0.512**
Avg. DP ⇑	0.671	0.693	0.713	**0.719**	**0.723**	**0.736**
Avg. FPS ⇑	26	**58**	**45**	24	**51**	40

The success plot is used to visualize the ratio of the number of frames where the IoU exceeds a given threshold. According to [40], trackers are ranked by the distance precision (DP) at a CLE threshold of 20 pixels in the precision plot, and the area under the curve (AUC) is applied to rank the SR of trackers in the success plot. In addition, frames per second (FPS) is used to evaluate the tracking speed.

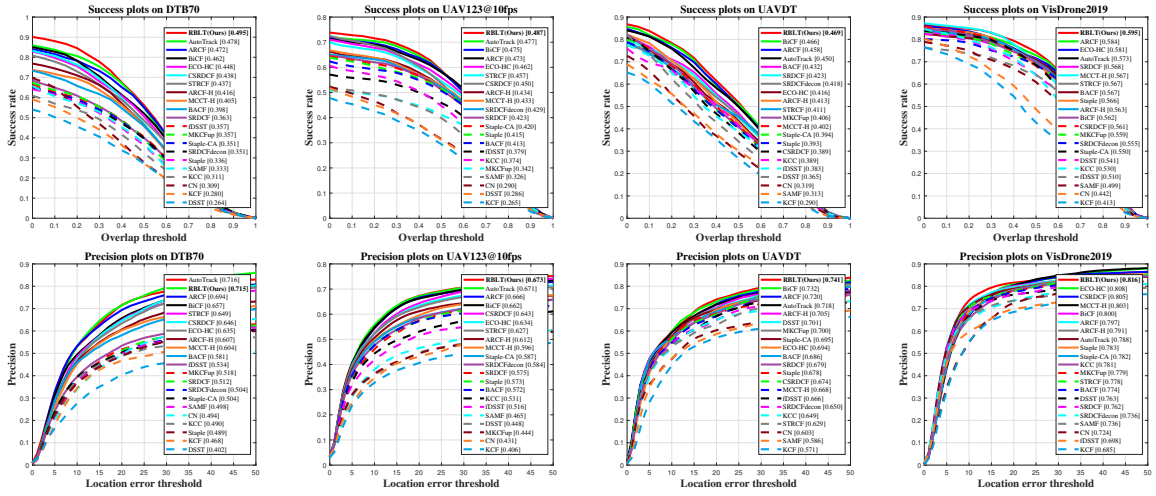
B. Comparison with CPU-based Trackers

The proposed RBLT is evaluated extensively on four UAV benchmarks with the other 20 CPU-based trackers, including KCF [8], SAMF [11], DSST [12], fDSST [13], CN [21], KCC [41], Staple [42], Staple-CA [33], SRDCF [9], SRDCFdecon [43], BACF [10], MCCT-H [25], MKCFup [19], ECO-HC [27], ARCF-H, ARCF [30], STRCF [28], BiCF [35], CSRDCF [44] and AutoTrack [45].

Fig. 4 shows that the proposed RBLT performs significantly better than other CPU-based trackers on all four UAV benchmarks. In terms of AUC, RBLT performs the best among all four benchmarks, surpassing the second-best tracker by 3.6%, 2.1%, 0.6%, and 1.9%, respectively. The RBLT achieves the second best DP score on DTB70, trailing behind AutoTrack by 0.1%. But on UAV123@10fps, UAVDT and VisDrone2019, RBLT has an advantage of 0.3%, 1.2% and 1% over the second-best trackers in terms of DP. TABLE I provides the average AUC, DP and speed of top-6 ranked trackers running on a single CPU over all four benchmarks. In terms of average AUC and DP, RBLT outperforms the second-ranked trackers by 3.0% and 1.8%, respectively. And RBLT has an average tracking speed of 40 FPS, meeting the real-time UAV tracking requirement.

C. Comparison with GPU-based Trackers

Comprehensive comparison is conducted on DTB70 and UAV123@10fps benchmarks between the proposed DeepRBLT and 16 other deep trackers, which consist of 8 CF-based trackers using deep features and 8 deep learning (DL)-based trackers. The CF-based trackers include KAOT [34], ECO [27], DeepSTRCF [28], CF2 [46], CCOT [26], MCCT [25], ASRCF [29], and CoKCF [47], while the DL-based trackers include DSiam [48], Ocean [49], TADT [50], DaSiamRPN [51], SiamFC [52], UpdateNet [53], UDTplus [54] and SE-SiamFC [55]. As show in Fig. 5, DeepRBLT achieved the highest AUC (0.530) and DP (0.790) on DTB70, surpassing the second-ranked tracker by 2.5% and 2.7%, respectively. And on UAV123@10fps, DeepRBLT achieves



(a) DTB70 (b) UAV123@10fps (c) UAVDT (d) VisDrone2019

Fig. 4. Success and precision plots of RBLT and other 20 CPU-based trackers on the four benchmarks. Note that AUC and DP are given in the legend.

the best DP of 0.742 and the highest AUC score of 0.524. It is worth noting that DeepRBLT has a tracking speed of ∼10 FPS, which is superior to most deep trackers.

TABLE II reports that RBLT outperforms other 15 state-of-the-art deep trackers in terms of both AUC and DP on the UAVDT benchmark. Moreover, as shown in TABLE III, on the VisDrone2019 benchmark, the CPU-based RBLT still outperforms most Siamese trackers based on large-scale offline training. In addition, the computational cost of RBLT is much lower than those deep trackers, requiring only one CPU to achieve real-time speed. The above results provide evidence supporting the effectiveness of the joint response and background learning approach proposed in this paper.

In order to further validate the generality of our method, we evaluate DeepRBLT on the well-known GOT10K benchmark [18]. Note that AO denotes the average of overlap rates between tracking results and GTs over all frames. TABLE IV reports that DeepRBLT demonstrates the best performance compared to the other four advanced trackers, which verifies that our method is not only applicable to UAV tracking but also to general tracking tasks.

D. Attribute-based Evaluation

DTB70 [14], UAV123@10fps [15], VisDrone2019 [16], and UAVDT [17] are fully annotated by 11, 12, 12, and 9 challenging attributes, respectively. Our trackers demonstrate the best performance under challenges such as fast motion, camera motion and background clutter, aligning with the design intention of our joint response and background learning method. More specifically, in case of background clutter, RBLT demonstrates an improvement of performance by 10.7% in UAVDT, and 15.3% in VisDrone2019 against ECO-HC [27]. The ability to cope with object/UAV motion is also promoted by 27.4% in DTB70, 18% in UAV123@10fps and 5.9% in UAVDT from the baseline SRDCF [9]. Thanks to the proposed coarse-to-fine scale search strategy, the ability of DeepRBLT to handle aspect ratio change even surpasses many state-of-the-art Siamese trackers.

TABLE II
PERFORMANCE COMPARISON OF RBLT WITH OTHER 15 STATE-OF-THE-ART DEEP TRACKERS ON UAVDT BENCHMARK. NOTE THAT THE SUPERSCRIPT * DENOTES GPU SPEED.

Trackers	AUC	DP	FPS	Trackers	AUC	DP	FPS
CF2 [46]	0.355	0.602	20*	CCOT [26]	0.406	0.656	1*
CoKCF [47]	0.319	0.605	21*	TADT [50]	0.431	0.677	31*
ECO [27]	0.454	0.700	16*	DSTRCF [28]	0.437	0.667	5*
IBCCF [56]	0.389	0.603	3*	UDTplus [54]	0.416	0.697	57*
MCPF [57]	0.403	0.660	0.6*	ASRCF [29]	0.437	0.700	14*
MCCT [25]	0.437	0.671	8*	SiamFC [52]	0.465	0.708	18*
ADNet [58]	0.319	0.605	7*	CFNet [59]	0.428	0.680	40*
DSiam [48]	0.457	0.704	16*	**RBLT**	**0.469**	**0.741**	**40**

TABLE III
PERFORMANCE COMPARISON OF RBLT WITH OTHER 11 STATE-OF-THE-ART SIAMESE TRACKERS ON VISDRONE2019.

Trackers	AUC	DP	FPS	Trackers	AUC	DP	FPS
DaSiamRPN [51]	0.536	0.763	20*	LightTrack [60]	0.579	0.754	19*
Ocean [49]	0.500	0.706	15*	SE-SiamFC [55]	0.544	0.734	5*
SiamBAN [61]	0.601	0.806	6*	SiamCAR [65]	0.630	0.838	6*
SiamFC++ [62]	0.609	0.788	17*	SiamGAT [66]	0.606	0.811	17*
SiamMask [63]	0.588	0.806	13*	UpdateNet [53]	0.562	0.790	12*
SiamAPN [64]	0.575	0.802	35*	**RBLT**	0.595	0.816	40

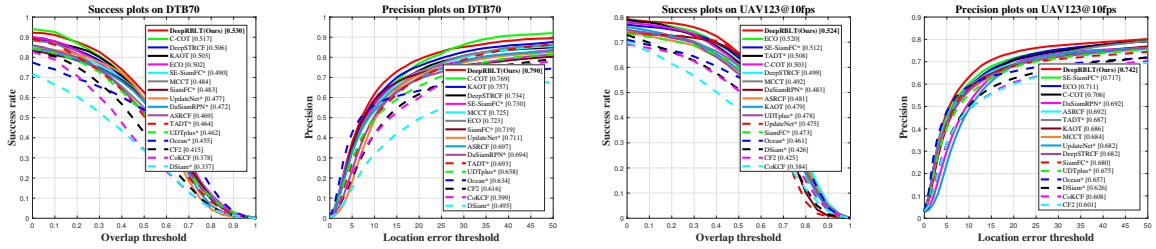
TABLE IV
EVALUATION OF 5 ADVANCED TRACKERS ON GOT10K.

	ECO	CCOT	SiamFC	CFNet	**DeepRBLT**
AO ⇑	0.316	0.325	**0.348**	**0.374**	**0.409**
$SR_{0.5}$ ⇑	0.309	0.328	**0.353**	**0.404**	**0.417**
$SR_{0.75}$ ⇑	**0.111**	0.107	0.098	**0.144**	**0.137**

E. Ablation Study

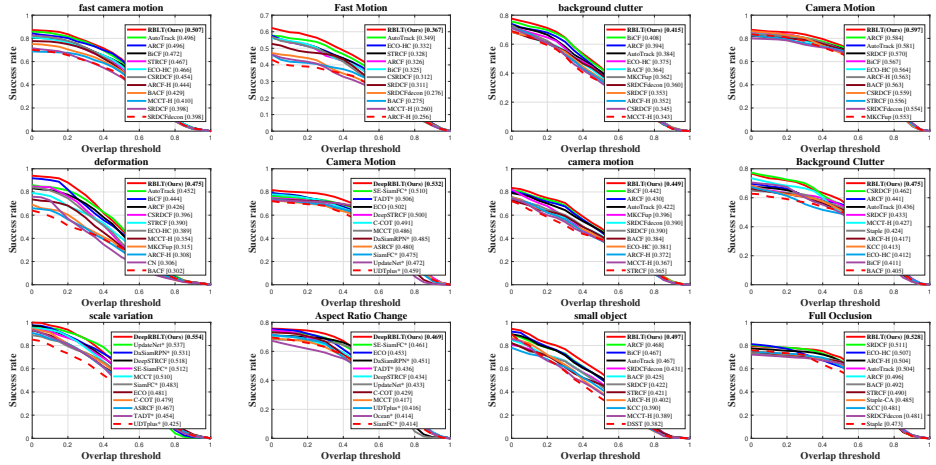
We evaluate the effect of each innovative modules in RBLT, including the response-consistent module (RC), response-reversible module (RR) and background-aware module (BA). The baseline (BL) is the SRDCF [9], i.e. the first and forth terms of Eq. (10), which is equipped with the same features and learning rate as the RBLT.

As shown in TABLE V, the various modules we proposed have all brought positive improvements to the performance



(a) DTB70 (b) UAV123@10fps

Fig. 5. Success and precision plots of DeepRBLT and other 16 GPU-based deep trackers. Note that DL-based trackers are marked with *.



(a) DTB70 (b) UAV123@10fps (c) UAVDT (d) VisDrone2019

Fig. 6. Attribute-based performance evaluation. Success plots of RBLT, DeepRBLT and other trackers on UAV-specific attributes from four benchmarks.

TABLE V

ABLATION STUDY OF THE PROPOSED RBLT.

	DTB70		UAV123@10fps	
	AUC	DP	AUC	DP
BL	0.468	0.687	0.442	0.631
BL+RC	0.476	0.690	0.459	0.638
BL+RC+RR	0.488	0.696	0.468	0.652
BL+RC+RR+BA	0.495	0.715	0.487	0.673

TABLE VI

ANALYSIS OF THE EFFECTIVENESS OF THE PROPOSED COARSE-TO-FINE SCALE SEARCH STRATEGY.

	DTB70		UAV123@10fps	
	AUC	DP	AUC	DP
DeepRBLT	0.524	0.790	0.519	0.732
DeepRBLT+CFSE	0.530	0.790	0.524	0.742
ECO	0.502	0.723	0.520	0.711
ECO+CFSE	0.511	0.740	0.528	0.729

of the tracker. To further validate the effectiveness of the proposed coarse-to-fine scale search strategy (CFSE), we conduct experiments on two benchmarks by comparing DeepRBLT, ECO [27] and their improved versions based on CFSE. TABLE VI reports that the improvement in performance due to CFSE is clearly evident.

F. Limitations

- When extracting features from deeper pre-trained networks (*e.g.* ResNet-50 [67]) for object representation, the speed of the tracker decreases significantly, but the improvement in performance is not substantial.
- Though DeepRBLT performs favorably in the situations of fast motion and background clutter, it is still limited when the object disappear for a long time. Therefore, it is imperative to equip the tracker with an efficient re-detection module.
- Compared to the commonly used scale regression methods of the Siamese trackers, the scale search strategy is still inefficient. Therefore, in the future, a stronger and faster tracker can be designed by combining CFs with scale regression methods.

VI. CONCLUSIONS

For UAV visual tracking, we propose joint response and background learning correlation filters which can suppress abnormal response and background noise. Extensive experiments on five authoritative benchmarks have validated our trackers perform favorably in precision, with enough speed for real-time applications. Furthermore, the coarse-to-fine scale searching strategy can be extended to other trackers to further enhance their performance. Employing the RBLT for UAV tracking in the real environment is also taken into consideration to conduct in the future. We believe that our methods can promote the development of UAV tracking.

ACKNOWLEDGMENT

This work was supported by NSFC (62376015, 61976013, U22B2038, 61973044 and 62073020).

REFERENCES

[1] G. J. Laguna and S. Bhattacharya, "Path planning with Incremental Roadmap Update for Visibility-based Target Tracking," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1159-1164.

[2] A. A. Meera, M. Popovic, A. Millane, and R. Siegwart, "Obstacle-Aware Adaptive Informative Path Planning for UAV-based Target Search," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 718-724.

[3] S. Lin, M. A. Garratt, and A. J. Lambert, "Monocular vision-based real-time target recognition and tracking for autonomously landing an uav in a cluttered shipboard environment," *Autonomous Robots*, vol. 41, no. 4, pp. 881-901, 2017.

[4] C. Fu, A. Carrio, M. A. Olivares-Mendez, and P. Campoy, "Online learning-based robust visual tracking for autonomous landing of Unmanned Aerial Vehicles," in *Proceedings of International Conference on Unmanned Aircraft Systems (ICUAS)*, 2014, pp. 649-655.

[5] B. Patel, T. D. Barfoot, and A. P. Schoellig, "Visual Localization with Google Earth Images for Robust Global Pose Estimation of UAVs," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6491-6497.

[6] C. Fu, A. Carrio, M. A. Olivares-Mendez, R. Suarez-Fernandez, and P. Campoy, "Robust Real-Time Vision-Based Aircraft Tracking from Unmanned Aerial Vehicles," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 5441-5446.

[7] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2010, pp. 2544-2550.

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

[9] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 4310-4318.

[10] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 1144-1152.

[11] Li, Yang and Zhu, Jianke, "A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration", in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2015, pp. 254-265.

[12] M. Danelljan, G. Hger, F. Shahbaz Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2014, pp. 1-11.

[13] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1561-1575, 2017.

[14] S. Li and D.-Y. Yeung, "Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models," in *AAAI Conference on Artificial Intelligence*, 2017, pp. 4140-4146.

[15] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 445-461.

[16] H. Fan, L. Wen, D. Du, P. Zhu, Q. Hu et al., "VisDrone-SOT2020: The Vision Meets Drone Single-Object Tracking Challenge Results," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 728-749.

[17] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 375-391.

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

[19] M. Tang, B. Yu, F. Zhang and J. Wang, "High-Speed Tracking with Multi-kernel Correlation Filters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4874-4883.

[20] H. K. Galoogahi, T. Sim and S. Lucey, "Multi-channel Correlation Filters," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3072-3079.

[21] M. Danelljan, F. S. Khan, M. Felsberg, and J. Van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1090-1097.

[22] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler, "Joint group feature selection and discriminative filter learning for robust visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7950-7960.

[23] T. Xu, Z. -H. Feng, X. -J. Wu and J. Kittler, "Learning Low-Rank and Sparse Discriminative Correlation Filters for Coarse-to-Fine Visual Object Tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3727-3739, Oct. 2020.

[24] J. Zhang, S. Ma, and S. Sclaroff, "MEEM: Robust tracking via multiple experts using entropy minimization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 188-203.

[25] N. Wang, W. Zhou, Q. Tian, R. Hong, M. Wang, and H. Li, "Multi-cue correlation filters for robust visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 4844-4853.

[26] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, "Beyond correlation filters: Learning continuous convolution operators for visual tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 472-488.

[27] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6931-6939.

[28] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial-temporal regularized correlation filters for visual tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 4904-4913.

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

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

[31] J. Ye, C. Fu, F. Lin, F. Ding, S. An and G. Lu, "Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6004-6014, June 2022.

[32] Y. Li, H. Zhang, Y. Yang, H. Liu and D. Yuan, "RISTrack: Learning Response Interference Suppression Correlation Filters for UAV Tracking," in *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1-5, 2023.

[33] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1387-1395.

[34] Y. Li, C. Fu, Z. Huang, Y. Zhang and J. Pan, "Keyfilter-Aware Real-Time UAV Object Tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 193-199.

[35] F. Lin, C. Fu, Y. He, F. Guo, and Q. Tang, "BiCF: Learning bidirectional incongruity-aware correlation filter for efficient UAV object tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2365-2371.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," in *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1-122, 2011.

[37] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. "The matrix cookbook." *Technical University of Denmark*, 7(15):510, 2008.

[38] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sep. 2010.

[39] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., "Return of the devil in the details: Delving deep into convolutional nets," 2014, arXiv:1405.3531.

[40] Y. Wu, J. Lim, and M.-H. Yang, "Object Tracking Benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848, 2015.

[41] C. Wang, L. Zhang, L. Xie, and J. Yuan, "Kernel cross-correlator," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 4179-4186.

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

[43] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1430-1438.

[44] A. Lukezic, T. Vojir, L. Cehovin Zajc, J. Matas, and M. Kristan, "Discriminative Correlation Filter with Channel and Spatial Reliability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6309-6318.

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

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

[47] L. Zhang and P. N. Suganthan, "Robust visual tracking via co-trained kernelized correlation filters," *Pattern Recognition*, vol. 69, pp. 82-93, Sep. 2017.

[48] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, "Learning dynamic siamese network for visual object tracking," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 1781-1789.

[49] Zhang, Z., Peng, H., Fu, J., Li, B., Hu, W., "Ocean: Object-Aware Anchor-Free Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 771-787.

[50] X. Li, C. Ma, B. Wu, Z. He, and M.-H. Yang, "Target-aware deep tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 1369-1378.

[51] 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)*, 2018, pp. 103-119.

[52] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2016, pp. 850-865.

[53] L. Zhang, A. Gonzalez-Garcia, J. V. D. Weijer, M. Danelljan, and F. S. Khan, "Learning the Model Update for Siamese Trackers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 4009-4018.

[54] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, "Unsupervised Deep Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1308-1317.

[55] I. Sosnovik, A. Moskalev, and A. Smeulders, "Scale Equivariance Improves Siamese Tracking," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2764-2773.

[56] F. Li, Y. Yao, P. Li, D. Zhang, W. Zuo, and M.-H. Yang, "Integrating boundary and center correlation filters for visual tracking with aspect ratio variation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2017, pp. 2001-2009.

[57] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 4819-4827.

[58] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-decision networks for visual tracking with deep reinforcement learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 1349-1358.

[59] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5000-5008.

[60] B. Yan, H. Peng, K. Wu, D. Wang, J. Fu, and H. Lu, "LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15175-15184.

[61] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese Box Adaptive Network for Visual Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6667-6676.

[62] Y. Xu, Z. Wang, Z. Li, Y. Yuan, and G. Yu, "SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12549-12556.

[63] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. Torr, "Fast Online Object Tracking and Segmentation: A Unifying Approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1328-1338.

[64] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, "Siamese Anchor Proposal Network for High-Speed Aerial Tracking," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 510-516.

[65] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6268-6276.

[66] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph Attention Tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9538-9547.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.