End-to-End Feature Decontaminated Network for UAV Tracking

Haobo Zuo, Changhong Fu*, Sihang Li, Junjie Ye, and Guangze Zheng

***Abstract*— Object feature pollution is one of the burning issues in vision-based UAV tracking, commonly caused by occlusion, fast motion, and illumination variation. Due to the contaminated information in the polluted object features, most trackers fail to precisely estimate the object location and scale. To address the above disturbing issue, this work proposes a novel end-to-end feature decontaminated network for efficient and effective UAV tracking, *i.e.*, FDNT. FDNT mainly includes two modules: a decontaminated downsampling network and a decontaminated upsampling network. The former reduces the interference information of the feature pollution and enhanced the expression of the object location information with two asymmetric convolution branches. The latter restores the object scale information with the super-resolution technology-based low-to-high encoder, achieving a further decontamination effect. Moreover, a novel pooling distance loss is carefully developed to assist the decontaminated downsampling network in concentrating on the critical regions with the object information. Exhaustive experiments on three well-known benchmarks validate the effectiveness of FDNT, especially on the sequences with feature pollution. In addition, real-world tests show the efficiency of FDNT with 31.4 frames per second. The code and demo videos are available at `https://github.com/vision4robotics/FDNT`.**

I. Introduction

Vision-based UAV tracking is a significant branch in intelligent object tracking for numerous practical applications [1]–[3]. Based on the chosen object in the initial frame, trackers offer a fast prediction of the object position and size as soon as a new frame is acquired by the UAV. Since the complex scenarios like occlusion, fast motion, and illumination variation are more common from the UAV perspectives, object feature pollution is a more serious issue in aerial tracking situations than it is in ordinary tracking scenes. Despite recent advances in improving tracking performance, UAV tracking is still plagued by the problem of feature pollution. Siamese tracking [4]–[7], regarded as a type of state-of-the-art (SOTA) approach, is typically capable of handling a wide range of conditions in UAV tracking [8]–[10]. Siamese trackers are characterized by the template matching with the convolutional neural networks (CNNs), providing robust tracking performance. Nevertheless, lightweight CNNs such as AlexNet [11] struggle to effectively identify the object features in the presence of pollution, making robust tracking performance difficult in the challenging tracking situations. While increasing the kernel size or the depth of the backbone [12] will ease some of the problems of feature pollution, tracking efficiency and practicability will suffer as a consequence, particularly in UAV tracking.

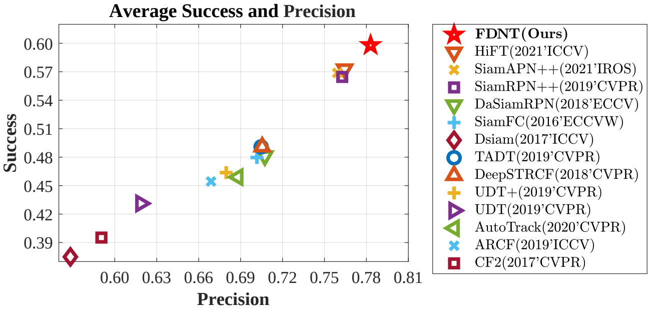

Fig. 1. An extensive comparison of the proposed FDNT with state-of-the-arts (SOTA) on three challenging UAV tracking benchmarks. The authoritative benchmarks are UAV123 [34], UAV123@10fps [34], and UAVTrack112_L [33], respectively. FDNT achieves more robust performance than other SOTA trackers in the average precision and success.

Recently, some image denoising and deblurring methods [13]–[16] utilize downsampling and upsampling to solve the problems respectively. Specifically, the denoising method [13] applies an invertible downsampling operation to expand the respective field. Furthermore, InvDN [14] downscales the input image to a latent representation, separating the low-resolution and high-frequency parts for further denoising. Moreover, downsampling can reduce the processing load on devices with limited computational capability, such as UAV platforms [17]. While the image deblurring can use the upsampling method to supplement specific details for the blurred image. BMDSRNet [15] and SRDN [16] utilize the super-resolution networks to handle the motion deblurring difficulties, achieving the significant effect. Noise and blur are the main parts of the object feature pollution issues, which are frequently produced by the challenging UAV tracking scenes. However, the current upsampling and downsampling techniques can only partially address the feature contamination problems. Besides, since the size of the downsampled image is reduced, the tracked object scale information is also changed. Thereby the trackers can be misled by the incorrect object information. While directly upsampling the image will increase additional pollution information unrelated to the tracked object. In the meantime, the image upsampling is hard to satisfy the real-time needs of practical applications for UAV tracking since the enlarged image expands the processing time.

Consequently, an efficient and effective tracker is required to cope with object feature pollution. This work first applies the downsampling technology to remove the interference in the object features, thereby enhancing the object location information. Furthermore, to exactly estimate the object size during the UAV tracking process, the upsampling technology is used to restore the object scale information changed by

Authors are with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China. changhongfu@tongji.edu.cn

*Corresponding author

the downsampling. Through the feature downsampling and upsampling, the clean object features can be obtained for precise tracking results.

In the proposed tracker, a novel end-to-end feature decontaminated network-based tracker (FDNT) for UAV tracking is proposed. FDNT mainly consists of two parts, a decontaminated downsampling network to strengthen the location information of the object and a decontaminated upsampling network to restore the object scale information. Specifically, the decontaminated downsampling network applies two asymmetric convolution branches to reduce the polluted object features. Through the downsampling process, the low-resolution object features effectively place importance on the overall position information of the tracked object, alleviating the influence of the feature pollution. Moreover, the low-resolution object features can reduce the amount of calculation for the practical requirement of UAV tracking. Subsequently, through the decontaminated upsampling network, the high-resolution features can be obtained with more scale information. Especially, the low-to-high (LTH) encoder in the upsampling network is proposed to remove the additional contamination information. In addition, the pooling distance (PD) loss is designed to enhance the decontamination capability of the decontaminated downsampling network. The proposed FDNT has efficiently achieved robust performance in three challenging aerial tracking benchmarks [33], [34], and the average results are as shown in Fig. 1, where FDNT has the most robust performance. The primary contributions of this work are as follows:

- A decontaminated downsampling network is proposed to decrease the polluted object features with the supervision of the carefully designed PD loss.
- A decontaminated upsampling network is presented to strengthen the scale information of the tracked object, with the LTH encoder for further decontamination.
- Exhaustive evaluations on three challenging aerial benchmarks prove the promising tracking performance of FDNT to SOTA trackers. Real-world experiments are carried out on a common aerial robot, validating the practicability of FDNT in real tracking applications.

II. Related Works

A. UAV Tracking with Correlation Filter

Due to the computational limitations of the UAVs, the tracking algorithms must be lightweight in order to provide real-time performance for the practical applications. With the aerial perspectives, UAV tracking is more likely to face challenges of difficult scenes, which makes robust tracking difficult. Hence, it is necessary to realize a favorable trade-off between speed and performance. Previously, the correlation filter (CF)-based trackers [18], [19] have drawn considerable interest due to their computational efficiency, making them well-suited for the real-time UAV applications [20]. However, the cyclically shifted samples for training the CF-based trackers are not real samples. Thus, the effects are not ideal when object feature pollution is present.

B. UAV Tracking with Siamese Network

Siamese-based tracker is another well-known branch of UAV tracking. Siamese networks [4]–[7], which solve the tracking problem through template matching, have made tremendous strides in recent years. As a pioneering study, SiamFC [4] demonstrates the Siamese framework's advantages by defining tracking as the similarity matching process between the template and the search patch. SiamRPN [7], inspired by the region proposal network (RPN) [21], divides tracking into two subtasks using classification and regression branches. Siamese trackers [8]–[10], based on SiamRPN, distinguish the tracked object characteristics in a variety of UAV settings with a balance of efficiency and effect. Although the existing UAV tracking approaches work well in a range of tracking scenes, when extracting robust search features is challenging owing to a lack of discriminative object information, the matching process is typically ineffective. Especially, object feature pollution which is caused by frequent problems in UAV tracking such as occlusion, fast motion, and illumination variation, will have a significant impact on the performance of trackers. To handle the problem of feature pollution effectively and efficiently, this work proposes a novel feature decontamination framework, achieving robust UAV tracking performance.

C. Image Denoising & Deblurring

Downsampling has been increasingly common for image denoising jobs in recent years. SDCNN [13] uses the BDCT-based downsampling operation to effectively improve the reconstruction quality by incorporating frequency information with spatial context. Additionally, InvDN [14] applies several DownScale Blocks to remove the noise from the noisy image. This approach samples a fresh latent variable from the previous distribution and combines it with the low-resolution image to reconstruct the information lost due to downsampling.

The technology of upsampling is frequently employed for the image deblurring process. BMDSRNet [15] applies the super-resolution technology to assist in the acquisition of dynamic spatio-temporal information from a single static motion-blurred image. SRDN [16] develops the space target super-resolution network to effectively extract the global feature, further rebuilding the high-resolution and clean images. However, the image downsampling and upsampling can just handle image denoising and deblurring respectively, only part of the contamination issues. They are unable to address the object feature pollution issues in UAV tracking, easily caused by the challenging scenarios, such as occlusion, fast motion, and illumination variation.

III. Methodology

This section introduces the proposed FDNT in depth. As depicted in Fig. 2, FDNT can be divided into four submodules, *Feature extraction network*, *Decontaminated downsampling network*, *Decontaminated upsampling network*, and *Classification & regression network*.

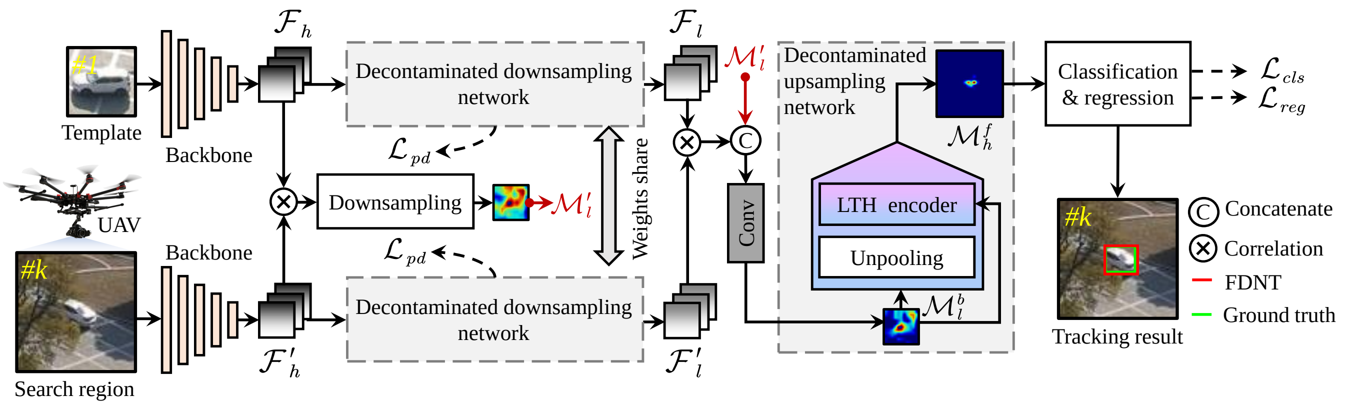

Fig. 2. Overview of the proposed FDNT. The components from the left to right are *Feature extraction network*, *Decontaminated downsampling network*, *Decontaminated upsampling network*, and *Classification & regression network*. Best viewed in color (the image frames from *person16* in UAV123@10fps [34]).

A. *Feature Extraction Network*

To meet the real-time applications onboard the embedded platform, FDNT uses a lightweight backbone known as AlexNet [11], which is used in both the template and search branches to extract the features. The last layer output feature maps of the search branch and template branch can be utilized in the subsequent process.

Remark 1: Since the template and search region features will pass the weight-sharing decontaminated downsampling network with the same process, the input and output of the network are uniformly represented by $\mathcal{F}_h \in \mathbb{R}^{W\times H\times C}$ and $\mathcal{F}_l \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$ in the following introduction (C, W, H represent the channel, width, and height of the feature maps respectively).

B. *Decontaminated Downsampling Network*

The decontaminated downsampling network consists of two asymmetric convolution branches. As shown in Fig. 3, the first branch is to downsample the high-resolution features $\mathcal{F}_h$, and then enhance the features with several convolution (Conv) layers. Thereby the preliminary low-resolution features $\mathcal{F}_l^a \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$ is formed:

$$\mathcal{F}_l^a = \mathrm{Conv}(\mathrm{Downsample}(\mathcal{F}_h)) \quad . \tag{1}$$

To better enhance the useful information, the second branch first applies some convolution layers to explore the valid information in $\mathcal{F}_h$, and then reduce the resolution to the output features $\mathcal{F}_l^b \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$. After the above two branches, $\mathcal{F}_l \in \mathbb{R}^{\frac{W}{2}\times\frac{H}{2}\times C}$ can be obtained with the concatenation (Cat) and Conv:

$$\begin{aligned} \mathcal{F}_l^b &= \mathrm{Downsample}(\mathrm{Conv}(\mathcal{F}_h)) \quad , \\ \mathcal{F}_l &= \mathrm{Conv}(\mathrm{Cat}(\mathcal{F}_l^a, \mathcal{F}_l^b)) \quad . \end{aligned} \tag{2}$$

Subsequently, the low-resolution features of the template and search are cross-correlated to the low-resolution $\mathcal{M}_l^a \in \mathbb{R}^{\frac{W_1}{2}\times\frac{H_1}{2}\times C}$. To reserve more object information, the high-resolution features of the template and search are cross-correlated to the initial high-resolution $\mathcal{M}_h' \in \mathbb{R}^{W_2\times H_2\times C}$. After the downsampling with a convolution layer, $\mathcal{M}_h'$ is converted to $\mathcal{M}_l' \in \mathbb{R}^{\frac{W_1}{2}\times\frac{H_1}{2}\times C}$. Thereby, $\mathcal{M}_l^a$ and $\mathcal{M}_l'$ are combined to $\mathcal{M}_l^b \in \mathbb{R}^{\frac{W_1}{2}\times\frac{H_1}{2}\times C}$ through Cat and Conv.

Remark 2: $\mathcal{M}_l^a$ strengthens the low-resolution features and $\mathcal{M}_l'$ enhances the high-resolution features. Afterward, $\mathcal{M}_l^a$ and $\mathcal{M}_l'$ are fused to obtain the robust low-resolution similarity maps, precisely expressing the location of the object.

C. *Pooling Distance Loss*

To better train the decontaminated downsampling network, this work proposes a novel PD loss to supervise the downsampling process. PD between the high-resolution features H and the low-resolution features L is defined as follows:

$$\mathcal{L}_{\mathrm{pd}} = \frac{\gamma}{S_H}\sum_{l=1}^{S_L}\min_{h\in S_H}\|L(l) - H(h)\|_2^2 \quad , \tag{3}$$

where γ, S_L, S_H, $L(l)$, and $H(h)$ represent the weight of this loss function, the multiplication between the width and height of L, the multiplication between the width and height of H, the l position values of all channels in L, and the h position values of all channels in H, respectively. By optimizing PD loss, low-resolution features obtained can effectively place importance on the object position information.

Remark 3: By adjusting the weight value in the loss function, the downsampling process can reach an ideal effect. In particular, the loss can reduce the interference information in the low-resolution features, hence improving the relevance to the object.

D. *Decontaminated Upsampling Network*

The decontaminated upsampling network utilizes the unpooling to obtain the high-resolution features $\mathcal{M}_h^a \in \mathbb{R}^{W_1\times H_1\times C}$ with more scale information, which has been reduced in the low-resolution features $\mathcal{M}_l^b$. Furthermore, the LTH encoder exploits $\mathcal{M}_l^b$ to ensure that the information recovered in the high-resolution features has no interference, illustrated in Fig. 3. Subsequently, The LTH encoder is introduced in detail.

Before encoding, the input high-resolution and low-resolution feature maps are reshaped to $\mathcal{M}_h^b \in \mathbb{R}^{W_1H_1\times C}$ and $\mathcal{M}_l^c \in \mathbb{R}^{\frac{H_1}{2}\frac{W_1}{2}\times C}$. Subsequently, the input of the

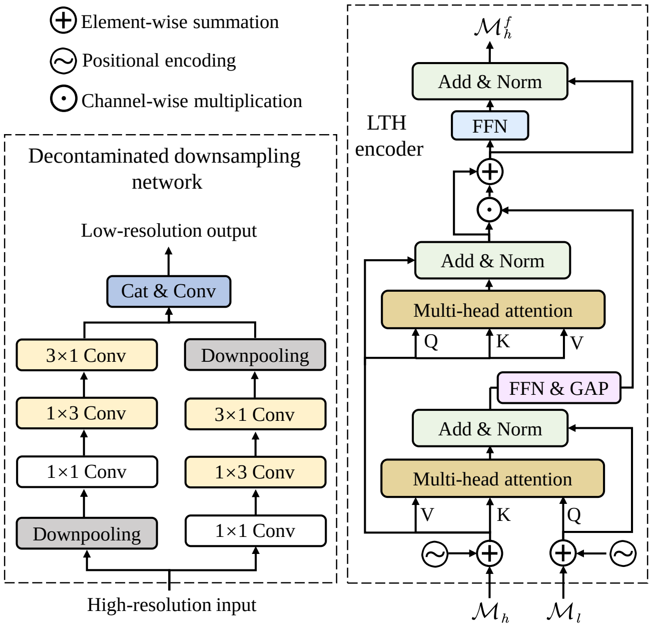

Fig. 3. Detailed workflow of the decontaminated downsampling network and the LTH encoder. The left sub-window illustrates the composition of the decontaminated downsampling network. The right one shows the structure of the LTH encoder. Best viewed in color.

encoder $\mathcal{M}_h \in \mathbb{R}^{W_1 H_1 \times C}$ and $\mathcal{M}_l \in \mathbb{R}^{\frac{H_1}{2}\frac{W_1}{2} \times C}$ can be acquired with the learnable positional encoding. The scaled dot-product attention (Att) can typically be stated by:

$$\begin{aligned} \mathrm{mAtt}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= (\mathrm{Cat}(\mathbf{a}_1,...,\mathbf{a}_N))\mathbf{W}^C \quad, \\ \mathbf{a}_n &= \mathrm{Att}(\mathbf{Q}\mathbf{W}_1^n,\mathbf{K}\mathbf{W}_2^n,\mathbf{V}\mathbf{W}_3^n) \quad, \end{aligned} \tag{4}$$

where $\mathbf{W}^C \in \mathbb{R}^{C\times C}$, $\mathbf{W}_1^n, \mathbf{W}_2^n, \mathbf{W}_3^n \in \mathbb{R}^{C\times C_d}$ can be considered the fully connected layer operator and $n \in 1, ..., N$ ($C_d = C/N$, N represents the amount of attention heads in parallel). Thereby, the result of the first multi-head attention module $\mathcal{M}_l^d \in \mathbb{R}^{\frac{H_1}{2}\frac{W_1}{2}\times C}$can be obtained by:

$$\mathcal{M}_l^d = \mathrm{mAtt}(\mathcal{M}_l, \mathcal{M}_h, \mathcal{M}_h) \quad . \tag{5}$$

After the normalization (Norm), $\mathcal{M}_l^e \in \mathbb{R}^{\frac{H_1}{2}\frac{W_1}{2}\times C}$ can be formulated as:

$$\mathcal{M}_l^e = \mathrm{Norm}(\mathcal{M}_l + \mathcal{M}_l^d) \quad . \tag{6}$$

To prevent fresh contamination, $\mathcal{M}_l^e$ strengthens the valid information of $\mathcal{M}_h \in \mathbb{R}^{W_1 H_1 \times C}$ through the feed-forward network (FFN) and the global average pooling (GAP). Thereby $\mathcal{M}_h^f \in \mathbb{R}^{W_1 H_1 \times C}$ is acquired by:

$$\begin{aligned} \mathcal{M}_h^c &= \mathrm{mAtt}(\mathcal{M}_h, \mathcal{M}_h, \mathcal{M}_h) \quad, \\ \mathcal{M}_h^d &= \mathrm{Norm}(\mathcal{M}_h + \mathcal{M}_h^c) \quad, \\ \mathcal{W} &= \mathrm{FFN}(\mathrm{GAP}(\mathcal{M}_l^e)) \quad, \\ \mathcal{M}_h^e &= \mathcal{M}_h^d + \gamma_1 * \mathcal{W} * \mathcal{M}_h^d \quad, \\ \mathcal{M}_h^f &= \mathrm{Norm}(\mathrm{FFN}(\mathcal{M}_h^e) + \mathcal{M}_h^e) \quad, \end{aligned} \tag{7}$$

where γ_1 and $*$ represent a learning weight and the channel-wise multiplication respectively. Owing to the LTH encoder, the high-resolution features contain more scale information of the object without adding new interference.

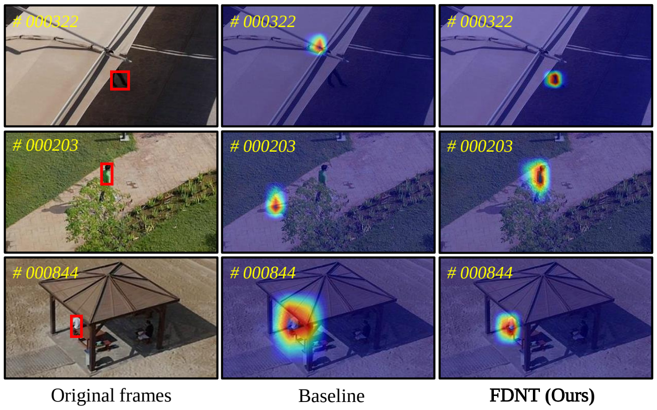

Fig. 4. Visualization of the confidence maps of the baseline and the proposed FDNT. In the original frames, red boxes are used to identify the objects. FDNT obtains better attention on tracked objects for UAV tracking (the image frames from *person12_2*, *person17_1*, and *group2_3* in UAV123@10fps [34]).

Remark 4: Through the downsampling and upsampling process, the object location and scale can be precisely estimated in the subsequent classification & regression network [10]. As shown in Fig. 4, the final confidence maps have been significantly improved by FDNT, compared to the baseline which is composed of the feature extraction network and the classification & regression network.

IV. Experiments

A. Implementation Details

FDNT was trained for 70 epochs totally, with the last two layers of AlexNet being regulated in the final 60 epochs and the first three layers being frozen. The learning rate is set to 0.005 and steadily drops from 0.005 to 0.0005 in the log space. Moreover, the sizes of the template and search are set to 127×127 and 287×287, respectively. We trained the proposed tracker with image pairs derived from COCO [22], ImageNet VID [23], GOT-10K [24], and LaSOT [25]. In addition, FDNT was trained on a PC equipped with an Intel i9-9920X CPU, 32GB of RAM, and two NVIDIA TITAN RTX GPUs.

B. Evaluation Metrics

The main metrics in the one-pass evaluation (OPE) metrics are precision, normalized precision, and success rate [26]. These three indicators can be used to assess tracker performance. The success rate is computed by the intersection over union (IoU) of the ground truth and estimated bounding boxes. The success plot (SP) indicates the percentage of frames that have an IoU greater than a predefined threshold. Additionally, the precision is determined by computing the center location error (CLE) between the estimated and ground truth locations. The precision plot (PP) illustrates the proportion of frames with CLEs falling inside a certain range. Besides, the normalized precision is derived by normalizing the precision across the size of the ground truth bounding

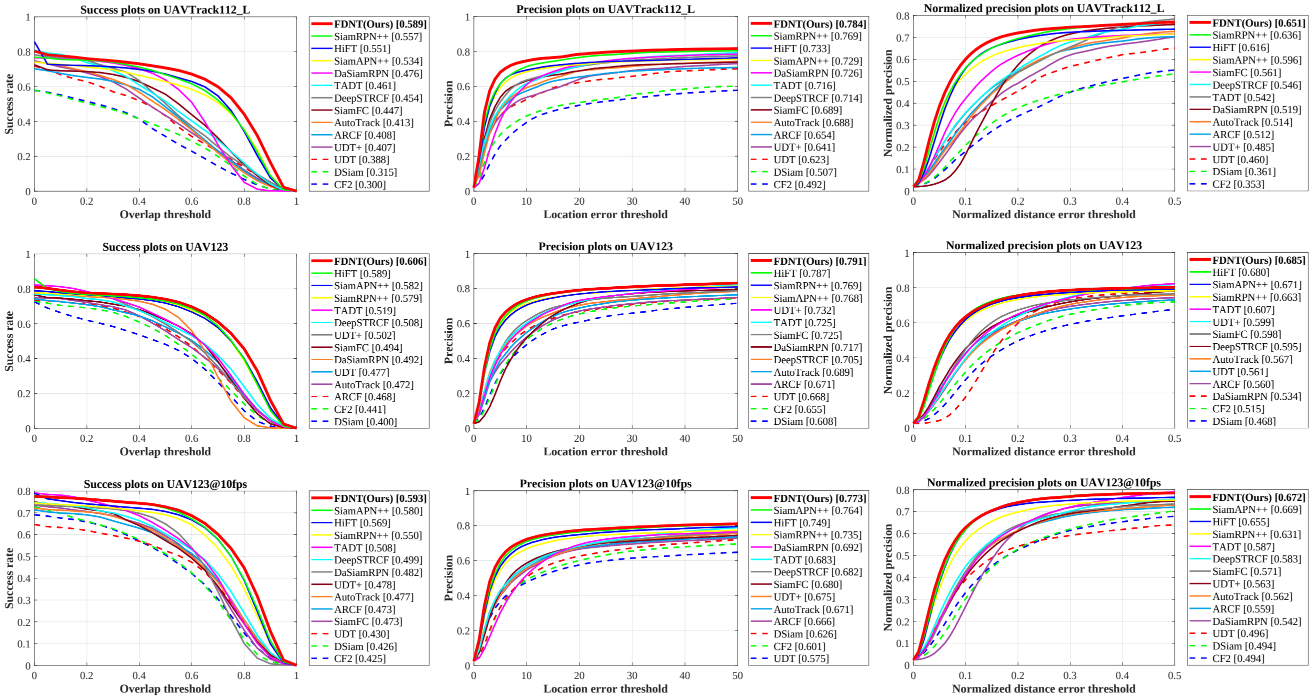

Fig. 5. Overall performance of FDNT and SOTA trackers on UAV123 [34] (the first row), UAV123@10fps [34] (the second row), and UAVTrack112_L [33] (the third row) benchmarks. The evaluation results indicate that the proposed approach, *i.e.*, FDNT, achieves superior performance on all benchmarks.

box, which is used to remove the effect of different object sizes on the precision. The normalized precision plot (NPP) is evaluated by the area under the curve (AUC). Meanwhile, the tracking methods are evaluated based on the SP's AUC, the NPP's AUC, and the PP at a 20-pixel threshold.

C. Evaluation on Aerial Benchmarks

1) Overall performance: FDNT is compared to other 13 SOTA trackers, including HiFT [10], SiamAPN++ [8], SiamRPN++ [12], SiamFC [4], DaSiamRPN [27], DSiam [28], AutoTrack [18], TADT [29], ARCF [19], UDT+ [30], UDT [30], CF2 [31], and DeepSTRCF [32], using three challenging and authoritative aerial tracking benchmarks.
Remark 5: For the justice, every Siamese-based tracker adopts the same lightweight backbone network AlexNet [11], pre-trained on ImageNet [23].

UAVTrack112_L [33]: To show the proposed framework's efficacy in lengthy tracking situations, we undertake assessments on UAVTrack112_L, a demanding long-term aerial tracking benchmark with more than 60k frames. As seen in Fig. 5, FDNT has progressed to the highest level. FDNT ranks the first place in all precision **(0.784)**, normalized precision **(0.651)**, and success rate **(0.589)**, followed by SiamRPN++ with a precision of 0.769, a normalized precision of 0.636 and a success rate of 0.557. The proposed feature decontaminated network is responsible for the effective performance of FDNT to handle feature pollution.

UAV123 [34]: UAV123 is a huge UAV benchmark consisting of 123 high-quality sequences totaling over 112K frames that cover a number of difficult aerial scenarios such as object occlusion, illumination variation, and out-of-view. Consequently, UAV123 is able to assist in conducting a thorough assessment. FDNT performs better than competing SOTA trackers in a variety of aerial tracking circumstances, indicating its higher robustness. Specifically, FDNT yields the best success score **(0.606)**, surpassing the second-best HiFT (0.589) and the third-best SiamAPN++ (0.582) by 2.9% and 4.1%, respectively. Similarly, in precision rate, FDNT achieves the best score (0.791), followed by HiFT (0.787) and SiamRPN++ (0.769). Besides, the normalized precision (0.685) of FDNT shows that our tracker has a robust overall performance. The promising performance demonstrates that FDNT is a satisfying decision in difficult aerial tracking scenes.

UAV123@10fps [34]: UAV123@10fps is a downsampled version of the original version. Therefore, the issue of forceful motion is exacerbated in UAV123@10fps compared to UAV123. As depicted in the second row of Fig. 5, FDNT can consistently achieve satisfactory results, attaining the highest success rate, precision, and normalized precision. In terms of success, FDNT ranks first with a success score of **0.593**, exceeding the second- and third-best SiamAPN++ (0.580) and HiFT (0.569) by 2.2% and 4.2% respectively. As for the precision, FDNT (0.773) also maintains a favorable performance at the best rate. While in terms of normalized precision, FDNT (0.672) also reaches the first place. To summarize, FDNT has a more preferable performance than other SOTA trackers, demonstrating the superior robustness in a variety of aerial tracking conditions.

2) Attribute-based comparison: Abundant object feature pollution in aerial tracking scenarios makes tracking more challenging. We assess the tracker performance on aerial-specific attributes to comprehensively study the robustness of FDNT, including partial occlusion, full occlusion, illumination variation, and fast motion as illustrated in Fig. 6.

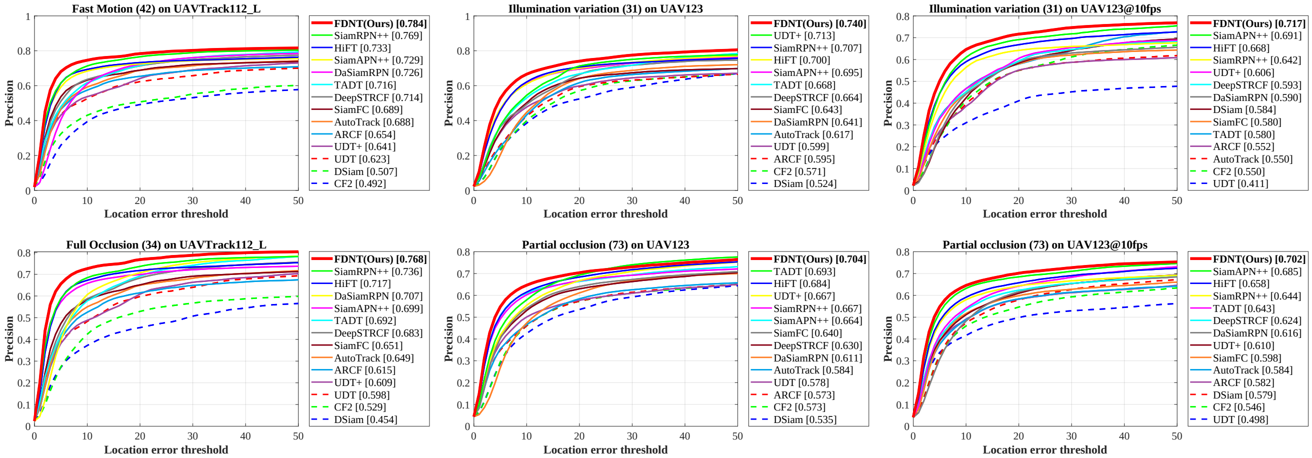

Fig. 6. Attribute-based evaluation decontamination performance of FDNT and other SOTA trackers on UAVTrack112_L [33], UAV123 [34], and UAV123@10fps [34] benchmarks. It shows that FDNT has the best performance in difficult situations where the object features are contaminated.

Compared with other trackers, FDNT outperforms the best performance in the above challenging tracking scenes, ranking first in terms of evaluation metrics. Especially, FDNT improves the second-best performance in fast motion and full occlusion by approximately 2% and 4% in UAVTrack112_L respectively, which is shown in Fig. 6.

Remark 6: The promising results demonstrate that FDNT can reduce the contaminated object features to overcome feature pollution issues. Particularly, when the object in UAV tracking is seriously occluded, FDNT can learn robust features to effectively estimate the location and scale of the object. Specifically, FDNT improves the performance significantly in occlusion scenarios.

3) Ablation study: On UAV123, in-depth investigations are conducted among FDNT having various modules enabled to confirm the efficacy of each module of the proposed method. For clarity, we first introduce the meaning of symbols used in TABLE I. This work considers Baseline as the model with only feature extraction network and classification & regression network. DD represents the decontamination downsampling network and NP represents the non-learnable pooling operation. DU indicates the decontamination upsampling network and NI symbolizes the non-learnable interpolation. Except for the examined module, each version of the tracker employs the same training strategy for the sake of equality. As shown in TABLE I, if the features are downsampled by the non-learnable pooling without reducing the interference of the feature pollution, it will cause performance degradation. This is due to the fact that the non-learnable pooling cannot enhance the location information of the object. While if the low-resolution features downsampled are not upsampled by the decontaminated upsampling network, but are directly input into the non-learnable interpolation, the effect will also be reduced. The key reason is that the non-learnable interpolation is unable to restore the object's scale information contrapuntally.

TABLE I

ABLATION STUDY OF VARIOUS PARTS OF THE PROPOSED FDNT ON UAV123 [34]. Δ SYMBOLIZES THE IMPROVEMENT OVER THE BASELINE TRACKER. PRE REPRESENTS THE PRECISION AND SUC REPRESENTS THE SUCCESS

Trackers	Pre	$\Delta_{pre}(\%)$	Suc	$\Delta_{suc}(\%)$
Baseline	0.709	-	0.495	-
Baseline+NP+DU	0.707	-0.3	0.467	-6.0
Baseline+DD+NI	0.740	+4.4	0.498	+0.6
Baseline+DD+DU (FDNT)	**0.791**	**+11.6**	**0.606**	**+22.4**

4) Key parameter analysis: Due to the fact that the weight of the first classification branch w directly affects the object location prediction in the test, it has an important influence on tracking performance. In order to assess the impact of w, w is set to various values for further investigation. It is set from 0.95 to 1.05. As presented in Fig. 7, excessively diminutive w makes the first classification branch ineffective. While excessively large w will enlarge the error of the first classification branch, impacting the determination of the object location. The AUC and precision of FDNT acquire the best results while w = 1.02. Therefore, we set w to 1.02 in all evaluations.

D. Comparison to Trackers with Deeper Backbone

FDNT exploits the feature decontaminated network to precisely predict the position and size of the object by reducing the contaminated features. Consequently, without introducing a large computational burden, FDNT achieves

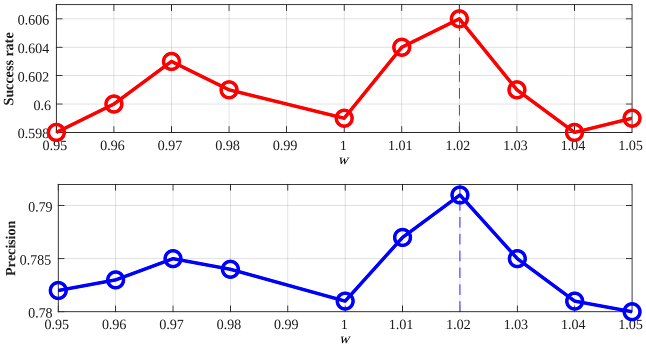

Fig. 7. Key parameter analysis of the first classification weight on UAV123@10fps [34]. While w = 1.02, FDNT obtains the optimal results.

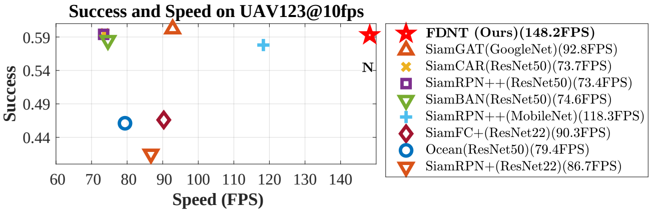

Fig. 8. Precision-speed research by quantitatively comparing FDNT and trackers with a deeper backbone on UAV123@10fps [34]. The proposed tracker achieves a superior trade-off on the benchmark.

SOTA performance. In order to further evaluate how effective it is, we compare FDNT with top-ranked trackers adopting deeper backbones. Trackers, including SiamRPN++ (ResNet-50) [12], SiamRPN++ (MobileNet) [12], Ocean (ResNet-50) [35], SiamCAR (ResNet-50) [36], SiamGAT (GoogleNet) [37], SiamBAN (ResNet-50) [38], SiamFC+ (ResNet22) [39], and SiamRPN+ (ResNet22) [39], are among the most advanced trackers. As shown in Fig. 8, despite utilizing the lightweight network AlexNet [11] as the backbone, FDNT has a desirable balance between tracking robustness and speed. Attributing to the decontaminated downsampling network and decontaminated upsampling network, FDNT is competent for effective and efficient object tracking in aerial tracking conditions.

Remark 7: For fairness, the speed of all compared trackers is measured on the PC equipped with an Intel i9-9920X CPU, 32GB of RAM, and two NVIDIA TITAN RTX GPUs.

E. Failure Cases

A few failure instances of the proposed approach are shown in Fig. 9. In these sequences, when the object disappears from view or the angle of view changes greatly, the continuous frame information needed by the trackers will provide the incorrect object appearance information. Thus, the process of feature downsampling and upsampling will be misled by the chaotic object information. Finally, the trackers are easy to lose the tracked object due to the influence of these tracking situations.

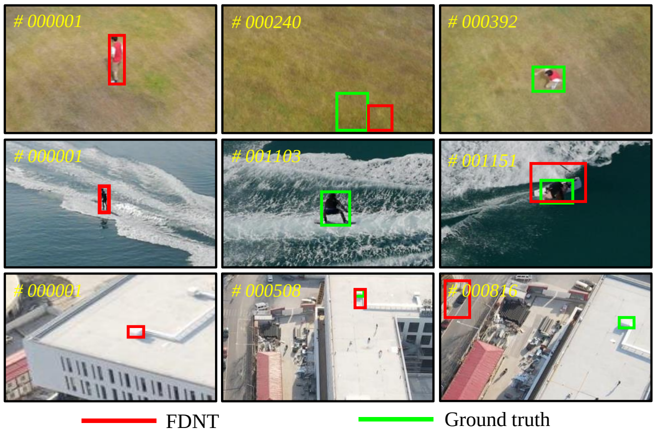

Fig. 9. Failure tracking scenarios of FDNT. The first, second, and third rows specifically display the tracking results on *person*7_1 from UAV123@10fps [34], *wakeboard*6 from UAV123 [34], and *electric box* from UAVTrack112_L [33], respectively.

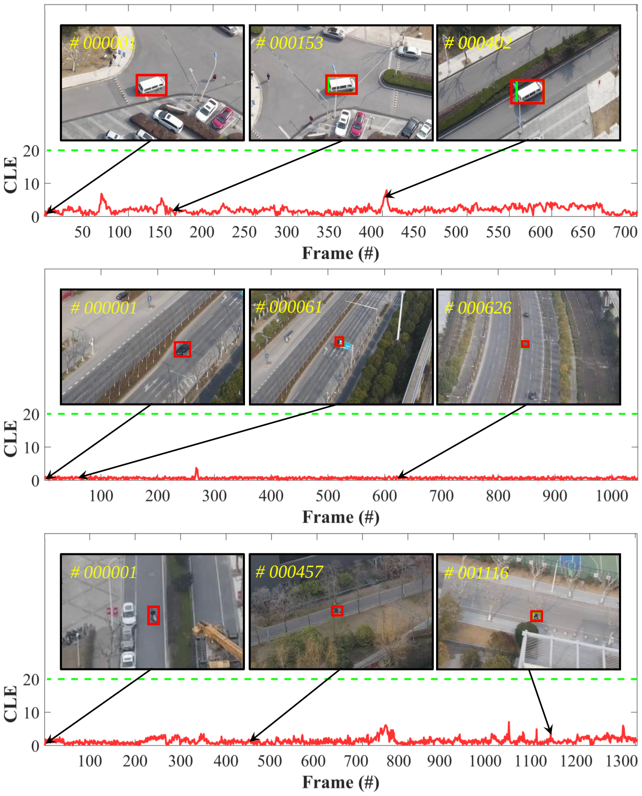

Fig. 10. Onboard tests of tracking in various UAV settings. With 31.4 frames per second, FDNT provides the robust performance. The red boxes represent the tracking result, whereas the green boxes represent the ground truth. In addition, the CLE score below the green dashed line is regarded as the success tracking outcome for the real-world tests. The experimental results verify FDNT's satisfactory tracking performance.

V. Real-World Tests

In order to prove the practicability in real-world applications, FDNT is further demonstrated in this section on a typical UAV robot equipped with an embedded onboard processor, namely the NVIDIA Jetson AGX Xavier. As shown in Fig. 10, three field tests are depicted, including different challenging scenes. The primary difficulties encountered in the tests are occlusion (the first row, the second row, and the third row), fast camera motion (the first row and the third row), and small object (the second row and the third row). FDNT maintains satisfying tracking robustness in a variety of challenging scenarios with an average speed of 31.4 frames per second during the tests. Hence, the real-world tests conducted onboard have effectively validated FDNT's superior performance and efficiency in a variety of UAV-specific scenarios.

VI. Conclusion

This work proposes a novel end-to-end feature decontaminated network for UAV tracking. The objective is to cope with the feature pollution issues, thereby obtaining robust object features while maintaining high effectiveness in challenging UAV tracking scenes. To enhance the location and scale information of the object, the decontaminated downsampling network supervised by PD loss and the decontaminated upsampling network are presented. Extensive experiments have established that FDNT is capable of achieving a superior performance-efficiency trade-off. In conclusion, we believe that FDNT will aid in the advancement of UAV tracking and anti-pollution visual object tracking.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62173249), the Natural Science Foundation of Shanghai (No. 20ZR1460100), and the Key R&D Program of Sichuan Province (No. 2020YFSY0004).

References

[1] R. Bonatti, C. Ho, W. Wang, S. Choudhury, and S. Scherer, “Towards a Robust Aerial Cinematography Platform: Localizing and Tracking Moving Targets in Unstructured Environments,” in *Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS)*, 2019, pp. 229–236.

[2] H. Yu, F. Zhang, P. Huang, C. Wang, and L. Yuanhao, “Autonomous Obstacle Avoidance for UAV based on Fusion of Radar and Monocular Camera,” in *Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS)*, 2020, pp. 5954–5961.

[3] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, “Multi-Regularized Correlation Filter for UAV Tracking and Self-Localization,” *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6004-6014, 2022.

[4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, “Fully-Convolutional Siamese Networks for Object Tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 850–865.

[5] H. Fan and H. Ling, “Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 7944–7953.

[6] Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, “Deformable Siamese Attention Networks for Visual Object Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6727-6736.

[7] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High Performance Visual Tracking with Siamese Region Proposal Network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8971–8980.

[8] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, “SiamAPN++: Siamese Attentional Aggregation Network for Real-Time UAV Tracking,” in *Proc. IEEE Int. Conf. Intell. Robots Syst. (IROS)*, 2021, pp. 3086-3092.

[9] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, “Siamese Anchor Proposal Network for High-Speed Aerial Tracking,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 510-516.

[10] Z. Cao, C. Fu, J. Ye, B. Li, and Y. Li, “HiFT: Hierarchical Feature Transformer for Aerial Tracking,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 15457-15466.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1097–1105.

[12] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, “SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4277–4286.

[13] J. Zhao, R. Xiong, J. Xu, F. Wu, and T. Huang, “Learning a Deep Convolutional Network for Subband Image Denoising,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2019, pp. 1420-1425.

[14] Y. Liu, Z. Qin, S. Anwar, P. Ji, D. Kim, S. Caldwell, and T. Gedeon, “Invertible Denoising Network: A Light Solution for Real Noise Removal,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13360-13369.

[15] W. Niu, K. Zhang, W. Luo, and Y. Zhong, “Blind Motion Deblurring Super-Resolution: When Dynamic Spatio-Temporal Learning Meets Static Image Understanding,” *IEEE Trans. Image Process.*, vol. 30, pp. 7101-7111, 2021.

[16] X. Yang, X. Wang, N. Wang, and X. Gao, “SRDN: A Unified Super-Resolution and Motion Deblurring Network for Space Image Restoration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-11, 2022.

[17] 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 *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2014, pp. 5441-5446.

[18] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11920–11929.

[19] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, “Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2891–2900.

[20] C. Fu, B. Li, F. Ding, F. Lin, and G. Lu, “Correlation Filters for Unmanned Aerial Vehicle-Based Aerial Tracking: A Review and Experimental Evaluation,” *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 1, pp. 125-160, 2022.

[21] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[22] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[24] L. Huang, X. Zhao, and K. Huang, “GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562-1577, 2021.

[25] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, “LaSOT: A High-quality Large-scale Single Object Tracking Benchmark,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5369-5378.

[26] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, “TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 300-317.

[27] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, “Distractor-Aware Siamese Networks for Visual Object Tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.

[28] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang, “Learning Dynamic Siamese Network for Visual Object Tracking,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1781–1789.

[29] X. Li, C. Ma, B. Wu, Z. He, and M. Yang, “Target-Aware Deep Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1369–1378.

[30] N. Wang, Y. Song, C. Ma, W. Zhou, W. Liu, and H. Li, “Unsupervised Deep Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 1308-1317.

[31] C. Ma, J. Huang, X. Yang, and M. Yang, “Hierarchical Convolutional Features for Visual Tracking,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3074–3082.

[32] F. Li, C. Tian, W. Zuo, L. Zhang, and M. Yang, “Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4904–4913.

[33] C. Fu, Z. Cao, Y. Li, J. Ye, and C. Feng, “Onboard Real-Time Aerial Tracking With Efficient Siamese Anchor Proposal Network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-13, 2022.

[34] M. Mueller, N. Smith, and B. Ghanem, “A Benchmark and Simulator for UAV Tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 445–461.

[35] Z. Zhang, H. Peng, J. Fu, B. Li, and W. Hu, “Ocean: Object-aware Anchor-free Tracking,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 771-787.

[36] D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, “SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6268–6276.

[37] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, “Graph Attention Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9538-9547.

[38] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, “Siamese Box Adaptive Network for Visual Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6667-6676.

[39] Z. Zhang and H. Peng, “Deeper and Wider Siamese Networks for Real-Time Visual Tracking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4586-4595.