Infrared UAV Target Tracking with Dynamic Feature Refinement and Global Contextual Attention Knowledge Distillation

Houzhang Fang, *Member, IEEE*, Chenxing Wu, Kun Bai, Tianqi Chen, Xiaolin Wang, Xiyang Liu, Yi Chang, *Member, IEEE*, Luxin Yan, *Member, IEEE*

***Abstract*—Unmanned aerial vehicle (UAV) target tracking based on thermal infrared imaging has been one of the most important sensing technologies in anti-UAV applications. However, the infrared UAV targets often exhibit weak features and complex backgrounds, posing significant challenges to accurate tracking. To address these problems, we introduce SiamDFF, a novel dynamic feature fusion Siamese network that integrates feature enhancement and global contextual attention knowledge distillation for infrared UAV target (IRUT) tracking. The SiamDFF incorporates a selective target enhancement network (STEN), a dynamic spatial feature aggregation module (DSFAM), and a dynamic channel feature aggregation module (DCFAM). The STEN employs intensity-aware multi-head cross-attention to adaptively enhance important regions for both template and search branches. The DSFAM enhances multi-scale UAV target features by integrating local details with global features, utilizing spatial attention guidance within the search frame. The DCFAM effectively integrates the mixed template generated from STEN in the template branch and original template, avoiding excessive background interference with the template and thereby enhancing the emphasis on UAV target region features within the search frame. Furthermore, to enhance the feature extraction capabilities of the network for IRUT without adding extra computational burden, we propose a novel tracking-specific target-aware contextual attention knowledge distiller (TCAKD). It transfers the target prior from the teacher network to the student model, significantly improving the student network's focus on informative regions at each hierarchical level of the backbone network. Extensive experiments on real infrared UAV datasets demonstrate that the proposed approach outperforms state-of-the-art target trackers under complex backgrounds while achieving a real-time tracking speed.**

***Index Terms*—UAV surveillance, infrared target tracking, feature fusion, knowledge distillation, attention mechanism.**

I. Introduction

Manuscript received xx xx, 2025.

This work was supported in part by the Open Research Fund of the National Key Laboratory of Multispectral Information Intelligent Processing Technology under Grant 61421132301, the National Natural Science Foundation of China under Grants 61971460 and 62101294. *(Corresponding author: Kun Bai; Houzhang Fang.)*

H. Fang, Y. Chang and L. Yan are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (houzhangfang@xidian.edu.cn, yichang@hust.edu.cn, yanluxin@hust.edu.cn).

C. Wu, T. Chen, X. Wang, and X. Liu are with the School of Computer Science and Technology, Xidian University, Xi'an 710126, China (e-mail: wcx1111111029@163.com, 18957384576@163.com, wxl@stu.xidian.edu.cn, xyliu@xidian.edu.cn)

K. Bai is with the Xi'an Modern Control Technology Research Institute, Xi'an 710065, China (baikundb@126.com)

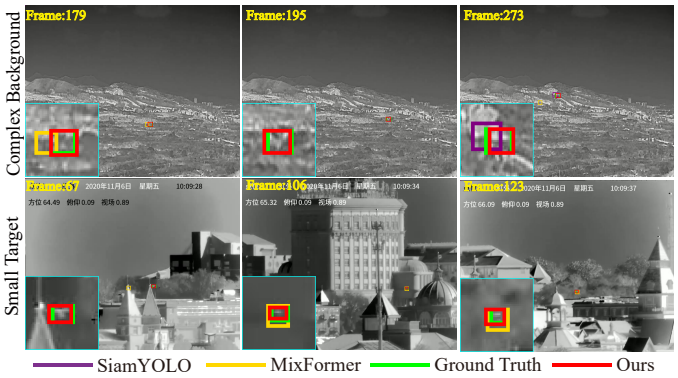

Fig. 1. Qualitative comparison of our method with the baseline SiamYOLO [1] and MixFormer [2] on two challenging video sequences (complex background and small target). The proposed Siamese tracker demonstrates superior performance in infrared UAV target tracking across various complex backgrounds due to its dynamic feature enhancement and global contextual attention-based knowledge distillation.

RECENTLY, unmanned aerial vehicles (UAVs) have been widely applied in many fields, such as precision agriculture, environmental monitoring, and aerial photography [3]–[7]. However, the wide use of UAVs can also potentially pose great threats to both aerial safety and public security. UAV surveillance based on infrared thermal imaging is capable of monitoring UAVs at a long range in both day and night scenarios [3]–[6]. As a key perception technology for UAV surveillance in the anti-UAV system, infrared UAV target (IRUT) tracking is a very challenging problem due to weak target features, small in scale, and distractors in complex backgrounds.

Many works have been developed for target tracking task. SiamFC [8] initially introduced deep learning into the target tracking task and utilized a correlation operation to obtain a similarity matrix for locating the target. This method is further enhanced by SiamRPN [9] and SiamRPN++ [10]. However, these methods are susceptible to distractors since the linear correlation operations between template and search frame. Recently, Transformer has been successfully applied in target tracking fields [2], [11]–[14] and significantly improved the tracking performance. However, these methods utilize all pair-wise attention of Transformer, lacking a focus on the primary areas. The IRUT size is relatively small, leading to a reduced importance weight for informative areas. Meanwhile, the cross-attention introduces excessive background into the

template frame, which undermines the features of small IRUT. This makes the subsequent enhancement of target features in the search frame more susceptible to interference from similar objects.

For the IRUT tracking task, Huang et al. [15] employed a C-C RPN to coarsely select candidate regions, which are combined with reliable historical predictions and sent into an S-L RCNN for accurate prediction. Fang et al. [7] constructed a contrast-enhanced block and a full spatial resolution attention mechanism to improve the representation ability of IRUT. Yu et al. [16] proposed a local-global tracker that employs a Transformer-based approach as the local tracker and utilizes a multi-region local tracking module to track potential target regions simultaneously. However, these methods fail to capture the local details of IRUT, which makes it difficult to highlight multi-scale target features in the search frames under complex scenarios.

Accurate and efficient target tracking is essential for practical applications. Some work [17]–[20] has been developed to address the trade-off between speed and performance. However, the performance still has a big gap compared with networks with larger parameter sizes. Knowledge distillation [21] is viewed as an effective method for enhancing the performance of lightweight models without affecting inference efficiency. Shen et al. [22] were the first to introduce knowledge distillation into Siamese trackers, designing a Siamese target response to effectively transfer tracking-specific knowledge from the teacher model to the student model. However, due to the small scale of the IRUT, the calculation of the Siamese target response relies on a local correlation process, which can easily result in a local optimum and is particularly susceptible to distractors. This leads to inaccurate knowledge transfer.

In this paper, we present a novel dynamic feature fusion Siamese network (SiamDFF) for infrared UAV target tracking. Specifically, we first introduce a selective target enhancement network (STEN) that integrates intensity-aware multi-head cross-attention to focus on the most relevant information in all pairwise relations and reduce the influence of excessive background information on the learning of correlation attention weights between pairs of pixels, thereby effectively enhancing the features of the target regions.

We then propose a dynamic spatial feature aggregation module (DSFAM) to improve the UAV target feature representation capability in the search frame. DSFAM employs a local-global complementary dual-branch structure, where the CNN-based branch effectively extracts local spatial details, and the Transformer-based branch captures global contextual information. After processing through STEN, the Transformer-based features provide a coarse representation of the target's shape. When combined with the CNN-based features, this allows for a complementary enhancement of target details, making them more focused and prominent. To adaptively utilize representations from both types of features, DSFAM facilitates feature interaction between the two branches, generating two spatial attention maps. These attention maps are then applied to each respective branch, enhancing the target-specific features across both branches. Finally, we introduce a dynamic channel feature aggregation module (DCFAM) to simultaneously integrate the original template and the mixed template generated from STEN in the template branch. DCFAM effectively prevents excessive contamination of the template by background elements. The dual-branch structure promotes target-related discriminative feature learning, enhancing mutual interaction with the search frame features and improving the representation of target region features within the search frame. DCFAM facilitates bidirectional feature interaction between the two types of templates, generating channel-wise attention maps through local cross-channel interactions and multiplying with the different branches to recalibrates and enhances the corresponding branches.

Furthermore, to enhance the feature extraction capability of the backbone without introducing additional computational burden, we propose the target-aware contextual attention knowledge distiller (TCAKD). TCAKD initially utilizes prior knowledge of the target mask from the template frame to focus on the relevant target features. It then employs a Transformer-based architecture to model long-range dependencies between the template and the search frame, resulting in a tracking-specific attention map that highlights the target region while suppressing background interference. The teacher model can better extract IRUT features, which improves the precision of template matching process and leads to a high response to the target region features. This refined knowledge is then transferred to the student network, assisting it in learning the template matching patterns established by the teacher network. The qualitative comparison of our method with the other two baseline trackers is shown in Fig. 1.

The main contributions of this article are summarized as follows.

- We propose a novel framework SiamDFF for consistently and accurately tracking IRUTs in complex backgrounds. It utilizes STEN to dynamically adjust the weights of pairwise relations in the Transformer to enhance the informativeness of the feature representation. We also introduce a DSFAM to effectively combine local detail information and global contextual information of IRUT to improve the feature representation ability in the search frame. Furthermore, to mitigate the template contamination issue, we design a DCFAM that utilizes the inherent information of the target template to complement the target details.
- We design a knowledge distiller TCAKD to enhance tracking performance while maintaining high inference speed by transferring a target-focused spatial attention map to the student model to learn the matching patterns derived from the teacher model.
- We conduct a comprehensive comparison of various tracking methods on real-world IRUT dataset. Extensive experiments demonstrate that our method achieves superior tracking performance against other SOTA methods, and realizes real-time tracking.

II. Related Work

A. Infrared Target Tracking Methods

In recent years, infrared target tracking has gained increasing attention. Liu et al. [23] introduced a dual-level

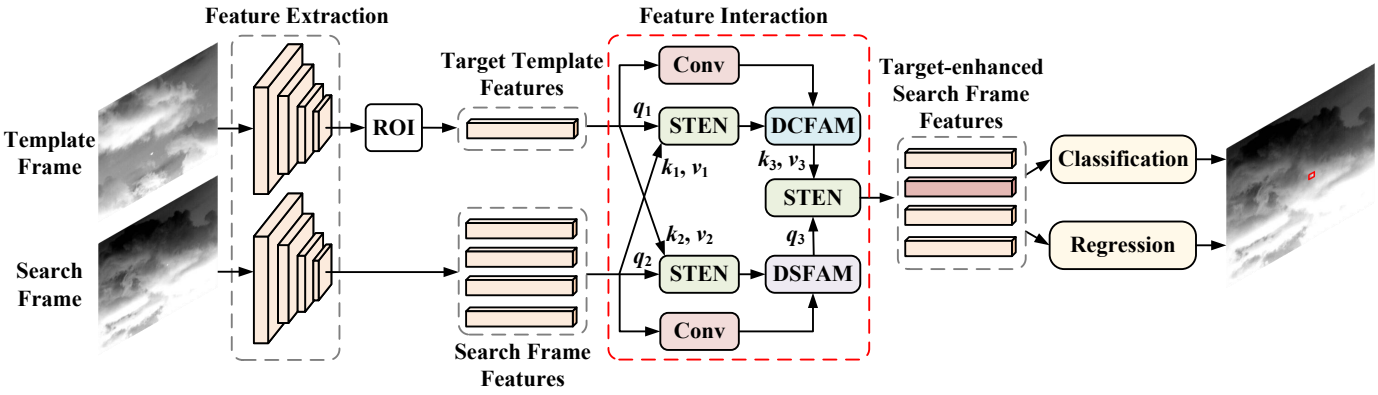

Fig. 2. Overview of the proposed SiamDFF. The feature interaction module enhances search frame features to improve the ability of the model for target classification and localization, which includes the proposed STEN, DSFAM, and DCFAM. The ROI denotes region of interest.

feature model that integrates discriminative features with fine-gained correlation features to enhance infrared object representation. Liu et al. [23] proposed a multi-level similarity model combined with global semantic similarity and local structural similarity for infrared object tracking. Fang et al. [1] developed an infrared UAV tracker based on global search, employing spatial-temporal information to counter distractions and using attention mechanisms to fuse features from different levels, thus enhancing target feature representation. Jiang et al. [24] constructed a large-scale anti-UAV dataset and proposed a dual-flow semantic consistency training strategy to differentiate tracked instances from distractors. Another class of long-term tracking methods combines global detection [5], [25]–[28] with local tracking, such as TOPIC [29]. The detection methods [5], [25]–[28] have demonstrated strong effectiveness in detecting infrared small targets. This type of tracking method performs well on targets in simple scenarios, but still face challenges when dealing with small-scale IRUTs in complex backgrounds and under real-time processing requirements. Despite recent advancements, the small size and weak features of IRUTs continue to pose significant challenges for tracking methods, making it difficult to effectively extract discriminative features and limiting their salient representation in the search frame. While these methods can track infrared targets, they tend to overlook the feature interaction phase between template and search frame, which is crucial for enhancing target features in the search frame. Additionally, there is insufficient focus on improving the feature extraction capabilities of the backbone, which impacts the subsequent template matching process and leads to inaccuracies in tracking. In light of this, we focus on enhancing the feature interaction process and improving the feature extraction capability for IRUTs.

B. Knowledge Distillation for Target Tracking

Knowledge distillation is a model compression technique that improves model performance by transferring knowledge from a teacher network to a lightweight student network. For target tracking task, Wang et al. [30] developed a fidelity loss and a correlation tracking loss to effectively compress the network while transferring its focus from object recognition to visual tracking. Shen et al. [22] proposed a Siamese target response learning algorithm that facilitates the transfer of tracking-specific knowledge to the student network. However, these distillation strategies focus solely on local matching information related to tracking-specific loss, neglecting contextual feature information. Consequently, distractor information is included in the feature response maps transferred to the student network, which hinders its ability to effectively learn the target region. In this study, we propose the TCAKD, which transfers the global contextual attention generated by the teacher model to the student network, enhancing the ability of the student network to perceive the target during feature extraction and improving its robustness against interference.

III. Methodology

A. SiamDFF Framework

In this subsection, we introduce the proposed SiamDFF framework, as depicted in Fig. 2. We employ ResNet18 [31] as the backbone network for feature extraction, ensuring the efficiency of our framework. The feature interaction module comprises STEN, DSFAM, and DCFAM for enhancing the representation capability of the target feature within the search frame. Finally, the target-enhanced search frame features are sent to the anchor-free head for the final prediction.

1) Selective Target Enhancement Network: The feature interaction process is a crucial step in the Siamese network for target tracking. Transformer-based interaction has become a mainstream paradigm in current tracking framework due to its ability to capture long-range dependencies and integrate contextual information. However, the cross-attention performs a one-to-all spatial similarity calculation for feature aggregation, which can introduce redundant information and hinder the ability of the model to establish accurate long-range dependencies. For small IRUTs occupying only a small portion of the image, dependence on pairwise similarity increases the weights of the background, resulting in inaccurate enhancement of the target features. Existing works [32], [33] aim to adjust the correlation computation weight map in the self-attention module to reduce background interference. However, their modules require substantial computational resources, significantly reducing the efficiency of the model.

To achieve accurate and focused feature interaction in cross-attention for IRUT, we propose STEN, which combines

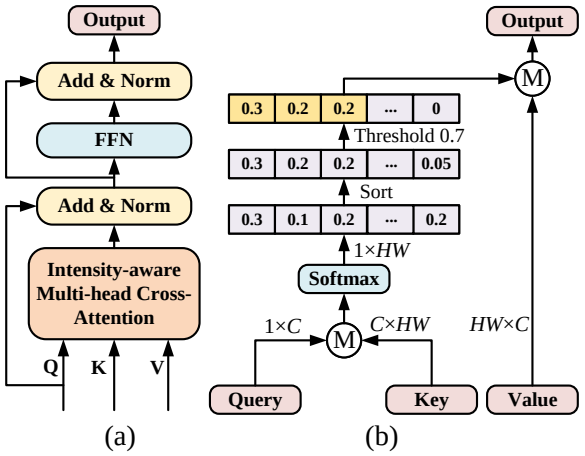

Fig. 3. (a) Structure of selective target enhancement network (STEN). (b) Structure of intensity-aware multi-head cross-attention (IMC). Ⓜ denotes the matrix multiplication.

intensity-aware multi-head cross-attention (IMC) that aggregates the most useful global context to each pixel. IMC first calculates all pairwise relations between the query and key to form a similarity map. The aggregated global context for each query focus varies at different positions within an image. Therefore, we establish a threshold that enables dynamic adjustment. Specifically, for each pixel in feature map, we utilize the values of the top k percent of positions based on their similarity to all other pixels to identify which regions should be considered for feature aggregation. As a result, enhancing the ability of the model to exploit global features while primarily focusing on the most relevant areas.

The constructed STEN is shown in Fig. 3(a). We first send Q, K, and V into IMC to precisely model the relation between the template and search frame feature. The calculation of STEN can be formulated as follows:

$$\begin{aligned} X &= \mathrm{Norm}\left(\mathrm{IMC}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)+\boldsymbol{Q}\right), \\ Y &= \mathrm{Norm}\left(\mathrm{FFN}\left(\boldsymbol{X}\right)+\boldsymbol{X}\right), \end{aligned} \tag{1}$$

where FFN and Norm denote the feed forward network and layer normalization, respectively. X and Y denote the outputs from IMC and STEN.

The IMC is shown in Fig. 3(b). We begin by computing the all pairwise relations between the query and key. Next, we sort the values in the similarity matrix in descending order. We then sequentially retain the top-ranked larger values until the cumulative sum of the selected values reaches the specified threshold T, ensuring that we retain the most informative elements. We compute the cumulative sum sequentially from the first element. For low attention weight elements, we set their values to zero. The calculation of IMC can be formulated as follows:

$$\mathrm{IMC} = \mathrm{softmax}\left(\mathcal{F}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\right)\mathbf{V}, \tag{2}$$

where d represents the dimension of the key. $\mathcal{F}$ is a function that determines which elements to discard:

$$[\mathcal{F}(\boldsymbol{E})]_{ij} = \begin{cases} E_{ij}, & \text{if } S_{ij} \geq t_i, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

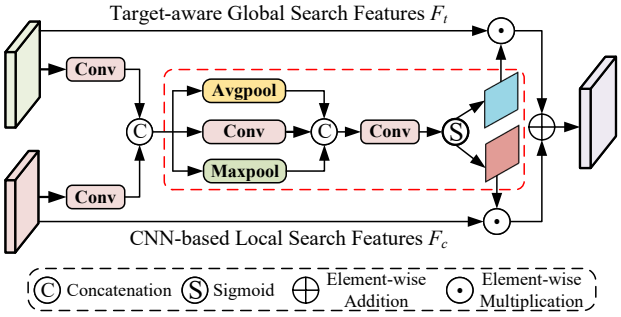

Fig. 4. Architecture of the proposed DSFAM.

where t_i is the smallest element that enables the cumulative sum to reach the specified threshold T in the i-th row of $\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}$. The selection of the specified threshold T is presented in Table IV of the blation study.

In the search branch, after passing through the STEN module, irrelevant context is effectively discarded, thereby enhancing the search frame's ability to perceive the target. Meanwhile, for template branch, STEN improves the template frame's discriminative capability between the target and background, facilitating the decoder's further emphasis on the features of the target region.

2) Dynamic Spatial Feature Aggregation Module: The Transformer-based feature interaction effectively leverages global contextual information to model dependencies and encodes spatial relationships, enhancing the robustness of the model against interference and assisting in capturing the shape details of targets across various scales. However, global features often lack the local detail necessary for precise IRUT representation. Meanwhile, the global perspective of Transformers is advantageous for detecting large-scale targets, their ability to extract features from small-scale targets is limited [34]. The local feature can provide fine-gained details essential for accurate target tracking. By effectively combining the strengths of both, we can significantly enhance the network's capability to track multi-scale IRUTs.

We propose a DSFAM, which aggregates Transformer-based features and CNN-based features using a spatial attention mechanism. Simply adding or concatenating the two branches leads to insufficient correlation between the features. Instead, we fuse their features and apply a transformation to generate spatial attention weight maps that contain target-aware spatial information. DSFAM allows the model to dynamically adjust the weights of each branch, precisely enhancing the target representation and resulting in a more effective fusion way. Unlike convolution operations that indiscriminately extract features of a certain type without considering the specific target, DSFAM leverages the target perception capabilities of the Transformer branch while selectively utilizing the relevant features extracted by the CNN branch. By employing an attention-guided fusion strategy, DSFAM achieves a more fine-gained integration of features, maximizing the complementary advantages of both branches and improving the overall performance in target tracking.

As shown in Fig 4, we first element-wise add the features from both branches. We then employ spatial attention maps

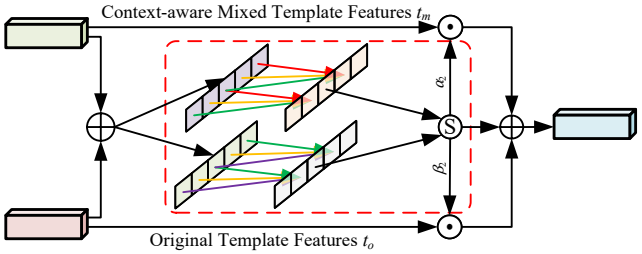

Fig. 5. Architecture of the proposed DCFAM.

to determine the importance of each branch. Given the input target-aware global search feature is $F_t \in \mathbb{R}^{C\times H\times W}$ and CNN-based local feature is $F_c \in \mathbb{R}^{C\times H\times W}$. We first squeeze their channel to $\frac{C}{2}$ by 1×1 convolution to reduce the computational burden of subsequent operations. Then we concatenate the output feature and obtain the fused feature denoted as $F_f \in \mathbb{R}^{C\times H\times W}$. The output of F_f can be formulated as

$$F_f = \mathrm{Cat}\left(\mathrm{Conv}_1(F_t), \mathrm{Conv}_1(F_c)\right), \tag{4}$$

where Conv_1 and Cat denote the 1×1 convolution and concatenation operation, respectively.

Next, we apply convolution, average pooling, and max pooling to obtain the attention response in the spatial dimension. The global pooling operations are employed to capture global spatial statistical characteristics, while the convolution component is used to integrate local spatial information. Then the three obtained feature maps concatenate in channel-wise and then utilize 3×3 convolution to obtain w_t and w_c which imply the spatial coefficients for each branch. Afterward, a sigmoid function is applied to separately emphasize the importance of spatial region. The formulation can be denoted as

$$w_c = \mathrm{Conv}_3(F_f), \tag{5}$$

$$w_m = \mathrm{MaxPool}(F_f), \tag{6}$$

$$w_a = \mathrm{AvgPool}(F_f), \tag{7}$$

$$\alpha_1, \beta_1 = \sigma\left(\mathrm{Conv}_3\left(\mathrm{Cat}\left(w_c, w_m, w_a\right)\right)\right), \tag{8}$$

where Conv_3 and σ denote the 3×3 convolution and sigmoid function, respectively. The final output on the right-hand side of Eq. (8) has dimensions of $H\times W\times 2$, with the first channel corresponding to α_1 and the second to β_1.

Finally, the modulated feature maps F_m are obtained by the Transformer-based feature and CNN-based feature based on the calculated α_1 and β_1. The feature fusion operation can be denoted as

$$F_m = \alpha_1 * F_t + \beta_1 * F_c. \tag{9}$$

3) Dynamic Channel Feature Aggregation Module: The quality of the template significantly influences tracking performance. Incorporating contextual information from the search frame into the template is beneficial, as it enables the template to adapt to variations in the target's shape and scale by leveraging the global context provided by the search frame. However, by interacting the template with the search frame, the mixed template incorporates a substantial amount of background area, which can lead to potential contamination by distractors [12].

We propose DCFAM which prevents potential degradation issues while providing a discriminative target-oriented fused template feature. DCFAM employs channel attention maps to adjust the importance of the two type templates. Due to the typically small scale of UAV targets, directly compressing the channel dimension when calculating channel attention can inevitably lead to the loss of small target information. In DCFAM, we introduce channel attention without dimension compression [35], allowing for a more accurate calculation of the importance weights for each branch.

As shown in Fig 5, assume the mixed template is $t_m \in \mathbb{R}^{C\times 1\times 1}$ and origin template is $t_o \in \mathbb{R}^{C\times 1\times 1}$. We first fuse the two type templates and get fused template $t_f \in \mathbb{R}^{C\times 1\times 1}$. The process is denoted as

$$t_f = t_m + t_o. \tag{10}$$

We then employ 1D convolution to obtain local cross-channel attention vectors for each branch. Afterward, a sigmoid function is applied to generate finally weight vectors. The procedure can be denoted as

$$\alpha_2, \beta_2 = \sigma\left(\mathrm{Conv1_m}(\mathrm{t_m}), \mathrm{Conv1_o}(\mathrm{t_o})\right), \tag{11}$$

where $\mathrm{Conv1}_m(t_m)$ and $\mathrm{Conv1}_o(t_o)$ denote 1D convolution used for mixed template and 1D convolution used for origin template, respectively.

Finally, the fused template t_g is generated by the mixed template and origin template combined with α_2 and β_2. The feature fusion can be expressed as

$$t_g = \alpha_2 * t_m + \beta_2 * t_o. \tag{12}$$

B. Target-aware Contextual Attention Knowledge Distiller

Knowledge distillation is a widely used model compression technique that helps the lightweight model balance between speed and performance. Existing works [22], [36], [37] utilize attention maps from the teacher network to guide the student network in focusing on key regions. FGD [36] helps the student network in learning the most critical information by distilling channel and spatial attention maps. However, this distillation method is not well-suited for tracking tasks, as detection tasks require attention to focus on all targets in the image, while tracking is a target-specific task that only should focus on one target. For target tracking tasks, DST [22] utilizes the attention weight map generated from the correlation operation between the template frame and the search frame, applying it to the feature map to suppress background information in the search frame and make the student mimic the response map. However, the similarity response map obtained from the correlation operation between the template frame and the search frame of the teacher network only matches from a local perspective, leading the student network to learn information that is influenced by similar distractors.

To accurately transfer teacher knowledge to student, we propose TCAKD, which distills attention map knowledge from a

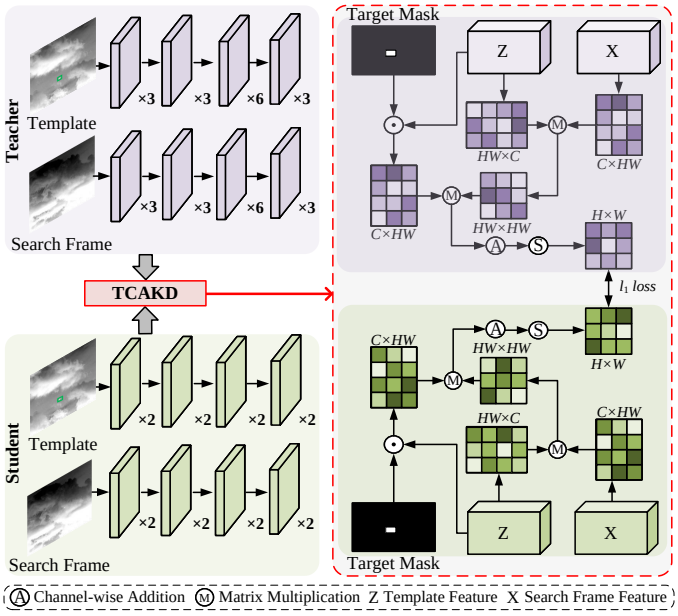

Fig. 6. Overview of the proposed TCAKD. The teacher network and student network are wrapped in purple and green boxes, respectively. The output feature maps of both template and search frame from each matching stage will all be computed for attention-based distillation. The target mask is generated from template, where the black area signifies the background and the white area indicates the target.

global perspective, reducing the impact of distracting features on the learning process of the student network. Instead of relying on local correlation operations to generate an attention map, TCAKD employs a Transformer-based structure to model the long-range dependencies between the template and the search frame, resisting distractor interference and enhancing target salience. The features are aggregated into a response map through summation, guiding the student to mimic this map based on the assumption that higher activation means greater importance.

The constructed module is shown in Fig 6. The output from the corresponding stages will be used for attention-based distillation. Let $F_z^t \in \mathbb{R}^{C\times H\times W}$ and $F_x^t \in \mathbb{R}^{C\times H\times W}$, represent the teacher template feature map and search frame feature map from the backbone network, respectively. $F_m \in \mathbb{R}^{H\times W}$ denotes the target mask where the target region pixel value equals one and the background region pixel value equals zero. We demonstrate the calculation of attention map of teacher model as an example. We first multiply the target mask $F_m \in \mathbb{R}^{H\times W}$ with the template frame features $F_z^t \in \mathbb{R}^{C\times H\times W}$ channel by channel to focus on the target region feature and ignore the background information. Then, we transform the $F_z^t \in \mathbb{R}^{C\times H\times W}$ as $K \in \mathbb{R}^{HW\times C}$ and $V \in \mathbb{R}^{C\times HW}$. Then, we convert the search frame feature $F_x^t \in \mathbb{R}^{C\times H\times W}$ into $Q \in \mathbb{R}^{C\times HW}$. We get the similarity matrix between query and key by matrix multiplication. After, we calculate target-aware feature map by multiplying the similarity matrix with the value. Finally, to get teacher target-aware contextual attention map $A_t \in \mathbb{R}^{H\times W}$ we gather the feature map by summation and pass through a sigmoid function. The important region will have high value and irrelevant region will have low value. Similarly, we can also compute the attention map $A_s \in \mathbb{R}^{H\times W}$ of the student network. The formulation can be defined as

$$A_t = \sigma\left(\sum_{i=1}^{C} V \otimes \mathrm{Softmax}\left(\frac{K \otimes (Q \cdot F_m)}{\sqrt{d}}\right)\right), \quad (13)$$

where σ and $\otimes$ denote sigmoid function and matrix multiplication, respectively.

After get $A_t \in \mathbb{R}^{H\times W}$ and $A_s \in \mathbb{R}^{H\times W}$, the learning process can be defined as

$$\mathcal{L}_{TCAKD} = \mathcal{D}\left(A_t \in \mathbb{R}^{H\times W}, A_s \in \mathbb{R}^{H\times W}\right), \quad (14)$$

where $\mathcal{D}$ is defined as a l_1-norm loss.

IV. Experiments

A. Datasets

We select ten infrared image sequences, each of which represents common challenges encountered in anti-UAV scenarios, such as heavy clouds and similar backgrounds. Seq. 1-Seq. 5 and Seq. 7 are collected from public datasets and can be accessed online [38], [39]. The rest are collected by ourselves and carefully labeled. We use a total of 19392 images that cover targets of various scales. The division ratio of the training and testing sets is set to 7:3.

B. Evaluation Metrics and Experimental Setups

The proposed network is trained by an SGD optimizer with a momentum of 0.9 and a weight decay of 0.0001. We use a cosine learning rate schedule, where warmup epochs are 50 and the base learning rate is set to 0.003. The total training epoch and batch are 250 and 12, respectively. We use random flipping for data augment. The network is trained from scratch without loading pre-training parameters. For training TCAKD, we use SiamCAP [7] as teacher model and ResNet18 [31] as a student model. The framework is implemented on a server with a NVIDIA GeForce RTX 3090 GPU accelerated by CUDA 11.1. For software implementation, we use Python 3.8 and Pytorch 1.8.1. The size of the search image is 384×384.

We select GlobalTrack [40], SiamYOLO [1], Unicorn [41] SimTrack [18], HCAT [17], ROMTrack [12], HIT [19], SMAT [20], ODTrack [14], and MixFormer [2] for comparison with our method. SiamYOLO is designed for IRUT tracking task. SimTrack, HCAT, HIT, and SMAT belong to efficient tracking methods. GlobalTrack, SiamYOLO, and Unicorn are global search based tracking methods. ROMTrack, ODTrack, and MixFormer are high performance Transformer-based methods. All experiments are trained on the same infrared UAV datasets and tuned to the optimal.

We use the precision plot [42], success plot [42], normalized precision (NPrecision) plot [43], and state accuracy (SA) [24] to evaluate the performance of each method. We employ the precision at a 5 pixels threshold as a key metric for each tracker, which effectively assesses the tracking performance of small IRUTs [44]. The normalized precision is derived by adjusting the precision relative to the dimensions of the ground truth bounding box. To rank each tracker in the normalized precision plot, we utilize the area under the curve (AUC) as a representative score. Additionally, the success plot computes

TABLE I
QUANTITATIVE COMPARISONS OF THE PROPOSED METHOD AND OTHER METHODS. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN, AND BLUE FONTS, RESPECTIVELY.

# Seq.	Metrics	GlobalTrack (AAAI'20)	SiamYOLO (ICCVW'21)	Unicorn (ECCV'22)	HCAT (ECCV'22)	SimTrack (ECCV'22)	ROMTrack (ICCV'23)	SMAT (WACV'24)	HIT (ICCV'23)	ODTrack (AAAI'24)	MixFormer (TPAMI'24)	Ours
1	SA	0.661	**0.700**	**0.716**	0.235	0.215	0.379	0.213	0.216	0.584	0.416	**0.757**
	Success	0.644	**0.682**	**0.698**	0.296	0.271	0.478	0.269	0.272	0.636	0.525	**0.790**
	Precision	0.869	**0.886**	**0.909**	0.285	0.280	0.488	0.277	0.285	0.741	0.533	**0.941**
	NPrecision	0.787	**0.801**	**0.817**	0.269	0.253	0.442	0.253	0.255	0.681	0.485	**0.889**
	FPS	9.2	**60.3**	28.9	52.4	**63.1**	51.1	50.3	**62.1**	40.1	28.4	52.8
2	SA	0.415	**0.596**	0.544	0.046	0.037	0.350	0.017	0.044	0.037	**0.580**	**0.590**
	Success	**0.402**	**0.580**	0.528	0.045	0.037	0.345	0.017	0.044	0.083	**0.572**	**0.580**
	Precision	0.621	0.789	**0.813**	0.059	0.045	0.480	0.028	0.059	0.112	**0.795**	**0.827**
	NPrecision	0.532	**0.709**	0.691	0.055	0.048	0.423	0.029	0.054	0.104	**0.696**	**0.740**
	FPS	9.7	**60.0**	29.1	52.6	**65.4**	52.3	55.4	**64.0**	44.2	28.7	53.0
3	SA	0.668	**0.691**	**0.688**	0.055	0.091	0.185	0.141	0.146	0.156	0.160	**0.728**
	Success	0.652	**0.679**	**0.673**	0.068	0.112	0.230	0.173	0.179	0.191	0.198	**0.751**
	Precision	**0.851**	0.835	**0.848**	0.067	0.131	0.253	0.184	0.189	0.205	0.224	**0.899**
	NPrecision	**0.786**	0.769	**0.773**	0.067	0.114	0.225	0.166	0.169	0.185	0.195	**0.844**
	FPS	9.6	**61.4**	28.9	52.8	**69.2**	52.9	50.3	**61.9**	43.2	29.2	53.1
4	SA	0.646	**0.665**	**0.653**	0.175	0.168	0.234	0.156	0.166	0.158	0.173	**0.667**
	Success	0.629	**0.662**	**0.636**	0.171	0.164	0.229	0.153	0.163	0.155	0.169	**0.652**
	Precision	**0.885**	0.846	**0.906**	0.209	0.214	0.300	0.209	0.220	0.204	0.220	**0.856**
	NPrecision	**0.799**	0.772	**0.796**	0.199	0.193	0.271	0.186	0.196	0.186	0.201	**0.810**
	FPS	9.7	**59.4**	28.9	52.7	**66.9**	52.4	54.4	**63.5**	43.5	29.2	52.5
5	SA	0.357	**0.456**	**0.530**	0.045	0.063	0.093	0.064	0.052	0.133	0.087	**0.597**
	Success	0.347	**0.443**	**0.514**	0.044	0.060	0.089	0.0613	0.050	0.129	0.084	**0.579**
	Precision	0.531	**0.667**	**0.811**	0.067	0.107	0.144	0.104	0.077	0.219	0.139	**0.856**
	NPrecision	0.455	**0.563**	**0.680**	0.061	0.086	0.121	0.087	0.068	0.181	0.117	**0.749**
	FPS	9.8	**59.6**	29.1	52.7	**69.5**	52.4	52.8	**63.1**	42.7	28.9	50.4
6	SA	0.733	0.731	0.733	**0.756**	0.724	0.718	0.734	0.716	0.732	**0.734**	**0.742**
	Success	0.713	0.711	0.713	**0.736**	0.704	0.699	**0.714**	0.696	0.712	0.714	**0.723**
	Precision	0.987	0.979	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**	**0.997**	**1.000**	**1.000**
	NPrecision	**0.909**	0.904	**0.906**	0.903	0.897	0.890	0.900	0.890	0.897	0.900	**0.910**
	FPS	9.6	**63.0**	29.0	52.8	**65.1**	51.9	53.5	**64.6**	43.0	27.8	50.6
7	SA	0.856	**0.866**	**0.872**	0.266	0.379	0.612	0.265	0.702	0.789	0.515	**0.917**
	Success	0.836	**0.847**	**0.850**	0.257	0.368	0.595	0.256	0.684	0.769	0.500	**0.897**
	Precision	**0.996**	0.948	**0.994**	0.276	0.402	0.672	0.292	0.752	0.866	0.566	**0.992**
	NPrecision	**0.947**	**0.914**	**0.947**	0.289	0.425	0.682	0.292	0.770	0.868	0.573	**0.964**
	FPS	9.6	**60.6**	28.9	52.4	**67.5**	51.7	47.6	**61.6**	42.3	27.5	52.1
8	SA	**0.816**	0.806	**0.815**	0.230	0.224	0.227	0.225	0.551	0.810	0.279	**0.818**
	Success	**0.796**	0.787	**0.795**	0.225	0.219	0.221	0.219	0.537	0.790	0.272	**0.798**
	Precision	**0.998**	0.964	**1.000**	0.268	0.270	0.272	0.270	0.686	0.983	0.333	**0.987**
	NPrecision	**0.943**	0.911	**0.931**	0.258	0.253	0.255	0.253	0.635	0.924	0.315	**0.949**
	FPS	9.5	**62.0**	28.8	52.6	**65.2**	52.5	54.4	**60.8**	41.1	28.5	52.7
9	SA	0.523	**0.613**	**0.612**	0.090	0.103	0.114	0.070	0.128	0.104	0.114	**0.792**
	Success	0.507	**0.597**	**0.594**	0.086	0.098	0.108	0.067	0.122	0.099	0.108	**0.772**
	Precision	**0.800**	0.781	**0.917**	0.197	0.225	0.242	0.156	0.272	0.236	0.242	**1.000**
	NPrecision	0.650	**0.720**	**0.773**	0.178	0.162	0.182	0.122	0.207	0.170	0.183	**0.922**
	FPS	9.5	**63.0**	29.0	52.6	**69.8**	52.7	51.5	**63.2**	43.1	27.6	53.7
10	SA	0.383	0.600	**0.653**	0.411	0.602	0.403	0.396	0.601	0.596	**0.606**	**0.746**
	Success	0.373	0.583	**0.633**	0.400	**0.583**	0.392	0.385	0.582	0.578	**0.587**	**0.727**
	Precision	0.515	0.793	0.962	0.574	0.988	0.574	0.574	**0.994**	0.979	**0.997**	**1.000**
	NPrecision	0.469	0.712	0.813	0.524	0.803	0.502	0.504	**0.815**	**0.815**	**0.814**	**0.905**
	FPS	9.5	**61.1**	28.8	52.6	**62.0**	50.2	50.7	**62.1**	43.2	28.4	52.8

the proportion of instances where the intersection over union (IoU) between the predicted and ground-truth bounding boxes exceeds a specified threshold. We apply the AUC as a ranking score for each tracker.

C. Quantitative Results

Table I presents a quantitative comparison between our method SiamDFF and other ten methods across ten infrared image sequences. It is evident that SiamDFF surpasses the competing methods in terms of SA, success, precision, and nprecision in most of the test sequences, especially for Seqs. 5, 9, and 10, where the target scale is relatively small, while maintaining a high inference speed of over 50 FPS. This demonstrates that our method can effectively track small targets in complex backgrounds, while other methods struggle to accurately capture these targets due to their inability to enhance infrared target features. In comparison to efficient tracking methods like SimTrack, HCAT, HIT, and SMAT, our SiamDFF significantly exceeds their performance across all metrics while still maintaining efficiency within an acceptable range. This can be attributed to the proposed TCAKD, which enhances the network's feature extraction capabilities without compromising inference speed. Specifically, compared with Unicorn which achieves the overall second-best quantitative results, our SiamDFF has improved mSA which generated from average the SA across all sequences by 5.3%.

Furthermore, we present the success plot, precision plot, and normalized precision plot in Fig. 7. A larger area under the curve means better tracking performance. Our method consistently demonstrates superior location accuracy and scale estimation compared to other methods. In particular, our method outperforms the second-best method Unicorn by 5.1%

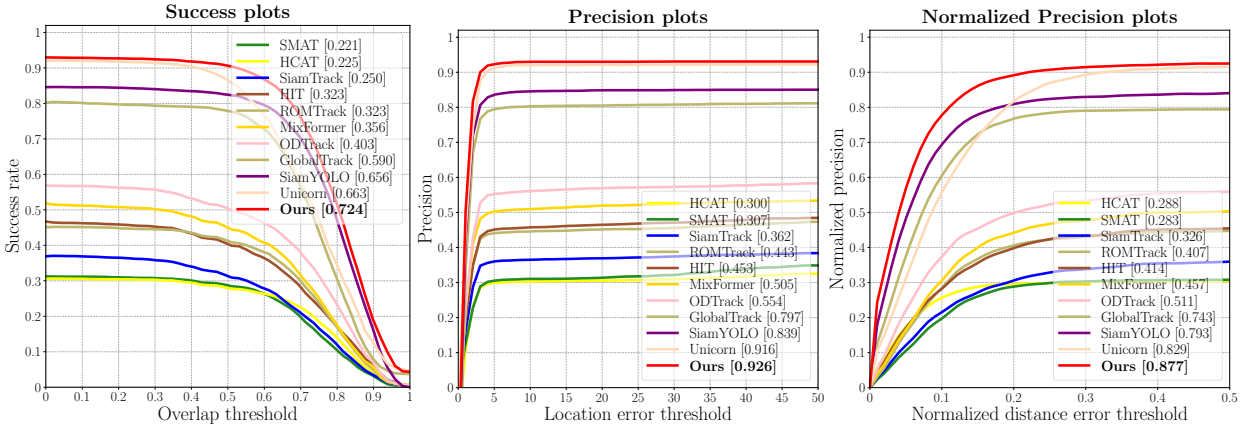

Fig. 7. Success plot, precision plot, and normalized precision plot of eleven trackers.

TABLE II
COMPUTATIONAL COMPLEXITY OF DIFFERENT METHODS

Metrics	GlobalTrack	SiamYOLO	Unicorn	HCAT	SimTrack	ROMTrack	SMAT	HIT	ODTrack	MixFormer	Ours
Params (M)	58.76	7.12	121.42	6.27	0.28	25.40	2.59	37.58	92.12	35.97	6.82
FLOPs (G)	95.12	30.20	59.30	9.75	75.60	92.10	1.56	14.18	73.10	33.18	58.48

in success, by 1% in precision, and by 4.8% in nprecision.

We also present the number of parameters and FLOPs for all comparative methods in the Table II. It can be observed that the proposed method has a moderate number of parameters and FLOPs compared to the other methods.

More experimental results can be found in the supplementary material.

D. Qualitative Results

Figure 8 shows the infrared target tracking results of different methods from Seq. 4, Seq. 5, Seq. 7 Seq. 9, and Seq. 10. In Seq. 4, the UAV is small in scale and moves through a complex forest background. Local tracking methods can hardly track the target once a tracking failure occurs since UAV targets move fast and they lack a global search mechanism that limits the search range. SiamYOLO can only track the part of the target and mis-tracks the target during motion blur. In contrast, our method can consistently track the target. This is attributed to the DSFAM, which aggregates both global and local features to enhance the representational capability of the model. Additionally, the spatial attention mechanism within DSFAM effectively suppresses background features while enhancing the target features, thereby improving tracking performance in complex backgrounds. In Seq. 5, the UAV exhibits small and weak features, making it challenging for HIT and MixFormer to fully locate the target due to their limited feature extraction capabilities. In contrast, our model benefits from TCAKD, which enables the student model to effectively mimic the feature extraction abilities of the teacher model and allows our model to extract target features with greater precision. In Seq. 7(c) and (d), the UAV is large in size with a distinct shape. Although all methods can identify the UAV target correctly, our approach achieves superior localization accuracy. In Seq 7(a), (b), and (e), where the target appears blurred, our method still effectively estimates its scale. In Seq. 9(c), the target is completely obscured by bright clouds, and only our method successfully tracks the target, demonstrating the effectiveness of our approach in extracting IRUT features. In Seq. 10(c) and (d), as the UAV enters a bush, the target is blended with the background. Although HIT and MixFormer manage to locate the target to some degree, they struggle to estimate its size accurately. By employing both the original and mixed templates, DCFAM enhances the robustness of the model against distractors during the feature interaction process.

TABLE III
EFFECTIVENESS OF INTEGRATING STEN, DSFAM, DCFAM, AND TCAKD

Row	STEN	DSFAM	DCFAM	TCAKD	Metrics			
					mSA	Success	Precision	NPrecision
1	-	-	-	-	0.674	0.670	0.862	0.817
2	✓	-	-	-	0.686	0.682	0.877	0.832
3	-	✓	-	-	0.692	0.686	0.882	0.836
4	-	-	✓	-	0.687	0.683	0.875	0.831
5	-	-	-	✓	0.707	0.702	0.906	0.858
6	✓	✓	-	-	0.700	0.695	0.889	0.843
7	✓	✓	✓	-	0.709	0.702	0.897	0.851
8	✓	✓	✓	✓	**0.735**	**0.724**	**0.926**	**0.877**

E. Ablation Study

1) Effectiveness of Integrating All Sub-Component: Table III presents the quantitative results of our tracking method with and without the proposed components: STEN, DFFAM, DCFAM, and TCAKD. The results in rows 1 and 8 demonstrate that compared to the baseline, mSA, success, precision, and nprecision have increased by 6.1%, 5.4%, 6.4%, and 6%, re-

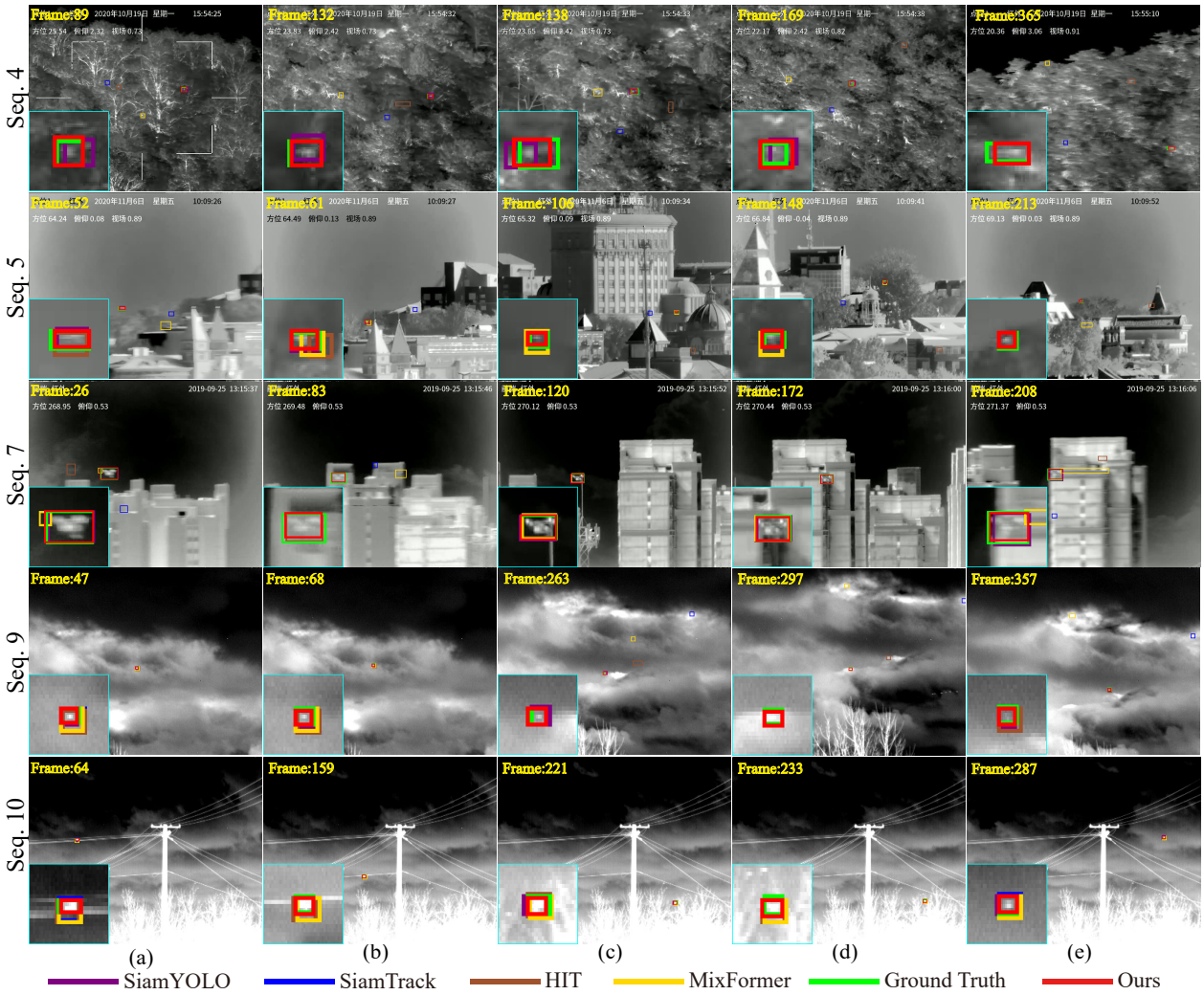

Fig. 8. Qualitative tracking results of five methods from Seq. 4, Seq. 5, Seq. 7, Seq. 9, and Seq. 10. Each row presents a sequence with five representative images. Close-ups are given for better visualization.

TABLE IV
ABLATION STUDY OF DIFFERENT THRESHOLDS

Threshold	mSA	Success	Precision	NPrecision
0.5	0.676	0.673	0.866	0821
0.6	0.684	0.681	0.875	0.830
0.7	**0.686**	**0.682**	**0.877**	**0.832**
0.8	0.679	0.675	0.869	0.824
1.0	0.674	0.670	0.862	0.817

spectively. This clearly indicates that incorporating any of the proposed components enhances tracking performance relative to the baseline, highlighting that each component contributes to improving the capability of the model to track infrared UAV targets in complex backgrounds. Furthermore, the integration of all components yields superior results compared to single or paired combinations, suggesting that each component provides complementary advantages.

2) Effectiveness of Integrating STEN: As shown in rows 1 and 2 of Table III, changing cross-attention to STEN brings an improvement of mSA by 1.2%, success by 1.2%, precision by 1.5%, and nprecision by 1.5%. This indicates that minimizing irrelevant context effectively enhances the features of the target regions. Unlike cross-attention which computes similarities across all query-key pairs, our proposed STEN focuses solely on the most relevant regions, thereby reducing background interference.

The key parameter for STEN is the specified threshold T, which significantly affects the considered part of background. Table IV comprehensively shows the performance change brought by T. When T =1, STEN becomes native cross-attention. We find that STEN consistently outperforms cross-attention and achieves the best results when the threshold T is carefully and manually selected as 0.7. We argue that when T is small, only a limited amount of contextual information is utilized, which fails to effectively highlight the features of the target regions. Conversely, when T is large, an excess of background information can affect the learning of the cross-attention values. We also attempt to set the threshold T as a learnable parameter and found that the learned value, $T = 0.6426$, yields tracking performance very close to that of the manually selected optimal threshold $T = 0.7$. Further experimental details can be found in the supplementary

TABLE V
ABLATION STUDY FOR DIFFERENT FUSION STRATEGIES OF DSFAM

Fusion Strategy	mSA	Success	Precision	NPrecision
Concatenate	0.679	0.676	0.865	0.820
Sum	0.682	0.678	0.869	0.823
Ours	**0.692**	**0.686**	**0.882**	**0.836**

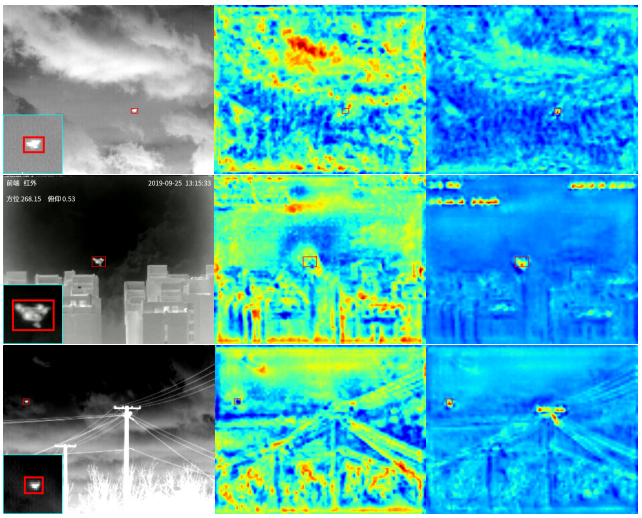

Fig. 9. Visualization of response maps from three representative sequences before and after applying DSFAM. The features of the target region are significantly enhanced after leveraging DSFAM.

TABLE VI
ABLATION STUDY FOR DIFFERENT FUSION STRATEGIES OF DCFAM

Fusion Strategy	mSA	Success	Precision	NPrecision
Concatenate	0.671	0.667	0.854	0.809
Sum	0.677	0.673	0.865	0.821
Ours	**0.687**	**0.683**	**0.875**	**0.831**

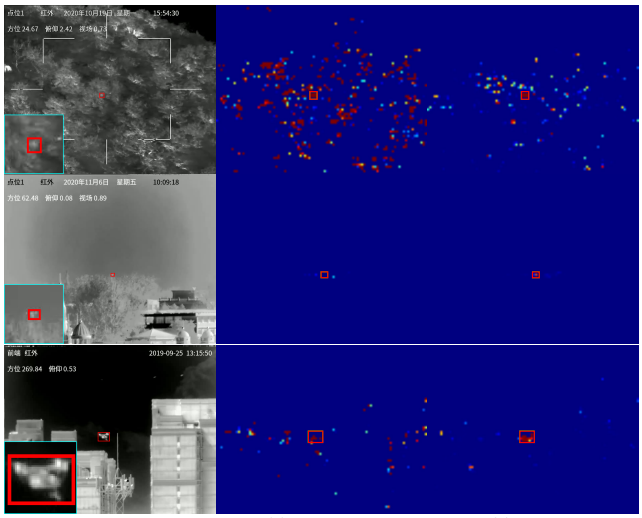

Fig. 10. Visualization of target prior from three representative sequences before and after applying TCAKD. The red box indicates the tracking target. The network can better focus on the target region after leveraging TCAKD.

material.

3) Effectiveness of Integrating DSFAM: As shown in rows 1 and 3 of Table III, the mSA, success, precision, and nprecision improve by 1.8%, 1.6%, 2%, and 1.9%, respectively. The Transformer-based branch focuses on extracting global features that capture contour details. In contrast, DSFAM integrates local features, which enrich the representation of the IRUT in the search frame. The combination of local and global features significantly enhances tracking performance.

Table V provides a comprehensive evaluation of the performance achieved through various fusion methods. Both direct addition and concatenation can enhance performance compared to baseline. This indicates that simultaneously integrating local and global features provides a more effective representation of target characteristics in complex scenarios. When employing attention-guided fusion, DSFAM achieves the highest performance. The ability to dynamically adjust the weights of different branches in response to various input images is emphasized, recognizing that different feature types contribute uniquely to targets of varying sizes.

Figure 9 visualizes the feature maps before and after applying DSFAM from Seq. 6, Seq. 7, and Seq. 10. It is evident that before adding DSFAM, the network fails to focus on the target region. However, after incorporating local features, the target area is significantly enhanced. Additionally, the attention mechanism effectively mitigates background interference, leading to more salient target features. By leveraging this dual branch, the network is better equipped to accurately classify and locate the target in subsequent processing.

4) Effectiveness of Integrating DCFAM: From rows 1 and 4 of Table III, we can observe that the incorporation of DCFAM results in an enhancement of 1.3% in mSA, 1.3% in success, 1.3% in precision, and 1.4% in nprecision. This indicates that relying solely on a mixed template of contextual information is insufficient for addressing the challenges in complex backgrounds. By utilizing the original template and the mixed template, DCFAM effectively minimizes the influence of excessive background information. Additionally, it allows the model to leverage the global contextual information from the search frame, enhancing its ability to adapt to new target shapes in the template frame.

Table VI presents a thorough evaluation of performance across various fusion methods. It is clear that simple concatenation can actually diminish model performance, emphasizing the importance of a well-designed fusion strategy to enhance overall effectiveness. Meanwhile, utilizing an addition way yields only a slight improvement in performance. In contrast, the use of DCFAM leads to a significant performance improvement. This demonstrates that employing channel attention based on local channel interactions effectively guides the fusion process, enriching the template information.

5) Effectiveness of Integrating TCAKD: From rows 1 and 8 of Table III, we can observe that the incorporation of DCFAM results in an enhancement of 3.3% in mSA, 3.2% in success, 4.4% in precision, and 4.1% in nprecision, indicating TCAKD significantly enhances the backbone feature extraction capability by effectively transferring knowledge from the teacher network to the student network.

Figure 10 visualizes the target prior [41] calculated between template and search frame in backbone network from Seq. 4, Seq. 5, and Seq. 7. In Seq. 4, the target is completely obscured by the background. Without TCAKD, the template matches with many similar objects, preventing the network from focusing on the target area. However, with TCAKD, the network can accurately concentrate on the target while significantly suppressing the background. In Seq. 5, the target is small and has weak features. The absence of TCAKD results in the network's failure to identify the target, whereas its inclusion allows for accurate tracking. Sequence 7 presents a large-scale target scenario. Without TCAKD, the network can only partially locate the target. However, with TCAKD, the network fully focuses on it. Overall, these observations highlight that TCAKD can be applied across various scales while maintaining strong focus, even in complex backgrounds.

V. Further Discussion and Analysis

A. Comparative Analysis of Our SiamDFF and Conventional Siamese Architectures

Traditional Siamese networks rely on global similarity computation between the template and search frame, making them highly vulnerable to background interference when tracking low-resolution and weak infrared targets. To address this, our proposed SiamDFF architecture introduces three dedicated modules. STEN employs intensity-aware attention to retain only the top-k of salient query-key pairs, effectively suppressing irrelevant context. DSFAM combines global Transformer features and local CNN details via spatial attention, enhancing multi-scale target representation. DCFAM adaptively fuses original and context-enhanced templates using channel-wise attention without dimensional compression, preserving semantic integrity. Overall, this design enhances feature discrimination, improves robustness under low-contrast conditions, and increases adaptability to scale variations, offering a principled solution for challenging infrared UAV tracking scenarios.

B. Analysis of Intensity-aware Multi-head Cross-attention (IMC) Module

The proposed Intensity-aware Multi-head Cross Attention (IMC) module enhances low-contrast infrared targets by amplifying salient regions and suppressing irrelevant context. Unlike standard cross-attention, IMC applies a top-k thresholding strategy to retain only the most informative query-key pairs, producing a sparse and focused attention map. This not only strengthens weak yet meaningful features but also filters out distractive background noise—particularly vital for thermal imagery with weak textures. Ablation results (Tables II and III) confirm that IMC consistently outperforms standard cross-attention, validating its effectiveness in boosting target saliency and tracking performance.

C. Failure Cases and Limitations of Our Tracking Method

As presented in Fig. 11, the UAV target features are extremely weak, our tracking method fails to track the targets correctly in frames 357, 361, and 365 of Seq. 4. This means that the proposed method may encounter tracking failure when the target features are extremely weak, the target undergoes deformation, or historical templates are not used. This highlights a limitation that the proposed method should address in future work.

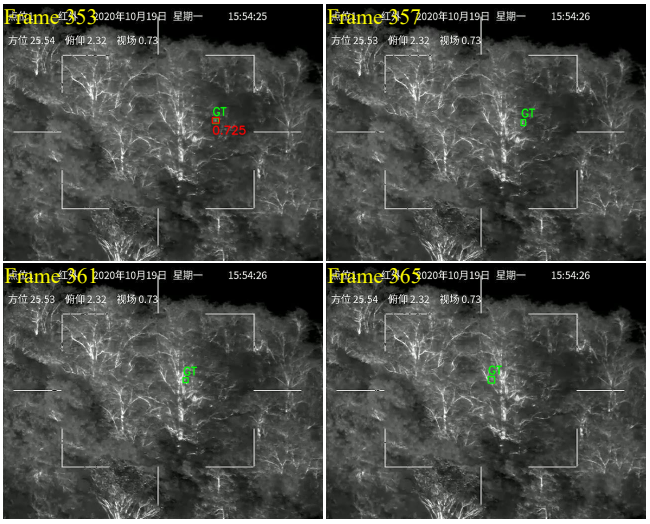

Fig. 11. Failure tracking results of our method on Seq. 4. Ground truth is shown in green, and tracking results are shown in red.

VI. Conclusion

In this paper, we introduce a dynamic feature fusion Siamese network (called SiamDFF) for real-time IRUT tracking in complex backgrounds. First, we present a novel module STEN designed to minimize the influence of excessive background information on cross-attention values. Second, we propose a module DSFAM that effectively combines global and local features using a spatial attention mechanism, enhancing the feature representation capability in the search frame. Third, we introduce a module DCFAM that integrates the original template while utilizing a mixed template to improve the contextual awareness of the target in the search frame, aiming to reduce the impact of distractions on the information delivered by the template. To further enhance tracking performance, we propose a knowledge distiller TCAKD to strengthen the feature extraction ability of the backbone by transferring the global contextual attention map produced by the teacher model. Extensive experiments demonstrate the effectiveness of our approach, achieving high tracking performance and enabling real-time tracking.

Acknowledgments

The authors would like to thank Chenhui Zhang for his assistance with the experiments, and the reviewers for their valuable and insightful suggestions.

References

[1] H. Fang, X. Wang, Z. Liao, Y. Chang, and L. Yan, "A real-time anti-distractor infrared UAV tracker with channel feature refinement module," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV) Workshops*, October 2021, pp. 1240–1248.

[2] Y. Cui, C. Jiang, L. Wang, and G. Wu, "MixFormer: End-to-end tracking with iterative mixed attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4129–4146, 2024.

[3] H. Fang, X. Wang, Z. Li, L. Wang, Q. Li, Y. Chang, and L. Yan, "Detection-friendly nonuniformity correction: A union framework for infrared UAV target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2025, pp. 11 898–11 907.

[4] H. Fang, L. Ding, L. Wang, Y. Chang, L. Yan, and J. Han, "Infrared small UAV target detection based on depthwise separable residual dense network and multiscale feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–20, 2022.

[5] H. Fang, Z. Liao, X. Wang, Y. Chang, and L. Yan, "Differentiated attention guided network over hierarchical and aggregated features for intelligent UAV surveillance," *IEEE Trans. Ind. Informat.*, vol. 19, no. 9, pp. 9909–9920, 2023.

[6] H. Fang, Z. Liao, L. Wang, Q. Li, Y. Chang, L. Yan, and X. Wang, "DANet: Multi-scale UAV target detection with dynamic feature perception and scale-aware knowledge distillation," in *Proc. 31th ACM Int. Conf. Multimedia (ACMMM)*, October 29-November 3 2023, pp. 2121–2130.

[7] H. Fang, C. Wu, X. Wang, F. Zhou, Y. Chang, and L. Yan, "Online infrared UAV target tracking with enhanced context-awareness and pixel-wise attention modulation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024.

[8] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshops*, October 2016, pp. 850–865.

[9] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with Siamese region proposal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2018, pp. 8971–8980.

[10] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2019, pp. 4282–4291.

[11] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2021, pp. 10 448–10 457.

[12] Y. Cai, J. Liu, J. Tang, and G. Wu, "Robust object modeling for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2023, pp. 9589–9600.

[13] Q. Yu, K. Fan, and Y. Zheng, "Domain adaptive transformer tracking under occlusions," *IEEE Trans. Multimedia*, vol. 25, pp. 1452–1461, 2023.

[14] Y. Zheng, B. Zhong, Q. Liang, Z. Mo, S. Zhang, and X. Li, "ODTrack: Online dense temporal token learning for visual tracking," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, February 2024, pp. 7588–7596.

[15] B. Huang, Z. Dou, J. Chen, J. Li, N. Shen, Y. Wang, and T. Xu, "Searching region-free and template-free Siamese network for tracking drones in TIR videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[16] Q. Yu, Y. Ma, J. He, D. Yang, and T. Zhang, "A unified transformer based tracker for anti-UAV tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, June 2023, pp. 3036–3046.

[17] X. Chen, B. Kang, D. Wang, D. Li, and H. Lu, "Efficient visual tracking via hierarchical cross-attention transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, October 2022, pp. 461–477.

[18] B. Chen, P. Li, L. Bai, L. Qiao, Q. Shen, B. Li, W. Gan, W. Wu, and W. Ouyang, "Backbone is all your need: A simplified architecture for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, October 2022, pp. 375–392.

[19] B. Kang, X. Chen, D. Wang, H. Peng, and H. Lu, "Exploring lightweight hierarchical vision transformers for efficient visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, October 2023, pp. 9612–9621.

[20] G. Y. Gopal and M. A. Amer, "Separable self and mixed attention transformers for efficient object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, January 2024, pp. 6708–6717.

[21] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. Hoi, "Distilled Siamese networks for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, 2022.

[23] Q. Liu, X. Li, Z. He, N. Fan, D. Yuan, and H. Wang, "Learning deep multi-level similarity for thermal infrared object tracking," *IEEE Trans. Multimedia*, vol. 23, pp. 2114–2126, 2021.

[24] N. Jiang, K. Wang, X. Peng, X. Yu, Q. Wang, J. Xing, G. Li, Q. Ye, J. Jiao, Z. Han, and J. Zhao, "Anti-UAV a large-scale benchmark for vision-based UAV tracking," *IEEE Trans. Multimedia*, vol. 25, pp. 486–500, 2023.

[25] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "ISNet: Shape matters for infrared small target detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 867–876.

[26] M. Zhang, H. Bai, J. Zhang, R. Zhang, C. Wang, J. Guo, and X. Gao, "RKformer: Runge-kutta transformer with random-connection attention for infrared small target detection," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, p. 1730–1738.

[27] M. Zhang, R. Zhang, J. Zhang, J. Guo, Y. Li, and X. Gao, "Dim2Clear network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[28] M. Zhang, Y. Wang, J. Guo, Y. Li, X. Gao, and J. Zhang, "IRSAM: Advancing segment anything model for infrared small target detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, p. 233–249.

[29] X. Cao, Y. Zheng, Y. Yao, H. Qin, X. Cao, and S. Guo, "TOPIC: A parallel association paradigm for multi-object tracking under complex motions and diverse scenes," *IEEE Trans. Image Process.*, vol. 34, pp. 743–758, 2025.

[30] N. Wang, W. Zhou, Y. Song, C. Ma, and H. Li, "Real-time correlation tracking via joint model compression and transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 6123–6135, 2020.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 770–778.

[32] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, October 2022, pp. 146–164.

[33] Z. Fu, Z. Fu, Q. Liu, W. Cai, and Y. Wang, "SparseTT: Visual tracking with sparse transformers," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, July 2022, pp. 905–912.

[34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2021.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2020, pp. 11 531–11 539.

[36] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2022, pp. 4643–4652.

[37] X. Yang, J. Huang, Y. Liao, Y. Song, Y. Zhou, and J. Yang, "Light Siamese network for long-term onboard aerial tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[38] "2023 Anti-UAV challenge dataset," Available online: https://anti-uav.github.io/dataset/ (accessed on 23 Augest 2023).

[39] J. Zhao, G. Wang, J. Li, L. Jin, N. Fan, M. Wang, X. Wang, T. Yong, Y. Deng, Y. Guo *et al.*, "The 2nd anti-UAV workshop & challenge: Methods and results," *arXiv preprint arXiv:2108.09909*, 2021.

[40] L. Huang, X. Zhao, and K. Huang, "GlobalTrack: A simple and strong baseline for long-term tracking," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, February 2020, pp. 11 037–11 044.

[41] B. Yan, Y. Jiang, P. Sun, D. Wang, Z. Yuan, P. Luo, and H. Lu, "Towards grand unification of object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, October 2022, p. 733–751.

[42] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2013, pp. 2411–2418.

[43] 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)*, September 2018, pp. 300–317.

[44] R. Kou, C. Wang, Y. Yu, Z. Peng, F. Huang, and Q. Fu, "Infrared small target tracking algorithm via segmentation network and multi-strategy fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, no. 5612912, pp. 1–12, 2023.

Houzhang Fang received the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, China, in 2014.

He is currently an Associate Professor. His research interests include target recognition, target tracking, and image restoration.

Chenxing Wu received the B.S. and M.S. degrees from the School of Computer Science and Technology at Xidian University, Xi'an, China, in 2022 and 2025, respectively.

He is currently an engineer at Alibaba Group, Hangzhou, China. His research interests include computer vision and deep learning.

Kun Bai received the B.S. degree in physics and the Ph.D. degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 2011 and 2017, respectively.

He is currently a Senior Engineer with the Xi'an Modern Control Technology Research Institute. His research interests include image target tracking, target detection, and image processing.

Tianqi Chen is currently pursuing a B.S. degree at Xidian University, Xi'an, China. His research interests include object detection and deep learning.

Xiaolin Wang received the B.S. and M.S. degrees from the School of Computer Science and Technology at Xidian University, Xi'an, China, in 2020 and 2023, respectively. He is currently pursuing a Ph.D. degree at Xidian University. His research interests include object tracking, object detection, and infrared image processing.

Xiyang Liu received the B.S. degree in software engineering from the Xidian University, in 1992, and the M.S. degree in software engineering in 1995, and the Ph.D. degree in software engineering in 2007, both from the Xidian University, Xi'an, China.

He is currently a Professor with the School of Computer Science and Technology, Xidian University. He is also the director of the Software Engineering Institute, Xidian University. He is the Member of Software Engineering Professional Committee of China Computer Society and the Member of the Medical Artificial Intelligence Branch of the Chinese Society of Biomedical Engineering. His research interests include clinically applicable machine learning for high performance medical image interpretation (3D breast ultrasound and PET volumes), computational pathology (hepatocellular and gastric carcinoma) and clinical decision support in cases of diagnostic uncertainty and complexity (women's pain related diseases), industrial intelligent visual inspection, and high trustworthy software technology for aerospace systems.

Yi Chang received the B.S. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2011, and the M.S. degree in pattern recognition and intelligent systems in 2014 and the Ph.D. degree in control science and engineering in 2019, both from the Huazhong University of Science and Technology (HUST), China.

He is currently an Associate Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include image processing, computer vision, and machine learning.

Luxin Yan received the B.S. degree in electronic communication engineering and the Ph.D. degree in pattern recognition and intelligence system from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2001 and 2007, respectively.

He is currently a Professor with the School of Artificial Intelligence and Automation, HUST. His research interests include multispectral image processing, pattern recognition, and real-time embedded systems.