This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

StrongSiamTracker: A Siamese Tracker with Dynamic Global Detection for Robust Anti-UAV Tracking

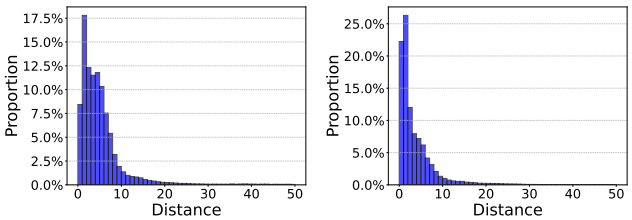
Xiaolong Cui, Liu Wan, Lingqi Kong, Jimin Li,
Chaojie Zhang, Ruohan Zhao, Panlong Wu*, Shan He
Nanjing University of Science and Technology
xlcui1201@163.com, plwu@163.com

Abstract

In the field of anti-UAV tracking, researchers often face challenges such as scale variations, background interference, and rapid movement, all of which significantly impair tracking performance. To address these issues, we propose a Strong Siamese Drone Tracker that dynamically integrates global detection and local tracking branches. In the initial tracking phase, we have designed a detector specifically tailored for small-sized targets to swiftly locate the drone's position. This detector incorporates a dynamic detection head to enhance the feature representation of small targets and employs a loss function based on Normalized Wasserstein Distance to more accurately assess the similarity between small targets. To mitigate short-term target loss caused by distractors in the tracking branch, we have integrated a trajectory-guided Spatial Constraint Proposal Suppression strategy into the baseline framework. This strategy selectively retains candidate regions that are spatially consistent with the target's historical motion trajectory. Additionally, we propose an adaptive Kalman filter that predicts target states during tracking failures. Finally, we establish dynamic switching rules between the two branches. Extensive experiments conducted on the 4th Anti-UAV Challenge and the Anti-UAV410 dataset demonstrate the superiority of our method. Notably, we achieved fourth and third places in Track 1 and Track 2 of the 4th Anti-UAV Challenge, respectively.

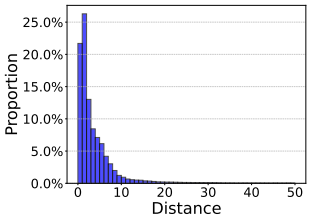
1. Introduction

In recent years, the rapid advancement of Unmanned Aerial Vehicles (UAVs) technology has led to extensive applications in logistics delivery, security patrols, and military reconnaissance[43]. However, the widespread use of UAVs has also highlighted significant inadequacies in regulatory mechanisms, resulting in an increase in illegal flights, covert surveillance, and potential attacks. These issues pose substantial challenges to public safety and national defense. Consequently, the real-time and accurate detection and tracking of UAVs are of considerable research significance and practical value.

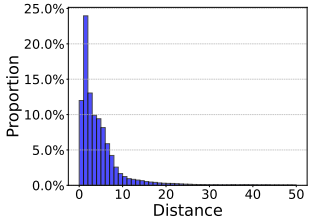


(a) 1st Anti-UAV challenge

(b) 3rd Anti-UAV challenge



(c) 4th Anti-UAV challenge



(d) Anti-UAV410

Figure 1. The distance distribution of consecutive frames of targets across different datasets. The Distance is defined as the Euclidean distance between the centers of two bounding boxes.

With the extensive application of deep learning in computer vision, a multitude of deep learning-based tracking methods has emerged, primarily classified into Siamese-based[2, 25, 42] and transformer-based[5, 6, 48, 52] approaches. Most existing object tracking methods are predominantly designed for visible light tracking. Although these methods can be directly applied to infrared image tracking, their heavy reliance on features such as color and texture makes them insensitive to thermal radiation characteristics, leading to suboptimal tracking performance when directly transferred. This paper focuses on the challenge of anti-UAV target tracking in infrared imagery.

Recently, target tracking in infrared images has attracted

significant attention from researchers. Traditional methods, such as mean-shift[55], sparse representation[28], particle filters[40, 60], and correlation filters[45, 54], predominantly depend on handcrafted features, which impose considerable limitations on the generalizability of tracking. In contrast, deep learning-based approaches, similar to visible light tracking, can be classified into Siamese-based[20, 62] and transformer-based[48] trackers.

Currently, most trackers are primarily employed for traditional target tracking applications, such as monitoring pedestrians and vehicles; however, they are less effective in infrared anti-UAV scenarios. This ineffectiveness arises from several challenges inherent to infrared anti-UAV target tracking: (1) the presence of complex background interference in the tracking environment, including cloud cover, mountainous terrain, forests, and thermal noise; (2) the minimal temperature difference between the drone and its background, which results in low target contrast; (3) the small size of drone targets, leading to a scarcity of discernible features.

Existing infrared anti-UAV target trackers predominantly rely on prior positional information of the target from the initial frame to facilitate subsequent tracking[21, 22, 36], which limits their effectiveness to scenarios where the target's initial position is known. Furthermore, these trackers struggle to recover once the target is lost and lack robust interference suppression, making them susceptible to disruptive elements. To address the reliance on prior information for successful tracking, we have developed a Strong Siamese Drone Tracker (StrongSiamDT) based on the SiamDT[22].

StrongSiamDT integrates a global detection framework with a Siamese tracker to enhance the stability of anti-UAV target tracking. In a stable tracking state, the position and scale of the target between consecutive frames do not undergo abrupt changes, as illustrated in Fig. 1. Consequently, the distance between consecutive frames are highly concentrated. To mitigate unnecessary interference, we developed a Spatial Constraint Proposal Suppression (SCPS) module to filter out significantly deviating proposals. Finally, we introduced an Adaptive Kalman Filter (AKF) for post-processing the tracking results, enabling effective management of brief instances of target loss.

Regarding the global detection model, the limited features available for distinguishing small targets result in insufficient sensitivity from the detector. To address this issue, we have incorporated a dynamic detection head[13] and a loss function based on Normalized Wasserstein Distance (NWD)[44] to improve the accuracy of small target detection. Moreover, to facilitate precise and effective transitions, we established a dynamic switching mechanism between the Siamese tracker and the global detector. Through these enhancements to the baseline model, SrongSiamDT has achieved exceptional performance in anti-UAV tasks.

The main contributions of this work are summarized as follows:

- A joint anti-UAV tracking method has been developed that integrates a global detector with a Siamese tracker, specifically optimized for tracking small UAV targets.
- A spatial constraint proposal suppression module and an adaptive Kalman filter have been proposed to effectively address issues such as background cluster interference and brief occlusions or losses of the target.
- Our method secured fourth place in Track 1 and third place in Track 2 of the 4th Anti-UAV Challenge.

2. Related Work

2.1. Visual single object tracking

Visual object tracking aims to locate and continuously follow a specified object in real time within video sequences. Initial research primarily relied on raw pixel features[4], while the KCF[19] was the first to introduce multi-channel HOG features to enhance tracking performance significantly. Inspired by KCF, numerous approaches utilizing HOG features and employing multi-feature fusion strategies have proliferated[15, 29].

SiamFC[2], as a pioneering Siamese-based tracker, was the first to reframe the tracking problem as a similarity matching challenge. Building on the foundation established by SiamFC, numerous enhanced Siamese-based trackers[9, 25, 26, 42, 58, 64] have emerged, many of which explore the use of more robust feature extractors. By employing deeper and wider backbone networks, such as Siam RPN++[26] and SiamDW[58], these trackers have demonstrated superiority across various benchmark tests. Following this progression, Stark[48], SiamCAR[9], and GlobalTrack[23] continue to investigate the potential of Siamese networks. With the successful application of transformers in computer vision[17, 41], a multitude of one-stage trackers based on transformers has gradually emerged, including ostrack[52], MixFormer[10, 11], and others[6, 47].

2.2. Infrared target detection and tracking.

In the context of small target detection in infrared imaging, a range of specialized methods has been developed, focusing on both detection and segmentation models. Among the model-driven approaches, notable techniques include background suppression[16], HVS-based methods[18, 38], and optimization strategies[57]. Data-driven methods, particularly Infrared Small Target Detection(IRSTD), have seen remarkable advancements in recent years. ACMNet[14] has introduced an asymmetric attention feature fusion module to extract multi-scale infrared small target features, while EFLNet[49] enhances the model's sensitivity to infrared small targets by integrating a dynamic detection

head[13] with YOLOv7. Methods such as DNANet[27] and ISNet[56] have refined UNet to create precise infrared small target segmentation models. Transformer-based methods[31, 46] for IRSTD have also been extensively researched. Currently, approaches utilizing weak label information for IRSTD are emerging as a promising direction for future development[8, 53].

The methods employed for infrared target tracking exhibit similarities to those used in optical target tracking. CF-based methods[45, 54] are applicable to infrared target tracking, with the majority concentrating on the extraction of target features and the integration of multiple characteristics. Contemporary infrared target trackers have developed a range of trackers based on Siamese networks[32, 33]. In addition to Siamese-based approaches, several other representative methods[1, 34, 50] have been specifically designed for infrared target tracking.

In the field of infrared anti-UAV tracking, this process can be viewed as a specific instance of conventional visual target tracking. Due to the significant imaging distance, UAV typically appear small within the field of view. In response to various anti-UAV benchmarks[22, 24, 61, 63], numerous algorithms specifically designed for tracking drone targets have been proposed. SiamSTA[21] incorporates spatiotemporal motion information of the target to achieve robust UAV tracking. Some studies[30, 35, 39, 51] indicate that the effectiveness of tracking is heavily dependent on the performance of UAV target detection, utilizing YOLO series detectors[37] alongside straightforward tracking logic to monitor small targets. The Anti-UAV410[22] not only introduces a comprehensive anti-UAV benchmark but also presents a tracker, SiamDT, which integrates global search with background suppression. Furthermore, several studies[59, 63] have advocated for a parallel approach that involves detection and tracking branches to achieve stable UAV tracking.

3. Method

In this section, we first present the motivation and overall framework of StrongSiamDT, followed by a detailed explanation of the tracking branch, detection branch, and switching mechanism.

3.1. Motivation

In infrared backgrounds, clutter from heat sources, such as clouds and vehicles, exhibits thermal radiation characteristics similar to those of drone targets. Relying solely on a tracker often results in tracking drift, complicating recovery efforts. The introduction of a global detection model enhances robustness against prolonged occlusions, target loss, and large-scale movements. When the drone occupies only a few pixels, the tracker's response map is susceptible to noise interference, leading to missed or erroneous tracking. To address this issue, a dynamic switching mechanism is implemented to activate the global detector for relocalization when tracking confidence falls below a certain threshold or when the current tracking result significantly deviates from the target's previous position. Furthermore, during short-term occlusions or transiently low tracking responses, a single tracker is likely to lose the target. To mitigate this risk, a motion model is established that integrates the tracker's spatial constraint proposal suppression with motion estimation, thereby enhancing tracking stability and reducing interference from false targets.

3.2. Overall Framework

The overall tracking framework of StrongSiamDT is depicted in Fig. 2, which includes the siamese tracking branch, the global detection branch, and a dynamic switching mechanism. The global detection branch identifies potential targets within the entire image and is primarily responsible for re-localizing targets when prior location information is unavailable or when targets are lost. In contrast, the tracking branch focuses on local tracking, ensuring the continuous localization of targets across successive input frames. Additionally, we have integrated AKF for post-processing the tracking results, which helps mitigate issues related to target drift, missed detections, and tracking failures. Within this framework, dynamic switching between the tracking and detection branches is achieved through a comprehensive evaluation of both tracking and detection confidence thresholds.

3.3. Siamese tracker

We utilized SiamDT[22] as the baseline for the tracking branch and subsequently implemented a series of enhancements and refinements. To address the challenge of tracking in scenarios where the first frame lacks prior information, we introduced a robust global UAV detector that operates in parallel with the tracking branch. Additionally, to account for the temporal consistency of the drone's position and scale[21], we integrated the motion characteristics of the targets to impose spatial constraints on the global proposals of the tracking branch, thereby mitigating the effects of global interference. Consequently, we streamlined the baseline by eliminating both the background interference suppression branch and the foreground detection branch, further enhancing tracking efficiency. A schematic representation of the tracking branch is presented in Fig. 2(a).

In the tracking branch, StrongSiamDT is based on the principles of SiamDT[22], GlobalTrack[23], using the template features of the target to be tracked as a query. It calculates similarity with the input instances of the Region Proposal Network (RPN) and the Region-based Convolutional Neural Network (RCNN), thereby improving the sensitivity of the feature map in the search frame towards the target.

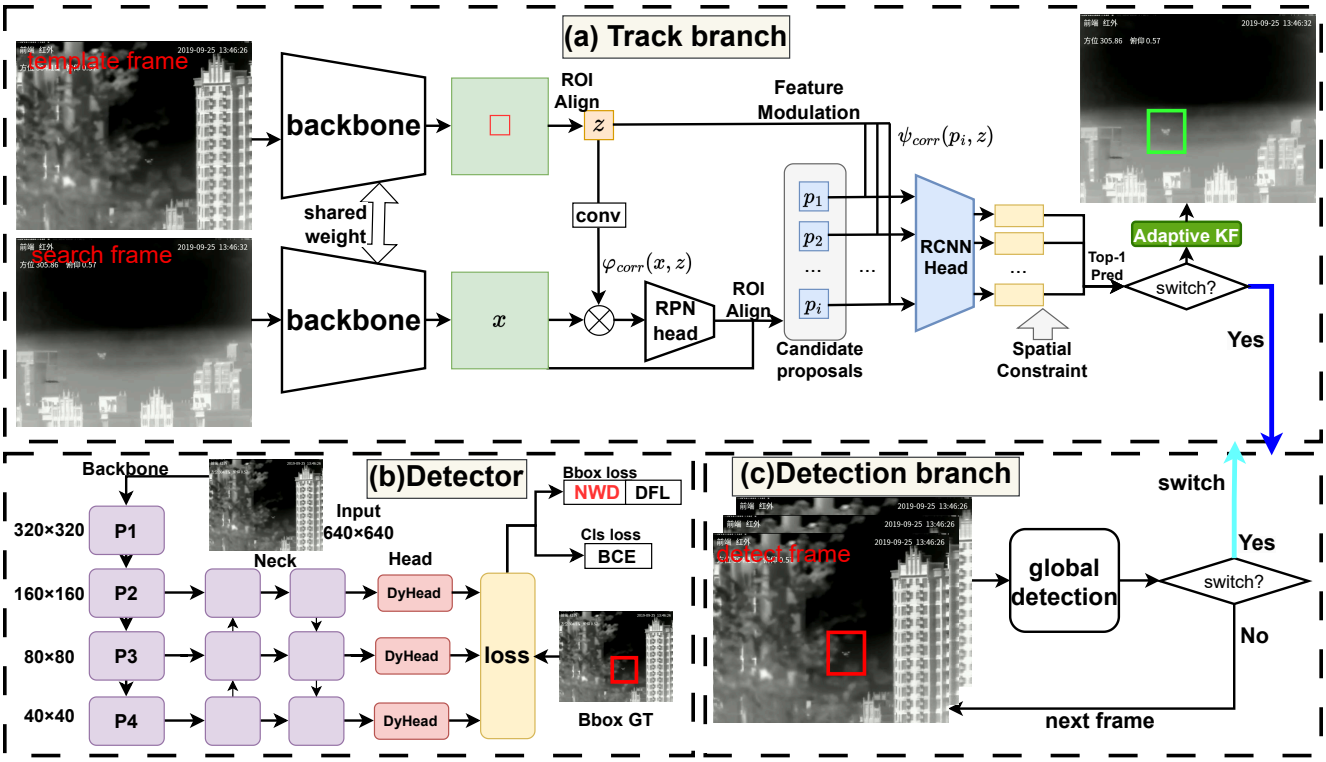


Figure 2. The overall framework of StrongSiamDT. The framework is primarily divided into two branches: tracking and detection. The tracking branch is constructed using a Siamese network, while the detection branch is specifically designed for drone targets. Dynamic switching between the two branches is facilitated

For the input feature map x of the RPN, it is necessary to compute the similarity between the Region of Interest (ROI) features z of the template and the feature map x of the image being searched. The similarity response feature is denoted as $\varphi_{corr}(x, z)$. Let $z \in \mathbb{R}^{k\times k\times c}$ represent the ROI features of the target area in the tracking template, and let $x \in \mathbb{R}^{h\times w\times c}$ denote the feature map of the image to be searched, where k, h, and w are the dimensions of the respective feature maps.

$$\varphi_{corr}(x, z) = f_{out}(f_{3\times3}(x) \otimes f_{k\times k}(z)) \tag{1}$$

where $\otimes$ denotes the convolution operation used to compute the similarity between z and x. $f_{k\times k}(\cdot)$ represents a convolutional layer with a kernel size of $k \times k$ and no padding, while $f_{3\times3}(\cdot)$ indicates a convolutional layer with a 3×3 kernel and padding of 1. $f_{out}(\cdot)$ is a 1×1 convolutional layer that ensures the output retains the same number of channels c as the input. $\varphi_{corr}(x, z)$ has the same dimensions as x, allowing both to be concatenated and subsequently fed into the RPNHead to generate candidate proposals.

After the RPN generates candidate proposals, the RCNN further refines the labeling and bounding box (bbox) predictions based on the ROI features of these proposals. This phase enhances the semantic information of the genuine target proposal features by calculating the similarity between the candidate proposals and the target template ROI features, thereby improving the distinguishability between genuine target proposals and interference proposals. The ROI features of the i-th proposal can be denoted as $p_i \in \mathbb{R}^{k\times k\times c}$. For a given $z \in \mathbb{R}^{k\times k\times c}$, we compute the similarity response map $\psi_{\text{corr}}(p_i, z)$ for each template and proposal individually, calculated as follows:

$$\psi_{corr}(p_i, z) = h_{out}(h_{3\times3}(p_i) \odot h_{3\times3}(z)) \tag{2}$$

where $h_{3\times3}(\cdot)$ represents a convolutional layer with a 3×3 kernel and a padding of 1. The symbol $\odot$ denotes element-wise multiplication , while $h_{out}(\cdot)$ indicates a 1×1 convolutional layer, ensuring that the output maintains a channel count of c. Finally, the proposal B_i corresponding to the i-th similarity response map is obtained through the RCNNHead.

$$B_i = RCNNHead(\psi_{corr}(p_i, z)) \tag{3}$$

The proposal boxes are represented in set notation as $P_g = \left\{B_i \,\middle|\, B_i = (c_x^i, c_y^i, w^i, h^i, s_i)\right\}$, where $s_i \in [0, 1]$ denotes the confidence score associated with the candidate proposal B_i. To ensure temporal consistency in the target's

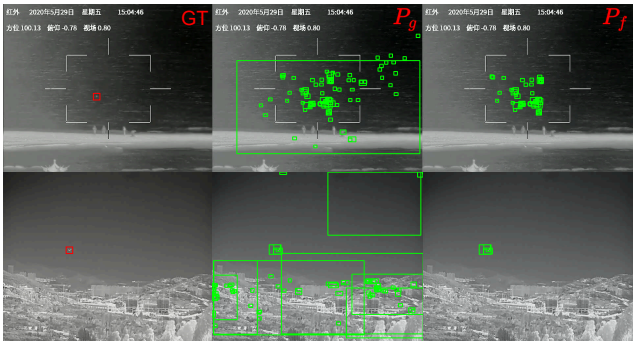


Figure 3. The SCPS effectively suppresses global proposals, thereby mitigating interference from proposals that significantly deviate from the target established in the previous frame.

position and scale during stable tracking, we focus on retaining only the proposal boxes associated with the region surrounding the target box from the previous frame. Consequently, we introduce a SCPS module.

Based on the target state of the previous frame under stable tracking, denoted as $B_{t-1} = (c_x^{t-1}, c_y^{t-1}, w^{t-1}, h^{t-1})$, the SCPS primarily involves positional hard suppression and scale soft suppression. The former defines a local region constrained by $\Delta_x = \frac{W}{8}$ and $\Delta_y = \frac{H}{8}$, where (W, H) represent the size of the image. Initially, candidate proposals P_l that meet the positional constraints are filtered.

$$P_l = \left\{ B_i \in P_g \,\middle|\, |c_x^i - c_x^{t-1}| \leq \Delta_x \ , \ |c_y^i - c_y^{t-1}| \leq \Delta_y \right\} \tag{4}$$

Subsequently, scale soft suppression is applied to adjust the confidence scores s_i of the candidates within P_l, with the revised confidence s_i' calculated as follows:

$$s_i' = s_i \cdot \underbrace{\exp\left(-\frac{\Delta_w^2 + \Delta_h^2}{2\sigma_s^2} - \frac{\Delta_r^2}{2\sigma_r^2} \right)}_{\text{scale weight } w_{\text{scale}}} \tag{5}$$

where $\Delta_w = \log(w^i / w^{t-1})$, $\Delta_h = \log(h^i / h^{t-1})$,$\Delta_r = \left| \log\left((w^i/h^i)/(w^{t-1}/h^{t-1}) \right) \right|$

The final proposal is as follows:

$$P_f = \left\{ \left(c_x^i, c_y^i, w^i, h^i, s_i' \right) \,\middle|\, B_i \in P_l \right\} \tag{6}$$

As illustrated in Fig. 3, the global proposals P_g are distributed across the entire image, whereas the local proposals P_f are restricted to the vicinity of the target from the previous frame, preserving a scale and aspect ratio that closely resemble the historical state. By eliminating candidate boxes that do not conform to the assumption of spatiotemporal consistency, we select the proposal with the highest confidence as the final output.

In comparison to the baseline, the tracking branch presented in this study retains only the most essential template features and performs multi-stage similarity calculations with the features of the frames being searched. Concurrently, through the use of SCPS, the target search space is reduced from a global scale to a local neighborhood, thereby enhancing the efficiency of the tracking branch while preserving its tracking capabilities.

3.4. Global detector

In the context of anti-UAV tracking scenarios that lack prior tracking knowledge, we have developed a strong detector specifically designed for small to medium-sized drones, based on YOLOv8s, as illustrated in Fig. 2(b). Initially, we removed the P5 large target detection layer from the existing backbone, enabling the Neck module to receive features exclusively from P2, P3, and P4. This modification enhances the retention of features associated with small targets.

We subsequently integrated a dynamic detection head module that incorporates scale, spatial, and task-aware attention into the existing detection head. This enhancement enables adaptive fusion of multi-level features, thereby improving the model's recall rate for smaller targets. The feature output from the Neck, denoted as $F \in \mathbb{R}^{L \times S \times C}$, is processed by the dynamic detection head, which consists of multiple interconnected attention blocks, expressed as follows:

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \tag{7}$$

where $\pi_L(\cdot)$ represents scale attention, $\pi_S(\cdot)$ denotes spatial attention, and $\pi_C(\cdot)$ signifies task attention. For a comprehensive introduction, please refer to[13].

To address the issue of the IoU loss function in the original YOLOv8s being overly sensitive to positional deviations of small targets, we have introduced the NWD[44] as a metric for measuring the similarity between the predicted bbox and the ground truth bbox. This approach effectively mitigates the effects of pixel offsets and scale sensitivity associated with small targets. For the predicted bbox $P = [x_p, y_p, w_p, h_p]$ and the ground truth bbox $G = [x_g, y_g, w_g, h_g]$, along with their respective Gaussian distributions N_p and N_g, the loss function based on NWD, denoted as L_{NWD}, can be expressed as follows:

$$\begin{aligned} W_2^2(N_p, N_g) = \Big\| \Big(& [x_p, y_p, \tfrac{w_p}{2}, \tfrac{h_p}{2}]^\top, \\ & [x_g, y_g, \tfrac{w_g}{2}, \tfrac{h_g}{2}]^\top \Big) \Big\|_2^2 \\ L_{\text{NWD}} = 1 - \exp & \left(- \frac{\sqrt{W_2^2(N_p, N_g)}}{C} \right) \end{aligned} \tag{8}$$

Where C denotes a normalization constant. By utilizing the smoothing properties of NWD regarding positional deviations, this approach effectively mitigates the challenges associated with IoU sensitivity, rendering it especially suitable for small object detection tasks.

3.5. Tracking Optimization

In the tracking branch, factors such as occlusion, deformation, and motion blur can negatively impact this confidence. Rather than directly outputting the region proposal with the highest confidence as the result for the current frame, we have developed an AKF that dynamically adjusts the observation noise covariance $\mathbf{R}$ during post-processing. The prediction and update steps of the AKF are detailed in Eq. (9), where $\mathbf{F}$, $\mathbf{H}$, $\mathbf{Q}$, and $\mathbf{R}$ denote the state transition matrix, observation matrix, process noise covariance matrix, and measurement noise covariance matrix, respectively. The state variable $\mathbf{x}$ remains consistent with that in SORT[3].

$$\begin{aligned}
&\text{predict}\begin{cases}\hat{\mathbf{x}}_{t|t-1} = \mathbf{F}\hat{\mathbf{x}}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}^\top + \mathbf{Q}\end{cases} \\
&\text{update}\begin{cases}\tilde{\mathbf{R}} = \begin{cases}\alpha \cdot \mathbf{R}, & \text{if } c > \tau \\ (1-c)\cdot \mathbf{R}, & \text{otherwise}\end{cases} \\ \mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}^\top(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}^\top + \tilde{\mathbf{R}})^{-1} \\ \hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}\hat{\mathbf{x}}_{t|t-1}) \\ \mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_{t|t-1}\end{cases}
\end{aligned} \tag{9}$$

The output of the proposed tracking branch provides measurement information for the AKF. Since confidence reflects the quality of tracking, and the observation noise covariance matrix $\mathbf{R}$ in the linear Kalman filter is fixed and unable to adapt to variations in tracking quality, relying solely on measurements with high noise during periods of low confidence can lead to trajectory drift. Consequently, the AKF we designed dynamically adjusts the observation noise covariance $\mathbf{R}$, enabling an adaptive fusion of Kalman prediction and tracking branch output. This adjustment is illustrated in the update step of Eq. (9), where $\tilde{\mathbf{R}}$ is defined, with c representing the current proposal confidence score from the tracking branch, $\mathbf{R}$ as the baseline noise covariance matrix, and α as a small value approaching zero to prevent numerical issues that may arise from $\tilde{\mathbf{R}} = 0$. When the confidence of the measurement proposal ($c > \tau$) is high, the filtering results align with the proposal to prevent prediction drift. Conversely, when confidence is low ($c < \tau$), the AKF utilizes kinematic priors to maintain short-term tracking.

Although the AKF exhibits a degree of predictive capability that enables it to temporarily manage measurement loss and drift, extended occurrences of these issues can result in a gradual increase in the prediction error of the Kalman Filter. Consequently, the AKF generates predictions for a maximum of three frames; if no targets with sufficiently high confidence are detected during this period, it automatically shifts to the detection branch. In this branch, any targets identified with a confidence level exceeding a predetermined threshold are initially classified as candidates. If two consecutive frames detect targets associated with these candidates, the tracking branch can then be activated. In addition to the proactive transition from the tracking branch to the global detection branch, we also periodically activate the detection branch to ensure that the current detection results are consistent with the tracking outcomes.

4. Experiments

4.1. Experimental Setup

StrongSiamDT operates on a server equipped with two NVIDIA RTX 4090 GPUs, utilizing a software framework that includes Python 3.7 and PyTorch 1.8.1.

Datasets: To evaluate the efficacy of StrongSiamDT, we conducted tracking tests using both the 4th Anti-UAV Challenge and the Anti-UAV410[22] dataset. Each dataset presents challenging drone videos characterized by dynamic backgrounds, small-scale targets, and complex movements. The 4th Anti-UAV dataset serves as the official dataset for the 4th Anti-UAV Challenge and comprises three subsets: a train dataset with annotated data, a test dataset for Track 1 that includes only the initial frame's target annotations, and a test dataset for Track 2 that contains no annotations. We trained our model on the train dataset and submitted the results from Tracks 1 and 2 to the official testing platform of the 4th Anti-UAV Challenge. The Anti-UAV410 dataset is currently the largest publicly available infrared Anti-UAV benchmark, encompassing a total of 410 videos featuring drone targets. Due to the unavailability of labels for the 4th Anti-UAV Challenge test dataset, we utilized the Ant-UAV410 dataset to evaluate the tracking performance of the proposed method in comparison to other algorithms.

Evaluation metrics: We choose tracking accuracy, referred to as Acc, as our evaluation metric, which is defined as follows:

$$\begin{aligned}
Acc = &\sum_{t=1}^{T}\frac{IoU_t \times \delta(\nu_t > 0) + p_t \times (1 - \delta(\nu_t > 0))}{T} \\
&- 0.2 \times \left(\sum_{t=1}^{T^*}\frac{p_t \times \delta(\nu_t > 0)}{T^*}\right)^{0.3}
\end{aligned} \tag{10}$$

where IoU_t represents the intersection over union between the predicted bbox and the ground truth bbox. The variable p_t denotes the predicted visibility flag, which takes a value of 1 when the predicted bbox is empty and 0 otherwise. The term v_t signifies the ground truth visibility flag of the target, and the indicator function $\delta(v_t > 0)$ equals 1 when $v_t > 0$ and 0 otherwise. Overall, the first half of Eq. (10) is used to calculate the average IoU across all frames, while the second half functions as a penalty term to address missed tracking instances.

Table 1. Comparison of Acc metrics for different methods on the Anti-UAV410 dataset.

Tracker	Correlation filter-based			Siamese-based			Transformer-based			Anti-UAV task		
	KCF	STRCF	ASRCF	SiamBAN	SiamCAR	GlobalTrack	Stark	Ostrack	MixFormerV2	SiamDT	EDTC	Ours
Initial Bbox	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×
Acc	0.1267	0.3285	0.2908	0.3757	0.3748	0.6050	0.4944	0.5129	0.5222	0.6281	0.6489	**0.6673**

Table 2. The final rankings of the 4th Anti-UAV Challenge.

	track1		track2	
1	chainey	0.7323	qwqwq	0.6676
2	prettyjob	0.7308	chainey	0.5712
3	result	0.7145	**cxl**	**0.5530**
4	**cxl**	**0.6715**	qiuqiuren	0.5313
5	hbut_zhang	0.6699	hbut_zhang	0.5084
6	shan666	0.6473	shan666	0.5076
7	autograd	0.6433	stephenx24	0.4971
8	YUJI	0.6421	ihrd_2025	0.4635
9	zjq21271139	0.6354	zjq21271139	0.4489
10	kust_kk	0.6333	McNu1ty	0.4429

Table 3. Ablation studies on components of StrongSiamDT. GD:Global Detection,AKF:Adaptive Kalman Filter,SCPS:Spatial Constraint Proposal Suppression

Component	baseline			
GD	-	✓	✓	✓
AKF	-	-	✓	✓
SCPS	-	-	-	✓
track1	0.6295	0.6641	0.6694	**0.6715**
track2	-	0.5295	0.5481	**0.5530**
Anti-UAV410	0.6281	0.6572	0.6621	**0.6673**

4.2. Experimental Results

Performance on the 4th Anti-UAV dataset: StrongSiamDT is specifically designed to tackle the challenge of tracking drone targets without any prior knowledge, particularly addressing the issues presented in Track 2 of the 4th Anti-UAV Competition. The results of our method in both tracks of the competition are presented in Tab. 2, where we achieved rankings of 4th and 3rd in Track 1 and Track 2, respectively, with Acc of 0.6715 and 0.5530.

Performance on the Anti-UAV410: In the Anti-UAV410 test dataset, we selected a range of typical trackers for a comparative experiment. The trackers compared include KCF[19], [29], ASRCF[12], SiamBAN[7], SiamCAR[9], GlobalTrack[23], [48], Ostrack[52], MixFormer V2[11], SiamDT[22], and EDTC[63]. Most methods require the initial bbox from the first frame to facilitate subsequent tracking; however, similar to our proposed method, EDTC can achieve target tracking without relying on prior information from the initial frame. The tracking accuracy of various methods is presented in Tab. 1. Our method achieved an accuracy of 0.6673 on the Anti-UAV410 dataset, representing a 3.92% improvement over the baseline. Notably, our approach attains optimal tracking accuracy without depending on prior information from the first frame.

Qualitative Evaluation: We present several representative complex scenarios from the tracking sequences for visual comparison, as illustrated in Fig. 4. We selected the most effective tracker from each category of tracking algorithms for our comparative visualization. Our method demonstrates commendable tracking performance in typical scenarios characterized by rapid changes in target scale, target-background blending, occlusion by trees, small targets amidst significant global interference, and swift movements of both the target and the camera. This success can be attributed to several key factors: (1) the integration of a joint framework that combines detection and tracking branches, allowing the tracker to quickly re-establish focus on the target through the detection branch after it disappears and reappears; (2) the design of a detector specifically for small drone targets, which effectively identifies targets in complex environments; (3) the utilization of SCPS and AKF, which mitigates global interference while addressing short-term loss issues.

4.3. Ablation Study

To validate the contributions of various modules within StrongSiamDT to tracking performance, we conducted a series of ablation experiments. As illustrated in Tab. 3, both the AKF and SCPS modules enhance the tracker's performance, resulting in a modest improvement in tracking accuracy. Notably, the global detection branch significantly increases the framework's efficacy, yielding improvements of 3.46% and 2.91% on the Track1 and Anti-UAV410 datasets, respectively. This enhancement addresses the challenges encountered in resuming tracking after losing sight of the target before the implementation of the global detection

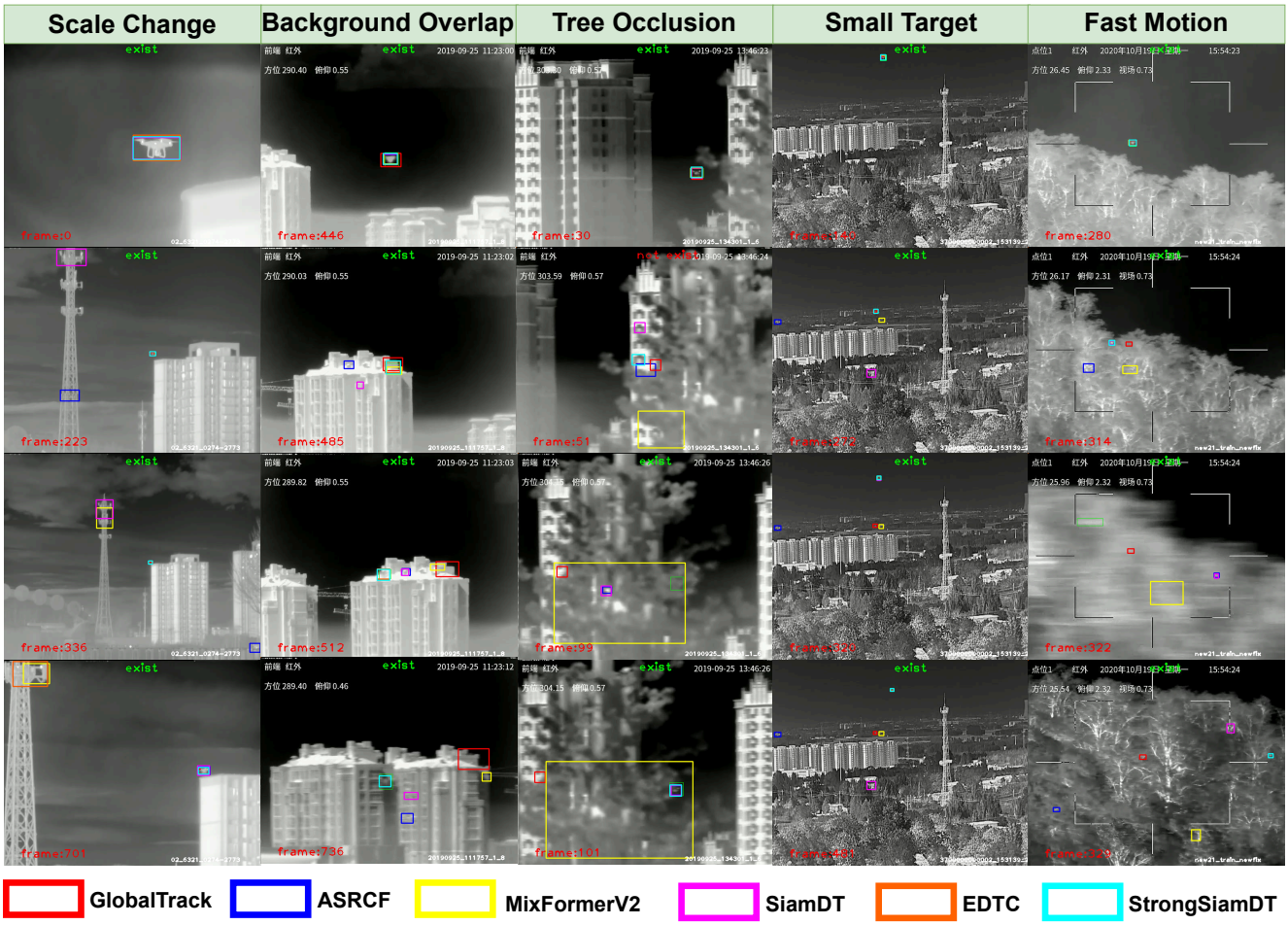


Figure 4. Visual comparison of different methods. StrongSiamDT effectively addresses challenging scenarios, including scale variations, occlusion, and background interference, while demonstrating superior tracking performance compared to the algorithms used for comparison.

Table 4. Ablation studies on components of Global Detector

	DyHead	NWD	P	R	mAP50
YOLO v8s	-	-	0.9197	0.7417	0.8215
	✓	-	0.9801	0.9220	0.9452
	✓	✓	**0.9824**	**0.9256**	**0.9489**

branch; its introduction facilitates the timely restoration of previously interrupted tracking.

To validate the effectiveness of the NWD loss utilized in the detection branch and the dynamic detection head, we conducted an independent ablation study on the detector. The results, presented in Tab. 4, demonstrate that the incorporation of the dynamic detection head led to increases of 6.04%, 18.03%, and 12.37% in precision, recall, and mAP50, respectively, thereby significantly enhancing detection performance. Furthermore, the NWD loss effectively optimizes small object detection, further improving the detector's sensitivity to smaller targets.

5. Conclusion

In this paper, we propose a strong siamese drone tracker for Anti-UAV tracking. We first developed a powerful detector tailored for small targets, which significantly enhances the recall rate of diminutive objects. Subsequently, we implemented global proposal suppression and adaptive Kalman filtering to facilitate tracking. Finally, we devised a dynamic switching logic for the detection and tracking branches. The StrongSiamDT has demonstrated exceptional performance across multiple datasets. Future research may explore the design of a more lightweight tracker while maintaining optimal performance.

References

[1] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. Channel coded distribution field tracking for thermal infrared imagery. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pages 9–17, 2016. 3

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part II 14*, pages 850–865. Springer, 2016. 1, 2

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. Ieee, 2016. 6

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2544–2550. IEEE, 2010. 2

[5] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8126–8135, 2021. 1

[6] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14572–14581, 2023. 1, 2

[7] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020. 7

[8] Xiaolong Cui, Xingxiu Li, Panlong Wu, Shan He, and Ruohan Zhao. Weakly semi-supervised infrared small target detection guided by point labels. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 3

[9] Ying Cui, Dongyan Guo, Yanyan Shao, Zhenhua Wang, Chunhua Shen, Liyan Zhang, and Shengyong Chen. Joint classification and regression for visual tracking with fully convolutional siamese networks. *International Journal of Computer Vision*, pages 1–17, 2022. 2, 7

[10] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022. 2

[11] Yutao Cui, Tianhui Song, Gangshan Wu, and Limin Wang. Mixformerv2: Efficient fully transformer tracking. *Advances in neural information processing systems*, 36:58736–58751, 2023. 2, 7

[12] Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li. Visual tracking via adaptive spatially-regularized correlation filters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4670–4679, 2019. 7

[13] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 2, 3, 5

[14] Yimian Dai, Yiquan Wu, Fei Zhou, and Kobus Barnard. Asymmetric contextual modulation for infrared small target detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 950–959, 2021. 2

[15] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 4310–4318, 2015. 2

[16] Suyog D Deshpande, Meng Hwa Er, Ronda Venkateswarlu, and Philip Chan. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets 1999*, pages 74–83. SPIE, 1999. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[18] Jinhui Han, Yong Ma, Bo Zhou, Fan Fan, Kun Liang, and Yu Fang. A robust infrared small target detection algorithm based on human visual system. *IEEE Geoscience and Remote Sensing Letters*, 11(12):2168–2172, 2014. 2

[19] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 2, 7

[20] Yangguang Hu, Mingqing Xiao, Shaoyi Li, and Yao Yang. Aerial infrared target tracking based on a siamese network and traditional features. *Infrared Physics & Technology*, 111: 103505, 2020. 2

[21] Bo Huang, Junjie Chen, Tingfa Xu, Ying Wang, Shenwang Jiang, Yuncheng Wang, Lei Wang, and Jianan Li. Siamsta: Spatio-temporal attention based siamese tracker for tracking uavs. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1204–1212, 2021. 2, 3

[22] Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Zhao, and Tingfa Xu. Anti-uav410: A thermal infrared benchmark and customized scheme for tracking drones in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2852–2865, 2023. 2, 3, 6, 7

[23] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11037–11044, 2020. 2, 3, 7

[24] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: A large-scale benchmark for vision-based uav tracking. *T-MM*, 2021. 3

[25] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018. 1, 2

[26] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019. 2

[27] Boyang Li, Chao Xiao, Longguang Wang, Yingqian Wang, Zaiping Lin, Miao Li, Wei An, and Yulan Guo. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing*, 32:1745–1758, 2022. 3

[28] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 2

[29] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4904–4913, 2018. 2, 7

[30] Yifan Li, Dian Yuan, Meng Sun, Hongyu Wang, Xiaotao Liu, and Jing Liu. A global-local tracking framework driven by both motion and appearance for infrared anti-uav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3026–3035, 2023. 3

[31] Fangcen Liu, Chenqiang Gao, Fang Chen, Deyu Meng, Wangmeng Zuo, and Xinbo Gao. Infrared small and dim target detection with transformer under complex backgrounds. *IEEE Transactions on Image Processing*, 32:5921–5932, 2023. 3

[32] Qiao Liu, Xin Li, Zhenyu He, Nana Fan, Di Yuan, Wei Liu, and Yongsheng Liang. Multi-task driven feature models for thermal infrared tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11604–11611, 2020. 3

[33] Qiao Liu, Xin Li, Zhenyu He, Nana Fan, Di Yuan, and Hongpeng Wang. Learning deep multi-level similarity for thermal infrared object tracking. *IEEE Transactions on Multimedia*, 23:2114–2126, 2020. 3

[34] Qiao Liu, Di Yuan, Nana Fan, Peng Gao, Xin Li, and Zhenyu He. Learning dual-level deep representation for thermal infrared tracking. *IEEE Transactions on Multimedia*, 25:1269–1281, 2022. 3

[35] Yanyi Lyu, Zhunga Liu, Huandong Li, Dongxiu Guo, and Yimin Fu. A real-time and lightweight method for tiny airborne object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3016–3025, 2023. 3

[36] Kun Qian, Shou-jin Zhang, Hong-yu Ma, and Wen-jun Sun. Siamist: Infrared small target tracking based on an improved siamrpn. *Infrared Physics & Technology*, 134:104920, 2023. 2

[37] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[38] Xiaopeng Shao, Hua Fan, Guangxu Lu, and Jun Xu. An improved infrared dim and small target detection algorithm based on the contrast mechanism of human visual system. *Infrared Physics & Technology*, 55(5):403–408, 2012. 2

[39] Zongheng Tang, Yulu Gao, Zizheng Xun, Fengguang Peng, Yifan Sun, Si Liu, and Bo Li. Strong detector with simple tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3047–3053, 2023. 3

[40] Mengchu Tian, Zhimin Chen, Huifen Wang, and Linyan Liu. An intelligent particle filter for infrared dim small target detection and tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 58(6):5318–5333, 2022. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[42] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 1, 2

[43] Bingshu Wang, Qiang Li, Qianchen Mao, Jinbao Wang, CL Philip Chen, Aihong Shangguan, and Haosu Zhang. A survey on vision-based anti unmanned aerial vehicles methods. *Drones*, 8(9):518, 2024. 1

[44] Jinwang Wang, Chang Xu, Wen Yang, and Lei Yu. A normalized gaussian wasserstein distance for tiny object detection. *arXiv preprint arXiv:2110.13389*, 2021. 2, 5

[45] Xiaotian Wang, Kai Zhang, Shaoyi Li, Yangguang Hu, and Jie Yan. An optimal long-term aerial infrared object tracking algorithm with re-detection. *IEEE Access*, 7:114320–114333, 2019. 2, 3

[46] Tianhao Wu, Boyang Li, Yihang Luo, Yingqian Wang, Chao Xiao, Ting Liu, Jungang Yang, Wei An, and Yulan Guo. Mtunet: Multilevel transunet for space-based infrared tiny ship detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 3

[47] Fei Xie, Lei Chu, Jiahao Li, Yan Lu, and Chao Ma. Videotrack: Learning to track objects via video transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22826–22835, 2023. 2

[48] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10448–10457, 2021. 1, 2, 7

[49] Bo Yang, Xinyu Zhang, Jian Zhang, Jun Luo, Mingliang Zhou, and Yangjun Pi. Eflnet: Enhancing feature learning network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. 2

[50] Chao Yang, Qiao Liu, Gaojun Li, Honghu Pan, and Zhenyu He. Learning diverse fine-grained features for thermal infrared tracking. *Expert Systems with Applications*, 238: 121577, 2024. 3

[51] Xin Yang, Gang Wang, Weiming Hu, Jin Gao, Shubo Lin, Liang Li, Kai Gao, and Yizheng Wang. Video tiny-object

detection guided by the spatial-temporal motion information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3054–3063, 2023. 3

[52] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022. 1, 2, 7

[53] Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, and Shilin Zhou. Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15528–15538, 2023. 3

[54] Di Yuan, Xiu Shu, Qiao Liu, Xinming Zhang, and Zhenyu He. Robust thermal infrared tracking via an adaptively multi-feature fusion model. *Neural Computing and Applications*, 35(4):3423–3434, 2023. 2, 3

[55] Sungmin Yun and Sungho Kim. Robust infrared target tracking using thermal information in mean-shift. In *Pattern Recognition and Tracking XXX*, pages 52–57. SPIE, 2019. 2

[56] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. Isnet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 877–886, 2022. 3

[57] Tianfang Zhang, Zhenming Peng, Hao Wu, Yanmin He, Chaohai Li, and Chunping Yang. Infrared small target detection via self-regularized weighted sparse model. *Neurocomputing*, 420:124–148, 2021. 2

[58] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4591–4600, 2019. 2

[59] Zhihao Zhang, Lei Jin, Shengjie Li, JianQiang Xia, Jun Wang, Zun Li, Zheng Zhu, Wenhan Yang, PengFei Zhang, Jian Zhao, et al. Modality meets long-term tracker: A siamese dual fusion framework for tracking uav. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1975–1979. IEEE, 2023. 3

[60] Dong Zhao, Lin Gu, Kun Qian, Huixin Zhou, Tingwu Yang, and Kuanhong Cheng. Target tracking from infrared imagery via an improved appearance model. *Infrared Physics & Technology*, 104:103116, 2020. 2

[61] Jian Zhao, Jianan Li, Lei Jin, Jiaming Chu, Zhihao Zhang, Jun Wang, Jiangqiang Xia, Kai Wang, Yang Liu, Sadaf Gulshad, et al. The 3rd anti-uav workshop & challenge: Methods and results. *arXiv preprint arXiv:2305.07290*, 2023. 3

[62] Liying Zheng, Shuo Zhao, Yanbo Zhang, and Lei Yu. Thermal infrared pedestrian tracking using joint siamese network and exemplar prediction model. *Pattern Recognition Letters*, 140:66–72, 2020. 2

[63] Xue-Feng Zhu, Tianyang Xu, Jian Zhao, Jia-Wei Liu, Kai Wang, Gang Wang, Jianan Li, Qiang Wang, Lei Jin, Zheng Zhu, et al. Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-uav system. *arXiv preprint arXiv:2306.15767*, 2023. 3, 7

[64] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018. 2