electronics

MDPI

Article

Anti-UAV Target Tracking with Motion Association Integration

Yaofu Cao [1], Xiaoyong Sun [2,*], Runze Guo [2], Zhaoyang Dang [2], Shaojing Su [2] and Desen Bu [2]

[1] School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, China; 202321511129@smail.xtu.edu.cn
[2] College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China
* Correspondence: sunxiaoyong14@nudt.edu.cn

Abstract

While the rapid development and widespread application of drone technology have brought about significant advancements, they have also introduced security challenges, making anti-UAV technology a key research focus. However, existing methods still face severe challenges when dealing with UAV tracking in complex scenarios. To address this, this paper proposes an integrated Motion-associated Detection and Tracking Collaboration (MDTC) system for anti-UAV applications. To better handle the perception of target existence states, we designed a motion association module that dynamically senses the presence of targets and responds quickly to target disappearance. Simultaneously, to address the issue of feature degradation in small targets, we optimized the detection branch to enhance robust perception of multi-scale targets. Additionally, the proposed verification matching mechanism can infer the integrity and reliability of targets in occluded scenarios, ensuring stable tracking. Compared to existing methods, our approach achieves superior performance across three benchmark datasets. On Anti-UAV600, it attains IoU, ACC, and SR scores of 0.525, 0.427, and 0.641, respectively—surpassing the second-best method, GlobalTrack, by 6.2%, 6.4%, and 5.9%. These gains highlight the method's strengths in prompt target response, scale adaptability, and occlusion awareness, underscoring its reliability and practicality for real-world deployment.

Keywords: anti-UAV system; object detection; visual tracking

1. Introduction

The rapid development and widespread application of UAV technology in recent years have epitomized the convergence and innovation of information technologies. Their utilization in agriculture, logistics, urban security, and other fields is accelerating the intelligent transformation of traditional industries. However, their low cost, high mobility, and strong concealment also pose significant security governance challenges. Malicious actors exploit UAVs for illegal surveillance and attacks, severely threatening personal privacy and public safety. In this context, research on anti-UAV systems holds substantial practical significance. Among these technologies, computer vision-based target detection and tracking is regarded as a critical component of active defense systems due to its non-contact nature and all-weather capabilities [1]. In the field of UAV perception, researchers have conducted numerous studies on anti-UAV technologies to effectively address challenges posed by small-sized targets and complex environmental factors. Among these, some studies frame anti-UAV tasks as single-object tracking (SOT) problems, such as SiamDT [2],

Check for updates

Academic Editor: Felipe Jiménez

Received: 5 December 2025
Revised: 26 January 2026
Accepted: 11 February 2026
Published: 15 February 2026

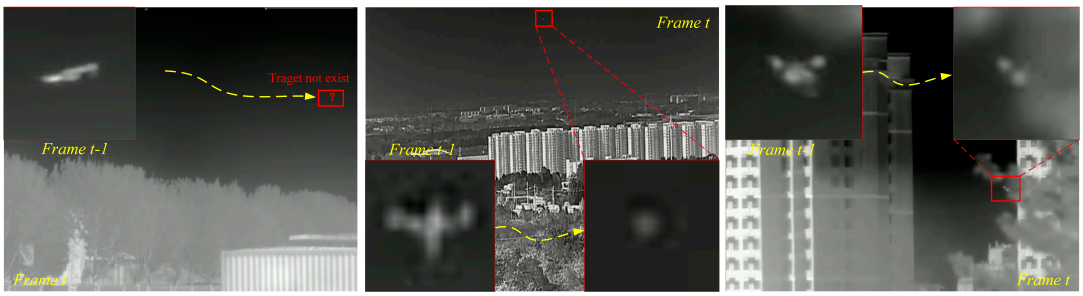
Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

ATOM [3], and GlobalTrack [4]. However, SOT typically requires the initial state information of the UAV during the initialization phase. In contrast, in real-world scenarios, UAVs may appear and disappear at any time, thereby limiting the applicability of this method in practical environments.

To address the perception problem of UAV targets lacking prior knowledge in complex scenarios, Zhu et al. [5] proposed an Evidence Detection and Tracking Collaboration (EDTC) framework, which reformulates the anti-UAV task as a novel joint detection and tracking paradigm. This framework significantly enhances the perception performance of single UAV targets in unknown environments through a synergistic optimization mechanism between a global detector and a local tracker. Liu et al. [6] introduced an adaptive detection-tracking collaboration mechanism and incorporated data augmentation methods, improving tracking robustness and accuracy.

Although deep learning has achieved significant progress in the field of object detection and tracking, UAV tracking still faces numerous challenges in practical applications. UAV targets are typically small in size, making them difficult to identify at long distances or against complex backgrounds. Infrared imaging can enhance target recognition in low-light conditions; however, its effectiveness is limited by low resolution and interference from background noise, which increases the difficulty of tracking [7].

Figure 1 illustrates several challenges encountered in UAV tracking tasks from an infrared perspective. Figure 1a demonstrates the target disappearance problem—during UAV movement, traditional single-object tracking (SOT) algorithms fail to determine the target's presence quickly. Even after the target has left the field of view, predicted bounding boxes continue to be generated, significantly increasing the false alarm rate. Figure 1b highlights the scale adaptability challenge—as the distance between the UAV and the camera increases, the target's imaging scale sharply decreases, accompanied by significant image blurring effects. Combined with background clutter and noise interference, this leads to reduced target discriminability. Figure 1c presents an occlusion scenario—UAV may be partially or fully obscured by background elements such as buildings, trees, or clouds in complex environments. Among these, occlusion-induced local feature visibility issues are particularly prominent. Among these, the issue of visibility of local features caused by occlusion is particularly prominent. We explore the problem of local feature perception in targets.



(**a**) (**b**) (**c**)

Figure 1. Schematic diagram of challenging scenarios in anti-UAV tracking.

To address the aforementioned challenges, this paper proposes an integrated motion association target detection and tracking collaboration (MDTC) for anti-UAV task, which integrates YOLOv8s and a Kalman filter as the detection and tracking branches, respectively. First, to address the target presence perception problem, we design a motion association module that dynamically assesses the target's existence through a multi-level confidence evaluation mechanism. This enables real-time responses to target disappearance, effectively mitigating the false alarm issue caused by delayed judgments in traditional SOT methods.

Second, to resolve feature degradation due to UAV scale variations in complex scenarios, we enhance the detection branch. The improved network strengthens the perception capability for small and blurred targets, ensuring robust performance across different scales. Finally, for target perception under occlusion, we propose a grayscale-based verification matching mechanism. By validating the local features of the unconcluded regions of the target and incorporating historical trajectory priors, this mechanism infers the target's integrity and reliability, thereby maintaining stable tracking even in occlusion scenarios.

The proposed method was evaluated on three public anti-UAV datasets: Anti-UAV [8], Anti-UAV410 [2], and AntiUAV600 [5]. Experimental results demonstrate that the proposed algorithm not only exhibits robust tracking performance for UAV targets but also effectively identifies target disappearance states and promptly terminates trajectories. More importantly, in scenarios where targets are occluded, the method can maintain continuous tracking through local information perception. These results fully validate the reliability and effectiveness of the proposed approach in anti-UAV applications. The main contributions of this paper can be summarized as follows:

- An integrating motion association target detection and tracking collaboration for anti-UAV tasks is proposed.
- A motion association module is designed, which dynamically assesses target presence through a multi-level confidence evaluation mechanism and rapidly responds to target disappearance.
- To address target perception under occlusion, a grayscale-based verification matching mechanism is introduced, enabling stable tracking even in occluded conditions.

2. Materials and Methods

2.1. Object Detection

In the field of visual detection, deep learning methods have demonstrated remarkable effectiveness in image object detection tasks. Among them, the region proposal-based two-stage object detection framework R-CNN [9] laid the research foundation for this domain. The core processing pipeline of such algorithms consists of three key steps: first, a selective search region generation strategy is employed to construct an initial set of candidate regions; then, feature extraction and classification are performed for each candidate region; finally, the bounding box coordinates are refined through regression analysis. For instance, Chen et al. [10] proposed an infrared small target detection method that enhances target features and suppresses background clutter by leveraging image contrast and gradient information.

The regression-based single-stage detection framework (represented by the YOLO series [11]) has garnered significant attention due to its high computational efficiency and concise network architecture. This algorithm divides the input image into grid cells and directly completes bounding box localization and probability prediction in a single forward pass. This end-to-end processing effectively avoids the computational overhead of candidate region generation in traditional methods. For example, Wang et al. [12] improved the YOLOv10 model by introducing the contextual anchor attention mechanism, significantly enhancing the model's inference efficiency. Qin et al. [13] adjusted the feature fusion strategy of YOLOv7 by adding a small target detection layer, thereby improving detection efficiency. Ye et al. [14] constructed a global-to-local feature enhancement network that effectively integrates multi-scale feature information. BBGFA-YOLO [15] introduced a multi-scale deformable convolution-based attention enhancement module, significantly improving detection accuracy in infrared image object detection tasks. ESG-YOLO [16] incorporated the stride-free convolution module SPD [17] to modify the convolutional network, mitigating the performance degradation caused by fine-grained information loss.

2.2. Visual Tracking

The core task in the field of visual object tracking (VOT) is to predict the spatial position and scale changes in a target in subsequent frames based on its state in the initial frame of a video sequence. In recent years, thanks to the rapid development of deep learning technology and its in-depth application in computer vision, VOT research has achieved significant progress. Among discriminative model-based tracking methods, the algorithms proposed in references [3,18–20] exhibit superior performance in terms of tracking accuracy and robustness. ATOM [3] was the first to decouple target classification and bounding box estimation into two independent branches, achieving precise localization through offline network training. DiMP [18] introduced a discriminative learning mechanism on the basis of ATOM, iteratively optimizing target model parameters via backpropagation, significantly enhancing tracking robustness. PrDiMP [19] probabilized the regression classification task, employing a KL divergence loss function to mitigate imbalance issues during training. KYS [20] constructed a spatiotemporal semantic reasoning module to dynamically adapt by learning the interaction between the target and its environment through online learning. These methods optimize the feature representation of targets and backgrounds by learning discriminative classifiers, achieving tracking performance superior to traditional approaches.

Additionally, Siamese network-based tracking methods represent an important branch in this field. These methods combine the classical template-matching concept with deep learning, forming a comprehensive tracking paradigm. Traditional algorithms in this category include SiamFC [21], SiamRPN [22], SiamRPN++ [23], and SiamCAR [24], which focus on integrating diverse ideas into the Siamese framework. Early representative works include: SiamFC [21], which first established an end-to-end Siamese network architecture; SiamRPN [22], which integrated a region proposal network into the Siamese framework; SiamRPN++ [23], which enhanced feature representation by introducing the deep residual network ResNet; and SiamCAR [24], which improved target localization accuracy by adding a centrality evaluation branch.

In recent years, effective improvements based on Siamese networks include: SiamDT [2], which enhanced the semantic discriminability of target regions by constructing a complementary feature extraction mechanism and a background interference suppression module. SiamCAP [25] integrated a context-aware feature enhancement module with a pixel-level attention mechanism, achieving hierarchical enhancement of feature representation. FocTrack [26] incorporated a selective feature update-based local template optimization mechanism into the Siamese network, improving adaptability to target appearance changes.

2.3. Anti-UAV Detection and Tracking Techniques

Early anti-UAV research was designed as an independent research direction, primarily focusing on UAV target detection or tracking tasks. The DUT Anti-UAV [27] dataset was proposed for visible-light-based anti-UAV tasks, including detection and tracking subsets. It features rich target and background diversity, with over 35 types of UAV targets and seven complex outdoor backgrounds.

For infrared modality datasets, notable works include [2,5,8]. The Anti-UAV [8] dataset contains 318 video sequences in both visible and infrared modalities. The team proposed DFSC, which employs a dual-stream semantic consistency strategy to learn multi-modal feature data, effectively improving UAV target tracking performance. The Anti-UAV410 [2] dataset consists of 410 thermal infrared video sequences, further divided into six challenging subsets. Notably, the team's proposed SiamDT algorithm achieved outstanding performance in UAV target tracking under complex environments. The Anti-UAV600 dataset and the EDTC method were simultaneously proposed by Zhu et al. [5].

EDTC, based on evidence-theoretic collaborative perception, integrates detection and tracking modules to achieve autonomous state awareness and continuous tracking of UAV targets without relying on prior initial state information.

Regarding detection and tracking collaborations, existing methods primarily fuse target features, motion information, and spatial location data to compute similarity, then employ optimization algorithms for matching. For example, Bewley et al. [28] proposed the Simple Online and Realtime Tracking (SORT) algorithm, which incorporates target position and motion information into a Kalman filter to predict the target's next-frame location, using Intersection over Union between predictions and detections as a similarity metric. Sun et al. [29] fully leveraged the Kalman filter's advantages in short-term prediction and combined it with more effective recovery strategies to achieve more stable and reliable bounding box predictions. Xu et al. [30] proposed a FlexiLength Network (FLN) that integrates trajectory data of varying observation lengths to address the "observation length shift" problem in trajectory prediction tasks. SiamYOLO [31] detects UAV targets through global perception and predicts global candidate boxes by computing spatiotemporal information. Liu et al. [6] proposed a data augmentation strategy and applied it to a "detection-verification-tracking" framework, providing new insights for anti-UAV tracking.

3. Methodology

3.1. Overall Framework

The proposed MDTC method draws key inspiration from the case study of frogs capturing flying insects. The frog's continuous perception of flying insect targets is based on the neural processing mechanism of its visual system, which includes specialized "pest detector" neurons for detecting small, dark, moving insects, as well as pathways responsible for processing detailed visual information [32]. To mimic this dual-information flow processing mechanism, this study proposes an MDTC method based on a detection-verification-tracking paradigm.

As shown in Figure 2, MDTC first identifies candidate targets through an object detector and generates bounding box sets with confidence scores. Subsequently, it determines whether to activate the motion association module based on confidence evaluation. When detection confidence is low, the method performs target information association by analyzing motion and feature relationships, achieving cross-frame association matching of potential targets and optimal tracking result selection.

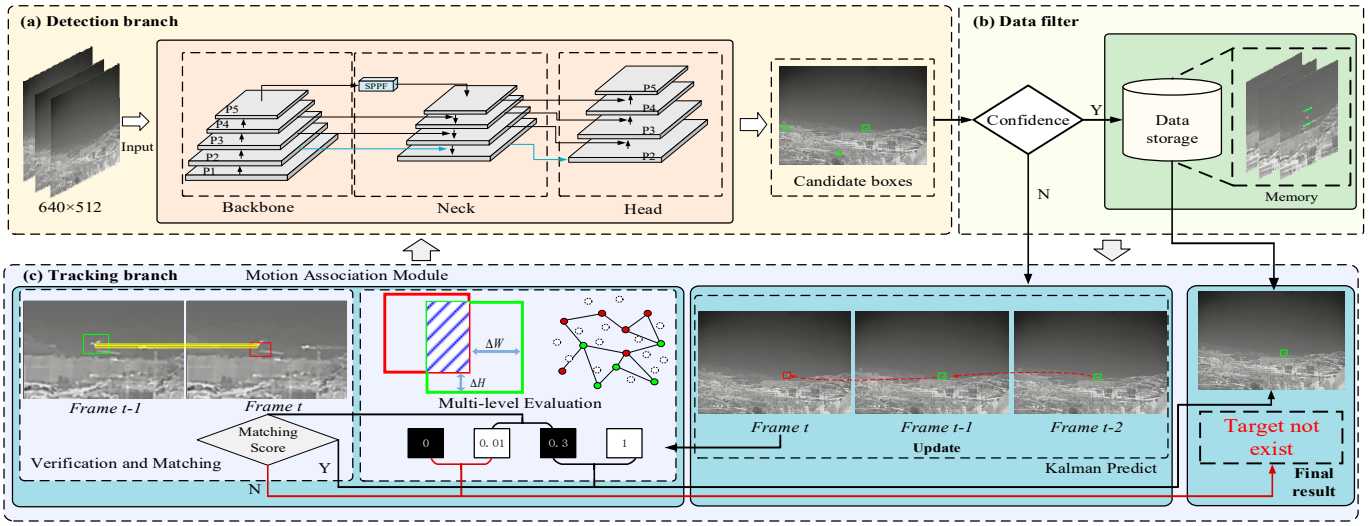


Figure 2. Overall framework of the MDTC method. The framework consists of three core modules: (**a**) detection branch, (**b**) data filter, and (**c**) tracking branch.

3.2. Detection Branch

This study employs YOLOv8s as the global detector for MDTC. In video sequences, the significant scale variation in the UAV target has a significant impact on the detector, resulting in a decline in the accuracy of the baseline YOLOv8s model when encountering such scale changes. To address this issue, this study proposes improvements to the YOLOv8s model, with its network structure illustrated in Figure 3. Specifically, to enhance the model's adaptability to complex scenarios, we introduce the Contextual Anchor Attention (CAA) module [33] into the Backbone network to achieve precise capture of key feature regions. Additionally, to mitigate performance degradation caused by the loss of fine-grained information in low-resolution images, we adjust the convolutional structure by incorporating the Space-to-Depth (SPD) module [17] to preserve more fine-grained details. Finally, to strengthen the model's capability in detecting small targets, we add a small target detection head to the P2 feature layer.

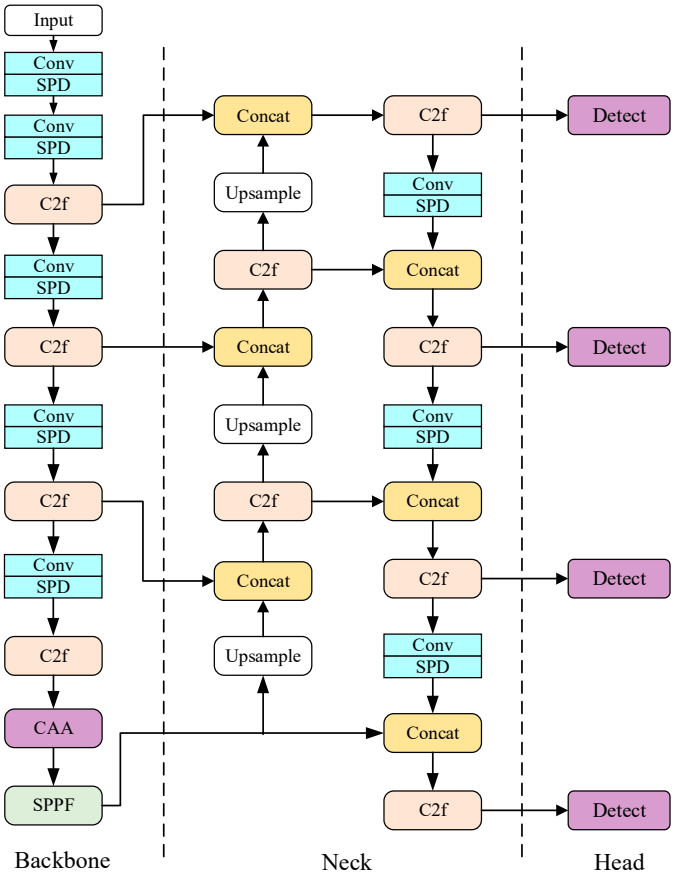


Figure 3. The improved algorithm network structure based on YOLOv8s in this paper.

3.2.1. Contextual Anchor Attention

The optimization of deep learning network architectures based on attention mechanisms achieves enhanced feature representation capabilities through dynamic weight allocation strategies. Classical attention mechanisms, such as Squeeze-and-Excitation Networks (SE) [34] and Convolutional Block Attention Module (CBAM) [35], demonstrate significant effectiveness in cross-dimensional attention weight allocation and integrated feature learning. However, these conventional attention mechanisms still present opportunities for optimization in balancing computational complexity and feature representation efficiency when applied to drone target detection tasks in complex backgrounds. To address these issues, this study introduces the Contextual Anchor Attention (CAA) mechanism.

CAA establishes a dynamic association model between semantic context and local features through an anchor positioning strategy, enabling precise capture of key feature regions.

Specifically, this study incorporates the CAA mechanism into the Backbone network of YOLOv8s to enhance the network's representation capability for key regions in input feature maps. Compared with conventional attention methods, the innovation of CAA lies in constructing multiple sub-feature groups through channel-wise partitioning. Secondly, feature tensor reshaping is employed to strengthen the model's perception of multiscale features. Additionally, the cross-branch fusion architecture designed in CAA facilitates interactive enhancement of multiscale features, enriching feature representation. CAA effectively balances computational efficiency and feature learning capability. By concentrating computational resources on key regions and integrating cross-dimensional salient information, the model's learning ability for complex feature patterns is significantly improved, while avoiding substantial increases in computational costs. The specific implementation steps of the CAA module are as follows:

$$F_{\text{pool}} = \text{Conv}_{1\times1}\left(P_{\text{avg}}(X)\right) \tag{1}$$

In the equation, P_{avg} represents the global average pooling operation, and X denotes the input feature map.

$$F_{\text{strip}}^{\text{h}} = \text{DWConv}_{1\times k_b}\left(F_{\text{pool}}\right) \tag{2}$$

$$F_{\text{strip}}^{\text{v}} = \text{DWConv}_{k_b\times1}\left(F_{\text{strip}}^{\text{h}}\right) \tag{3}$$

Equations (2) and (3) indicate the adoption of two sets of orthogonal one-dimensional strip convolution kernels, which extract long-range features along the spatial horizontal and vertical directions respectively, where k_b represents the convolution kernel size.

$$A = \sigma\left(\text{Conv}_{1\times1}\left(F_{\text{strip}}^{\text{v}}\right)\right) \tag{4}$$

Equation (4) represents the fusion of orientation-aware convolution extracted directional features with original features to generate attention weights, where s denotes the Sigmoid activation function.

$$F_{\text{attn}} = A \odot X + X \tag{5}$$

Equation (5) demonstrates the weighted enhancement of the original features, where denotes element-wise multiplication.

CAA enables the precise capture of global semantic information and key regional features, thereby enhancing the model's ability to extract salient features of drone targets in complex backgrounds while suppressing the adverse effects of background noise. Additionally, it provides more discriminative feature representations for subsequent Neck and Head modules.

3.2.2. Space-to-Depth

To address the issue of fine-grained feature degradation in traditional convolutional neural networks when processing low-resolution images and tiny targets, this paper introduces a strideless convolutional network as a novel feature extraction architecture. This framework employs a dual-module combination of Space-to-Depth (SPD) layer and Non-Stride Convolution (NSD) layer, replacing conventional stride convolutions and pooling operations. Compared to traditional methods, this design effectively preserves the high-resolution characteristics of feature maps, thereby mitigating detail loss caused by down-sampling and significantly enhancing model performance in low-resolution object detection tasks.

The SPD layer adopts a spatial-to-channel transformation strategy, reorganizing the spatial structure information of feature maps into channel-wise structural information. This approach reduces the spatial resolution of feature maps without information loss, as detailed below:

Given an intermediate feature map X with width S, height S, and channel number C (denoted as $S \times S \times C$), the SPD module first partitions the feature map X, calculated as follows:

$$X_0 = X[0:S:2, 0:S:2] \tag{6}$$

$$X_1 = X[1:S:2, 0:S:2] \tag{7}$$

$$X_2 = X[0:S:2, 1:S:2] \tag{8}$$

$$X_3 = X[1:S:2, 1:S:2] \tag{9}$$

In the equation, each sub-feature map has a shape of $\frac{S}{2} \times \frac{S}{2} \times C$. The sub-feature maps are sequentially concatenated to form a new feature map with increased channel dimensionality, with the concatenation process expressed as:

$$Y = Concat(X_0, X_1, X_2, X_3) \tag{10}$$

The NSD layer subsequently processes the feature maps. The NSD layer performs convolution operations while maintaining the spatial dimensions of the feature maps unchanged, allowing for the refined extraction of local features. This architectural design effectively mitigates information loss caused by excessive down-sampling, preserves more fine-grained details, and provides richer spatial feature information for subsequent processing.

3.2.3. Small Object Detection Head

Due to the relatively low pixel proportion of some sequential UAV targets, as shown in Figure 4, the UAV targets exhibit feature blurring after continuous downsampling operations. To address this issue, this study introduces a dedicated feature extraction branch (P2 layer) for small target detection into the YOLOv8s network. By constructing a multi-scale feature fusion mechanism, the feature representation capability for small-sized targets is enhanced, thereby effectively improving the model's localization and recognition performance for small targets.

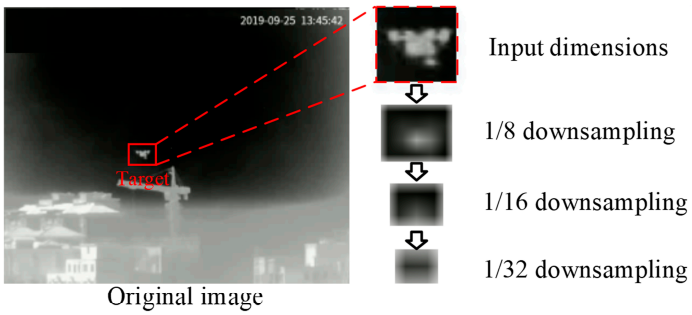


Figure 4. Example of UAV target downsampling, where the effective information of the target becomes extremely weak after the 1/8 downsampling process.

Specifically, the P2 small target detection head is based on a collaborative architecture combining the Feature Pyramid Network (FPN) and the Path Aggregation Network (PAN). The P2 layer feature maps originate from the shallow stages of the network, possessing higher spatial resolution. However, due to their low-level semantic hierarchy in the network, their semantic representation capability for targets is relatively weak. To mitigate the degradation in representation ability caused by multi-scale target variations, this

detection head enhances fine-grained feature extraction, alleviating the issue of small target feature disappearance and thereby improving detection accuracy.

3.3. Tracking Branch

Multiple candidate bounding boxes obtained through detector computation are processed by MDTC using a confidence score-based data filtering mechanism, as shown in Figure 2c. These filtered results are temporarily stored in a position repository, ensuring that historical detection data can be promptly retrieved during motion association analysis, thereby enabling continuous tracking of target trajectories.

Considering the scale variations and occlusions that UAV targets may encounter during motion, these factors can lead to degraded detection performance and loss of predicted bounding boxes. Additionally, accounting for the inherent motion characteristics of UAV targets, we first establish a state equation and observation equation for target motion to achieve minimum variance estimation of the target's state variables. During the recursive estimation process, a predictor-corrector dual-stage iterative mechanism is constructed to update the target's motion state continuously. Specifically, the motion state equation of the target is as follows:

$$x = \lceil u,\ v,\ s,\ r \rceil \tag{11}$$

In the equation, u and v represent the horizontal and vertical coordinates of the UAV target bounding box center point, respectively. r denotes the height of the predicted bounding box, and s indicates the aspect ratio of the bounding box. The Kalman filter recursively propagates the state from time step $k-1$ to k, with the state equation expressed as:

$$x_k = A_k x_{k-1} + B_k u_k + w_k \tag{12}$$

In the equation, x_k and x_{k-1} represent the state variables at time steps k and $k-1$ respectively; u_k denotes the external control input at time step k; w_k indicates the process noise at time step k; A_k represents the state transition matrix at time step k; and B_k corresponds to the control-input model at time step k. The observation equation is expressed as:

$$z_k = H_k x_k + v_k \tag{13}$$

In the equation, v_k represents the observation noise, while H_k denotes the observation model, for which we employ a constant velocity motion model. As illustrated in Figure 5, during the temporal update process of the Kalman filter [36], the target bounding box prediction at frame t relies on the detection box information from frame $t-1$. To enhance the robustness of target state estimation, this study employs a confidence threshold θ in the data filter to screen reliable image data, ensuring that the target information stored in the position database meets credibility requirements. Specifically, when the detector outputs multiple candidate bounding boxes in an image, the data filter selects the detection box with the highest confidence for evaluation. If its confidence exceeds q, it is retained and temporarily stored in the data memory while being output as the frame result. If the confidence is below θ, all candidate boxes are preserved and forwarded to the motion association module for further processing.

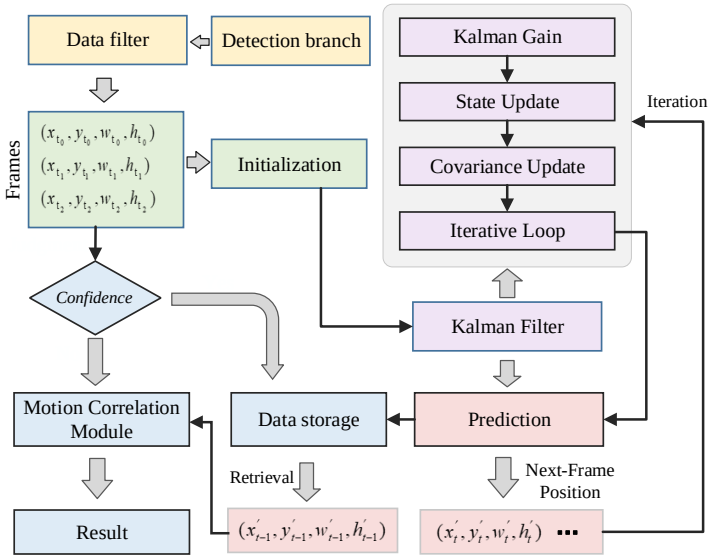


Figure 5. Target perception architecture based on Kalman filter.

3.4. Motion Association Module

In anti-UAV tracking, when the confidence score of a candidate bounding box falls below the predefined threshold θ_d, it indicates that the candidate box contains both the actual target detection box and false detection boxes caused by interference factors. Influenced by complex background interference factors such as target occlusion and scale variations, the target features become attenuated. In such cases, the confidence scores of false detection boxes may even exceed those of true target detection boxes. To address this challenge, as shown in Figure 6, this study proposes a motion association module to enhance the robustness and accuracy of the system.

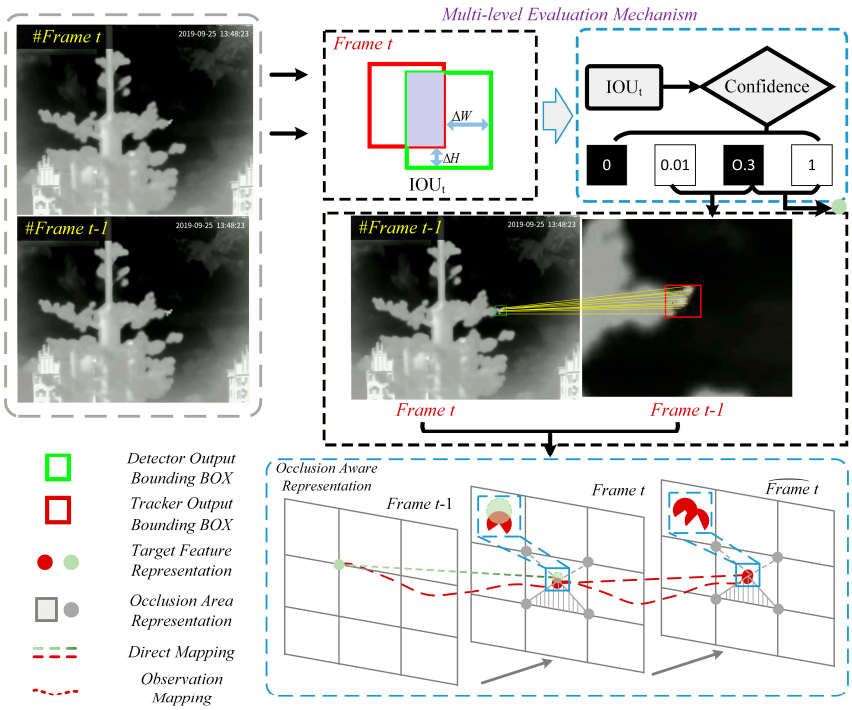


Figure 6. Motion association module architecture, with a detailed legend in the lower-left corner. In the occlusion awareness schematic, the incomplete red pattern represents the partial features of the target under occlusion. Through the direct mapping and observation mapping of the detection bounding box in Frame $t-1$, our method can infer the target position in Frame t and its position in the next frame.

3.4.1. Multi-Level Evaluation Mechanism

Since the Kalman filter is based on a linear prediction model, it struggles to handle the nonlinear motion changes in targets. Therefore, this study introduces an additional verification step by constructing a multi-level evaluation mechanism to compute state parameters. This mechanism infers the reasonable position of the target from the results of all candidate bounding boxes. The verification process effectively enhances the system's adaptability to complex motion patterns and ensures the accuracy of target tracking.

To be specific, the set of candidate boxes C for the t-th frame can be expressed as:

$$C_t = \{c_{t_1}, c_{t_2}, \ldots, c_{t_i}\} \tag{14}$$

where $c_{t_i} = (x_{t_i}, y_{t_i}, w_{t_i}, h_{t_i})$. Next, retrieve the high-confidence detection result box $c_{t-1} = (x_{t-1}, y_{t-1}, w_{t-1}, h_{t-1})$ from the $t-1$ frame in the position information database, and use the Kalman gain to calculate the predicted position $c'_{t-1} = (x'_{t-1}, y'_{t-1}, w'_{t-1}, h'_{t-1})$ the target in the t-th frame. Finally, compute the IoUk between each candidate box in sets c'_{t-1} and C_t to determine the association. Regarding the calculation of the state parameter dk in the verification stage, we adhere to the following:

$$\delta_k = \begin{cases} low & \mathrm{IoU}_k < \sigma_\mathrm{L} \\ medium & \sigma_\mathrm{L} \le \mathrm{IoU}_k < \sigma_\mathrm{M} \\ high & \text{otherwise} \end{cases} \tag{15}$$

In the equation, σ_L and σ_M are threshold coefficients, where σ_L is set to 0.01 and σ_M to 0.3 in this study. The state parameter δ_k represents the confidence condition of the predicted box c'_{t-1}. The "high" state indicates highly reliable prediction results that can be directly output; the "medium" state requires secondary verification; and the "low" state suggests unreliable prediction results that will be discarded. Therefore, the determination of σ_L is particularly critical. If threshold σ_L is set too high, it may lead to the omission of actual targets; conversely, if threshold σ_L is too low, the system will fail to respond effectively when no target is present in the field of view. Thus, we provide a detailed discussion on this in Section 4.5.

3.4.2. Verification Matching Mechanism

During UAV motion, the visual features of targets degrade due to interference from complex environmental backgrounds. Among these factors, occlusion has a particularly severe impact on target feature discernibility, resulting in targets being only partially visible in the image. This makes it difficult for conventional detection algorithms to accurately infer the integrity and authenticity of targets based on limited visual information from partial regions. To address this, we employ a normalized cross-correlation (NCC) [37] grayscale image matching method for secondary verification when state parameter δ_k is classified as "medium".

As shown in Figure 6, we separately extract the predicted box c'_{t-1} and detection box c_{t-1} for normalization processing. By defining the pixel grayscale value I, along with the extreme grayscale values I_max and I_min of the original image, and setting the target grayscale range I'_max and I'_min after normalization, the linearly normalized grayscale value In is calculated as follows:

$$I_\mathrm{n} = (I - I_\mathrm{min}) \frac{I'_\mathrm{max} - I'_\mathrm{min}}{I_\mathrm{max} - I_\mathrm{min}} + I'_\mathrm{min} \tag{16}$$

The detection box $T(x,y)$ is resized to match the dimensions of the prediction box $I(x,y)$ (both $m \times n$), after which the NCC value between the two images is calculated as follows:

$$NCC = \frac{\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\left[T(x,y)-\overline{T}\right]\left[I(x,y)-\overline{I}\right]}{\sqrt{\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\left[T(x,y)-\overline{T}\right]^2 \cdot \sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\left[I(x,y)-\overline{I}\right]^2}} \tag{17}$$

In the equation, $\overline{T}$ and $\overline{I}$ represent the mean pixel values of the local region centered at point (x,y) in the detection box image and prediction box image respectively, calculated as follows:

$$\overline{T} = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1} T(x,y) \tag{18}$$

$$\overline{I} = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1} I(x,y) \tag{19}$$

$NCC \in [-1, 1]$, it is used to measure the linear correlation between the grayscale distributions of two images, where a higher value indicates greater similarity between the images. This study evaluates the accuracy of the predicted bounding box by setting a matching threshold coefficient τ_{R}. When the NCC value exceeds the preset threshold τ_{R}, the predicted bounding box and the detection bounding box are determined to have high similarity. They can be output as valid tracking results. If the NCC value fails to meet the threshold, the target is considered absent from the field of view. In this case, the system suspends the Kalman prediction process until the detector recaptures a high-confidence target detection result, thereby initiating a new tracking cycle. Notably, our proposed method effectively addresses the issue of degraded feature discriminability caused by occlusion scenarios. As illustrated in Figure 6, we demonstrate how our method infers the integrity and authenticity of the target using partial information under occlusion conditions. The motion association algorithm based on multi-level evaluation proposed in this study is presented in Algorithm 1.

Algorithm 1: Motion Association Algorithm Based on Multi-level Evaluation.

Input: The t-th frame image ($\theta_d < 0.25$)
Output: Predicted bounding box in the t-th frame image
Obtain c_{t-1} and compute the Kalman gain to derive c'_{t-1};
Perform matching computation between c'_{t-1} and c_t to obtain IoU_k;
if $IoU_k \geq \sigma_{\mathrm{M}}$ then
  $c'_{t-1} = (x'_{t-1},\ y'_{t-1},\ w'_{t-1},\ h'_{t-1})$;
end
else if $\sigma_{\mathrm{L}} \leq IoU_k < \sigma_{\mathrm{M}}$ then
  Compute the NCC between c'_{t-1} and c_{t-1};
  if $NCC_t \geq \tau_R$ then $c'_{t-1} = (x'_{t-1}, y'_{t-1}, w'_{t-1}, h'_{t-1})$;
  else Output (0, 0, 0, 0);
end
else
  (0, 0, 0, 0)
end

4. Experiments

4.1. Dataset and Metrics

4.1.1. Dataset

The Anti-UAV [8] dataset comprises over 318 video sequences, each with a frame rate of 25 frames per second (FPS). It includes more than 580 k manually annotated bounding boxes, covering both visible-light and thermal-infrared modalities. The dataset presents diverse and complex scenarios, including clouds, mountains, and urban areas. For performance evaluation, this study focuses on validating the proposed method using the thermal-infrared modality from this dataset.

The Anti-UAV410 [2] dataset comprises a training subset (200 sequences), a validation subset (90 sequences), and a test subset (120 sequences), totaling 410 thermal-infrared video sequences with over 438,000 annotated frames. It features various complex backgrounds, including mountains, buildings, clouds, and water surfaces. Additionally, the dataset categorizes six challenging scenarios: target scale variation, fast motion, occlusion, out-of-view, dynamic background clutter, and thermal crossover.

The AntiUAV600 [5] dataset comprises 600 thermal-infrared video sequences, divided into a training subset (300 sequences), a validation subset (50 sequences), and a test subset (250 sequences). It includes diverse complex backgrounds such as mountains, buildings, clouds, and water surfaces, with all images having a resolution of 640×512 pixels. Compared to AntiUAV and AntiUAV410, AntiUAV600 contains a significant number of fast-motion sequences. It should be noted that the test set labels of AntiUAV600 are not publicly available; therefore, our experiments are conducted on the training and validation sets.

4.1.2. Metrics

To evaluate the performance of the proposed method, we employ three metrics: average Intersection over Union (IoU) [38], Accuracy (ACC) [5], and Success Rate (SR). The calculation of per-frame IoUt follows the procedure below:

$$\mathrm{IoU}_t = \begin{cases} 1, & A_t = \varnothing \cap G_t = \varnothing \\ 0, & A_t \neq \varnothing \oplus G_t \neq \varnothing \\ \frac{|A_t \cap G_t|}{|A_t \cup G_t|}. & A_t \neq \varnothing \cap G_t \neq \varnothing \end{cases} \tag{20}$$

In the equation, A_t represents the predicted bounding box at frame t, while G_t denotes the ground truth box at frame t. The model's perception capability is evaluated based on the existence conditions of A_t and G_t. Following the 3rd Anti-UAV Workshop & Challenge [39], this paper introduces a penalty term for calculating ACC in anti-UAV systems.

$$\mathrm{ACC} = \sum_{t=1}^{T} \mathrm{IoU}_t - \alpha \times \left(\sum_{t=1}^{T^*} \frac{p_t \times \delta(v_t > 0)}{T^*} \right)^{\beta} \tag{21}$$

In the equation, T and T^* represent the total number of frames and the number of target-present frames, respectively. P_t denotes the predicted visibility flag, where $P_t = 1$ when the prediction box is empty (indicating target absence), and zero otherwise. v_t is the target existence flag, with the indicator function $\delta(v_t > 0) = 1$ when $v_t > 0$, and 0 otherwise. Following reference [5], parameters a and b are set to 0.2 and 0.3, respectively. The ACC metric provides a comprehensive evaluation of tracking performance by integrating both IoU and the tracker's target existence perception capability. This metric not only measures tracking accuracy under normal conditions but also effectively reflects the model's performance when UAV targets are absent, making it particularly valuable for precisely quantifying a tracker's effectiveness and robustness in real-world scenarios.

For quantifying the success probability of UAV target perception, this paper introduces the Success Rate (SR), calculated as:

$$\mathrm{SR} = \frac{1}{T}\sum_{t=1}^{T} u_t(\mathrm{IoU}_t > \theta_s) \tag{22}$$

In the equation, u_t represents the prediction success flag, where $u_t = 1$ when $\mathrm{IoU}_t > \theta_s$ is satisfied, and zero otherwise. In this study, threshold θ_s is set to 0.5. This metric can objectively evaluate the target perception capability of the tracking system.

Additionally, we employ conventional evaluation metrics including average Precision (P), Recall (R), and F1-score. The Average Precision (AP) represents the proportion of correct samples among all results, while Recall is calculated as the ratio of correctly tracked targets to the ground truth total. Specifically:

$$P = \frac{TP}{TP + FP} \tag{23}$$

$$R = \frac{TP}{TP + FN} \tag{24}$$

In the equation, TP represents the number of true positives, FP denotes false positives, and FN indicates undetected targets. We set the IoU threshold at 0.5 as the criterion for distinguishing true positives from false positives. The F1-score, calculated as the harmonic mean of AP and R, is given by:

$$F1 = 2 \cdot \frac{AP \cdot R}{AP + R} \tag{25}$$

These evaluation metrics not only require the tracker to follow targets accurately, but also demand the timely determination of whether targets remain in the field of view, imposing strict requirements on target existence perception.

4.2. Implementation Details

Training. The proposed MDTC framework employs an improved YOLOv8s [40] as the detector. During model training, we set the weight decay to 0.0005 and the initial learning rate to 0.01, utilizing stochastic gradient descent (SGD) with a batch size of 32 for parameter optimization. Training logs indicate that the validation loss begins to increase while mAP50 starts decreasing after 30 epochs; therefore, we adopt the best weights obtained within these 30 epochs for inference. Our experimental system runs on Ubuntu 20.04 with an NVIDIA GeForce RTX 4090 GPU, and the detailed hardware configuration is presented in Table 1.

Table 1. Experimental Environment Configuration.

Environment	Configuration	Specifications
Software	Programming Environment	Python 3.8.20
	Deep Learning Framework	PyTorch 2.4.1 + Torchvision 0.19.1
	Operating System	Ubuntu 20.04 LTS
Hardware	CPU	Intel® Xeon® Platinum 8352V CPU @ 2.10 GHz
	GPU	NVIDIA GeForce RTX 4090
	GPU Memory	24,564 MiB

Inference. During the detection phase, we set the detector's confidence threshold θ_d to 0.25 to ensure high reliability of output bounding boxes. For the tracking verification stage, the association threshold coefficients σ_{L} and σ_{M} are set to 0.01 and 0.3, respectively,

while the matching threshold coefficient τ_R in the verification module is set to 0.7. These parameter settings are further analyzed in Section 4.5 through ablation studies.

4.3. Comparison with the State-of-the-Art Methods

To comprehensively evaluate the performance of our proposed MDTC framework, this study selects three representative categories of object tracking algorithms as baseline methods, including ATOM [3], DiMP [18], KYS [20], SiamCAR [24], ToMP [41], MixFormer [42], GlobalTrack [4], and EDTC [5]. These encompass three methodological paradigms: Siamese network-based trackers, Transformer architecture-based approaches, and detection-tracking integration methods.

It should be noted that all trackers except EDTC cannot detect target information in the first frame. To address the absence of initial frame bounding box annotations in some sequences of the Anti-UAV and AntiUAV600 datasets, we employ a pretrained YOLOv8s detector to detect first-frame targets uniformly. This ensures fair comparative experiments under identical initialization conditions for all trackers.

Anti-UAV Dataset. To validate the effectiveness of the proposed method, we conducted systematic experiments on the Anti-UAV dataset, utilizing its standard training and test sets for model training and evaluation, respectively. The results are presented in Table 2. MDTC achieves state-of-the-art performance on both ACC (0.626) and SR (0.875) metrics. Notably, it also attains a high IoU score of 0.672, ranking second with only a 0.004 margin from the top performer. In complex scenarios with weak target features, conventional trackers exhibit significantly degraded robustness, frequently suffering from tracking failures due to target loss. Without effective re-detection mechanisms, traditional approaches like ToMP, SiamCAR, DiMP, and MixFormer are prone to drift when confronted with background clutter or similar objects. The difference between ACC and IoU can effectively measure the accuracy of a tracker's judgment regarding the presence of a target. This is because the ACC metric incorporates a penalty mechanism, where misjudging or missing a target incurs significant penalties. Therefore, a minor difference indicates a lower likelihood of missed detections or false alarms, demonstrating the tracker's robustness. Notably, MDTC achieved the smallest difference of 0.046, outperforming all other methods. This result not only validates the reliability of MDTC in target tracking tasks but also proves its capability to accurately determine the presence state of targets within the field of view.

Table 2. Comparative experimental results of MDTC and other trackers on the Anti-UAV dataset (Best results in bold, second-best in red, and third-best in blue).

Dataset	Methods	IoU	ACC	SR
Anti-UAV	DiMP	0.521	0.433	0.662
	Super-DiMP	0.603	0.536	0.791
	SiamCAR	0.466	0.372	0.598
	ToMP	0.427	0.355	0.563
	MixFormer	0.572	0.497	0.737
	GlobalTrack	0.624	0.551	0.797
	EDTC	**0.676**	0.617	0.866
	MDTC	0.672	**0.626**	**0.875**

Anti-UAV410 Dataset. Table 3 presents the comparative experimental results between our method and other trackers on the Anti-UAV410 dataset. In the Anti-UAV410 test set, each tracking sequence provides initial bounding box annotations, enabling a comprehensive evaluation of performance differences between MDTC and other trackers under known initial target states. All trackers were tested using their original pre-trained weights and default parameter configurations. The experimental results demonstrate that MDTC

achieves outstanding overall performance, with IoU, ACC, and SR scores of 0.618, 0.544, and 0.802, respectively. These metrics are highly competitive with the top-performing method, GlobalTrack, with only a 0.002 difference in IoU. Regarding the difference between IoU and ACC, our method slightly underperforms GlobalTrack, which we attribute to MDTC's inability to detect the initial target in a few test sequences, resulting in increased penalty terms. Since the Anti-UAV410 test set contains a relatively low proportion of target disappearance and reappearance cases, MDTC significantly outperforms other trackers without relying on initial bounding box annotations. Notably, its ACC score surpasses that of the third-best tracker, Super-DiMP, by over 9%.

Table 3. Comparative experimental results of MDTC and other trackers on the Anti-UAV410 and AntiUAV600 datasets (Best results in bold, second-best in red, and third-best in blue).

Methods	Anti-UAV410			AntiUAV600			FPS
	IoU	ACC	SR	IoU	ACC	SR	
ATOM	0.482	0.388	0.624	0.359	0.236	0.459	110
DiMP	0.537	0.453	0.691	0.411	0.296	0.511	100
Super-DiMP	0.573	0.497	0.743	0.428	0.320	0.536	87
KYS	0.422	0.315	0.536	0.340	0.214	0.434	65
SiamCAR	0.439	0.342	0.567	0.299	0.162	0.381	125
ToMP	0.520	0.432	0.679	0.441	0.332	0.560	97
GlobalTrack	**0.620**	**0.554**	**0.808**	0.463	0.363	0.582	30
MDTC	0.618	0.544	0.802	**0.525**	**0.427**	**0.641**	94

Anti-UAV600 Dataset. Table 3 presents the comparative experimental results between our method and other trackers on the Anti-UAV600 validation set. Even trackers with global search and verification capabilities are prone to being misled by complex sequences. Our method achieves the highest scores, significantly outperforming all competing trackers. Specifically, MDTC achieves 0.525, 0.427, and 0.641 in IoU, ACC, and SR, respectively. Notably, its ACC score surpasses the second-best tracker, GlobalTrack, by over 15%, and outperforms other trackers by more than 25%. Additionally, our method achieves the lowest IoU-ACC difference (below 0.1) among all competitors. It is worth emphasizing that the Anti-UAV600 dataset contains numerous target disappearance-reappearance scenarios, where most trackers struggle with tracking failures. For instance, ATOM, KYS, and SiamCAR exhibit IoU-ACC differences exceeding 1.2. In contrast, MDTC demonstrates superior stability in tracking continuity and robustness, validating its strong adaptability to complex scenarios.

General evaluation metrics. To comprehensively evaluate the performance of the MDTC algorithm, we employed three metrics—P, R and F1-score to provide a more holistic assessment. As shown in Table 4, the results on the Anti-UAV410 dataset demonstrate that MDTC ranks within the top two across all metrics. Although its R (0.802) is slightly lower than that of GlobalTrack (0.825), it outperforms GlobalTrack by over 1.9% in both P and F1. On the Anti-UAV600 dataset, our method achieves the highest scores among all competitors, with P (0.711), R (0.647), and F1 (0.670). Overall, MDTC, equipped with an integrated detector, exhibits strong adaptability in target disappearance scenarios and significantly reduces tracking failure rates in complex environments, ensuring robust and stable long-term tracking performance.

Table 4. General evaluation metrics of MDTC and other trackers on the Anti-UAV410 and AntiUAV600 datasets (Best results in bold, second-best in red, and third-best in blue).

Methods	Anti-UAV410			AntiUAV600		
	P	R	F1	P	R	F1
ATOM	0.634	0.638	0.636	0.488	0.503	0.494
DiMP	0.700	0.706	0.703	0.543	0.560	0.550
Super-DiMP	0.752	0.759	0.755	0.571	0.588	0.578
KYS	0.544	0.548	0.545	0.460	0.474	0.466
SiamCAR	0.575	0.579	0.577	0.404	0.415	0.409
ToMP	0.688	0.693	0.690	0.595	0.612	0.602
GlobalTrack	0.814	**0.825**	0.819	0.615	0.635	0.623
MDTC	**0.892**	0.802	**0.835**	**0.711**	**0.647**	**0.670**

4.4. Qualitative Comparisons

To comprehensively evaluate the actual performance of the proposed method across various scenarios and complex backgrounds, as illustrated in Figures 7 and 8, we conducted comparative visualization experiments on two benchmark datasets: Anti-UAV410 and AntiUAV600.

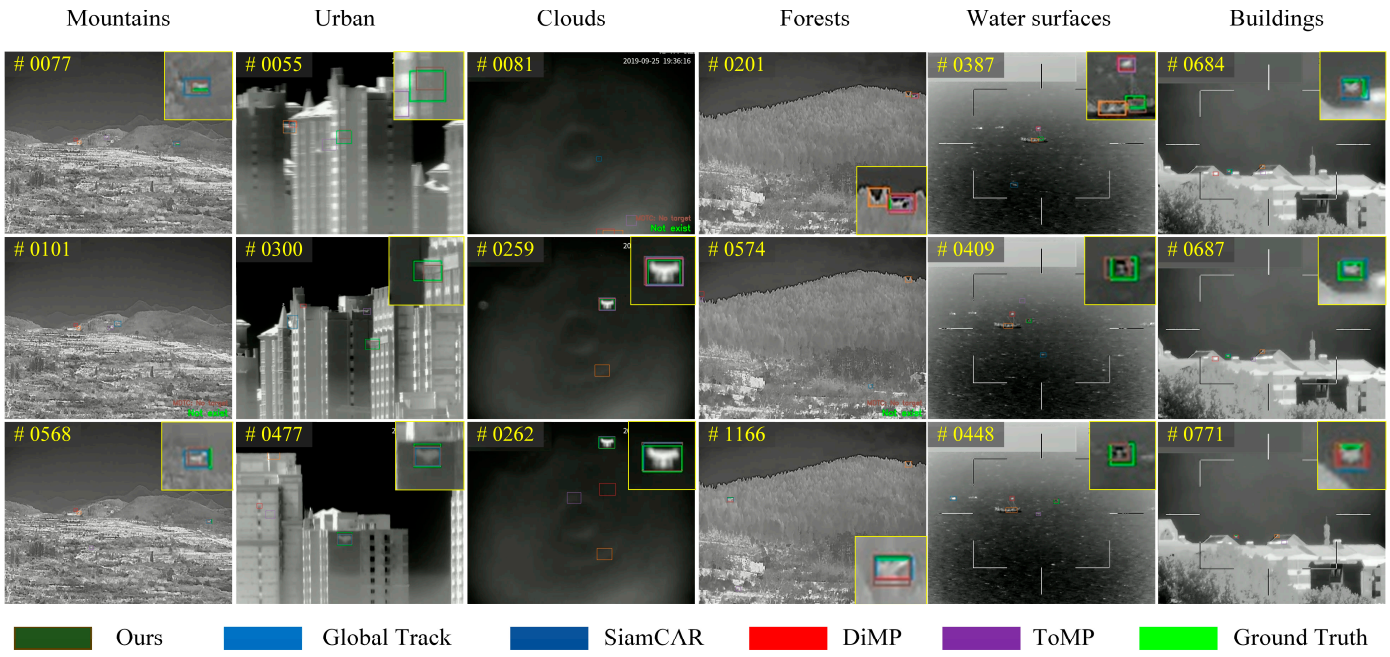


Figure 7. Visualized comparison results between the proposed method and other trackers on the Anti-UAV410 test set. Different colors represent different trackers, with the top-right and bottom-right corners showing enlarged views of the target region. Six background types were selected for comparison: mountains, urban, clouds, forests, water surfaces, and buildings.

Visualization on Anti-UAV410 Dataset. In scenarios with rich texture noise and complex background interference (e.g., mountains, buildings, and forests), it is crucial for trackers to accurately distinguish targets from distractors and effectively identify moments of target appearance and disappearance. As shown in Figure 7, the proposed method demonstrates robust capability in overcoming complex background interference, precisely locating target positions, and reliably determining target visibility states. Conventional tracking algorithms primarily rely on target appearance features, making them particularly susceptible to tracking failures when targets become blurred or degraded. This issue is especially prominent in urban scenarios. Due to the urban heat island effect, thermal radiation values between targets and background environments tend to converge, resulting

in highly similar temperature distributions in infrared imagery that severely degrade the performance of traditional trackers. In contrast, our proposed method significantly mitigates such interference effects and enhances tracking robustness. In relatively uniform backgrounds, such as clouds and water surfaces, conventional trackers typically maintain satisfactory performance. However, these environments present another challenge: abrupt spatial displacements caused by frame drops, which critically compromise the robustness of most tracking algorithms. To address this, our proposed detection-verification-tracking framework effectively alleviates tracking failures induced by abrupt target movements, substantially improving algorithmic stability.

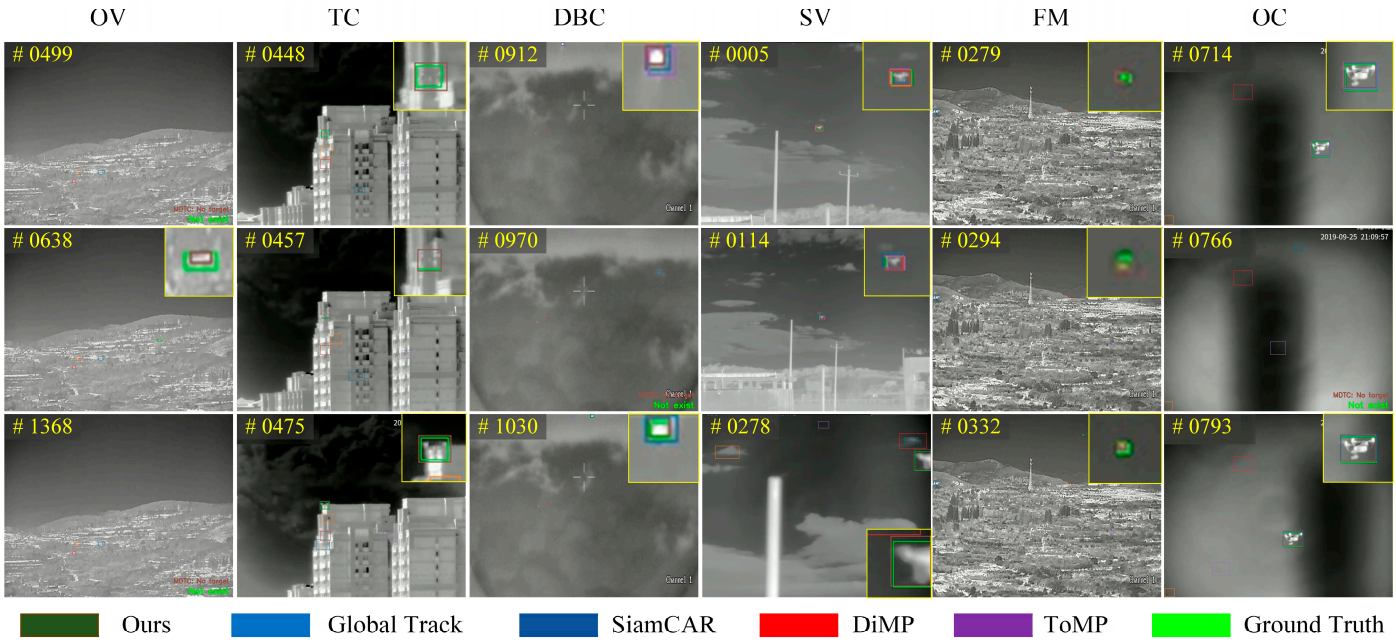


Figure 8. Visualized comparison results between the proposed method and other trackers on the AntiUAV600 validation set. Different colors represent different trackers, with the top-right and bottom-right corners displaying enlarged views of the target region. Six challenging scenarios were selected for evaluation: OV (Out-of-View), TC (Thermal Crossover), DBC (Dynamic Background Clutter), SV (Scale Variation), FM (Fast Motion), and OC (Occlusion).

Visualization on AntiUAV600 Dataset. Following the scene attribute categorization scheme of Anti-UAV410, we systematically classified the scenarios in the AntiUAV600 validation set, with visualization results presented in Figure 8. The AntiUAV600 validation set contains a substantial number of occlusion (OC) and out-of-view (OV) scenarios, characterized by targets temporarily or permanently disappearing from the field of view. When targets disappear, conventional methods often maintain tracking boxes in the image without a timely response to target absence, consequently increasing penalty terms in ACC evaluation. The proposed method addresses this limitation through a motion association module and verification matching mechanism, significantly enhancing the capability to discern target presence states and enabling timely adjustments based on target availability. In TC (Thermal Crossover) and FM (Fast Motion) scenarios, target features are simultaneously affected by thermal radiation differential blurring and motion blur effects, resulting in a degraded visual representation. Traditional vision-based tracking algorithms, particularly Siamese network architectures employing fixed initial template matching strategies, exhibit severely constrained performance. This limitation primarily stems from the nonlinear evolution of target features under thermal crossover effects and high-speed motion, while the static template matching mechanism fails to adaptively track such feature variations, resulting in tracking drift. Notably, the proposed tracking method

also demonstrates superior performance in challenging scenarios, including DBC (Dynamic Background Clutter) and SV (Scale Variation).

To further evaluate the actual performance of the proposed algorithm under different challenging scenarios, we conducted scenario-specific benchmark tests on both the Anti-UAV410 test set and the AntiUAV600 validation set. As shown in Figure 9, which presents comparative histograms of different metrics, we selected the top three currently best-performing baseline algorithms for comparison. The proposed method demonstrates significant performance advantages across multiple key challenging scenarios.

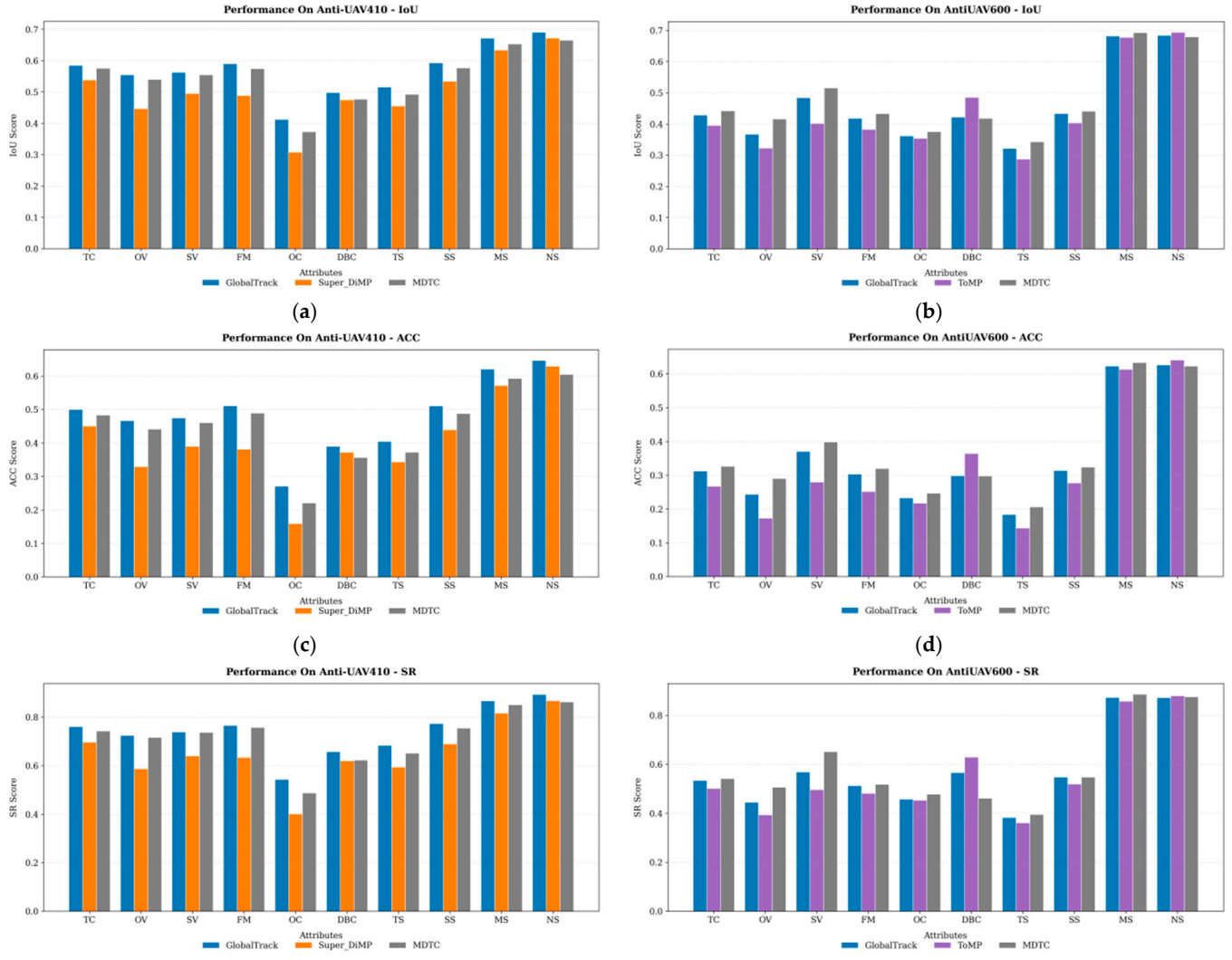


Figure 9. Multi-attribute performance comparison. Subfigures (**a**,**c**,**e**) present results on the Anti-UAV410 dataset, while subfigures (**b**,**d**,**f**) show results on the AntiUAV600 dataset. Attribute explanations: TC (Thermal Crossover), SV (Scale Variation), OC (Occlusion), OV (Out-of-View), DBC (Dynamic Background Clutter), FM (Fast Motion), TS (Tiny Size), SS (Small Size), MS (Medium Size), NS (Normal Size).

Analysis on Anti-UAV410 Dataset. As evidenced by the comparative experimental results in Table 3, MDTC exhibits slightly inferior overall performance to GlobalTrack in the comprehensive evaluation on the Anti-UAV410 test set. This discrepancy is visually represented in the histogram distributions of Figure 9a,c,e, where MDTC's bar heights consistently fall between those of GlobalTrack and Super-DiMP. Specifically, in NS scenarios, MDTC demonstrates optimal performance, with IoU, ACC, and SR metrics reaching

0.665, 0.605, and 0.863, respectively. Notably, it shows marginally lower ACC scores than GlobalTrack in TC and FM scenarios.

Analysis on AntiUAV600 Dataset. In the comprehensive evaluation on the AntiUAV600 validation set, MDTC displays significant advantages over both GlobalTrack and ToMP algorithms. As illustrated in Figure 9b,d,f, the proposed method maintains leading performance across all challenging scenarios. Notably, given AntiUAV600's richer TS sequences and smaller average target sizes, MDTC achieves particularly outstanding results, with its ACC metric surpassing the suboptimal GlobalTrack by over 12 percentage points. Under OC conditions, the method secures top rankings in all three metrics (IoU, ACC, and SR), which not only validates the effectiveness of the motion association module design but also highlights the unique advantages of the verification matching mechanism in addressing target occlusion challenges.

4.5. Quantitative Comparisons

In this section, we conduct a systematic evaluation of the key modules in the proposed MDTC framework using both qualitative and quantitative approaches. Additionally, we perform parameter sensitivity analysis to investigate the impact of different parameter configurations on model performance.

4.5.1. Component Effectiveness Analysis

To verify the impact of each component on MDTC's performance, we conducted ablation studies on both the Anti-UAV410 and Anti-UAV datasets. Table 5 presents the actual performance of the model when incorporating the Improved Detection Branch (IDB), the Motion Association Module (MAM) without a validation matching strategy, and the Validation Matching Strategy (VMS).

Table 5. Ablation study results (best results shown in bold).

Dataset	Baseline	IDB	MAM	VMS	IoU	ACC	SR
Anti-UAV410	✔				0.6057	0.5341	0.7886
	✔	✔			0.6079	0.5403	0.7954
	✔	✔	✔		0.6162	0.5412	**0.8018**
	✔	✔	✔	✔	**0.6177**	**0.5432**	**0.8018**
Anti-UAV	✔	✔			0.6646	0.6197	0.8678
	✔	✔	✔		0.6718	0.6263	**0.8751**
	✔	✔	✔	✔	**0.6719**	**0.6264**	0.8749

The Baseline configuration refers to the performance achieved without any additional modules, using only the original YOLOv8s as the detection branch with simple Kalman filter association. On the Anti-UAV410 dataset, this Baseline achieved an ACC score of 0.5341. When replacing the detector with our improved detector from Section 3.2 while maintaining the Baseline configuration, the model obtained scores of 0.6079 (IoU), 0.5403 (ACC), and 0.7954 (SR). Upon integrating the Motion Association Module (MAM) into this configuration, the model's ACC score improved to 0.5412. The complete proposed MDTC framework, incorporating all modules, achieved scores of 0.6177 (IoU), 0.5432 (ACC), and 0.8018 (SR), demonstrating an overall improvement exceeding 1.6% compared to the Baseline, thereby validating the effectiveness of our methodological enhancements. On the Anti-UAV dataset, our complete MDTC framework achieved scores of 0.6719 (IoU), 0.6264 (ACC), and 0.8749 (SR), showing consistent improvements across all metrics compared to the Baseline Improved Detection Branch (IDB) configuration.

4.5.2. Hyper-Parameters Analysis

To investigate the impact of different hyperparameters on experimental results, we first examined the values of θ_d (detection confidence threshold), as shown in Table 6. We conducted comparative experiments with various confidence thresholds on the Anti-UAV test set. Analysis revealed that an excessively low confidence threshold introduces substantial background noise, while an overly high threshold significantly degrades the system's detection performance for small targets. Through a comprehensive performance evaluation, we ultimately determined 0.25 as the optimal confidence threshold under which the system achieved peak performance on the Anti-UAV dataset.

Table 6. Impact of different θ_d values on MDTC performance (results from Anti-UAV dataset, best values shown in bold).

θ_d	0.2	0.25	0.3	0.35
IoU	0.6696	**0.6719**	0.6666	0.6703
ACC	0.6208	**0.6264**	0.6164	0.6226
SR	0.8683	**0.8749**	0.8655	0.8724

We subsequently conducted comparative experiments on the AntiUAV600 validation set to investigate the impact of threshold parameter σ_L on MDTC's performance. As shown in Table 7, the experimental results demonstrate that when σ_L is set too high, the motion association module incorrectly discards genuine targets. Conversely, when σ_L is too low, erroneous associations occur, introducing additional penalty terms that degrade the ACC metric. After a comprehensive evaluation, we selected 0.01 as the optimal value for parameter σ_L, under which MDTC achieves peak performance.

Table 7. Impact of different σ_L values on MDTC performance (Results from AntiUAV600 Dataset, Best Values Shown in Bold).

σ_L	0.005	0.01	0.015	0.02	0.025
IoU	0.5243	**0.5253**	0.5243	0.5243	0.5244
ACC	0.4265	**0.4273**	0.4265	0.4265	0.4266
SR	0.6402	**0.6411**	0.6402	0.6402	0.6402

Finally, we conducted comparative experiments on the matching threshold coefficient τ_R in the verification matching strategy using the Anti-UAV test set. As shown in Table 8 extensive ablation studies demonstrated that setting the matching threshold coefficient to 0.7 effectively suppresses false matching occurrences, thereby achieving optimal model performance. Consequently, we established 0.7 as the operational value for matching threshold coefficient τ_R.

Table 8. Impact of different τ_R values on MDTC performance (Results from Anti-UAV Dataset, Best Values in Bold).

τ_R	0.8	0.7	0.6	0.5	0.3
IoU	0.6696	**0.6719**	0.6696	0.6696	0.6695
ACC	0.6209	**0.6264**	0.6208	0.6208	0.6209
SR	0.8684	**0.8749**	0.8683	0.8683	0.8683

4.5.3. Performance Comparison of Different Detection Algorithms

As the core component of MDTC, the detector critically affects target perception capability. To evaluate the effectiveness of our proposed enhancements to YOLOv8s for

small-object detection, we conducted comparative experiments on the "tiny" and "small" subsets of the Anti-UAV410 dataset, using standard metrics: AP50 (average precision at IoU = 0.5) and AP50–95 (average precision averaged over IoU thresholds from 0.5 to 0.95). After 50 training epochs, the results in Table 9 show that our method consistently outperforms competing approaches. Specifically, it improves AP50 by 3.5% and AP50–95 by 3.7% over the baseline YOLOv8s, demonstrating its superior performance in detecting infrared low-altitude dim-and-small targets and providing MDTC with a robust, efficient detection module.

Table 9. Comparative experimental results of different detection algorithms.

Model	AP50 (%)	AP50–95 (%)
YOLOv8s	83.3	49.4
YOLOv10s	85.2	51.8
YOLOv12s	85.7	52.1
Faster R-CNN	79.3	48.2
FoveaBOX	76.0	42.6
Deformable DETR	86.5	47.4
Our	86.8	53.1

5. Discussion

The proposed method still exhibits certain limitations when addressing the challenge of dynamic background clutter. As shown in Figure 10, the first sequence comes from "02_6319_1500-2999" in the Anti-UAV410 test set. Observations reveal that at frames 93 and 131, intense cloud background interference significantly reduces the target's feature saliency, leading the tracker to misjudge the target's loss. The second sequence is from "3700000000002_133828_2" in the Anti-UAV410 test set. Analysis indicates that at frames 517 and 551, the highly similar texture features between the target and background substantially increase the difficulty of discrimination. To address such issues, future work should focus on achieving robust tracking under dynamic background clutter conditions.

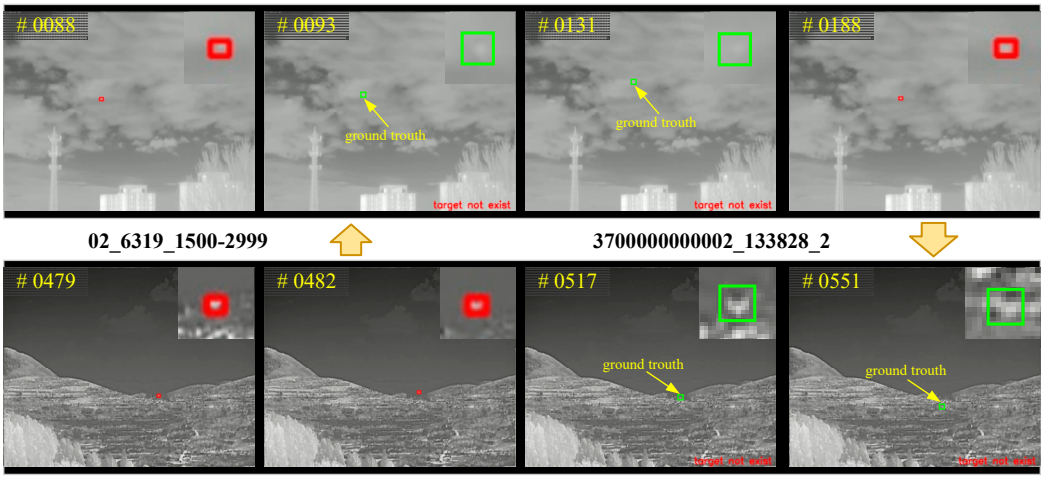


Figure 10. Example sequences demonstrating limitations of our method.

6. Conclusions

In this paper, we propose an integrated motion association target detection and tracking collaboration for anti-UAV tasks. First, to address the target presence perception problem, we design a motion association module that dynamically evaluates target existence through a multi-level confidence assessment mechanism, enabling rapid response

to target disappearance. Second, to solve the feature degradation caused by UAV scale variations in complex scenarios, we improve the detection branch. The enhanced network demonstrates improved perception capability for small and blurred targets, ensuring robust detection performance across different scales. Finally, for target perception under occlusion, we propose a grayscale-based verification matching mechanism. By matching and validating local features in the unconcluded regions of the target and incorporating historical trajectory priors, this mechanism infers target integrity and reliability, thereby maintaining stable tracking even under occlusion conditions. We conduct comprehensive evaluations on three public anti-UAV datasets: Anti-UAV, Anti-UAV410, and AntiUAV600. Experimental results demonstrate that the proposed algorithm achieves superior tracking accuracy and success rate. These findings fully validate the reliability and effectiveness of our method in anti-UAV application scenarios.

Author Contributions: Conceptualization, Y.C. and X.S.; methodology, R.G. and S.S.; validation, X.S. and Z.D.; formal analysis, D.B. and Z.D.; investigation, S.S. and Z.D.; resources, D.B. and Y.C.; data curation, R.G.; software implementation, Y.C.; writing—original draft preparation, X.S.; writing—review and editing, X.S. and R.G.; supervision, Y.C. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

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

References

1. Zhang, C.; Gao, Q.; Shi, R.; Yue, M. LDHD-Net: A Lightweight Network with Double Branch Head for Feature Enhancement of UAV Targets in Complex Scenes. *Int. J. Intell. Syst.* **2024**, *2024*, 7259029. [CrossRef]
2. Huang, B.; Li, J.; Chen, J.; Wang, G.; Zhao, J.; Xu, T. Anti-UAV410: A Thermal Infrared Benchmark and Customized Scheme for Tracking Drones in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2852–2865. [CrossRef]
3. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ATOM: Accurate Tracking by Overlap Maximization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4655–4664.
4. Huang, L.; Zhao, X.; Huang, K. GlobalTrack: A Simple and Strong Baseline for Long-Term Tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11037–11044. [CrossRef]
5. Zhu, X.-F.; Xu, T.; Zhao, J.; Liu, J.-W.; Wang, K.; Wang, G.; Li, J.; Wang, Q.; Jin, L.; Zhu, Z.; et al. Evidential Detection and Tracking Collaboration: New Problem, Benchmark and Algorithm for Robust Anti-UAV System. *arXiv* **2023**, arXiv:2306.15767. [CrossRef]
6. Liu, S.; Xu, T.; Zhu, X.-F.; Wu, X.-J.; Kittler, J. Learning Adaptive Detection and Tracking Collaborations with Augmented UAV Synthesis for Accurate Anti-UAV System. *Expert Syst. Appl.* **2025**, *282*, 127679. [CrossRef]
7. Tong, X.; Sun, X.; Zuo, Z.; Su, S.; Wu, P.; Wei, J.; Guo, R. GSFNet: Gyro-Aided Spatial-Frequency Network for Motion Deblurring of UAV Infrared Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5003718. [CrossRef]
8. Jiang, N.; Wang, K.; Peng, X.; Yu, X.; Wang, Q.; Xing, J.; Li, G.; Guo, G.; Ye, Q.; Jiao, J.; et al. Anti-UAV: A Large-Scale Benchmark for Vision-Based UAV Tracking. *IEEE Trans. Multimed.* **2023**, *25*, 486–500. [CrossRef]
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
10. Chen, Y.; Zhang, G.; Ma, Y.; Kang, J.U.; Kwan, C. Small Infrared Target Detection Based on Fast Adaptive Masking and Scaling with Iterative Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.

12. Wang, Y.; Song, J.; Wang, Y.; Wang, R.; Chen, H. Object Detection Method of Inland Vessel Based on Improved YOLO. *J. Mar. Sci. Eng.* **2025**, *13*, 697. [CrossRef]
13. Qin, Z.; Chen, D.; Wang, H. MCA-YOLOv7: An Improved UAV Target Detection Algorithm Based on YOLOv7. *IEEE Access* **2024**, *12*, 42642–42650. [CrossRef]
14. Ye, T.; Qin, W.; Li, Y.; Wang, S.; Zhang, J.; Zhao, Z. Dense and Small Object Detection in UAV-Vision Based on a Global-Local Feature Enhanced Network. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2515513. [CrossRef]
15. Wang, M.; Sheng, D.; Yuan, P.; Jin, W.; Li, L. Infrared Imaging Detection for Hazardous Gas Leakage Using Background Information and Improved YOLO Networks. *Remote Sens.* **2025**, *17*, 1030. [CrossRef]
16. Wu, W.; Zhang, J.; Zhou, G.; Zhang, Y.; Wang, J.; Hu, L. ESG-YOLO: A Method for Detecting Male Tassels and Assessing Density of Maize in the Field. *Agronomy* **2024**, *14*, 241. [CrossRef]
17. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In *Machine Learning and Knowledge Discovery in Databases*; Amini, M.-R., Canu, S., Fischer, A., Guns, T., Kralj Novak, P., Tsoumakas, G., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 13715, pp. 443–459. ISBN 978-3-031-26408-5.
18. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Learning Discriminative Model Prediction for Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6181–6190.
19. Danelljan, M.; Van Gool, L.; Timofte, R. Probabilistic Regression for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020; pp. 7181–7190.
20. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Know Your Surroundings: Exploiting Scene Information for Object Tracking. In *Computer Vision–ECCV 2020, Proceedings of the 6th European Conference, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12368, pp. 205–221.
21. Dong, X.; Shen, J. Triplet Loss in Siamese Network for Object Tracking. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11217, pp. 472–488, ISBN 978-3-030-01260-1.
22. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8971–8980.
23. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4277–4286.
24. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 14–19 June 2020; pp. 6268–6276.
25. Fang, H.; Wu, C.; Wang, X.; Zhou, F.; Chang, Y.; Yan, L. Online Infrared UAV Target Tracking with Enhanced Context-Awareness and Pixel-Wise Attention Modulation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5005417. [CrossRef]
26. Tao, J.; Chan, S.; Shi, Z.; Bai, C.; Chen, S. FocTrack: Focus Attention for Visual Tracking. *Pattern Recognit.* **2025**, *160*, 111128. [CrossRef]
27. Zhao, J.; Zhang, J.; Li, D.; Wang, D. Vision-Based Anti-UAV Detection and Tracking. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25323–25334. [CrossRef]
28. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
29. Sun, L.; Zhang, J.; Yang, Z.; Fan, B. A Motion-Aware Siamese Framework for Unmanned Aerial Vehicle Tracking. *Drones* **2023**, *7*, 153. [CrossRef]
30. Xu, Y.; Fu, Y. Adapting to Length Shift: FlexiLength Network for Trajectory Prediction. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16 June 2024; pp. 15226–15237.
31. Fang, H.; Wang, X.; Liao, Z.; Chang, Y.; Yan, L. A Real-Time Anti-Distractor Infrared UAV Tracker with Channel Feature Refinement Module. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Online, 11–17 October 2021; pp. 1240–1248.
32. Golding, B.; Pouchelon, G.; Bellone, C.; Murthy, S.; Di Nardo, A.A.; Govindan, S.; Ogawa, M.; Shimogori, T.; Lüscher, C.; Dayer, A.; et al. Retinal Input Directs the Recruitment of Inhibitory Interneurons into Thalamic Visual Circuits. *Neuron* **2014**, *81*, 1057–1069. [CrossRef] [PubMed]
33. Cai, X.; Lai, Q.; Wang, Y.; Wang, W.; Sun, Z.; Yao, Y. Poly Kernel Inception Network for Remote Sensing Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16 June 2024; pp. 27706–27716.

34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
35. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 3–19, ISBN 978-3-030-01233-5.
36. Khodarahmi, M.; Maihami, V. A Review on Kalman Filter Models. *Arch. Computat. Methods Eng.* **2023**, *30*, 727–747. [CrossRef]
37. Yoo, J.-C.; Han, T.H. Fast Normalized Cross-Correlation. *Circuits Syst. Signal Process.* **2009**, *28*, 819–843. [CrossRef]
38. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
39. Zhao, J.; Li, J.; Jin, L.; Chu, J.; Zhang, Z.; Wang, J.; Xia, J.; Wang, K.; Liu, Y.; Gulshad, S.; et al. The 3rd Anti-UAV Workshop & Challenge: Methods and Results. *arXiv* **2023**, arXiv:2305.07290. [CrossRef]
40. Yaseen, M. What Is YOLOv8: An in-Depth Exploration of the Internal Features of the next-Generation Object Detector. *arXiv* **2024**, arXiv:2409.07813.
41. Mayer, C.; Danelljan, M.; Bhat, G.; Paul, M.; Paudel, D.P.; Yu, F.; Van Gool, L. Transforming Model Prediction for Tracking. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 8721–8730.
42. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. MixFormer: End-to-End Tracking with Iterative Mixed Attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 13608–13618.

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