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.

DLST: Dual-Template Co-Evolution Learning for Robust Long-Term Drone Tracking in Dynamic Environments

Jiahao Zhang
Beijing University of Technology
Beijing, China
13008262771@163.com

Yixin Wei
Beijing University of Technology
Beijing, China
yixinwei@emails.bjut.edu.cn

Jinli Zhang*
Beijing University of Technology
Beijing, China
jlzhangcs@bjut.edu.cn

Zongli Jiang
Beijing University of Technology
Beijing, China
jiangzl@bjut.edu.cn

Peiwen Yu
Fudan University
Shanghai, China
important2002@163.com

Yufei Ma
Beijing University of Technology
Beijing, China
zhangjl_happy@126.com

Runan Jin
Beijing University of Technology
Beijing, China
54804282@qq.com

Abstract

To address the challenges of dynamic appearance variations, background interference, and long-term stability in drone small target tracking under complex environments, we propose a Dual Branch Learnable Siamese Tracker (DLST). The framework synergizes static and dynamically updated templates through three core modules: the Dual-Template Region Proposal Network (DT-RPN) for adaptive proposal generation via cross-correlation feature fusion, the Dual-Template R-CNN (DT-RCNN) for localization refinement guided by bidirectional template interactions, and the Dual-Template Background Suppression (DT-BS) module for distractor filtering through similarity scoring. By enabling bidirectional information flow between static and dynamic branches, DLST achieves robust adaptation to rapid appearance changes while maintaining computational efficiency. Extensive experiments on the Anti-UAV410 dataset demonstrate state-of-the-art performance, with a 68.81% State Accuracy (SA) surpassing existing methods. The framework exhibits exceptional robustness in scenarios involving dynamic backgrounds, fast motion, and occlusions, validated through comprehensive evaluations.

*Corresponding author

1. Introduction

With the rapid development of drone technology, their widespread applications in aerial photography, logistics, and security have brought societal convenience while also posing severe challenges to airspace safety and privacy protection. Drones may intrude into sensitive areas, disrupt communication networks, and even threaten aviation security, making real-time monitoring of drone locations and trajectories a critical requirement for airspace management. Infrared imaging technology, leveraging its all-weather capability, has become a key tool for drone tracking. However, civilian drones are typically small in size (often less than 20×20 pixels) and operate in complex, dynamic environments such as urban landscapes or forests, resulting in weak target signals, strong background interference, and frequent occlusions or deformations. Existing tracking algorithms struggle to address these challenges, particularly in complex scenarios where they are prone to distractions from similar objects and fail to achieve long-term stable tracking.

Current mainstream tracking methods primarily rely on Siamese network frameworks, which localize targets by matching features between templates and search regions. Algorithms like SiamRPN[14] and TransT[4] perform well in general object tracking but are not designed to account for the characteristics of tiny targets. These methods exhibit limited capability in extracting features from low-resolution targets and depend on fixed initial templates, making them

ill suited to adapt to appearance changes during long-term tracking. Additionally, traditional algorithms often assume temporal and spatial continuity in target motion, narrowing the search scope through localized detection and motion smoothing constraints. However, such assumptions frequently fail in real-world scenarios where drones exhibit rapid displacements or temporary disappearances, leading to tracking failures. Although recent studies have attempted to introduce dynamic template update mechanisms, e.g., the background suppression branch in SiamDT[19], the fundamental issue of error accumulation during updates remains unresolved, particularly when targets undergo drastic pose changes, where erroneous templates further degrade tracking accuracy.

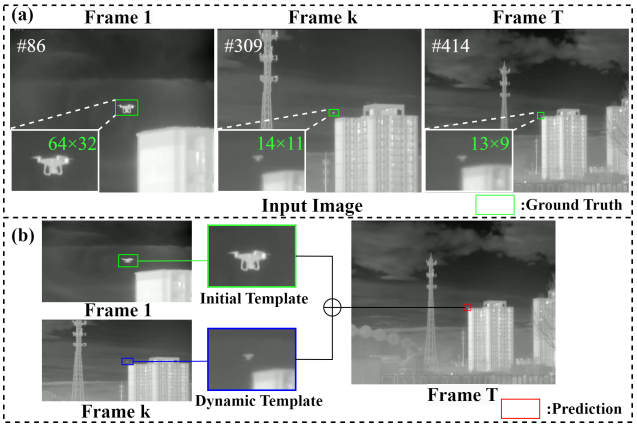


Figure 1. (a) Target size variation and background dynamics during long-term tracking. (b) Collaborative update process between Dynamic Template and Initial Template, demonstrating template adaptation capability to target posture changes.

To address the challenges of long-term tracking for small targets in complex backgrounds, this study proposes a novel adaptive tracking framework that achieves stable performance through multi-scale feature perception and dynamic template collaboration. The framework overcomes the limitations of conventional localized search strategies by employing a full-image multi-scale perception approach and temporal dynamic modeling to enhance deformation adaptability. Specifically, the spatiotemporal feature interaction iteratively optimizes historical trajectories and real-time observations to precisely model target appearance evolution. An adaptive confidence filtering mechanism dynamically adjusts detection thresholds based on quality assessment, activating global re-detection upon target disappearance to significantly reduce false alarms.Experimental results demonstrate significant performance improvements on mainstream benchmarks. For example, the proposed method achieves a state accuracy (SA) of 68.81% on the Anti-UAV410 dataset[11], outperforming the best existing algorithm by 0.62%. These findings highlight that the synergistic design of global search and dynamic updates substantially enhances tracking robustness in complex scenarios, offering new technical insights for drone surveillance.

In summary, our contributions are as follows.

- A dual-template fusion mechanism enabling joint static-dynamic feature modeling to enhance adaptability to target deformations and occlusions.
- A background-aware suppression module that combines historical and real-time background information for efficient interference filtering.
- A multi-stage training strategy to alleviate error propagation during template updates.

2. Related work

2.1. Target Tracking

Early Siamese tracking algorithms [1, 15] relied primarily on convolutional neural networks (CNNs) [9] with shared parameters and structures as their backbone to extract features from template and search regions. In order to integrate these features more deeply, algorithms often introduce lightweight relational modeling networks. However, this simple relational modeling network has limitations, which can not realize the bidirectional flow of information between template image and search area image.Benefiting from the Transformer [20] architecture, researchers [3, 24] started to investigate the utilization of stacked Transformer layers to replace conventional lightweight networks for capturing feature interdependencies more effectively.Among them,the TransT [3] algorithm achieves enhanced tracking performance by iteratively fusing features through meticulously designed stacks of self-attention and cross-attention layers.While these Transformer-based models perform well, this complex design also results in a drop in reasoning speed.

To overcome the speed challenges posed by complex structures, some advanced trackers [23, 26] have effectively enhanced inference speed and tracking performance by integrating feature extraction and relational modeling into a unified pipeline. The OSTrack algorithm [26] has made innovative strides in this area by merging feature extraction and relational modeling for the first time, allowing free information exchange between templates and search regions with lower computational costs. This mutually guiding approach can generate target-specific features, not only bolstering tracking performance but also elevating inference speed.

However, while conventional target tracking algorithms rely on template matching between search regions and targets, they inherently struggle with spectral-spatial interference in drone small target tracking scenarios. Due to the drones' compact size and susceptibility to motion blur or occlusions, their appearance often exhibits a high similar-

ity to background textures, leading to severe performance degradation in complex environments.

2.2. Drone Tracking

Anti-UAV tracking aims to monitor and track the activities of drones to address potential security and privacy threats posed by drones. Existing drone-based target tracking networks primarily utilize Siamese networks as the foundational architecture, leading to significant advances in the field of target tracking.

Jiang et al. [13] proposed the Dual-Flow Semantic Consistency (DFSC) drone tracker, which modulates the semantic flow of video sequences, enabling the tracker to learn more robust class-level semantic information and gain more discriminative instance-level features. Zhang et al. [28] proposed the SiamFusion tracker, which integrates internal target features, surrounding environmental information, and multimodal data, combining feature fusion and decision fusion to enhance target perception capabilities.Huang et al. [10] proposed a customized multi-branch tracking framework named SiamDT. By modeling the relationship between target regions and templates, as well as the probability of foreground object presence, they established a dual-semantic extraction mechanism and incorporated a background interference suppression branch, enhancing the tracker's ability to identify dynamic background interference. Despite these innovations, fundamental limitations persist in long-term tracking scenarios. First, the static template initialization mechanism fails to adapt to drastic viewpoint changes during UAV acrobatics. Second, error accumulation in dynamic template updates exacerbates tracking drift when targets temporarily disappear. Third, existing local search strategies [3] struggle to recover targets after rapid displacement that exceeds predefined regions.

3. Method

3.1. Overview

We introduce a novel adaptive dual-template tracking framework tailored for robust long-term tracking of tiny drones in complex environments. By synergizing initial template with dynamic template adaptation, our method addresses the limitations of conventional single-template trackers in handling temporal appearance variations and occlusions caused by moving backgrounds. As illustrated in Figure 2, the framework comprises three core modules: (1) Dual-Template RPN (DT-RPN) for adaptive proposal generation through cross-correlation feature fusion (2) Dual-Template R-CNN (DT-RCNN) for adaptive bounding box refinement with dynamic feature enhancement to handle deformations (3) Dual-Template Background Suppression (DT-BS) for suppressing background interference via similarity scoring based on patch matching with pre-learned templates. In the following sections, we will provide a detailed introduction to each module.

3.2. Dual-Template RPN

Tracking UAVs in complex environments presents unique challenges, including appearance variations and background clutter. Traditional single-template-based trackers, which rely only on the first frame's template, often fail to adapt to appearance changes, leading to drift or failure. To address this, we propose a Dual-Template Region Proposal Network (DT-RPN), which introduces a combination of a static template (first frame) and a dynamically updated template to enhance tracking robustness.

Given a template image z and a search image x, traditional RPN networks establish associations between the target in z and the search area x. Our Dual-Template approach introduces an additional dynamically updated template z_d that captures recent target appearance variations. Let the feature representations of the static and dynamic templates be $\phi_s(z)$ and $\phi_d(z_d)$, respectively, and the search region feature be $\phi(x)$. We define two semantic relationships: the static matching semantics capturing consistency between the first frame template and the target in the search image, formulated as:

$$\phi_{\text{static}}(x, z) = \text{Conv}_{\text{out}}\left(\text{Conv}_x(\phi(x)) * \text{Conv}_z(\phi_s(z))\right) + b \tag{1}$$

where $*$ denotes the depth-wise convolution operator used to compute cross-correlation between $\phi_s(z)$ and $\phi(x)$. Conv_z applies a $k \times k$ convolution layer to $\phi_s(z)$ to generate a learnable kernel, while Conv_x applies a 3×3 convolution with 1 pixel padding to $\phi(x)$. Conv_{out} converts the channel dimension back to c after cross-correlation.

The dynamic matching semantics capturing appearance variations using an updated template is defined as:

$$\phi_{\text{dynamic}}(x, z_d) = \text{Conv}_{\text{out}}\left(\text{Conv}_x(\phi(x)) * \text{Conv}_z(\phi_d(z_d))\right) + b \tag{2}$$

The final dual-template feature representation is obtained by concatenating both semantic features:

$$\Phi_{DT}(x, z, z_d) = \text{Concat}\left(\phi_{\text{static}}(x, z), \phi_{\text{dynamic}}(x, z_d)\right) \tag{3}$$

Based on the extracted features, the Dual-Template RPN (DT-RPN) generates proposals P_{DT} as:

$$P_{DT} = \text{RPNHead}\left(\Phi_{DT}(x, z, z_d)\right) \tag{4}$$

Finally, an ROI pooling layer is exploited to obtain aligned global RPN proposals:

$$P_{DT}^{\text{aligned}} = \text{RoIPooling}\left(P_{DT}\right) \tag{5}$$

Similar to prior works [8], we define the DT-RPN loss

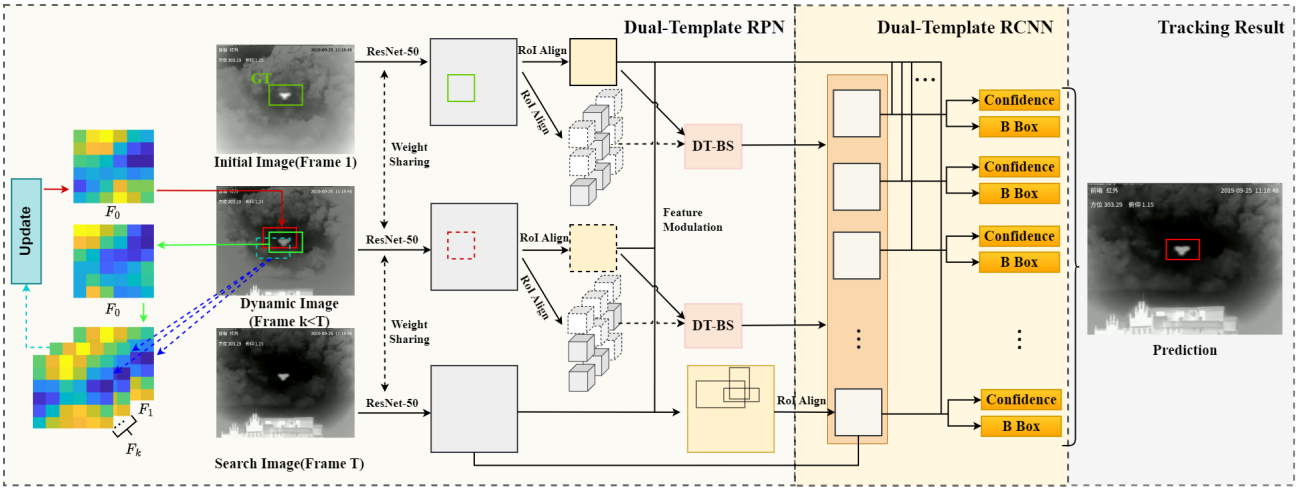


Figure 2. The overall architecture of DLST. This framework effectively addresses the limitations of traditional single-template trackers in handling target appearance variations and motion occlusions through a collaborative mechanism integrating a static reference template and a dynamically updated template. Its innovative architecture comprises three core modules: 1) The Dual-Template Region Proposal Network (DT-RPN) based on cross-correlation feature fusion, which generates adaptive candidate regions through dual semantic information integration; 2) The Dual-Template R-CNN (DT-RCNN) incorporating dynamic feature enhancement techniques, enabling robust localization refinement against target deformation; and 3) The Dual-Template Background Suppression (DT-BS) module leveraging pre-learned background templates, which filters background interference in real-time via patch-level similarity matching.

function as:

$$L_{DT-RPN} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(b_i, b_i^*) \tag{6}$$

where $p_i^* \in \{0, 1\}$ denotes the ground truth label for the i-th anchor, $p_i \in [0, 1]$ is the predicted foreground probability, b_i^* and b_i represent the ground truth and predicted bounding box coordinates, respectively, and λ balances classification and regression losses.

3.3. Dual-Template R-CNN

To further refine the proposals generated by DT-RPN, we introduce the Dual-Template R-CNN (DT-RCNN), which enhances target discrimination by leveraging both static and dynamic templates. Unlike conventional R-CNN modules that operate on a single-template feature, our DT-RCNN integrates two complementary template features to improve the classification and localization accuracy of UAV tracking.

Given a set of aligned proposals P_{DT}^{aligned} from DT-RPN, we extract region features and encode them with static and dynamic template information. Let the pooled region features be $\Psi(x) \in \mathbb{R}^{N \times k \times k \times c}$, where N is the number of proposals. The static-template encoded features are computed as follows:

$$\Psi_{\text{static}}(x, z) = \text{Conv}'_{\text{out}} \left(\text{Conv}'_x(\Psi(x)) \odot \text{Conv}'_z(\phi_s(z)) \right) + b \tag{7}$$

Here, $\odot$ denotes the element-wise Hadamard product that fuses template and region features. Conv'_x applies a 3×3 convolution to $\Psi(x)$, while Conv'_z generates a learnable kernel from $\phi_s(z)$. $\text{Conv}'_{\text{out}}$ converts the channel dimension back to c.

Similarly, the dynamic-template encoded features are obtained by:

$$\Psi_{\text{dynamic}}(x, z_d) = \text{Conv}'_{\text{out}} \left(\text{Conv}'_x(\Psi(x)) \odot \text{Conv}'_z(\phi_d(z_d)) \right) + b \tag{8}$$

We concatenate the extracted features to form the final representation:

$$\Psi_{DT}(x, z, z_d) = \text{Concat} \left(\Psi_{\text{static}}(x, z), \Psi_{\text{dynamic}}(x, z_d) \right) \tag{9}$$

The classification and regression heads of DT-RCNN then operate on this dual-template representation, yielding enhanced classification scores p_i and refined bounding box coordinates b_i. The loss function for DT-RCNN is defined as:

$$L_{DT-RCNN} = \frac{1}{N_{prop}} \sum_i \left(L_{cls}(p_i, p_i^*) + \lambda p_i^* L_{reg}(b_i, b_i^*) \right) \tag{10}$$

where N_{prop} is the number of proposals, and p_i^* indicates whether the i-th proposal contains the target. This formulation ensures that each proposal is independently evaluated using dual-template information, improving discriminative power in cluttered backgrounds.

By integrating static and dynamic template information, DT-RCNN effectively enhances the discriminative power of

the tracker, enabling more precise UAV localization even in challenging environments.

3.4. Dual-Template Background Suppression

In target tracking scenarios, background clutter often poses significant challenges to accurately identifying and following the target. To address this issue, we introduce two distinct background distractor sets: one derived from an initial template (D_0) and the other from a dynamic template (D_1). The initial template captures the background information at the start of the tracking process, providing a baseline representation of the static background elements. Meanwhile, the dynamic template adapts to changes in the background over time, such as the movement of non - target objects or variations in lighting conditions. By leveraging these two complementary background distractor sets, we can more effectively suppress background interference and improve the tracking performance.

$$\begin{cases} D_0 = \mathcal{F}_0(I_0, GT_0) \\ D_1^{(t)} = \mathcal{F}_1(I_t, PR_t) \end{cases} \tag{11}$$

where I_0 is the first frame, and I_t is the dynamic template frame. GT_0 represents the ground - truth bounding boxes in the initial frame I_0, and PR_t represents the predicted bounding boxes in the frame I_t.

Let P be the set of candidate proposals for the target. For a candidate proposal $p_i \in P$, we calculate its similarity to the two background distractor sets using a similarity function G. The similarity between p_i and D_0 is $G(D_0, p_i)$, and the similarity between p_i and $D_1^{(t)}$ is $G(D_1^{(t)}, p_i)$.

To combine the information from these two similarity scores, we use a weighted sum. Let α ($0 \leq \alpha \leq 1$) be a weighting factor. The final suppression score $G_{\text{final}}(p_i)$ for the candidate proposal p_i is given by:

$$G_{\text{final}}(p_i) = \alpha \cdot G(D_0, p_i) + (1 - \alpha) \cdot G(D_1^{(t)}, p_i) \tag{12}$$

This weighted combination allows us to balance the influence of the static background information from the initial template and the dynamic background information from the dynamic template, enabling more accurate background distractor suppression in various tracking scenarios.

3.5. Multi-stage training

To address the discrepancy between training with ground-truth labels and inference with predicted bounding boxes, we propose a multi-stage training strategy inspired by the work of Zhang et al. [27]. This approach iteratively refines the dynamic template generation process to better generalize to real-world tracking scenarios.

The training process is divided into K sequential stages, where each stage k refines the dynamic template generator $\mathcal{F}_1$ using the outputs from the previous stage. The overall framework is formulated as:

$$\begin{cases} \mathcal{F}_1^{(1)} = \text{Initial training with ground-truth} \\ \mathcal{F}_1^{(k)} = \text{Training with } \mathcal{F}_1^{(k-1)} \text{ predictions}, \quad \forall k > 1 \end{cases}$$

4. Experiments

We perform a quantitative comparison between our tracker and state-of-the-art trackers on the Anti-UAV410 dataset. Through detailed analysis of each tracker's performance under different attribute challenges, we systematically evaluate their capabilities and discuss their contributions to advancing anti-UAV tracking in real-world scenarios.

4.1. Experimental Settings

Implementation Details. The experiment was carried out on a platform based on the Ubuntu operating system, with the algorithm implemented on a NVIDIA GeForce RTX 4090. We use Faster-RCNN with ResNet-50 backbone as our base model. The channel number of the backbone features is c = 256. We employ ROI align to extract ROI features, where the output feature size is set to k = 7. We set the IoU thresholds for positive and negative samples at 0.7 and 0.3, respectively. The localization loss weight is set to $\lambda = 1$. We preprocessed the frame difference of the experimental data set.

Evaluation Metrics. The performance evaluation of trackers is conducted through One-Pass Evaluation (OPE) using three core metrics: precision plot, success plot, and state accuracy. Precision plot quantifies positional tracking reliability by measuring the percentage of frames where the predicted target coordinates reside within 20 pixels of ground truth annotations, with this threshold serving as the primary ranking criterion. Success plot evaluates bounding box alignment quality by calculating the proportion of frames where the Intersection-over-Union (IoU) between predicted and ground-truth boxes exceeds thresholds ranging from 0 to 1, with trackers ranked based on the Area Under the Curve (AUC) of these threshold-IoU relationships. State accuracy, as defined in Anti-UAV [13], extends beyond spatial precision by integrating target presence detection: it computes the average IoU across all sequences while requiring trackers to not only localize targets accurately in position and scale but also determine the target's presence state, including identifying when the target exits the field of view.

4.2. Experiments on Anti-UAV410

To verify the effectiveness of our method, we compared our method with the latest state-of-the-art tracking models on the large-scale dataset, Anti-UAV410.

According to Table 1, our proposed method demonstrates superior performance on the Anti-UAV410 dataset

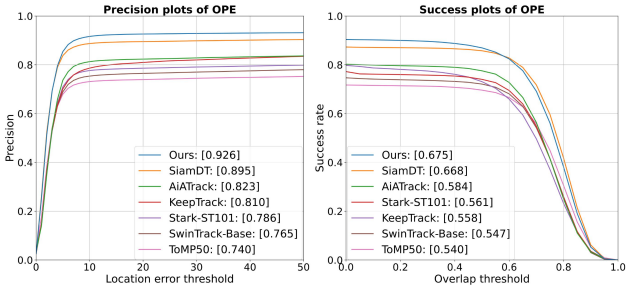


Figure 3. Precision and success plots on the test set of Anti-UAV410.

Methods	Source	SA
GlobalTrack[12]	AAAI20	66.45
PrDiMP50[6]	CVPR20	54.69
ROAM[25]	CVPR20	43.03
Siam R-CNN[21]	CVPR20	63.00
SiamBAN[5]	CVPR20	47.32
KYS[2]	ECCV20	44.90
KeepTrack[17]	ICCV21	56.80
AiATrack[7]	ECCV22	59.56
ToMP50[18]	CVPR22	55.09
ToMP101[18]	CVPR22	55.10
SwinTrack-Tiny[16]	NIPS22	53.15
SwinTrack-Base[16]	NIPS22	55.74
DropTrack[22]	CVPR23	60.15
SiamDT[10]	TPAMI24	68.19
DLST		**68.81**

Table 1. Comparison of State Accuracy for DLST and state-of-the-art trackers on the Anti-UAV410 dataset.

compared to existing state-of-the-art approaches. Specifically, it achieves an SA score of 68.81, surpassing the previous best result of 68.19 by 0.62%. The precision and success rate plots for the Anti-UAV410 dataset are shown in Figure 3. In the precision plot, our method's 0.926 surpasses others, highlighting superior target localization accuracy. For the success plot, our 0.675 success rate leads, proving better overlap consistency with the ground truth compared to competitors.

4.3. Ablation Studies

To validate the effectiveness of different components in our algorithm, we conducted extensive ablation experiments and detailed analyses. All experiments were evaluated on the Anti-UAV410 test set, with State Accuracy (SA) serving as the primary metric to assess tracking performance.

According to the results presented in Table 2. The experimental results demonstrate the incremental contributions of

DT-RPN	DT-RCNN	DT-BS	SA
-	-	-	66.45
✓	-	-	67.32
-	✓	-	67.12
✓	✓	-	67.84
-	-	✓	67.43
✓	✓	✓	68.81

Table 2. Ablation study results on the Anti-UAV410 Dataset

individual modules to the tracking performance. The baseline model achieves 66.45% SA without any proposed modules. Introducing the RPN with dynamic template matching (DT-RPN) improves SA to 67.32%, likely due to enhanced proposal generation by adaptively leveraging temporal context. Similarly, employing dynamic template-based RCNN (DT-RCNN) alone yields 67.12%, suggesting its effectiveness in refining tracking accuracy through template-guided feature alignment. Combining both DT-RPN and DT-RCNN further boosts performance to 67.84%, indicating their complementary roles in proposal generation and instance-level refinement. The background suppression module (DT-BS), which reduces false positives by computing similarity between proposals and a background set, independently elevates SA to 67.43%, validating its capability to mitigate background distractions. The full integration of all three modules achieves the highest SA of 68.81%, underscoring the synergistic effect of combining robust proposal generation (DT-RPN), precise detection (DT-RCNN), and background-aware filtering (DT-BS) to address challenges in UAV tracking scenarios. The progressive improvements highlight the necessity of jointly optimizing spatial-temporal modeling and discriminative feature learning for complex aerial targets.

Update(k)	SA
1	66.72
2	67.54
3	68.81
4	68.64
5	68.43

Table 3. SA Metrics under Multi-stage Training with Different update(k) Values

4.4. Multi-stage training Analysis

To identify the optimal number of training stages for dynamic template adaptation, we evaluated the proposed multi-stage training strategy across five iterations on the Anti-UAV410 dataset. The results demonstrate that increasing the number of stages initially improves tracking perfor-

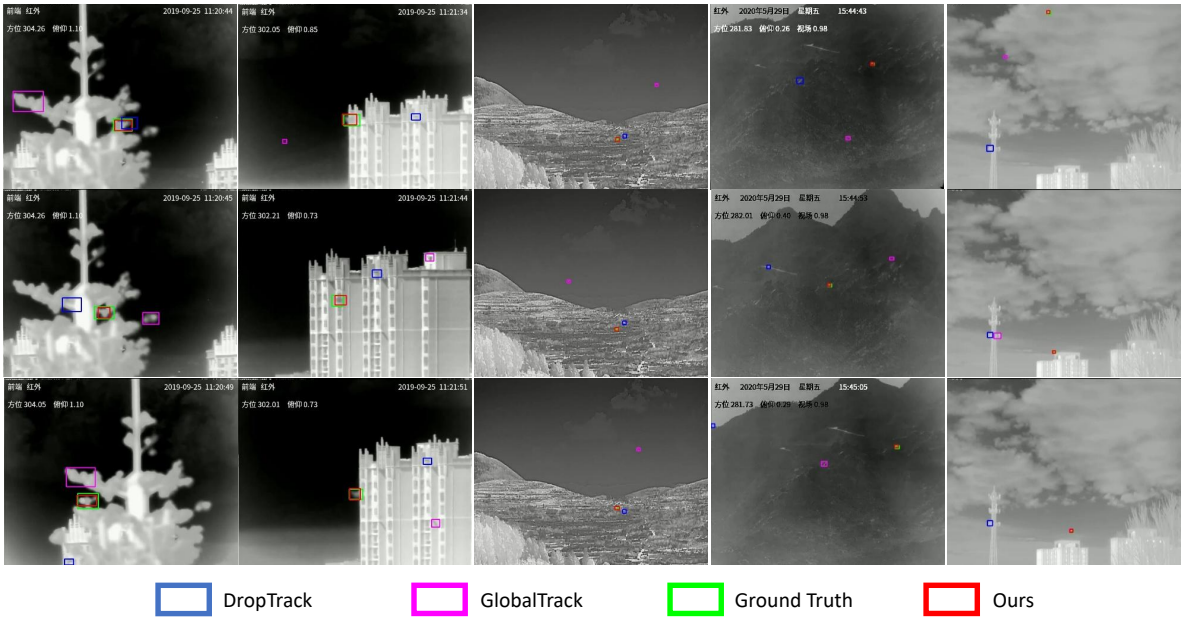


Figure 4. Qualitative comparison between our method and other trackers on the Anti-UAV410 dataset. Our approach demonstrates superior tracking performance for tiny drones across diverse complex backgrounds,

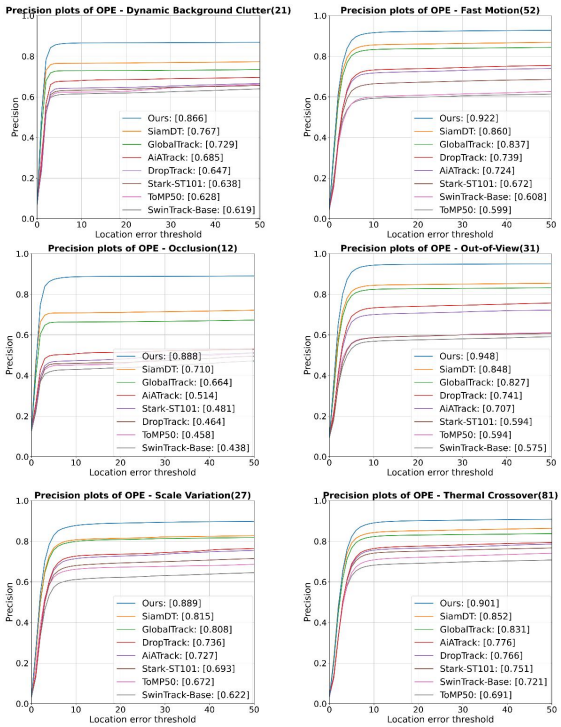


Figure 5. Precision plots for different attributes on the Anti-UAV410 dataset.

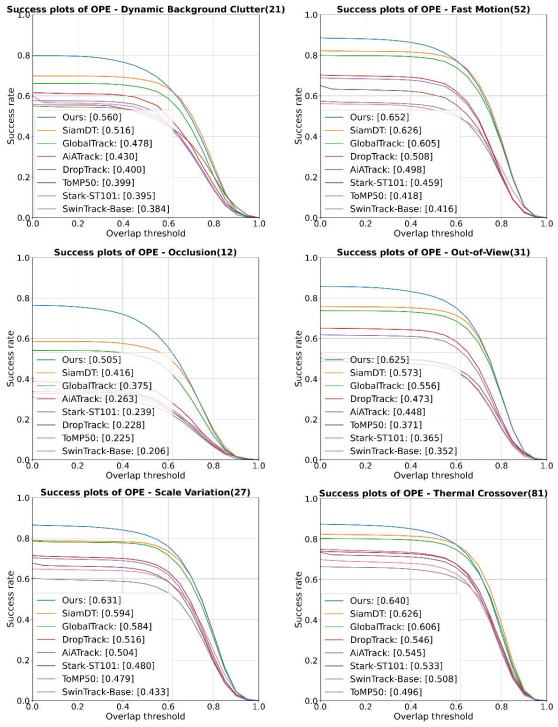


Figure 6. Success plots for different attributes on the Anti-UAV410 dataset.

mance, reaching a peak State Accuracy (SA) of 68.81% at stage 3. This improvement aligns with the iterative refinement mechanism described in the methodology, where each stage refines the dynamic template generator using predic-

tions from the previous stage. Beyond stage 3, the SA decreases slightly to 68.64% and 68.43% at stages 4 and 5, respectively, suggesting diminishing returns due to potential overfitting or cumulative errors in template updates. These findings validate the effectiveness of the multi-stage approach in balancing adaptation and stability, with stage 3 achieving the best trade-off between robustness and computational efficiency.

4.5. Attribution Evaluation

To validate the effectiveness of our method across diverse tracking challenges, we analyzed performance on attribute subsets defined in [10], using precision (fig.5) and success (fig.6) metrics. The proposed framework demonstrates superior robustness in dynamic background clutter, fast motion, and occlusion scenarios. These improvements stem from the DT-BS module's ability to suppress background interference via dual-template similarity scoring, while the DT-RPN and DT-RCNN modules dynamically adapt to rapid appearance changes. Notably, our method outperforms competitors in out-of-view recovery due to the global search mechanism integrated in the rchitecture, which maintains target re-detection capability even during temporary occlusions. The results highlight the framework's ability to balance discriminative feature learning and temporal consistency across complex environmental attributes.

5. Conclusion

This paper presents the Dual Branch Learnable Siamese Tracker (DLST), a novel framework that effectively addresses challenges such as abrupt appearance changes, background interference, and long-term instability in UAV tracking. By integrating static and dynamic branches, DLST leverages DT-RPN for adaptive proposal generation, DT-RCNN for precise localization refinement, and DT-BS for background-aware filtering. The dual-branch design facilitates bidirectional interaction between templates, enabling robust feature modeling under severe appearance variations. Extensive experiments on the Anti-UAV410 dataset demonstrate DLST's superiority, achieving a record SA of 68.81% and outperforming state-of-the-art trackers in dynamic clutter and occlusion recovery. Future research will focus on refining the adaptive mechanisms for dynamic template updates to improve robustness against extreme motion variations and prolonged occlusions. Additionally, we aim to explore advanced attention-based architectures that strengthen discriminative feature learning in highly cluttered environments, enabling more precise target-background separation. Further investigations will also extend the framework's applicability to multi-drone tracking scenarios while maintaining real-time performance.

References

[1] 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. 2

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European conference on computer vision*, pages 205–221. Springer, 2020. 6

[3] 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. 2, 3

[4] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking, 2021. 1

[5] 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. 6

[6] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020. 6

[7] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, pages 146–164. Springer, 2022. 6

[8] R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[10] 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*, 2023. 3, 6, 8

[11] 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, 2024. 2

[12] 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. 6

[13] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking. *arXiv preprint arXiv:2101.08466*, 2021. 3, 5

[14] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region pro-

posal network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 1

[15] 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. 2

[16] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances in Neural Information Processing Systems*, 35:16743–16754, 2022. 6

[17] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13444–13454, 2021. 6

[18] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8731–8740, 2022. 6

[19] Jiantong Shi and Yu Wang. Dual-template siamese network with attention feature fusion for object tracking. *Radioengineering*, 32, 2023. 2

[20] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[21] 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. 6

[22] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023. 6

[23] Zizheng Xun, Shangzhe Di, Yulu Gao, Zongheng Tang, Gang Wang, Si Liu, and Bo Li. Linker: Learning long short-term associations for robust visual tracking. *IEEE Transactions on Multimedia*, 2024. 2

[24] 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. 2

[25] Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B Chan. Roam: Recurrently optimizing tracking model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6718–6727, 2020. 6

[26] 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. 2

[27] Lichao Zhang, Abel Gonzalez-Garcia, Joost Van De Weijer, Martin Danelljan, and Fahad Shahbaz Khan. Learning the model update for siamese trackers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4010–4019, 2019. 5

[28] 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