TEMPORAL GEOMETRY-AWARE LOCAL SEARCH FOR THERMAL INFRARED ANTI-UAV TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Thermal infrared (TIR) UAV tracking is challenged by small targets, weak appearance cues, and large inter-frame displacement. For bounded local trackers that restrict the search to a crop around the previous target estimate, once the target leaves the search crop, appearance modeling alone cannot recover the missing visual evidence. Expanding the search region may introduce more background clutter, while ambiguous local cues make it difficult to decide where to search. We therefore view bounded recovery as two distinct decisions: where to acquire additional evidence and which hypothesis to trust. Based on this view, we propose TGTrack, which decouples search allocation from visual selection. A time-aware Temporal Geometry Branch models sparsely sampled trajectory states in global coordinates by encoding center and scale evolution, visual confidence, and temporal information, and predicts target geometry and localization reliability. When the predicted center leaves the normal search support, TGTrack allocates one bounded geometry-guided Recrop. We further introduce Gate–Hann, which modulates the within-Recrop spatial prior according to predicted reliability. Final cross-region selection remains appearance-driven using raw visual confidence. Extensive experiments show that TGTrack achieves state-of-the-art results among the compared local-search trackers, improves the controlled LoRAT-L (Lin et al., 2024) baseline by 2.86 AUC points on AntiUAV410 while running at 292.01 FPS, and by 4.03 points on AntiUAV under direct transfer without target-domain adaptation. Code will be released upon acceptance.

1 INTRODUCTION

In TIR anti-UAV tracking, UAV targets are typically small and provide limited appearance information, while thermal imagery further weakens discriminative visual cues. Scale variation, occlusion, and out-of-view events further increase localization difficulty (Jiang et al., 2023; Huang et al., 2024). Recent Transformer trackers have substantially improved target representation and matching through stronger template–search interaction (Chen et al., 2021; Yan et al., 2021; Cui et al., 2022; Ye et al., 2022). As a representative work, LoRAT further exploits large pretrained Vision Transformers for efficient and effective tracking (Lin et al., 2024). However, local search still imposes a spatial limitation: target-scaled search regions provide limited absolute coverage for small UAVs, making the tracker vulnerable to rapid target or camera motion.

Existing anti-UAV trackers mainly address this issue by extending spatial coverage through global search or adaptive search-region enlargement (Huang et al., 2024; Wang et al., 2025). However, expanding image-based search introduces more background clutter, making the responses of small TIR targets more easily overwhelmed, and increases computational cost, while weak or ambiguous local appearance cues can make it difficult to determine where additional search should be allocated. We therefore exploit trajectory geometry as a complementary cue to guide additional search.

However, conventional trajectory-based motion priors, such as constant-velocity or fixed-dynamics models, can be limited when target motion becomes nonlinear or highly variable (Kim et al., 2024). Rapid maneuvers, camera motion, scale variation, and occlusion can induce highly variable target trajectories, while unreliable visual responses can make reliable observations sparse and temporally irregular. Accordingly, we learn a geometric prior from sparse global-coordinate histories encoding center and scale evolution, visual confidence, and temporal information, to account for complex and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

irregular target motion. This joint encoding enables trajectory-guided additional search with reduced reliance on ambiguous local visual responses.

Based on this idea, we propose TGTrack, which decouples search allocation from visual selection. The proposed Temporal Geometry Branch predicts target geometry and localization reliability from sparse trajectory states. When the predicted center falls outside the normal search support, it places one bounded geometry-guided Recrop to obtain an additional visual observation. Within the Recrop, the proposed Gate–Hann module adaptively adjusts the strength of the Hann spatial prior according to the predicted localization reliability. The normal and Recrop hypotheses, i.e., the candidate boxes selected from the two search regions, are compared using raw visual scores, leaving temporal geometry to guide search allocation rather than directly determine the final prediction.

The main contributions are summarized as follows:

- We cast recovery in bounded local tracking as a decoupled search-allocation and visual-selection problem, where temporal geometry determines where additional evidence is acquired and appearance confidence determines which regional hypothesis is retained.
- We develop TGTrack with a time-aware trajectory model that predicts target geometry and localization reliability from sparsely sampled global-coordinate states. The prediction conditionally places one bounded Recrop, while Gate–Hann provides reliability-aware within-region regularization.
- TGTrack improves the controlled LoRAT-L baseline by 2.86 AUC points on AntiUAV410 while running at 292.01 FPS, and by 4.03 points on AntiUAV under direct transfer without target-domain adaptation.

2 Related Work

2.1 Thermal Infrared Anti-UAV Tracking

TIR anti-UAV tracking is challenged by small targets, weak appearance, occlusion, and out-of-view events, as reflected in benchmarks such as Anti-UAV, AntiUAV410, DUT Anti-UAV, and CST Anti-UAV (Jiang et al., 2023; Huang et al., 2024; Zhao et al., 2022; Xie et al., 2025). Existing methods improve motion coverage through full-frame search, local-to-global recovery, or adaptive search-region expansion (Huang et al., 2024; 2021; Wang et al., 2025; Yuan et al., 2026). TGTrack instead uses temporal geometry to allocate one additional bounded search region when needed, while leaving final cross-region selection to visual confidence.

2.2 Transformer-Based Visual Tracking

Transformer-based trackers have advanced target representation and template–search interaction through feature fusion, mixed attention, and unified one-stream architectures (Chen et al., 2021; Yan et al., 2021; Cui et al., 2022; Ye et al., 2022). More recently, LoRAT(Lin et al., 2024) adapts large pretrained Vision Transformers to tracking through low-rank adaptation and serves as the visual foundation of TGTrack. Several recent methods further exploit historical observations, including historical prompts in HIPTrack, gated visual memory in MemLoTrack, and temporal prompts in TAPTrack (Cai et al., 2024; Park & Han, 2025; Cui et al., 2026). These approaches primarily use temporal information to enhance appearance representation or template–search interaction, whereas TGTrack models low-dimensional target states in global image coordinates and uses their temporal evolution to guide subsequent search allocation.

2.3 Motion-Prior and Trajectory-Guided Tracking

Motion priors provide temporal cues when appearance-based localization becomes unreliable. Classical methods such as SORT and OC-SORT use explicit kinematic models for state prediction and temporal consistency (Bewley et al., 2016; Cao et al., 2023), while recent approaches combine motion priors with learned representations through motion-aware attention, state-space dynamics, or kinematic constraints (Lv et al., 2026; Fu et al., 2025; Singh et al., 2025). In contrast, TGTrack does not use temporal geometry to directly update the target state. Instead, its auxiliary branch predicts

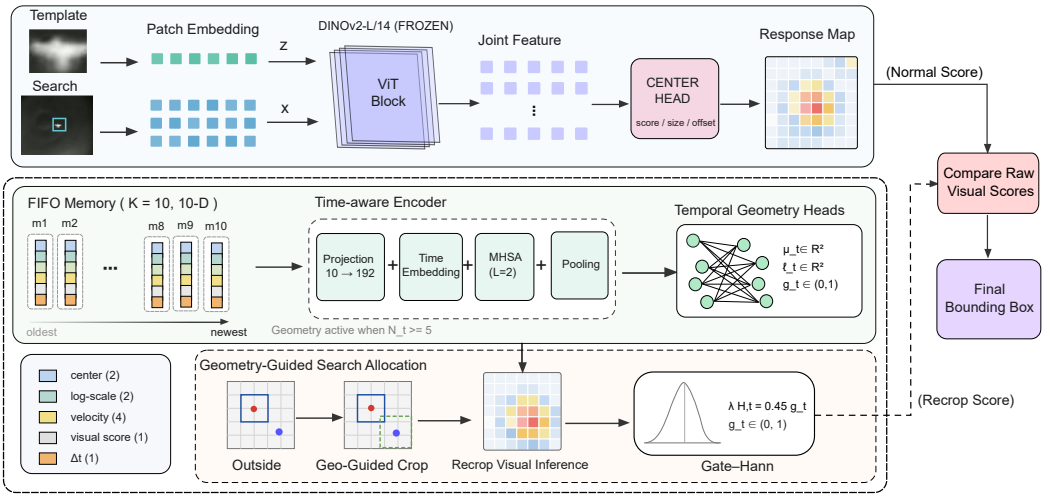


Figure 1: Overview of TGTrack. TGTrack decouples search allocation from visual selection: the Temporal Geometry Branch predicts where additional evidence should be acquired through a conditional Recrop, while the final regional hypothesis is selected by raw visual confidence. Gate–Hann modulates the spatial prior within the Recrop according to predicted localization reliability.

where an additional visual observation should be acquired and allocates a single bounded Recrop when the prediction leaves the normal search support. Temporal geometry therefore serves as a search cue rather than a direct state update, while final regional selection remains appearance-driven.

3 METHOD

Given an initial template z, LoRAT-L(Lin et al., 2024) localizes the target from a normal search crop x_t^{n}, whose support in normalized full-image coordinates is denoted by Ω_t^{n}. Large inter-frame displacement or accumulated localization error may move the target beyond this bounded region.

As illustrated in Fig. 1, TGTrack augments LoRAT-L with a Temporal Geometry Branch that predicts the target center $\boldsymbol{\mu}_t$, log-scale $\boldsymbol{\ell}_t$, and localization reliability g_t from recently admitted trajectory states. When the branch is inactive or $\boldsymbol{\mu}_t \in \Omega_t^{\mathrm{n}}$, TGTrack uses the normal-search hypothesis. Otherwise, a valid geometry prediction triggers one Recrop x_t^{r} centered at $\boldsymbol{\mu}_t$ with extent determined by $\boldsymbol{\ell}_t$.

The normal search uses a fixed Hann prior, whereas the Recrop uses g_t-controlled Gate–Hann modulation. Each region is localized from its Hann-modulated response, while cross-region selection compares the raw visual confidences. The selected raw confidence is used for Memory admission.

3.1 TEMPORAL GEOMETRY BRANCH

The detailed architecture of the Temporal Geometry Branch is illustrated in Appendix Fig. 3.

3.1.1 TRAJECTORY MEMORY AND TIME-AWARE ENCODING

As illustrated in the light-green block of Fig. 1, the Temporal Geometry Branch maintains a FIFO trajectory Memory and encodes its temporal, motion, and confidence information to predict the current target geometry and localization reliability. Before frame t, the Memory contains at most $K = 10$ visually admitted observations,

$$\mathcal{M}_t = \{\mathbf{m}_i\}_{i=1}^{N_t}, \qquad \mathbf{m}_i = [\mathbf{p}_i, \boldsymbol{\ell}_i, \mathbf{v}_i^p, \mathbf{v}_i^\ell, s_i, \delta_i] \in \mathbb{R}^{10}, \tag{1}$$

where $\mathbf{p}_i = (c_{x,i}/W_i, c_{y,i}/H_i)$ and $\boldsymbol{\ell}_i = (\log(w_i/W_i), \log(h_i/H_i))$ denote the normalized full-image center and log-scale. The finite-difference terms $\mathbf{v}_i^p$ and $\mathbf{v}_i^\ell$ describe center and scale variation

over the inter-observation gap δ_i, while s_i is the raw visual confidence. Using global image coordinates decouples the trajectory representation from individual search crops.

The first-frame ground truth initializes the Memory. During tracking, an observation is admitted only if its confidence satisfies $s_i > \tau_m = 0.4$ and its bounding box is geometrically valid. Because rejected observations create irregular temporal gaps, TGTrack models the age of each Memory state:

$$a_{t,i} = \Delta_t + \sum_{j=i+1}^{N_t} \delta_j, \qquad \bar{a}_{t,i} = \mathrm{clip}\left(\frac{a_{t,i}}{A_{\max}}, 0, 1\right), \tag{2}$$

where Δ_t is the prediction horizon from the latest admitted state to frame t, and $A_{\max} = 50$ is the age-normalization horizon.

Each 10-D state is projected to a 192-D token and augmented with learned FIFO-slot and age embeddings. Its contribution is weighted by visual confidence as

$$\omega_i = \omega_{\min} + (1 - \omega_{\min})\,\mathrm{clip}(s_i, 0, 1), \tag{3}$$

We set $\omega_{\min} = 0.1$, and the projected token is scaled by ω_i. The weighted tokens are processed by a two-layer, four-head Transformer with a learned relative-age bias derived from $|a_{t,i} - a_{t,j}|$, while invalid Memory slots are masked.

3.1.2 Trajectory Pooling and Horizon Fusion

A learned query aggregates the valid encoded Memory states using four-head attention. Let $d_h = 192/4 = 48$ denote the feature dimension of each attention head. For head h, the attention logit for Memory state i is

$$A_{h,i} = \frac{\mathbf{q}_h \mathbf{k}_{h,i}^\top}{\sqrt{d_h}} + b_{h,i}^{\mathrm{age}} + b_{h,i}^{\mathrm{score}}, \qquad b_{h,i}^{\mathrm{score}} = \gamma_h \log(\omega_i + \epsilon), \tag{4}$$

where $\epsilon = 10^{-6}$ is used for numerical stability, γ_h is learnable and initialized to 0.5, and $b_{h,i}^{\mathrm{age}}$ denotes the relative-age bias. Raw visual confidence therefore influences temporal aggregation through both token scaling in Eq. 3 and the attention-score bias. The attended values from all heads are concatenated and projected to obtain the history representation $\mathbf{h}_t \in \mathbb{R}^{192}$.

The prediction horizon is embedded separately and fused with the history:

$$\tilde{\mathbf{h}}_t = F_{\mathrm{hor}}\left([\mathbf{h}_t; F_\Delta(\log(1 + \Delta_t))]\right), \qquad \tilde{\mathbf{h}}_t \in \mathbb{R}^{192}. \tag{5}$$

3.1.3 Horizon-Adaptive Geometric Prediction

TGTrack combines the latest center velocity with learned temporal correction. Let N_t index the latest admitted state. The horizon-dependent displacement is $\mathbf{d}_t^p = \mathbf{v}_{N_t}^p \Delta_t$, and an adaptive motion gain

$$\alpha_t = 1 + \tanh\left(f_\alpha(\tilde{\mathbf{h}}_t)\right) \tag{6}$$

controls its contribution. The predicted center is

$$\boldsymbol{\mu}_t = \mathbf{p}_{N_t} + \alpha_t \mathbf{d}_t^p + \tanh(f_p(\tilde{\mathbf{h}}_t)), \tag{7}$$

which combines a kinematic motion reference with a bounded learned residual.

Scale is predicted relative to the latest normalized log-scale:

$$\boldsymbol{\ell}_t = \boldsymbol{\ell}_{N_t} + \log(\gamma_s)\tanh(f_\ell(\tilde{\mathbf{h}}_t)), \qquad \gamma_s = 4, \tag{8}$$

where γ_s bounds the width/height ratio relative to the latest admitted state to $[1/\gamma_s, \gamma_s]$. The stored scale velocity remains available to the temporal encoder but is not explicitly extrapolated.

Finally, a reliability head operates on a stop-gradient copy of the temporal representation:

$$g_t = \mathrm{sigmoid}\left(f_g(\mathrm{sg}(\tilde{\mathbf{h}}_t))\right). \tag{9}$$

The Recrop-center reliability g_t controls Gate–Hann during inference.

3.2 Decoupled Conditional Recovery

3.2.1 Geometry-Guided Search Allocation

TGTrack first performs the standard LoRAT-L local search. Let $S_t^{\mathrm{n}}(\mathbf{u})$ denote the raw visual confidence at response location $\mathbf{u}$. A fixed Hann prior is used for within-region ranking:

$$R_t^{\mathrm{n}}(\mathbf{u}) = (1 - \lambda_h) S_t^{\mathrm{n}}(\mathbf{u}) + \lambda_h H(\mathbf{u}). \tag{10}$$

We set $\lambda_h = 0.45$, following the default inference setting of the controlled LoRAT-L baseline.

If the geometry-predicted center $\boldsymbol{\mu}_t \in \Omega_t^{\mathrm{n}}$, where Ω_t^{n} denotes the full-image support of the normal crop, TGTrack directly uses the normal-search hypothesis. Otherwise, if the predicted geometry is valid, one additional local Recrop is centered at $\boldsymbol{\mu}_t$, with its spatial extent determined by the predicted scale $\boldsymbol{\ell}_t$. The Recrop follows the same search geometry as the normal LoRAT-L search.

Let $S_t^{\mathrm{r}}(\mathbf{u})$ denote the raw visual confidence within the Recrop. Because the Recrop is centered at the geometry prediction, TGTrack modulates its Hann prior according to the predicted Recrop-center localization reliability g_t:

$$R_t^{\mathrm{r}}(\mathbf{u}) = (1 - g_t \lambda_h) S_t^{\mathrm{r}}(\mathbf{u}) + g_t \lambda_h H(\mathbf{u}). \tag{11}$$

Since $g_t \in [0, 1]$, the effective Hann weight $g_t \lambda_h$ is bounded between zero and the base Hann weight λ_h. Thus, the Recrop response interpolates between the raw visual response at low reliability and the standard Hann-modulated response at high reliability. Gate–Hann is applied only to the Recrop; the normal search retains the fixed Hann prior in Eq. 10.

This modulation reflects the role of g_t as Recrop-center reliability. When the geometric prediction is unreliable, reducing the Hann weight avoids overly biasing localization toward a potentially inaccurate Recrop center. As reliability increases, the Hann weight progressively approaches the standard LoRAT-L setting, providing stronger spatial regularization around the Recrop center.

3.2.2 Appearance-Based Cross-Region Selection and Memory Update

For each region $q \in \{\mathrm{n}, \mathrm{r}\}$, the candidate location is first selected using its Hann-modulated response:

$$\hat{\mathbf{u}}_t^q = \arg\max_{\mathbf{u}} R_t^q(\mathbf{u}), \qquad \hat{s}_t^q = S_t^q(\hat{\mathbf{u}}_t^q), \tag{12}$$

where $\hat{s}_t^q$ is the raw visual confidence evaluated at the Hann-selected location.

When a Recrop is evaluated, TGTrack compares the two candidates using their raw visual confidences:

$$(\hat{\mathbf{b}}_t, \hat{s}_t) = \begin{cases} (\hat{\mathbf{b}}_t^{\mathrm{r}}, \hat{s}_t^{\mathrm{r}}), & \hat{s}_t^{\mathrm{r}} > \hat{s}_t^{\mathrm{n}}, \\ (\hat{\mathbf{b}}_t^{\mathrm{n}}, \hat{s}_t^{\mathrm{n}}), & \text{otherwise.} \end{cases} \tag{13}$$

If no Recrop is evaluated, the normal candidate is used directly. This keeps cross-region selection appearance-driven: temporal geometry determines where additional evidence is acquired, while the raw visual score determines which regional hypothesis is retained.

The selected box $\hat{\mathbf{b}}_t$ updates the tracking state, while Memory admits only geometrically valid observations with $\hat{s}_t > \tau_m = 0.4$. Raw confidence is used only as a reliability cue for admission, so unreliable observations are skipped, yielding a sparse and temporally irregular history. The stored s_i is the raw confidence at the selected location; hence, Hann priors affect only within-region localization, whereas cross-region selection and Memory admission remain based on raw visual confidence.

3.3 Training Strategy

TGTrack is trained in two stages to separate visual adaptation from temporal geometry learning. In Stage 1, LoRAT-L is adapted to AntiUAV410 and then frozen. In Stage 2, the frozen tracker is rolled out autoregressively to build an offline trajectory cache, with each search anchored by the preceding prediction. This preserves tracker drift and irregular observation gaps. Cache admission follows the inference-time raw-confidence threshold $\tau_m = 0.4$ and geometric-validity check, using raw Stage-1 responses without Hann modulation or geometry-guided Recrop.

The Temporal Geometry Branch is trained from these histories with ground-truth current-frame supervision:

$$\mathcal{L} = \mathcal{L}_{\text{ctr}} + 0.5\,\mathcal{L}_{\text{scl}} + 0.5\,\mathcal{L}_{\text{gate}} + 0.25\,\mathcal{L}_{\text{trig}} + 0.25\,\mathcal{L}_{\text{cov}}. \tag{14}$$

The center and log-scale terms use Smooth-L_1 regression.

The Recrop-center reliability g_t is supervised by the predicted-center error relative to the Recrop extent. Let $\mathbf{p}_t^*$ be the ground-truth center and define $\mathbf{r}_t = \mathbf{r}(\boldsymbol{\ell}_t)$ as the normalized Recrop half-size constructed from the predicted log-scale $\boldsymbol{\ell}_t$:

$$e_t^g = \left\| (\mathbf{p}_t^* - \text{sg}(\boldsymbol{\mu}_t)) \oslash \text{sg}(\mathbf{r}_t) \right\|_2, \qquad q_t^g = \exp\left[-\frac{1}{2} (e_t^g)^2 \right], \tag{15}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\oslash$ denotes element-wise division. The Gate head is trained by

$$\mathcal{L}_{\text{gate}} = \text{BCE}(g_t, q_t^g). \tag{16}$$

Its temporal feature is also detached, so this loss does not affect center or scale regression.

Finally, $\mathcal{L}_{\text{trig}}$ preserves the ground-truth inside/outside relation with respect to the normal search region, while $\mathcal{L}_{\text{cov}}$ encourages the predicted Recrop to cover the target for outside-region samples.

4 Experiments

4.1 Experimental Setup

Implementation and data. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU (24 GB) using PyTorch 2.0.0, CUDA 11.8, and Python 3.8 on Ubuntu 20.04. Our implementation is based on the official LoRAT codebase and uses LoRAT-L with a pretrained DINOv2-L/14 backbone as the visual tracker. Experiments follow the predefined AntiUAV410 splits (Huang et al., 2024). Stage 1 is trained for 20 epochs, while the Temporal Geometry Branch is optimized for 10 epochs in Stage 2. Detailed optimization settings are provided in Appendix B.

Inference and evaluation. Both the normal search and optional Recrop use a 224×224 crop with search factor $\text{SF} = 4$. The base Hann weight and Memory admission threshold are set to $\lambda_h = 0.45$ and $\tau_m = 0.4$, respectively, and at most one geometry-guided Recrop is evaluated per frame.

Evaluation follows the official AntiUAV410 protocol. Under the legacy precision metric used by the compared results, target-absent frames are counted as correct, whereas they contribute zero to SA for our local tracker, which always outputs a bounding box.

4.2 Comparison with State-of-the-Art Trackers

4.2.1 AntiUAV410

We evaluate TGTrack on the official AntiUAV410 test set against representative global- and local-search trackers, as reported in Table 1. Since global-search methods localize over the full frame whereas local-search trackers operate within bounded search regions, we report the two groups separately and primarily compare TGTrack with other local-search methods.

Among local-search trackers, TGTrack achieves the best SA and AUC, while attaining the highest reported precision, tied with TAPTrack after rounding. Relative to the controlled LoRAT-L baseline, it improves SA, precision, and AUC by 2.87, 3.74, and 2.86 points, respectively; the per-sequence AUC gain has a 95% bootstrap CI of $[+1.61, +4.28]$ (Appendix E). Compared with FocusTrack, the corresponding gains are 3.57, 1.63, and 3.51 points. Compared with TAPTrack, TGTrack achieves higher SA (67.47% vs. 58.4%) and AUC (66.31% vs. 63.1%), while maintaining comparable precision (87.83% vs. 87.8%).

Despite not performing persistent full-frame localization, TGTrack also approaches the performance of global-search trackers. Its 66.31% AUC is comparable to SiamDT (66.3%) and CAMTracker (66.7%), and remains within 1.2 points of DLST (67.5%). Its 67.47% SA is likewise comparable to SiamDT (67.7%) and CAMTracker (67.1%). These results suggest that temporal geometry-guided search allocation can extend the effective coverage of local search without requiring full-frame localization at every frame.

Table 1: Comparison on the AntiUAV410 test set. Best and second-best results among local-search trackers are shown in **bold** and underlined, respectively. † denotes results reported in prior work, and ‡ denotes our controlled LoRAT-L evaluation.

Search	Method	Venue	SA↑	P↑	AUC↑
Global	GlobalTrack† Huang et al. (2020)	AAAI20	66.4	85.3	65.1
	SiamDT† Huang et al. (2024)	TPAMI24	67.7	87.6	66.3
	CAMTracker† Wang et al. (2024)	Remote Sens. 24	67.1	88.6	66.7
	DLST† Zhang et al. (2025)	CVPRW 25	68.8	92.6	67.5
Local	TCTrack† Cao et al. (2022)	CVPR 22	41.6	60.4	41.1
	OSTrack† Ye et al. (2022)	ECCV 22	52.3	70.4	51.4
	DropTrack† Wu et al. (2023)	CVPR 23	60.2	82.2	59.2
	ROMTrack† Cai et al. (2023)	ICCV 23	55.7	74.5	54.7
	ZoomTrack† Kou et al. (2023)	NeurIPS 23	59.4	81.2	58.4
	HIPTrack† Cai et al. (2024)	CVPR 24	57.8	77.6	56.6
	AQATrack† Xie et al. (2024)	CVPR 24	56.6	75.3	55.5
	ARTrackV2† Bai et al. (2024)	CVPR 24	52.5	70.4	51.5
	LoRAT-L‡ Lin et al. (2024)	ECCV 24	<u>64.6</u>	84.1	<u>63.5</u>
	FocusTrack† Wang et al. (2025)	TGRS 25	63.9	86.2	62.8
	TDAT† Xue et al. (2025)	TITS 25	59.3	79.8	58.1
	ASRC+BTMU† (Yuan et al., 2026)	CCAI 26	–	<u>86.7</u>	62.4
	TAPTrack† Cui et al. (2026)	TGRS 26	58.4	**87.8**	63.1
	TGTrack	–	**67.5**	**87.8**	**66.3**

Table 2: Direct AntiUAV410-to-AntiUAV transfer without target-domain adaptation. Best and second-best results are shown in **bold** and underlined. † denotes published results and ‡ our controlled LoRAT-L evaluation.

Method	Venue	AUC↑	P↑	SA↑
OSTrack† Ye et al. (2022)	ECCV 22	59.20	79.40	60.20
ROMTrack† Cai et al. (2023)	ICCV 23	59.40	78.90	60.50
ZoomTrack† Kou et al. (2023)	NeurIPS 23	63.50	86.00	64.50
DropTrack† Wu et al. (2023)	CVPR 23	64.20	85.80	65.20
SUTrack† Chen et al. (2025)	AAAI 25	58.70	77.40	61.00
LoRAT-L‡ Lin et al. (2024)	ECCV 24	66.44	85.88	67.61
FocusTrack† Wang et al. (2025)	TGRS 25	<u>67.70</u>	<u>90.90</u>	<u>68.90</u>
TGTrack	–	**70.47**	**93.19**	**71.70**

4.2.2 Cross-Dataset Generalization on AntiUAV

Table 2 reports direct transfer from AntiUAV410 to AntiUAV. All compared trackers use AntiUAV410 as the training or adaptation data and are evaluated on AntiUAV without target-domain adaptation. Precision follows the legacy AntiUAV protocol used by the compared results.

The controlled LoRAT-L baseline achieves 66.44% AUC, 85.88% precision, and 67.61% SA on AntiUAV, while TGTrack reaches 70.47% AUC, 93.19% precision, and 71.70% SA, corresponding to gains of 4.03, 7.31, and 4.09 points, respectively. The consistent improvement across all three metrics shows that the gains of TGTrack persist under direct cross-dataset transfer.

TGTrack outperforms FocusTrack by 2.77 AUC, 2.29 precision, and 2.80 SA points. Its margin over the controlled LoRAT-L baseline is larger on AntiUAV than on AntiUAV410 in both AUC and precision. Together, these results suggest that TGTrack's geometry-guided recovery remains useful under this dataset shift without target-domain adaptation.

Table 3: Efficiency on AntiUAV410. Recrop overhead is averaged using the observed 6.19% execution rate.

Method / Component	Params (M)	MACs (G)	FPS
LoRAT-L baseline	307.43	103.02	349.47
Temporal Geometry Branch	+1.23	+0.01	–
Conditional Recrop (avg.)	–	+6.38	–
TGTrack	308.66	109.41	292.01

Table 4: Component ablation on AntiUAV410. ΔAUC is relative to the LoRAT-L baseline.

Temporal Geometry Branch	Gate–Hann	AUC↑	ΔAUC
		63.45	–
✓		65.16	+1.71
✓	✓	**66.31**	**+2.86**

Table 5: Effect of search-allocation modeling on AntiUAV410.

Configuration	AUC↑	ΔAUC
LoRAT-L baseline	63.45	–
+ Constant-Velocity Geo Recrop	63.91	+0.46
+ Generic Learnable Geo	64.27	+0.82
+ Proposed Temporal Geometry	65.16	+1.71

4.3 Efficiency Analysis

As shown in Table 3, the Temporal Geometry Branch adds only 1.23M parameters (0.40%) and 0.01 GMACs. The dominant additional cost comes from the conditional Recrop; with an execution rate of 6.19%, it contributes only 6.38 GMACs on average. Consequently, the average computation increases from 103.02 GMACs for LoRAT-L to 109.41 GMACs for TGTrack. Under identical evaluation settings, TGTrack runs at 292.01 FPS versus 349.47 FPS for LoRAT-L, retaining 83.6% of the baseline throughput. By invoking additional visual search only when needed, TGTrack maintains high efficiency without the persistent cost of enlarged or full-frame search.

4.4 Ablation Studies

Component contributions. Table 4 evaluates the contributions of the Temporal Geometry Branch and Gate–Hann. Adding the Temporal Geometry Branch increases AUC from 63.45% to 65.16%, yielding a gain of 1.71 points. In this variant, the predicted geometry determines when and where the additional Recrop is evaluated, while the Recrop retains the fixed Hann weight $\lambda_h = 0.45$ used by the normal search. The improvement therefore isolates the benefit of geometry-guided search allocation.

Gate–Hann further increases AUC to 66.31%, contributing an additional 1.15 points. It leaves the Recrop placement and cross-region selection unchanged while modulating the Hann prior within the Recrop according to the predicted reliability g_t. Overall, the two components provide complementary gains, improving the LoRAT-L baseline by 2.86 AUC points.

The predicted reliability g_t also correlates well with geometry accuracy ($\mathrm{Spearman}(g_t, -e_t) = 0.7004$); detailed reliability and gate-target consistency analyses are provided in Appendix H.

Effect of search-allocation modeling. Table 5 compares *where-to-look* formulations. All geometric variants share the same LoRAT-L tracker, conditional Recrop mechanism, and fixed Hann weight $\lambda_h = 0.45$, differing only in how the Recrop location is predicted. We compare a constant-velocity prior, a generic learnable trajectory model, and the proposed Temporal Geometry formulation.

Constant-Velocity Geo Recrop extrapolates the target center from the latest estimated velocity, while Generic Learnable Geo uses a two-layer Transformer over the same 10-D Memory with FIFO

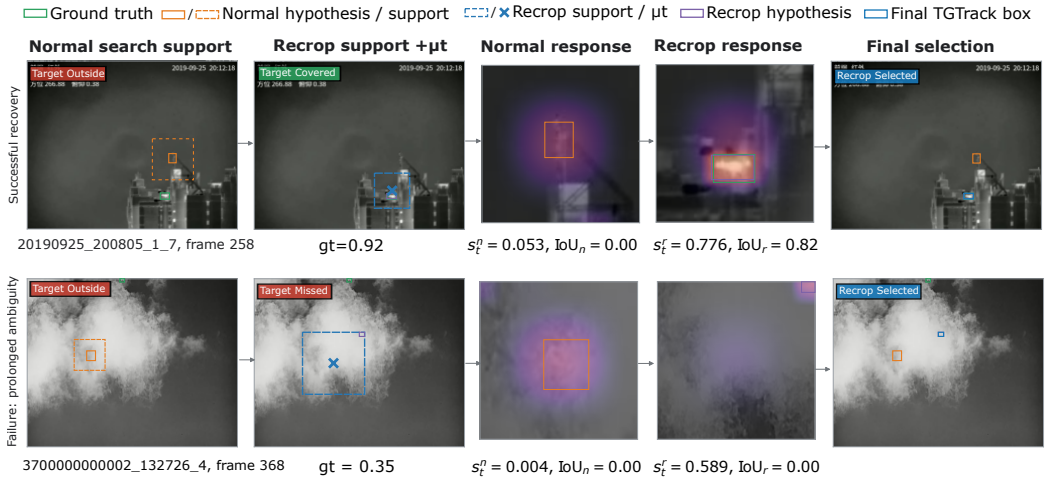


Figure 2: Visualization of geometry-guided recovery on AntiUAV410. Columns show the normal search support, geometry-guided Recrop support, normal response, Recrop response, and final selection, respectively. Dashed boxes denote search supports, $\boldsymbol{\mu}_t$ the predicted Recrop center, and scores the raw visual confidences used for cross-region selection.

positional embeddings and last-valid-token readout. The proposed Temporal Geometry further incorporates age- and score-aware encoding, pairwise temporal bias, learned-query pooling, horizon-conditioned fusion, and adaptive motion correction. The three variants improve the LoRAT-L baseline by 0.46, 0.82, and 1.71 AUC points, respectively, with the proposed model outperforming Generic Learnable Geo by 0.89 AUC points. This indicates that explicitly modeling temporal irregularity and observation reliability improves the allocation of the additional Recrop. Complementarily, raw-score selection agrees with the higher-IoU regional candidate in 90.57% of strictly ordered cases, with only 0.0077 average selection regret; analysis is provided in Appendix I.

Geometry-guided recovery mechanism. Figure 2 shows representative successful and failed recovery cases.

In the successful case, temporal geometry places the Recrop over a target missed by the normal search, restoring visual evidence that is then selected by its higher raw confidence. This illustrates the separation between geometry-guided search allocation and appearance-based selection. In the failure case, neither the normal search nor the predicted Recrop covers the target, so appearance-based selection and Gate–Hann cannot recover it. The low predicted reliability is consistent with the inaccurate Recrop placement.

Attribute analysis. TGTrack improves over the controlled LoRAT-L baseline across all ten AntiUAV410 subsets, with the largest AUC gains under fast motion (+5.55), occlusion (+5.24), out-of-view (+4.09), and small-target (+3.96) conditions. These gains are consistent with the intended role of temporal geometry in recovering targets under large displacement and unreliable observations; full results are provided in Appendix D.

5 Conclusion

We presented TGTrack, a temporal geometry-aware local tracker that decouples search allocation from visual selection. Temporal geometry guides a conditional Recrop, while raw appearance confidence selects the regional hypothesis and Gate–Hann modulates the Recrop according to predicted reliability. On AntiUAV410, TGTrack achieves 66.31% AUC at 292.01 FPS, outperforming the controlled LoRAT-L baseline by 2.86 points, and reaches 70.47% AUC under direct transfer to AntiUAV. These results suggest that targeted search allocation extends local-search coverage without persistent search enlargement or full-frame localization, while bounded recovery motivates longer-term extensions.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language polishing, and LaTeX formatting. Generative AI was not used to generate or fabricate any experimental results. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies visual tracking using established thermal infrared anti-UAV benchmarks and does not involve human subjects or private personal data. Anti-UAV tracking may have dual-use implications in security and surveillance applications. Any real-world deployment should comply with applicable legal, privacy, and safety requirements.

REFERENCES

Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. ARTrackV2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19048–19057, 2024. doi: 10.1109/CVPR52733.2024.01802.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pp. 3464–3468, 2016. doi: 10.1109/ICIP.2016.7533003.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. HIPTrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19258–19267, 2024.

Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9589–9600, 2023.

Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. TCTrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14798–14808, 2022. doi: 10.1109/CVPR52688.2022.01438.

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*, pp. 8126–8135, 2021.

Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. SUTrack: Towards simple and unified single object tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2239–2247, 2025. doi: 10.1609/aaai.v39i2.32223.

Xiaolong Cui, Xingxiu Li, Panlong Wu, Xinan Liu, and Shan He. TAPTrack: An efficient temporal-aware prompt tracker for infrared anti-UAV tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–14, 2026. doi: 10.1109/TGRS.2026.3682081.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. MixFormer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13608–13618, 2022. doi: 10.1109/CVPR52688.2022.01324.

Zenghuang Fu, Xiaofeng Han, Mingda Jia, Jin Ming Yang, Qi Zeng, Muyang Zahng, Changwei Wang, Weiliang Meng, and Xiaopeng Zhang. Dmtrack: Deformable state-space modeling for uav multi-object tracking with kalman fusion and uncertainty-aware association. *CoRR*, abs/2510.17860, 2025. doi: 10.48550/arXiv.2510.17860.

Bo Huang, Junjie Chen, Tingfa Xu, Ying Wang, Shenwang Jiang, Yuncheng Wang, Lei Wang, and Jianan Li. SiamSTA: Spatio-temporal attention based siamese tracker for tracking UAVs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1204–1212, 2021.

Bo Huang, Jianan Li, Junjie Chen, Gang Wang, Jian Xun 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. doi: 10.1109/TPAMI.2023.3335338.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. GlobalTrack: A simple and strong baseline for long-term tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):11037–11044, 2020. doi: 10.1609/aaai.v34i07.6758.

Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Guodong Guo, Qixiang Ye, Jianbin Jiao, Jian Zhao, and Zhenjun Han. Anti-UAV: A large-scale benchmark for vision-based UAV tracking. *IEEE Transactions on Multimedia*, 25:486–500, 2023. doi: 10.1109/TMM.2021.3128047.

Vitaliy Kim, Gunho Jung, and Seong-Whan Lee. Am-sort: Adaptable motion predictor with historical trajectory embedding for multi-object tracking. In *Pattern Recognition and Artificial Intelligence: 4th International Conference, ICPRAI 2024, Proceedings, Part I*, volume 14892 of *Lecture Notes in Computer Science*, pp. 92–107. Springer, 2024. doi: 10.1007/978-981-97-8702-9_7.

Yutong Kou, Jin Gao, Bing Li, Gang Wang, Weiming Hu, Yizheng Wang, and Liang Li. ZoomTrack: Target-aware non-uniform resizing for efficient visual tracking. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50959–50977. Curran Associates, Inc., 2023. doi: 10.52202/075280-2217.

Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets LoRA: Faster training, larger model, stronger performance. In *European Conference on Computer Vision (ECCV)*, pp. 300–318, 2024.

Mingjin Lv, Zelin Liu, Feifei Shao, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. Gatemot: Q-gated attention for dense object tracking. *CoRR*, abs/2604.26353, 2026. doi: 10.48550/arXiv.2604.26353.

Jae Kwan Park and Ji-Hyeong Han. MemLoTrack: Enhancing TIR Anti-UAV tracking with memory-integrated low-rank adaptation. *Sensors*, 25(23):7359, 2025. doi: 10.3390/s25237359.

Prithvi Raj Singh, Raju Gottumukkala, Anthony S. Maida, Alan B. Barhorst, and Vijaya Gopu. Physics-guided fusion for robust 3d tracking of fast moving small objects. *CoRR*, abs/2510.20126, 2025. doi: 10.48550/arXiv.2510.20126.

Ying Wang, Tingfa Xu, and Jianan Li. FocusTrack: A self-adaptive local sampling algorithm for efficient anti-UAV tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. doi: 10.1109/TGRS.2025.3562958.

Ziming Wang, Yuxin Hu, Jianwei Yang, Guangyao Zhou, Fangjian Liu, and Yuhan Liu. A contrastive-augmented memory network for anti-uav tracking in tir videos. *Remote Sensing*, 16(24):4775, 2024. doi: 10.3390/rs16244775.

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*, pp. 14561–14571, 2023.

Bin Xie, Congxuan Zhang, Fagan Wang, Peng Liu, Feng Lu, Zhen Chen, and Weiming Hu. CST Anti-UAV: A thermal infrared benchmark for tiny UAV tracking in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6216–6225, 2025.

Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19300–19309, 2024.

Yuanliang Xue, Guodong Jin, Tao Shen, Lining Tan, Nian Wang, Jing Gao, Ye Yu, and Siyuan Tian. Target-distractor aware uav tracking via global agent. *IEEE Transactions on Intelligent Transportation Systems*, 26(10):16116–16127, 2025. doi: 10.1109/TITS.2025.3581391.

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*, pp. 10448–10457, 2021.

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*, pp. 341–357. Springer, 2022. doi: 10.1007/978-3-031-20047-2_20.

Zhe Yuan, Wenting Cao, and Huiyu Luo. Adaptive search region control and template stabilization for anti-uav tracking. In *2026 6th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pp. 782–785. IEEE, 2026. doi: 10.1109/CCAI69603.2026.11642046.

Jiahao Zhang, Yixin Wei, Jinli Zhang, Zongli Jiang, Peiwen Yu, Yufei Ma, and Runan Jin. Dlst: Dual-template co-evolution learning for robust long-term drone tracking in dynamic environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 6707–6715, 2025.

Jie Zhao, Jingshu Zhang, Dongdong Li, and Dong Wang. Vision-based anti-uav detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25323–25334, 2022. doi: 10.1109/TITS.2022.3177627.

A Detailed Training Objectives

This section provides the detailed definitions of the Stage-2 objectives introduced in Sec. 3.3. During Stage 2, the LoRAT-L visual tracker remains frozen and only the Temporal Geometry Branch is optimized.

Center and scale regression. The center prediction is supervised using Smooth-L_1 regression:

$$\mathcal{L}_{\mathrm{ctr}} = \mathbb{E}_t \left[\sum_{d \in \{x,y\}} \rho_{\beta_{\mathrm{ctr}}} \left(\mu_{t,d} - p^*_{t,d} \right) \right], \tag{17}$$

where $\rho_\beta(\cdot)$ denotes the Smooth-L_1 penalty with transition parameter β, and $\beta_{\mathrm{ctr}} = 0.02$.

The normalized log-scale is supervised similarly:

$$\mathcal{L}_{\mathrm{scl}} = \mathbb{E}_t \left[\sum_{d \in \{w,h\}} \rho_{\beta_{\mathrm{scl}}} \left(\ell_{t,d} - \ell^*_{t,d} \right) \right], \tag{18}$$

where $\beta_{\mathrm{scl}} = 0.1$.

Boundary-aware trigger supervision. Let the center and half-size of the normal search support be

$$\mathbf{c}^{\mathrm{n}}_t = \frac{\mathbf{b}^-_t + \mathbf{b}^+_t}{2}, \qquad \mathbf{r}^{\mathrm{n}}_t = \frac{\mathbf{b}^+_t - \mathbf{b}^-_t}{2}, \tag{19}$$

and define the normalized crop-relative coordinates

$$\mathbf{o}^*_t = (\mathbf{p}^*_t - \mathbf{c}^{\mathrm{n}}_t) \oslash \mathbf{r}^{\mathrm{n}}_t, \qquad \hat{\mathbf{o}}_t = (\boldsymbol{\mu}_t - \mathbf{c}^{\mathrm{n}}_t) \oslash \mathbf{r}^{\mathrm{n}}_t. \tag{20}$$

For an outside-region sample, the dominant exit axis and direction are

$$a^*_t = \arg\max_{a \in \{x,y\}} |o^*_{t,a}|, \qquad d^*_t = \mathrm{sign}\left(o^*_{t,a^*_t} \right), \tag{21}$$

and the prediction is penalized if it does not cross the same boundary. The implementation uses zero boundary margin, giving

$$v_t = \begin{cases} \mathrm{ReLU}\left(1 - d^*_t \hat{o}_{t,a^*_t}\right), & \mathbf{p}^*_t \notin \Omega^{\mathrm{n}}_t, \\ \mathrm{ReLU}\left(\max_{a \in \{x,y\}} |\hat{o}_{t,a}| - 1\right), & \mathbf{p}^*_t \in \Omega^{\mathrm{n}}_t. \end{cases} \tag{22}$$

Table 6: Detailed training settings for TGTrack.

Setting	Stage 1	Stage 2
Optimized module	LoRAT-L	Temporal Geometry Branch
Epochs	20	10
Samples / epoch	39,936	39,936
Optimizer	AdamW	AdamW
Initial learning rate	5×10^{-5}	1×10^{-4}
Minimum learning rate	1×10^{-6}	1×10^{-6}
Weight decay	0.1	0.01
Effective batch size	128	128
LR schedule	Cosine	Cosine
Warm-up	1 epoch	1 epoch
Gradient clipping	1.0	1.0
Precision	AMP(FP16)	FP32

The per-sample trigger loss is

$$\ell_t^{\mathrm{trig}} = \rho_{\beta_{\mathrm{trig}}}(v_t), \tag{23}$$

where $\beta_{\mathrm{trig}} = 0.1$. Let $\mathcal{O}$ and $\mathcal{I}$ denote the outside- and inside-region samples in a mini-batch. When both classes are present, class balancing is implemented as

$$\mathcal{L}_{\mathrm{trig}} = \frac{1}{2}\left[\frac{1}{|\mathcal{O}|}\sum_{t\in\mathcal{O}}\ell_t^{\mathrm{trig}} + \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}}\ell_t^{\mathrm{trig}}\right]. \tag{24}$$

If only one class is present, the mean loss of that class is used directly. Unlike $\mathcal{L}_{\mathrm{gate}}$, this objective backpropagates through $\boldsymbol{\mu}_t$ and therefore directly trains the center predictor and the shared temporal representation.

Recrop coverage supervision. For samples whose ground-truth center lies outside the normal search region, let

$$\mathbf{h}_t^* = \frac{1}{2}\exp(\boldsymbol{\ell}_t^*)$$

denote the normalized half-size of the ground-truth box, and let $\mathbf{r}_t = \mathbf{r}_t(\boldsymbol{\ell}_t)$ denote the predicted Recrop half-size. We define the normalized coverage ratio as

$$E_t = \max_{d\in\{x,y\}} \frac{|p_{t,d}^* - \mu_{t,d}| + \frac{1}{2}\exp(\ell_{t,d}^*)}{r_{t,d}}. \tag{25}$$

The target box is fully covered when $E_t \le 1$. We therefore define the coverage loss as

$$\mathcal{L}_{\mathrm{cov}} = \mathbb{E}_{t:y_t^{\mathrm{out}}=1}\left[\rho_{\beta_{\mathrm{cov}}}\left(\max(E_t - 1,\, 0)\right)\right], \tag{26}$$

where $\rho_\beta(\cdot)$ denotes the Smooth-L_1 loss with transition parameter β, and $\beta_{\mathrm{cov}} = 0.1$. Thus, fully covered samples incur zero loss, while uncovered samples are penalized according to the excess coverage ratio $E_t - 1$.

B Additional Implementation Details

Table 6 summarizes the optimization settings used for the two training stages.

For Stage 1, the per-step batch size is 32 with four-step gradient accumulation, resulting in an effective batch size of 128. The warm-up starts from a learning rate of 10^{-7}. For Stage 2, center regression uses Smooth-L_1 with $\beta = 0.02$, while scale, trigger, and coverage objectives use $\beta = 0.1$.

C Temporal Geometry Branch Architecture

Figure 3 provides the detailed architecture of the Temporal Geometry Branch described in Sec. 3.

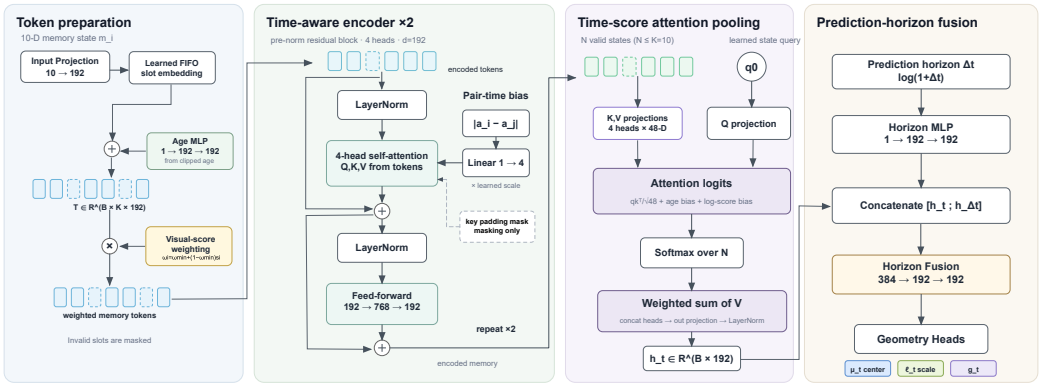


Figure 3: Detailed architecture of the Temporal Geometry Branch. Time-aware trajectory encoding and confidence-aware pooling summarize the Memory, which is fused with the prediction horizon to estimate target center, scale, and Recrop-center localization reliability.

Table 7: Success AUC (%) across AntiUAV410 challenge attributes and target-size categories. N denotes the number of sequences and Δ denotes the absolute gain over LoRAT-L.

Subset	N	LoRAT-L	TGTrack	Δ
TC	81	60.37	**63.90**	+3.53
OV	31	47.05	**51.14**	+4.09
SV	27	53.59	**57.14**	+3.55
FM	52	52.14	**57.69**	+5.55
OC	12	27.72	**32.96**	+5.24
DBC	21	47.17	**49.98**	+2.81
TS	33	47.78	**51.12**	+3.34
SS	87	57.68	**61.64**	+3.96
MS	84	67.62	**70.03**	+2.41
NS	32	73.24	**75.27**	+2.03
Overall	120	63.45	**66.31**	+2.86

D Full Attribute and Target-Size Analysis

Table 7 compares TGTrack with the controlled LoRAT-L baseline across the ten AntiUAV410 evaluation subsets. The six challenge attributes are thermal crossover (TC), out-of-view (OV), scale variation (SV), fast motion (FM), occlusion (OC), and dynamic background clutter (DBC). We additionally report four target-size categories: tiny size (TS), small size (SS), medium size (MS), and normal size (NS).

TGTrack improves over the baseline across all evaluated subsets. The largest gain occurs under fast motion (+5.55 AUC), followed by occlusion (+5.24), out-of-view sequences (+4.09), and small targets (+3.96). The strong gains under fast motion and out-of-view conditions are consistent with the intended role of temporal geometry in extending the recoverable range of bounded local search.

Dynamic background clutter also improves by +2.81 AUC, although this condition remains more dependent on appearance discrimination than on motion coverage. Occlusion shows a large sequence-level gain, but it remains the most difficult condition in absolute performance, suggesting that prolonged target disappearance is still challenging for the current bounded recovery mechanism.

E Qualitative Analysis

Sustained recovery examples. Figure 4 compares TGTrack with the controlled LoRAT-L baseline over sustained recovery segments. After LoRAT-L loses the target, TGTrack can re-acquire it and

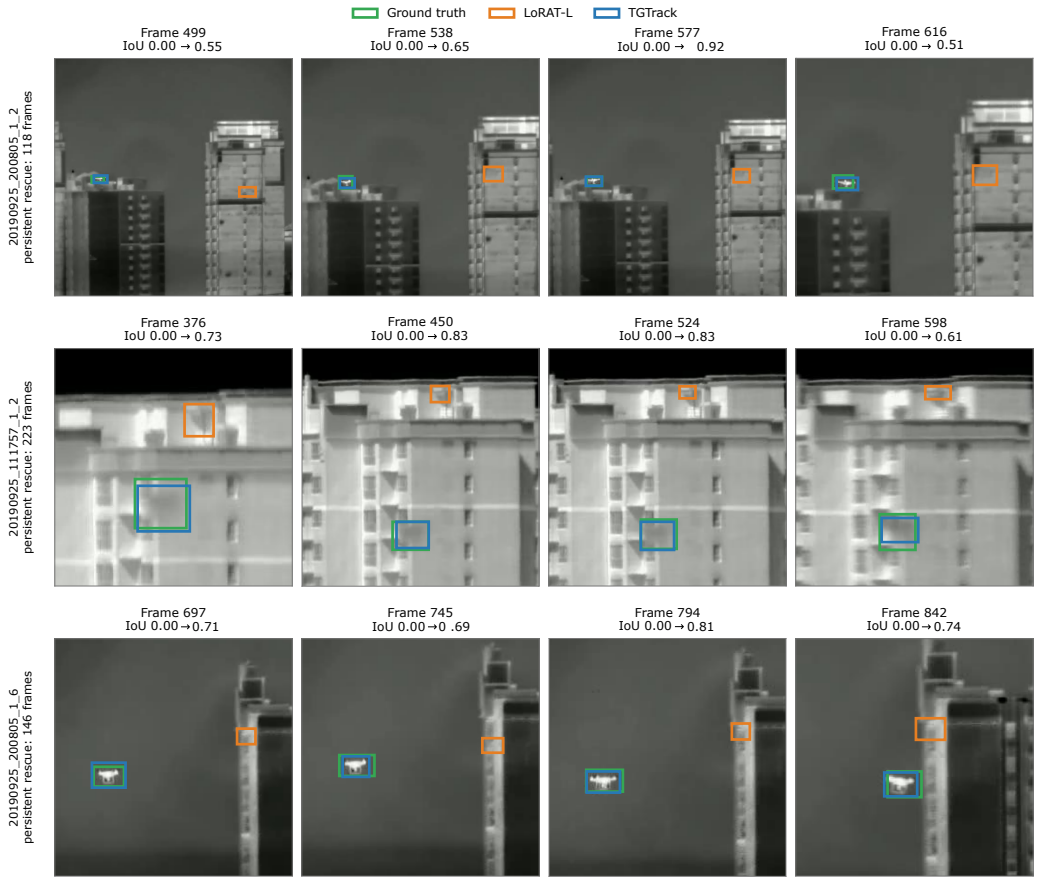


Figure 4: Qualitative recovery examples on AntiUAV410. Green, orange, and blue boxes denote ground truth, LoRAT-L, and TGTrack, respectively. IoU values are reported as LoRAT-L $\rightarrow$ TGTrack. Representative frames are sampled from sustained recovery segments in which the baseline loses the target while TGTrack re-acquires and continues tracking it.

maintain accurate localization over subsequent sampled frames. These examples are consistent with the quantitative improvements under fast motion, out-of-view, occlusion, and small-target conditions.

Sequence-level and closed-loop diagnostics. We compute the paired per-sequence AUC difference between TGTrack and LoRAT-L over the 120 AntiUAV410 test sequences. Resampling these sequence-level differences with replacement for 200,000 replicates gives a mean gain of 2.8662 AUC points and a 95% percentile-bootstrap confidence interval of $[1.6060, 4.2755]$ points.

We also measure how often the ground-truth center of a visible target falls outside each tracker's normal search support. The rate decreases from 17.16% for LoRAT-L to 12.67% for TGTrack, a difference of 4.49 percentage points. Because each tracker generates its search support from its own preceding predictions, this comparison describes their closed-loop outcomes rather than isolating the effect of an individual component.

F Effect of Stage-1 Trajectory Generation

Table 8 examines the effect of Stage-1 trajectory generation on AntiUAV410. Stage 2 learns temporal geometry from trajectories generated by the frozen Stage-1 tracker. We therefore examine whether applying the inference-time Hann prior during Stage-1 rollout improves the trajectory cache used for geometry learning. This comparison changes only the trajectory-generation rule; the Temporal

Table 8: Effect of Stage-1 trajectory generation on AntiUAV410. All Stage-2 settings are kept identical.

Trajectory	AUC↑	P↑	NormP↑	$SR_{0.5}$ ↑	$SR_{0.75}$ ↑	SA↑
Raw response	**66.31**	**87.83**	**85.43**	**84.20**	**61.31**	**67.47**
Hann-regularized	64.92	86.06	83.71	82.46	60.02	66.05

Table 9: Sensitivity to the Memory-admission threshold τ_m on AntiUAV410.

τ_m	AUC↑	P↑	NormP↑	$SR_{0.5}$ ↑	$SR_{0.75}$ ↑	SA↑
0.2	65.97	87.37	84.99	83.72	61.10	67.12
0.3	65.61	86.89	84.56	83.32	60.68	66.75
0.4	**66.31**	**87.83**	**85.43**	**84.20**	**61.31**	**67.47**
0.5	65.45	86.77	84.35	83.06	60.48	66.58
0.6	65.48	86.76	84.35	83.12	60.53	66.62

Geometry Branch, Stage-2 objective, optimization schedule, and 10-epoch training protocol remain identical.

Hann regularization produces a substantially cleaner Stage-1 cache but leads to weaker final tracking performance. The raw-response rollout contains 159,349 admitted states, including 26,511 off-target states, corresponding to a cache purity of 83.36%. With Hann regularization, the number of admitted states increases to 181,062, while off-target states decrease to 9,015 and cache purity rises to 95.02%. An admitted state is considered off-target when the target is visible but the selected box has IoU below 0.1, or when a state is admitted while the target is absent.

This result shows that cache purity is not equivalent to recovery utility. The Temporal Geometry Branch is trained from imperfect tracker-generated histories to predict where additional evidence should be acquired. Drifted and off-target observations therefore expose the model to precisely those states in which additional search allocation is most needed. Applying Hann during rollout suppresses many of these naturally occurring hard states and shifts the training cache toward easier local-tracking trajectories.

The same trend appears in the Stage-2 diagnostics. Relative to the raw-response cache, Hann regularization increases the validation center error from 0.0805 to 0.1179, Gate BCE from 0.3220 to 0.4011, and trigger loss from 0.1714 to 0.3331. On recovery-required frames, Recrop coverage decreases from 18.68% to 11.29%, while the geometry-recovery rate drops from 18.19% to 10.91%. The false-trigger rate also increases from 1.07% to 2.30%.

Consequently, final AUC decreases from 66.31 to 64.92 despite the higher purity of the Hann-regularized cache. We therefore use raw Stage-1 responses to construct the Stage-2 cache, preserving informative drift and recovery states for learning geometry-guided search allocation. Hann regularization remains enabled during online tracking, where it stabilizes within-region visual localization.

G Trajectory Memory Sensitivity

Memory-admission threshold We vary the Memory-admission threshold τ_m while keeping all other settings unchanged. As shown in Table 9, $\tau_m = 0.4$ achieves the strongest overall performance among the evaluated settings. A lower threshold admits more observations but also increases Memory contamination, whereas a higher threshold produces cleaner Memory at the cost of discarding useful trajectory states.

We additionally characterize admitted-state quality using the overlap between the final selected box $\hat{b}_t$ and the ground-truth box b_t^*. An admission is considered contaminated if the target is visible but $\mathrm{IoU}(\hat{b}_t, b_t^*) < 0.1$, or if a state is admitted while the target is absent. The contaminated-admission rate is

$$R_{\mathrm{cont}} = \frac{N_{\mathrm{write,misloc}} + N_{\mathrm{write,absent}}}{N_{\mathrm{write,total}}}. \tag{27}$$

Table 10: Effect of trajectory Memory capacity on AntiUAV410. The activation condition is fixed to $N_t \geq 5$ for all configurations.

K	AUC↑	P↑	NormP↑	$SR_{0.5}$ ↑	$SR_{0.75}$ ↑	SA↑
5	66.08	87.53	85.11	83.90	61.05	67.23
10	**66.31**	**87.83**	**85.43**	**84.20**	**61.31**	**67.47**
15	65.86	87.24	84.86	83.64	60.92	67.01
20	65.84	87.21	84.84	83.55	60.87	66.98

Admissions with IoU in $[0.1, 0.5)$ are treated as intermediate-quality states and are not counted as contaminated. This ground-truth-based measure is used only for post-hoc analysis of the online admission behavior.

At $\tau_m = 0.2$, the Memory write rate increases to 92.53%, while the contaminated-admission rate rises to 10.79%, indicating that overly permissive admission introduces more severely mislocalized or target-absent states. Increasing the threshold to 0.6 reduces the contaminated-admission rate to 6.85%, but also lowers the write rate to 86.69% and decreases AUC by 0.83 points relative to the default setting.

These results show that $\tau_m = 0.4$ provides a favorable balance between retaining sufficient temporal observations and limiting unreliable Memory admissions.

Memory capacity We vary the Memory capacity while fixing the activation condition to $N_t \geq 5$ for all configurations. As shown in Table 10, $K = 10$ achieves the strongest performance among the evaluated settings. A shorter Memory provides less temporal context, whereas further increasing the capacity does not yield additional benefit and may introduce older observations that are less relevant to the current motion. Overall, TGTrack remains relatively stable across the evaluated Memory capacities.

H Gate Reliability Analysis

We further analyze whether the predicted gate g_t provides a meaningful measure of the localization reliability of the Temporal Geometry Branch. The analysis is conducted on the 90 AntiUAV410 validation sequences using the Epoch-9 Stage-2 checkpoint corresponding to the final TGTrack configuration. We collect 92,043 valid geometry predictions from the Stage-1 tracker-admitted trajectory cache, with Memory capacity $K = 10$ and geometry activation requiring $N_t \geq 5$.

For each prediction, the localization error is measured by the normalized center distance

$$e_t = \sqrt{\left(\frac{\hat{x}_t - x_t^*}{W}\right)^2 + \left(\frac{\hat{y}_t - y_t^*}{H}\right)^2}, \tag{28}$$

where $(\hat{x}_t, \hat{y}_t)$ and (x_t^*, y_t^*) denote the predicted and ground-truth target centers, respectively, and W and H are the image width and height.

Localization reliability. Figure 5(a) shows the center-error distributions after grouping predictions according to g_t. Predictions with $g_t \in [0.8, 1.0]$ are substantially more accurate than those in the intermediate intervals, with a median normalized center error of only 0.00467 and a mean error of 0.01245. By comparison, the median errors for $g_t \in [0.2, 0.4)$, $[0.4, 0.6)$, and $[0.6, 0.8)$ are 0.17590, 0.19420, and 0.18353, respectively. No valid prediction falls in $[0, 0.2)$.

The same trend is reflected by the rank correlation between the predicted gate and localization accuracy:

$$\mathrm{Spearman}(g_t, -e_t) = 0.7004. \tag{29}$$

Although the intermediate intervals are not strictly ordered, larger gate values are strongly associated with lower localization error overall. In particular, g_t clearly separates the high-reliability predictions concentrated in $[0.8, 1.0]$ from less reliable predictions.

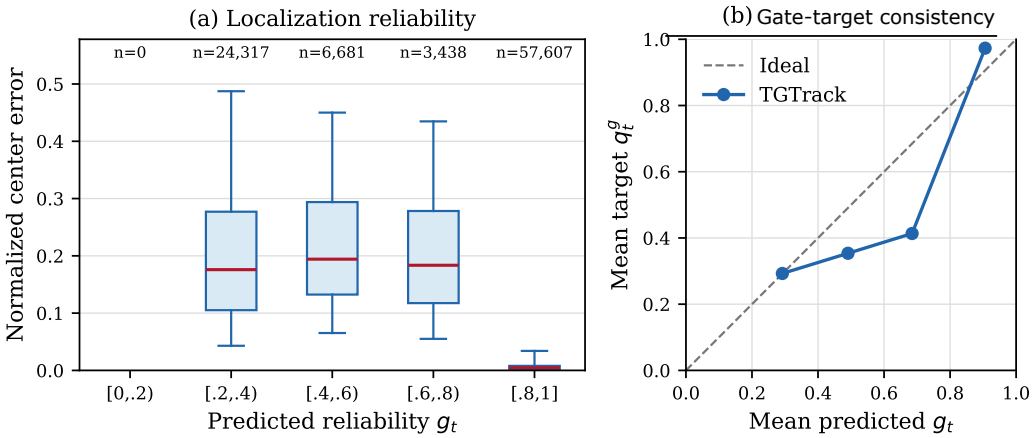


Figure 5: Reliability analysis of the predicted gate g_t on the AntiUAV410 validation set. (a) Distribution of normalized center error across fixed gate-score intervals. The number above each interval denotes the number of valid predictions. (b) Mean predicted gate versus the mean reliability target q_t^g within the same intervals. The dashed line denotes ideal agreement.

Gate-target consistency. Figure 5(b) compares the mean predicted gate with the mean reliability target q_t^g used to supervise $\mathcal{L}_{\text{gate}}$. The high-reliability interval closely follows its target, whereas larger deviations appear in the intermediate intervals. The sample-weighted mean absolute gap between bin-wise predicted and target reliability is 0.0622.

These results suggest that g_t is better interpreted as a relative reliability indicator than as a calibrated probability. This interpretation is consistent with its role in TGTrack: g_t does not determine the final target state or cross-region decision, but only adjusts the strength of the Hann prior within the geometry-guided Recrop. Final normal-versus-Recrop selection remains based on raw visual confidence.

I Decoupled Recovery Analysis

The main paper evaluates the *where-to-look* component by comparing different search-allocation models under the same visual tracker and Recrop mechanism. Here, we further analyze the complementary *what-to-trust* component, namely whether raw appearance confidence provides a reliable criterion for selecting between the normal and Recrop hypotheses.

Appearance-based selection quality. Let

$$I_t^q = \mathrm{IoU}(\hat{b}_t^q, b_t^*), \qquad q \in \{n, r\},$$

denote the IoU of the normal and Recrop candidates, and let I_t^{sel} denote the IoU of the candidate selected by the raw-score rule. We define the average selection regret as

$$R_{\text{sel}} = \frac{1}{N} \sum_t \left[\max(I_t^n, I_t^r) - I_t^{\text{sel}} \right]. \tag{30}$$

Among 3,424 visible frames for which the normal and Recrop candidates have a strict IoU ordering, the raw-score ordering agrees with the higher-IoU candidate in 90.57% of cases. The average selection regret is only 0.0077 IoU. Moreover, only 4.76% of executed Recrop frames satisfy $|\hat{s}_t^r - \hat{s}_t^n| < 0.01$, showing that ambiguous near-tie decisions are uncommon. These results support the use of raw appearance confidence for final cross-region arbitration after geometry has determined where additional evidence is acquired.

Margin sensitivity and temporal stability. We further test whether a conservative margin improves cross-region selection. Specifically, the Recrop candidate is accepted only if

$$\hat{s}_t^r > \hat{s}_t^n + \delta, \tag{31}$$

Table 11: Sensitivity to the cross-region selection margin δ on AntiUAV410. Recrop is accepted only when $\hat{s}_t^r > \hat{s}_t^n + \delta$. The Recrop win rate is computed only over frames in which Recrop is executed.

δ	AUC	P	P_{norm}	$SR_{0.5}$	$SR_{0.75}$	SA	Recrop Win. (%)
0	**66.31**	**87.83**	85.43	**84.20**	**61.31**	**67.47**	29.08
0.01	66.11	87.60	85.19	83.97	61.00	67.26	27.34
0.02	66.11	87.57	85.18	83.98	61.04	67.26	26.24
0.05	66.30	87.81	**85.45**	84.19	61.27	67.46	23.22

where $\delta \in \{0, 0.01, 0.02, 0.05\}$ and $\delta = 0$ corresponds to the default raw-score rule.

As shown in Table 11, introducing a positive margin does not improve tracking accuracy. The default $\delta = 0$ rule achieves the best AUC, precision, $SR_{0.5}$, $SR_{0.75}$, and SA, while increasing the margin mainly reduces the fraction of Recrop hypotheses retained as final outputs.

We also examine temporal switching. Under $\delta = 0$, the conditional flip rate is 17.21% over the 6,066 consecutive frame pairs in which Recrop is executed in both frames. These 1,044 flips account for only 0.81% of all 129,571 tracking frames. Increasing the margin to $\delta = 0.05$ reduces the conditional flip rate to 13.58%, but changes AUC only from 66.31 to 66.30. Thus, the simple raw-score rule already provides stable appearance-based cross-region selection without introducing an additional calibration margin.