PRMTRACK: PROGRESSIVE RELIABILITY MODELING FOR SATELLITE VIDEO OBJECT TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Satellite video object tracking remains challenging due to tiny object sizes, substantial aspect-ratio variations, frequent occlusions, and interference from similar objects in complex backgrounds, which often lead to template-search representation mismatch, response ambiguity, and dynamic template contamination. To address these challenges, we propose PRMTrack, a Siamese network-based progressive reliability modeling framework for satellite video object tracking that improves tracking reliability across three consecutive stages: object representation, response recalibration, and template renewal. Specifically, we develop a Multi-Scale Sparse Perception Network that strengthens structural object representations through multi-scale feature modeling and sparse interactions, alleviating feature mismatch caused by aspect-ratio variations. We further design an Adaptive Dual-Branch Response Recalibration Network that learns branch-specific Gamma mappings according to the distinct distributions of classification and centerness responses, enhancing genuine object responses while suppressing distracting peaks from similar objects. Finally, we introduce an Object-Ratio-Guided Reliable Template Renewal Mechanism that evaluates scale consistency using the area ratio between the current object and historical template and adaptively controls template renewal, thereby reducing template contamination caused by occlusions and abnormal scale estimation. Extensive experiments on SatSOT, OOTB, and SV248S demonstrate strong tracking performance, with PRMTrack achieving 74.9% precision and a 53.0% success rate on SatSOT. Source code is available at: https://anonymous.4open.science/r/Track-SVOT.

1 INTRODUCTION

With the rapid development of high-resolution Earth observation technologies, satellite videos enable continuous monitoring over large geographic areas, providing important support for applications such as traffic monitoring (Nocua et al., 2025), maritime surveillance (Lurbur et al., 2026), disaster response (Hu et al., 2025), and public safety. Satellite Video Object Tracking (SVOT) aims to continuously estimate the position and scale of an object in subsequent frames given its initial state in the first frame. Compared with natural-scene videos, satellite objects typically occupy only a few pixels and contain limited texture and semantic information. Meanwhile, geometric variations, frequent occlusions, and densely distributed similar objects further weaken object-background discrimination. If these uncertainties are not effectively controlled, erroneous representations may propagate across frames, contaminate dynamic templates, and eventually lead to persistent localization drift or even tracking failure (Zhao et al., 2022; Chen et al., 2024a).

Deep learning has greatly advanced single-object tracking. Bertinetto et al. (2016) formulate tracking as similarity matching between the template and search region, while Li et al. (2019) improve target representation with deeper networks and multi-level features. Later, Transformer-based trackers such as Chen et al. (2021), Yan et al. (2021), Ye et al. (2022), and Chen et al. (2026) enhance template–search interaction through attention. For SVOT, Yang et al. (2023) use dynamic templates and motion cues to improve stability, while Zhou et al. (2025a), Li et al. (2026), and Wang et al. (2026b) further exploit temporal cues, multi-temporal templates, and historical trajectories to improve accuracy. Despite these advances, preventing unreliable information from entering the tracker and accumulating during representation, localization, and template updating remains challenging.

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.

At the representation stage, targets such as trains and ships may undergo large aspect-ratio changes during rotation or deformation (Chen et al., 2024a), causing structural mismatch between the current target and the template. Zhang et al. (2024) enhance target features through adaptive multi-scale modeling, Chen et al. (2024b) design a rotation-equivariant Siamese network to capture target variations, while Lu & Peng (2026) and Du et al. (2026) use structured noise and diffusion models, respectively, to model target appearance. However, these methods mainly focus on target–background discrimination or overall scale changes, with limited modeling of template–search mismatch caused by non-uniform geometric variations.

Even with improved representation, response ambiguity may persist. In complex backgrounds, similar vehicles or background regions can produce multiple competing peaks. Existing methods mainly improve target–background discrimination at the feature level. Lai et al. (2023) adopts dual-particle filtering, Lin et al. (2024) use spatiotemporal regularized correlation filtering, and Zhang et al. (2026) jointly model frequency, spatial, and temporal cues learn spatial distributions to improve discrimination. However, these methods overlook the different distributions of classification and centerness responses, whose combination may still cause response diffusion or false peaks.

Unreliable localization may further propagate through dynamic template updating. Zhou et al. (2025b) and Li et al. (2026) use temporal information to model historical target states. However, occlusion or abnormal predictions may introduce substantial background into the current target region. Once used for template updating, such observations can corrupt subsequent matching. Although Yang et al. (2024) use background suppression to reduce interference and Lv et al. (2025) propose spectral–spatial–frame feature fusion, Wang et al. (2026a) introduce inter-frame autoregressive sparsification to distinguish the foreground from the background, these methods lack sufficient constraints on whether the current prediction is reliable enough for template updating. Therefore, the key is not only to exploit temporal information and feature fusion, but also to assess the reliability of the current observation.

To address the above three issues, we propose PRMTrack, a Siamese network-based progressive reliability modeling framework for satellite video object tracking. Rather than treating these issues independently, PRMTrack constrains unreliable information across three consecutive stages: Representation, Recalibration, and Renewal. Overall, our main contributions are summarized as follows: (1) We design a Multi-Scale Sparse Perception Network (MSPN) that captures complementary object structural information through multi-scale modeling and suppresses redundant background correlations via sparse interaction, thereby alleviating template-search structural mismatch caused by aspect-ratio variations. (2) We develop an Adaptive Dual-Branch Response Recalibration Network (ADRR) that learns branch-specific Gamma mappings according to the distinct distributions of classification and centerness responses, preserving genuine object responses while suppressing distracting peaks induced by similar objects. (3) We introduce an Object-Ratio-Guided Reliable Template Renewal Mechanism (ORTR) that evaluates scale consistency using the area ratio between the current object and the historical template and adaptively constrains dynamic template renewal, thereby reducing template contamination caused by occlusion and abnormal scale estimation. (4) Extensive experiments on SatSOT, OOTB, and SV248S demonstrate the effectiveness of PRMTrack, which achieves 74.9% Precision on SatSOT.

2 Related Work

Attention and Transformer. Vaswani et al. (2017) introduced the Transformer, which models global dependencies through multi-head self-attention, while Vision Transformer (Dosovitskiy et al., 2020) further demonstrated its effectiveness in visual representation learning. In object tracking, Chen et al. (2021) employ self- and cross-attention for template-search interaction, Yan et al. (2021) adopt an encoder-decoder architecture for spatio-temporal modeling, Cui et al. (2022) unify feature extraction and object interaction through mixed attention, Xie et al. (2024) use autoregressive queries to model temporal appearance variations, and Gao et al. (2026) distinguishe the object based on semantic and geometric knowledge. For SVOT, Zhang et al. (2024) further employ adaptive multi-scale Transformers to capture object information. However, these methods mainly focus on overall scale variations and still lack targeted modeling of the feature mismatch caused by aspect-ratio variations . In contrast, our MSPN performs multi-scale structural modeling and sparse interaction to suppress redundant background correlations and reduce their interference with target representation.

Template Update. Dynamic templates continuously adapt object representations using current or historical observations. Zhang et al. (2019) learn a data-driven mapping among templates, Yang et al. (2023) introduce dynamic templates into SVOT, Wang et al. (2025) combine visual prompting with dynamic templates to adapt pretrained models, Zhou et al. (2025b) stores discriminative historical information through temporal memory, and Li et al. (2026) use multi-temporal templates to characterize long-term object states. Although dynamic templates in these methods improve adaptation to object appearance changes, they rarely consider that unreliable observations under occlusion or abnormal scale estimation may still contaminate the dynamic template. Accordingly, the proposed ORTR evaluates scale consistency using the area ratio between the current object and the historical reliable template, thereby constraining unreliable template renewal.

Satellite Video Object Tracking. Early satellite video object tracking methods mainly relied on correlation filters and motion modeling. Du et al. (2017) combine kernelized correlation filters with three-frame differencing, Shao et al. (2019) introduce velocity features to improve object-background discrimination, and Xuan et al. (2019) incorporate motion estimation and historical trajectories for more stable localization. With the development of deep learning, Ruan et al. (2022) combine Siamese representations with motion fitting. Recent methods have increasingly explored Transformer-based representations and historical information. Zhou et al. (2025b), Wang et al. (2025), Li et al. (2026), Zhang et al. (2026), Du et al. (2026), and Wang et al. (2026b) have substantially improved tracking performance. However, their adaptability to complex scenarios remains limited. Therefore, we propose a progressive reliability modeling framework across three stages, namely Representation, Recalibration, and Renewal, to improve tracking stability and robustness.

3 METHOD

We propose PRMTrack, a Siamese network-based progressive reliability modeling framework for satellite video object tracking, whose overall architecture is illustrated in Figure 1. Given the template and the search region of the current frame, a weight-sharing ResNet-50 (He et al., 2016) is first employed as the backbone for feature extraction. PRMTrack then progressively improves tracking reliability across three consecutive stages: Representation, Recalibration, and Renewal. Specifically, MSPN (Section 3.1) first enhances structural object representations across different spatial scales. At the prediction stage, ADRR (Section 3.2) performs branch-specific nonlinear recalibration according to the distinct distribution characteristics of the classification and centerness responses produced by the prediction head. Finally, after obtaining the current object prediction, ORTR (Section 3.3) further evaluates the scale consistency between the current prediction and the historical reliable template and adaptively controls dynamic template renewal accordingly.

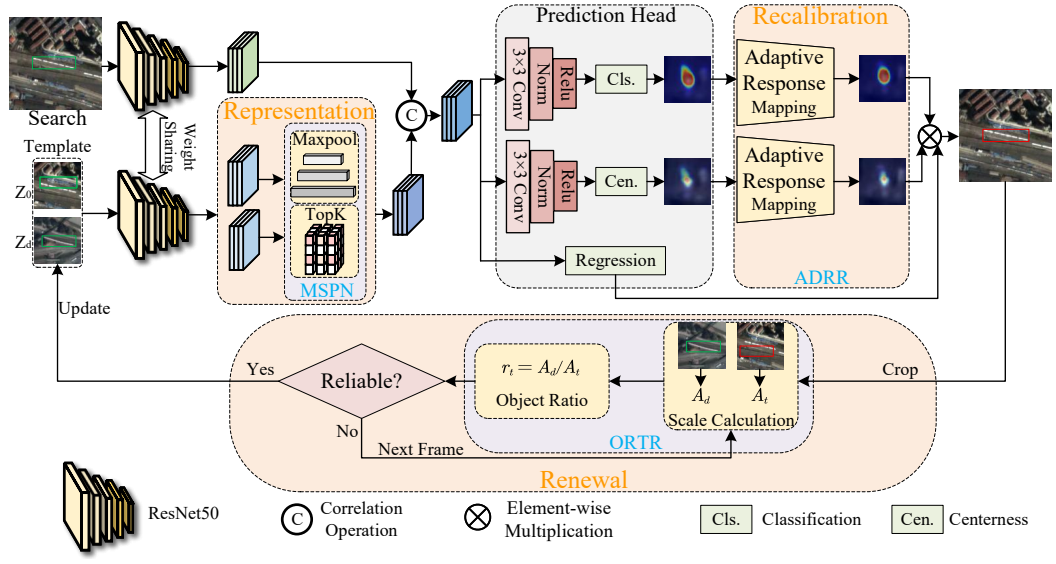


Figure 1: Overall architecture of PRMTrack.

3.1 MULTI-SCALE SPARSE PERCEPTION NETWORK

Objects in satellite videos typically occupy only a few pixels, making their effective structural information susceptible to background interference. More importantly, when slender objects such as trains and ships undergo rotation, deformation, or scale variations, their changes along the horizontal and vertical dimensions are often inconsistent, resulting in noticeable structural mismatch between the current object and the historical template. Conventional single-scale feature interaction struggles to simultaneously preserve local details and global structures, while dense attention may introduce a large number of object-irrelevant background relations. To address these issues, we propose a Multi-Scale Sparse Perception Network, as illustrated in Figure 2(a). The TopK sparse selection process is illustrated in Figure 2(b).

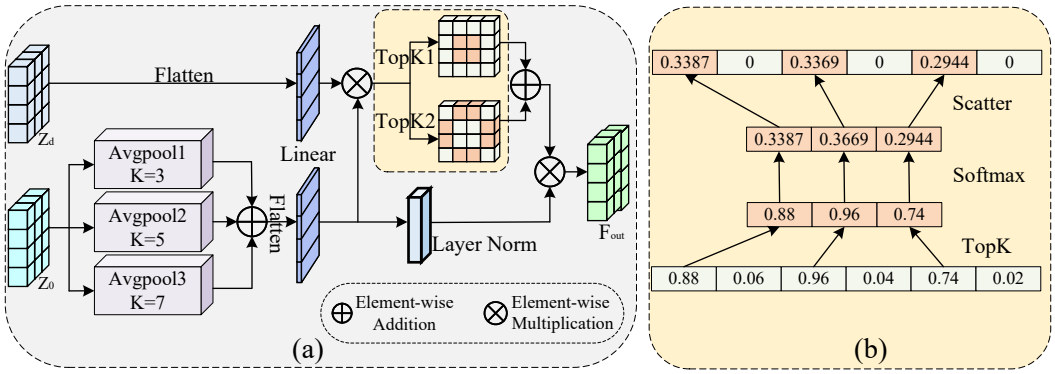


Figure 2: Architecture of the Multi-Scale Sparse Perception Network.

After feature extraction, the features of the initial template and the dynamic template are denoted as $Z_0, Z_d \in R^{Nz \times C}$, where N_z denotes the number of template tokens and C denotes the feature dimension. At initialization, both Z_0 and Z_d correspond to the template from the first frame. Let Z denote the input template feature. To capture contextual information over different ranges, one-dimensional average pooling operations with different kernel sizes are first employed to construct the multi-scale contextual representation representations $Z^{(s)}$:

$$Z^{(s)} = AvgPool_{k_s}(Z), \quad k_s \in \{3, 5, 7\} \tag{1}$$

where k_s denotes the pooling kernel size of the s-th scale branch. A smaller receptive field preserves local texture and edge information of the object, whereas a larger receptive field aggregates structural context over a broader range. The features obtained at different scales are then concatenated to form the multi-scale representation Z_{ms} :

$$Z_{ms} = Concat\left(Z^{(3)}, Z^{(5)}, Z^{(7)}\right) \tag{2}$$

compared with a single-scale representation, Z_{ms} simultaneously incorporates local and broader structural information, enabling the network to obtain complementary feature descriptions even when the object undergoes non-uniform aspect-ratio variations. The multi-scale features are subsequently linearly projected into query Q, key K, value V , and the token-wise correlation matrix A is then computed using the standard attention mechanism (Vaswani et al., 2017) :

$$Q = Z_{ms}W_Q, \quad K = Z_{ms}W_K, \quad V = Z_{ms}W_V, \quad A = QK^T/\sqrt{d} \tag{3}$$

where W_Q , W_K and W_v are learnable projection matrices, d denotes the feature dimension.For objects whose representations contain substantial background information, directly performing dense interactions among all tokens can introduce low-correlation background features into the object representation. Therefore, we further employ a TopK sparse selection strategy, retaining only highly correlated connections for each query position while setting the remaining connections to zero:

$$M_{ij}^{(b)} = \begin{cases} A_{ij}, & j \in TopK_{k_b}(A_{i,:}), \\ 0, & otherwise, \end{cases} \tag{4}$$

Different sparsity levels provide complementary interaction patterns. A higher sparsity level focuses on a small number of strongly correlated object structures, whereas a lower sparsity level preserves richer contextual information. We further introduce learnable fusion weights to adaptively combine the two sparse branches. Their normalized weights are defined as:

$$a_b = \exp(w_b) / \sum_{j=1}^{2} \exp(w_j), \qquad \sum_{b=1}^{2} a_b = 1 \tag{5}$$

where w_b is the learnable parameter associated with the b-th branch. Let F^b denote the output feature of each sparse branch. The final sparse fusion representation is obtained as $F_{sp} = \sum_{b=1}^{2} a_b F^{(b)}$. Finally, the fused representation is combined with the historical template features to obtain the final template representation F_{out} . Unlike approaches that only increase feature scales or directly employ dense attention, MSPN adjusts object information modeling along two complementary dimensions: receptive-field scale and feature-interaction sparsity. Consequently, when the object undergoes pronounced non-uniform geometric variations, MSPN can maintain more stable structural responses and provide a more reliable feature basis for subsequent template renewal and object localization.

3.2 Adaptive dual-branch response recalibration network

Similar objects and background clutter can make the genuine object response less prominent and produce multiple local response peaks in the prediction head. The classification and centerness branches describe object confidence and spatial center reliability, respectively, and exhibit distinct response distributions. Directly fusing these responses may therefore retain local spurious activations. To address this issue, we propose the Adaptive Dual-Branch Response Recalibration Network (ADRR), as illustrated in Figure 3.

Let the classification and centerness responses be denoted as $S_{cls}, S_{cen} \in R^{H \times W}$, respectively, where H and W denote the height and width. After normalization, they are expressed as:

$$\bar{S}_b = \frac{S_b - \min(S_b)}{\max(S_b) - \min(S_b) + \varepsilon}, \quad b \in \{cls, cen\} \tag{6}$$

where ε denotes a small constant introduced to prevent the denominator from becoming zero. To characterize the concentration and multi-peak uncertainty of the current response map, we first convert the normalized response of the b-th branch into a probability distribution:

$$P_{b,i} = (\bar{S}_{b,i} + \varepsilon) \Big/ \sum_{j=1}^{N} (\bar{S}_{b,j} + \varepsilon), \qquad N = H \times W \tag{7}$$

where $\bar{S}_{b,i}$ denotes the response value at the i-th position of the b-th branch, and N is the total number of response locations. The normalized response entropy is then defined as:

$$H_b = -\frac{1}{\log N} \sum_{i=1}^{N} P_{b,i} \log(P_{b,i} + \varepsilon) \tag{8}$$

where H_b denotes the response uncertainty of the current branch. Considering that the classification and centerness responses serve different functions, we construct two lightweight parameter mappings for the two branches. The adaptive Gamma parameters are defined as:

$$\gamma_{cls} = \sigma(\beta_c - Softplus(\alpha_c) H_{cls}), \gamma_{cen} = 1 + \sigma(\beta_m + Soft(\alpha_m) H_{cen}) \tag{9}$$

where $\sigma(\cdot)$ denotes the Sigmoid function, while Softplus is used to ensure that the entropy modulation coefficients remain positive. α_c, β_c, α_m, and β_m are learnable scalar parameters for the classification and centerness branches, respectively. Since the response recalibration operation and the subsequent tracking losses are differentiable, gradients can be directly propagated from the final training objective to these four parameters, allowing the network to learn appropriate recalibration strengths for different response distributions. As the response entropy increases, the concave mapping decreases γ_{cls}, while the convex mapping increases γ_{cen}, thereby strengthening classification evidence preservation and center-location constraints in the two branches, respectively. The adaptive Gamma transformation is then applied to the two branches as $\tilde{S}_{cls} = \bar{S}_{cls}^{\gamma_{cls}}$, $\tilde{S}_{cen} = \bar{S}_{cen}^{\gamma_{cen}}$.The two

transformed responses are normalized again to obtain the final recalibrated responses $\hat{S}_{\text{cls}}$and $\hat{S}_{\text{cen}}$. Combined with the scale and aspect-ratio penalty map computed from the bounding-box regression results, the final fused response R for object localization is formulated as:

$$R = P \otimes \hat{S}_{cls} \otimes \hat{S}_{cen} \tag{10}$$

where $\otimes$ denotes element-wise multiplication, and P denotes the candidate position.

Unlike directly fusing the original responses, ADRR dynamically adjusts the nonlinear mapping strength of the two prediction branches according to the response uncertainty of the current frame. By exploiting the complementary roles of the classification and centerness branches, ADRR preserves effective object responses while suppressing distracting peaks caused by backgrounds and similar objects.

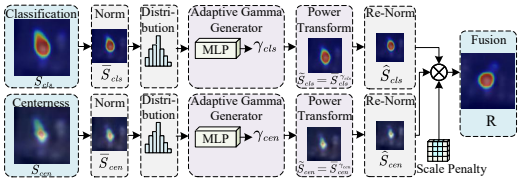


Figure 3: Architecture of the Adaptive Dual-Branch Response Recalibration Network.

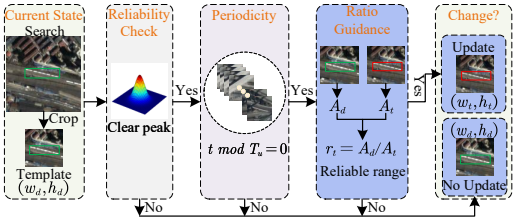


Figure 4: Object-Ratio-Guided Reliable Template Renewal Mechanism.

3.3 Object-Ratio-Guided Reliable Template Renewal Mechanism

Dynamic templates can incorporate recent object states to complement temporal appearance information, but their effectiveness highly depends on the reliability of the samples used for renewal. In satellite videos, when an object is occluded by buildings, bridges, or clouds, the predicted region may contain only a partial observation of the object, while localization drift or background interference may cause the predicted bounding box to expand abnormally. Directly updating the template with such predictions may propagate erroneous appearance information over time and eventually lead to template contamination. To address this issue, we propose the Object-Ratio-Guided Reliable Template Renewal Mechanism (ORTR), as illustrated in Figure 4. ORTR introduces an object-scale consistency constraint before dynamic template renewal, allowing only object states that satisfy the reliability conditions to participate in subsequent template updating.

Compared with strategies that renew dynamic templates solely at fixed temporal intervals, ORTR additionally constrains the template writing process according to the scale consistency between the current prediction and the historical reliable state. When occlusion or abnormal scale estimation causes a noticeable deviation in the object-area ratio, the renewal operation is actively suppressed. In this way, ORTR reduces the temporal accumulation of unreliable object information and improves the stability of dynamic templates. Please refer to Appendix B for the specific procedure.

3.4 Loss Function

We adopt a multi-task learning strategy and follow the loss design (Guo et al., 2020). Specifically, the classification branch employs a cross-entropy loss L_{cls} to distinguish the object from the background, the centerness branch uses a binary cross-entropy loss L_{cen} to supervise centerness prediction for positive samples within the object region, and the regression branch adopts an IoU loss L_{reg} to constrain the spatial overlap between the predicted and ground-truth bounding boxes. The total loss is defined as :

$$L = \lambda_{cls} L_{cls} + \lambda_{cen} L_{cen} + \lambda_{reg} L_{reg} \tag{11}$$

where λ_{cls}, λ_{cen}, and λ_{reg} denote the weights of the classification, centerness, and bounding-box regression losses, respectively, and are set to 1, 1, and 3 in our experiments.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Our method is implemented using Python 3.8.20 and PyTorch 1.13.0, and the model is trained on two NVIDIA RTX 4090 GPUs. The network is initialized with pretrained weights (Guo et al., 2020). During tracking, the template region is resized to 127×127, while the search region is resized to 255×255. The experimental results reported in the following sections are primarily based on our reproduced results.

Training data. Like most trackers,the training data consist of ImageNet VID (Russakovsky et al., 2015), LaSOT (Fan et al., 2019), and GOT-10K (Huang et al., 2019).

Test data. We evaluate the performance of the tracking methods on the following datasets:

- **SatSOT.**(Zhao et al., 2022) SatSOT contains 105 video sequences with 27,664 frames, covering typical object categories such as vehicles, trains, airplanes, and ships. The dataset provides frame-by-frame annotations using horizontal bounding boxes and defines 11 challenging attributes
- **SV248S.**(Li et al., 2022) SV248S is a relatively large-scale satellite video object tracking dataset with diverse challenge attributes. It is constructed from six satellite videos and contains 248 tracking sequences.
- **OOTB.**(Chen et al., 2024a) OOTB contains 110 sequences with 29,890 frames and employs oriented bounding boxes (OBBs) for annotation.

Evaluation Metrics. We adopt precision(Prec.) and success rate(Succ.), two commonly used metrics in object tracking, as the primary evaluation criteria. Precision measures the localization error between the predicted object center and the ground-truth object center. Considering that objects in satellite videos are typically small, we use Precision at a 5-pixel threshold as the primary precision metric, where a higher value indicates more accurate object localization. Success Rate evaluates performance according to the overlap between the predicted and ground-truth bounding boxes. A higher Success Rate indicates that the tracker produces accurate object bounding-box predictions.

Table 1: Comparison with SOTA on SatSOT.

Method	Source	Prec.	Succ.
Ours	–	**74.9**	**53.0**
HTTrack	Wang et al. (2026b)	65.0	<u>52.9</u>
TAGET	Gao et al. (2026)	65.6	49.5
SPTrack	Zhang et al. (2026)	65.6	49.9
MTTrack	Du et al. (2026)	68.7	47.9
AVPTrack	Wang et al. (2025)	65.7	49.3
SiamTITP	Zhou et al. (2025a)	72.2	51.2
EHTracker	Yang et al. (2024)	61.6	48.2
ORTrack	Wu et al. (2025)	69.1	48.3
SiamS^2F	Lv et al. (2025)	62.1	48.6
HIPTrack	Cai et al. (2024)	61.9	42.3
MACF	Lin et al. (2024)	<u>73.1</u>	49.4
SeqTrack	Chen et al. (2023)	65.9	43.7
GRM	Gao et al. (2023)	60.9	39.9
TCTrack	Cao et al. (2022)	57.9	32.1
MixFormer-L	Cui et al. (2022)	52.8	37.4

Table 2: Comparison with SOTA on SV248S.

Method	Source	Prec.	Succ.
Ours	–	**89.6**	**55.4**
HTTrack	Wang et al. (2026b)	80.9	51.5
TAGET	Gao et al. (2026)	77.1	48.6
SPTrack	Zhang et al. (2026)	75.8	47.4
SGLATrack	Xue et al. (2025)	40.9	18.6
AVPTrack	Wang et al. (2025)	82.5	52.4
SiamTITP	Zhou et al. (2025a)	<u>87.6</u>	<u>54.0</u>
EHTracker	Yang et al. (2024)	74.3	49.0
ORTrack	Wu et al. (2025)	38.2	17.5
SiamS^2F	Lv et al. (2025)	76.3	50.6
HIPTrack	Cai et al. (2024)	36.7	15.4
SVLPNet	Wang et al. (2024)	76.9	47.1
SeqTrack	Chen et al. (2023)	43.4	16.0
GRM	Gao et al. (2023)	28.8	14.1
TCTrack	Cao et al. (2022)	45.4	13.3
MixFormer-L	Cui et al. (2022)	37.4	14.6

4.2 COMPARISON EXPERIMENTS

To comprehensively evaluate PRMTrack in satellite video object tracking tasks, we compared it against a series of recent state-of-the-art (SOTA) methods.

SatSOT. As shown in Table 1, PRMTrack achieves a Precision of 74.9% and a Success Rate of 53.0%, ranking first in both metrics. Compared with MACF, which achieves the second-best Precision, PRMTrack improves Precision by 1.8%. Compared with the recent tracker HTTrack, PRMTrack improves Success Rate and Precision by 0.1% and 9.9%, respectively. Meanwhile, relative to the Baseline, the two metrics are improved by 2.7% and 1.8%, respectively.

SV248S. As shown in Table 2, PRMTrack achieves a Precision of 89.6% and a Success Rate of 55.4%, outperforming the Baseline by 2.0% and 1.4%, respectively. In addition, compared with the recent AVPTrack, PRMTrack improves the precision by 7.1% . These results further demonstrate the stable tracking performance of PRMTrack under complex interference conditions.

Table 3: Comparison with SOTA on OOTB.

Method	TCTrack	MixFormer-L	SeqTrack	GRM	TATrans	HIPTrack	AQATrack	SiamTITP	SGLATrack	ORTrack	SiamS^{2}F	SPTrack	Ours
Source	Cao et al. (2022)	Cui et al. (2022)	Chen et al. (2023)	Gao et al. (2023)	Lai et al. (2023)	Cai et al. (2024)	Xie et al. (2024)	Zhou et al. (2025a)	Xue et al. (2025)	Wu et al. (2025)	Lv et al. (2025)	Zhang et al. (2026)	–
Prec.	57.9	52.8	56.6	64.5	66.4	61.9	87.4	<u>87.9</u>	70.9	66.8	80.7	83.4	**89.2**
Succ.	32.1	37.4	39.8	50.0	51.8	42.3	59.9	<u>63.2</u>	54.2	52.4	48.6	61.2	**64.1**

OOTB. As shown in Table 3, PRMTrack achieves a Precision of 89.2% and a Success Rate of 64.1%. Compared with SiamTITP, which delivers the closest performance, PRMTrack improves the two metrics by 1.3% and 0.9%, respectively. These results indicate that PRMTrack adapts well to the pronounced geometric and appearance variations in the OOTB dataset.

4.3 Ablation Studies

To verify the effectiveness and complementarity of the proposed components in PRMTrack, we conduct ablation studies on SatSOT and three representative challenging attributes, namely Aspect Ratio Change, Full Occlusion, and Similar Object. MSPN, ADRR, and ORTR are introduced separately, with the results reported in Table 4.

Table 4: Ablation study on SatSOT and three challenging attributes: Aspect Ratio Change, Full Occlusion, and Similar Object.

Components			SatSOT		Aspect Ratio Change		Full Occlusion		Similar Object	
MSPN	ADRR	ORTR	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.
			72.2	51.2	21.0	27.8	78.5	47.0	81.8	52.0
✓			73.0	51.7	23.6	29.3	79.7	47.8	81.5	51.7
	✓		73.7	51.8	23.8	28.5	76.1	45.3	84.9	53.7
		✓	72.7	51.7	21.5	28.4	80.6	48.5	81.0	51.6
✓		✓	73.1	51.9	23.7	29.6	80.8	48.8	81.9	51.9
✓	✓		74.2	52.0	25.4	29.3	76.0	45.3	84.8	53.7
	✓	✓	73.9	52.2	24.7	29.3	79.2	47.9	81.9	52.0
✓	✓	✓	**74.9**	**53.0**	**26.2**	**31.2**	**82.1**	**49.8**	**85.0**	**53.9**

In terms of overall performance, all three components bring consistent improvements over the Baseline. With MSPN, ADRR, and ORTR individually introduced, the Precision/Success Rate reaches 73.0%/51.7%, 73.7%/51.8%, and 72.7%/51.7%, respectively. Combining MSPN and ORTR yields 73.1%/51.9%, while the complete PRMTrack achieves 74.9%/53.0%, improving over the Baseline by 2.7% and 1.8%, respectively.

On the Aspect Ratio Change attribute, introducing MSPN improves Precision and Success Rate by 2.6% and 1.5%, respectively. The complete PRMTrack further improves them by 5.2% and 3.4% over the Baseline, indicating that multi-scale contextual modeling and sparse feature interaction enhance structural object representation under aspect-ratio variations.

On the Similar Object attribute, ADRR improves Precision and Success Rate from 81.8%/52.0% to 84.9%/53.7%, corresponding to gains of 3.1% and 1.7%, respectively. The complete PRMTrack further reaches 85.0%/53.9%, showing that adaptive response recalibration effectively suppresses distracting peaks induced by similar objects.

On the Full Occlusion attribute, ORTR improves Precision and Success Rate by 2.1% and 1.5%, respectively. Combining ORTR with MSPN yields 80.8%/48.8%, while the complete PRMTrack reaches 82.1%/49.8%, outperforming the Baseline by 3.6% and 2.8%, respectively. These results indicate that reliable template renewal can reduce the accumulation of unreliable information under full occlusion.

4.4 Qualitative Analysis

We conduct a qualitative analysis on SatSOT from three perspectives: frame-by-frame localization, IoU variation, and internal feature and response visualization. Figure 5 presents the tracking results of PRMTrack and the Baseline on representative sequences together with the corresponding IoU curves, where a higher IoU indicates more accurate localization. In the car_01 sequence, after an occlusion occurs at around frame 464, the IoU of the Baseline drops noticeably, whereas PRMTrack maintains relatively stable localization, indicating that reliability-constrained dynamic template renewal helps reduce the accumulation of unreliable information under occlusion. In the car_43 sequence, complex backgrounds and nearby similar objects cause evident localization drift for the Baseline, while PRMTrack continues to follow the true object. In the train_01 and train_20 sequences, PRMTrack exhibits smoother IoU curves and better localization consistency under pronounced scale and aspect-ratio variations. These results demonstrate that PRMTrack provides more stable continuous tracking under occlusion, similar-object interference, and geometric variations.

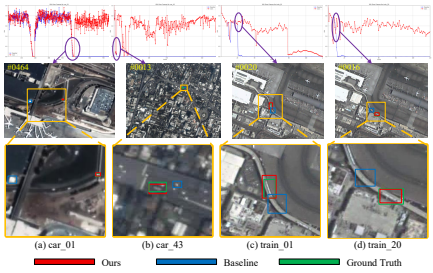


Figure 5: Comparison of IoU curves.

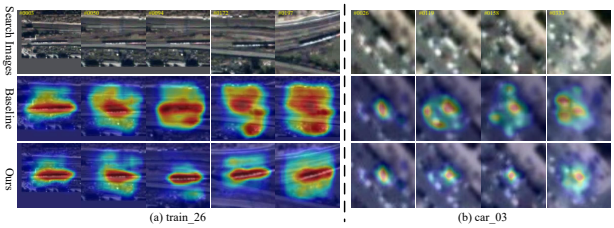


Figure 6: Comparison of tracking processes.

Figure 6 further visualizes the internal feature and response changes. Figure 6(a) shows the heatmaps before and after introducing MSPN. In the train_26 sequence, as the object undergoes rotation and aspect-ratio variation, the Baseline gradually exhibits dispersed attention and stronger background interference, whereas MSPN keeps the responses more concentrated on the true object and maintains a relatively stable attention region throughout the appearance changes, indicating that multi-scale contextual modeling and sparse feature interaction enhance structural object representation. Figure 6(b) shows the response maps before and after applying ADRR. In the car_03 sequence, similar objects tend to produce multiple competing response peaks. After adaptive nonlinear recalibration of the classification and centerness responses, the dominant response around the true object becomes more concentrated, while distracting responses in non-object regions are clearly suppressed, resulting in a sharper response peak and stronger object

5 Conclusion

This work presents PRMTrack, a Siamese network-based progressive reliability modeling framework for satellite video object tracking, to address structural representation mismatch, response ambiguity, and dynamic template contamination. Rather than improving these components independently, PRMTrack treats tracking reliability as a progressive process across Representation, Recalibration, and Renewal. MSPN strengthens structural representations under non-uniform geometric variations, ADRR recalibrates heterogeneous classification and centerness responses to reduce distracting peaks, and ORTR prevents unreliable predictions from contaminating dynamic templates. Extensive experiments on SatSOT, OOTB, and SV248S demonstrate strong and stable tracking performance. More importantly, these results indicate that improving satellite video tracking requires not only stronger representations or temporal modeling, but also controlling how unreliable information is generated and propagated throughout the tracking pipeline. This perspective provides a unified way to improve robustness against geometric variations, similar-object interference, and occlusion. Future work will further explore reliable long-term temporal modeling under prolonged full occlusion and more complex object-state variations.

REFERENCES

Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pp. 850–865. Springer, 2016.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19258–19267. IEEE, 2024.

Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14778–14788. IEEE, 2022.

Shih-Fang Chen, Jun-Cheng Chen, I hong Jhuo, and Yen-Yu Lin. GOT-edit: Geometry-aware generic object tracking via online model editing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aVa7etWnwF.

Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8122–8131. IEEE, 2021.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14572–14581. IEEE, 2023.

Yuzeng Chen, Yuqi Tang, Yi Xiao, Qiangqiang Yuan, Yuwei Zhang, Fengqing Liu, Jiang He, and Liangpei Zhang. Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:212–240, 2024a.

Yuzeng Chen, Yuqi Tang, Qiangqiang Yuan, and Liangpei Zhang. Reps: Rotation equivariant siamese network enhanced by probability segmentation for satellite video tracking. *International Journal of Applied Earth Observation and Geoinformation*, 128:103741, 2024b.

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.

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

Bo Du, Yujia Sun, Shihan Cai, Chen Wu, and Qian Du. Object tracking in satellite videos by fusing the kernel correlation filter and the three-frame-difference algorithm. *IEEE Geoscience and Remote Sensing Letters*, 15(2):168–172, 2017.

Guocai Du, Peiyong Zhou, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. Mttrack: A joint mamba-transformer framework with memory enhancement for real-time satellite remote sensing video object tracking. *Knowledge-Based Systems*, pp. 115572, 2026.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5369–5378. IEEE, 2019.

Long Gao, Haijun Chen, Xunfeng Zhao, Gang He, Yunsong Li, and Jifeng Ning. Target-aware gaussian distribution estimation for satellite video object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18686–18695. IEEE, 2023.

Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6268–6276. IEEE, 2020.

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*, pp. 770–778, 2016.

Yibiao Hu, You Zhou, Zhengqiang Zhu, Xi Yang, Han Zhang, Kun Bian, and Hong Han. Llvm-drone: A synergistic framework integrating large language models and vision models for visual tasks in unmanned aerial vehicles. *Knowledge-Based Systems*, pp. 114190, 2025.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019.

Pujian Lai, Meili Zhang, Gong Cheng, Shengyang Li, Xiankai Huang, and Junwei Han. Target-aware transformer for satellite video object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–10, 2023.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4277–4286. IEEE, 2019.

Yitong Li, Chen Wu, and Jia Shao. Tstrans: Temporal-sequence-driven transformer for single object tracking in satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Yuxuan Li, Licheng Jiao, Zhongjian Huang, Xin Zhang, Ruohan Zhang, Xue Song, Chenxi Tian, Zixiao Zhang, Fang Liu, Shuyuan Yang, et al. Deep learning-based object tracking in satellite videos: A comprehensive survey with a new dataset. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):181–212, 2022.

Bin Lin, Jinlei Zheng, Chaocan Xue, Lei Fu, Ying Li, and Qiang Shen. Motion-aware correlation filter-based object tracking in satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Yuwei Lu and Hongming Peng. Positive-incentive noise guided adaptive object tracking via motion state transition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

Moses Lurbur, Katherine C Wilson, and Noëlle Yochum. Towards automated bycatch monitoring: Optimizing and evaluating multi-object tracking of salmon in pollock trawls. *Ecological Informatics*, pp. 103674, 2026.

Pengyuan Lv, Xianyan Gao, and Yanfei Zhong. Siams^{2}f: Satellite video single-object tracking based on a siamese spectral-spatial-frame correlation network. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

Fredy Nocua, Wilson-Javier Perez-Holguin, and Camilo Pardo-Beainy. Urban traffic monitoring based on deep learning on an embedded gpu. *Expert Systems with Applications*, 273:126847, 2025.

Lu Ruan, Yujia Guo, Daiqin Yang, and Zhenzhong Chen. Deep siamese network with motion fitting for object tracking in satellite videos. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Jia Shao, Bo Du, Chen Wu, and Lefei Zhang. Tracking objects from satellite videos: A velocity feature based correlation filter. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10): 7860–7871, 2019.

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

Guijie Wang, Tong Lin, Yifan Bai, Anjia Cao, Shiyi Liang, Wangbo Zhao, and Xing Wei. FARTrack: Fast autoregressive visual tracking with high performance. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=lq7Zfr8kAS.

Jiahao Wang, Fang Liu, Licheng Jiao, Yingjia Gao, Hao Wang, Lingling Li, Puhua Chen, Xu Liu, and Shuo Li. Satellite video object tracking based on location prompts. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6253–6264, 2024.

Jiahao Wang, Fang Liu, Licheng Jiao, Hao Wang, Shuo Li, Yanbiao Ma, Lingling Li, Puhua Chen, Xu Liu, and Mengjia Wang. Adaptive visual prompting for effective satellite video tracking. *IEEE Transactions on Multimedia*, 2025.

Jiahao Wang, Fang Liu, Licheng Jiao, Hao Wang, Shuo Li, Xinyi Wang, Lingling Li, Puhua Chen, and Xu Liu. Httrack: Learning to perceive targets via historical trajectories in satellite video tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9856–9866, 2026b.

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17103–17113. IEEE, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19300–19309. IEEE, 2024.

Shiyu Xuan, Shengyang Li, Mingfei Han, Xue Wan, and Gui-Song Xia. Object tracking in satellite videos by improved correlation filters with motion estimations. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):1074–1086, 2019.

Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. Similarity-guided layer-adaptive vision transformer for uav tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6730–6740. IEEE, 2025.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 10428–10437. IEEE, 2021.

Jianwei Yang, Zongxu Pan, Ziming Wang, Bin Lei, and Yuxin Hu. Siammdm: An adaptive fusion network with dynamic template for real-time satellite video single object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023.

Jianwei Yang, Yuhan Liu, Yanxing Liu, Ziming Wang, Jiawei Li, Guangyao Zhou, Wenzhi Wang, and Yuxin Hu. Ehtracker: Toward fine-grained localization for satellite video target tracking. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:5582–5599, 2024.

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.

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*, pp. 4010–4019, 2019.

Xiaowen Zhang, Haijiang Sun, Hanqing Sun, Qiaoyuan Liu, Xinglong Sun, and Xinyi Yao. Synergistic frequencyspatial enhancement with temporal correlation for robust satellite video tracking. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2026.

Xin Zhang, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Wenping Ma, and Shuyuan Yang. Adaptive multi-scale transformer tracker for satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Manqi Zhao, Shengyang Li, Shiyu Xuan, Longxuan Kou, Shuai Gong, and Zhuang Zhou. Satsot: A benchmark dataset for satellite video single object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.

Jiawei Zhou, Yanni Dong, and Bo Du. Siam titp: Incorporating temporal information and trajectory prediction siamese network for satellite video object tracking. *IEEE Transactions on Image Processing*, 2025a.

Jiawei Zhou, Yanni Dong, Yuxiang Zhang, and Bo Du. Incorporating prior knowledge and temporal memory transformer network for satellite video object tracking. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:630–647, 2025b.

A Appendix

This appendix provides additional details of PRMTrack, including the workflow of the Object-Ratio-Guided Reliable Template Renewal Mechanism (ORTR), comparisons with state-of-the-art trackers across all SatSOT attributes, further experimental details, and qualitative comparisons of tracking results.

B The Workflow of ORTR

Let the predicted bounding box at frame t be $B_t = (x_t, y_t, w_t, h_t)$, where (x_t, y_t) denotes the object location, and w_t and h_t denote the width and height of the predicted bounding box, respectively. The area of the current predicted object is defined as $A_t = w_t h_t$. Let $\left(w_t^d, h_t^d\right)$ denote the bounding-box size associated with the most recent reliable template renewal, and its corresponding reference area be $A_t^d = w_t^d h_t^d$. To reduce the influence of absolute object scale, we measure scale consistency using the area ratio between the current prediction and the historical reliable state:

$$r_t = \frac{A_t}{A_t^d} = \frac{w_t h_t}{w_t^d h_t^d} \tag{12}$$

where r_t represents the relative area variation of the current prediction with respect to the historical reliable state. When r_t is close to 1, the two states exhibit high scale consistency. When occlusion or abnormal prediction occurs, the predicted bounding box may undergo noticeable contraction or expansion. Therefore, compared with directly using absolute pixel area, the area ratio provides a more stable characterization of relative state variations across objects of different scales. Based on statistical analysis, we set the reliable ratio interval to [0.7,1.5], with the lower and upper bounds denoted by τ_l=0.7 and τ_u=1.5, respectively. The corresponding object-ratio reliability criterion is defined as:

$$g_t^r = I\left(\tau_l \le r_t \le \tau_u\right) \tag{13}$$

where $I(\cdot)$ denotes the indicator function. When r_t falls within the reliable interval,g_t^r=1; otherwise, g_t^r=0, indicating that the current prediction does not satisfy the ratio-based reliability constraint.

Since scale consistency alone cannot guarantee that the current object state is reliable, ORTR further combines the area-ratio constraint with object visibility and periodic renewal conditions. Let $o_t \in \{0, 1\}$,where o_t=0 indicates that the object is considered visible, whereas o_t=1 indicates occlusion. The visibility state is determined according to the response-peak status. Let T_u denote the fixed template renewal interval. The final renewal decision is then defined as:

$$u_t = I\left(o_t = 0\right) \cdot I\left((t) \ mod \ T_u = 0\right) \cdot g_t^r \tag{14}$$

where t denotes the current frame index. Only when the object is not classified as occluded, the current frame satisfies the periodic renewal condition, and the area ratio meets the reliability constraint,

do we obtain u_t=1. Let $\Phi(I_t, B_t)$ denote the template region cropped from the current frame I_t according to the predicted bounding box B_t. The dynamic template renewal process is formulated as

$$z_t^d = \begin{cases} \Phi(I_t, B_t), & u_t = 1, \\ z_{t-1}^d, & u_t = 0. \end{cases} \tag{15}$$

where z_t^d and z_{t-1}^d denote the dynamic templates maintained at the current and previous time steps, respectively. Once the template is successfully renewed, the corresponding w_t and h_t are also stored as the reliable scale reference for subsequent area-ratio computation. Otherwise, the previous reliable dynamic template and its associated scale information are retained.

C Experimental Implementation Details

Our model is trained for 20 epochs, with 600,000 training samples in each epoch and a batch size of 96. We employ stochastic gradient descent (SGD) as the optimizer, with the momentum and weight decay set to 0.9 and 0.0001, respectively. A warm-up learning-rate strategy is adopted during training. The initial learning rate is set to 0.001 and gradually increased to 0.005 over the first 5 epochs, after which it is progressively decayed to 0.0005. To fully exploit the representation capability of the pretrained model while maintaining stable optimization, the backbone parameters are frozen during the first 10 epochs. During the remaining 10 epochs, the last three stages of the backbone are unfrozen and jointly optimized with the other network modules.

D Comparison with SOTA trackers across all attributes of SatSOT.

To further evaluate the robustness of PRMTrack under different challenging conditions, we conduct a fine-grained comparison across all challenge attributes of SatSOT, including Background Clutter (BC), Illumination Variation (IV), Low Quality (LQ), Rotation (ROT), Partial Occlusion (POC), Full Occlusion (FOC), Tiny Object (TO), Similar Object (SOB), Background Jitter (BJT), Aspect Ratio Change (ARC), and Deformation (DEF). The corresponding Precision and Success Rate results are shown in Figure 7 and Figure 8, respectively. Overall, PRMTrack demonstrates strong competitiveness across most challenge attributes. Specifically, it ranks first in Precision under FOC and ROT and second under TO, BC, LQ, POC, and SOB. In terms of Success Rate, PRMTrack achieves the best results on seven attributes, including BC, BJT, FOC, LQ, POC, ROT, and SOB, demonstrating strong stability across diverse challenging conditions.

For occlusion-related attributes, PRMTrack achieves 82.1% Precision and 49.8% Success Rate under FOC, ranking first in both metrics. Under POC, it reaches 63.7% Precision and 46.7% Success Rate, ranking second and first, respectively. These results are consistent with the proposed Object-Ratio-Guided Reliable Template Renewal Mechanism (ORTR). When the current observation is affected by partial or full occlusion, constraining abnormal object states from entering the dynamic template helps reduce the continuous accumulation of unreliable appearance information in subsequent frames. Under the SOB attribute, where interference from similar objects is particularly severe, PRMTrack achieves 85.0% Precision and 53.9% Success Rate, ranking second and first, respectively. This indicates that ADRR can enhance the discriminability of genuine object responses and suppress competing distracting peaks by adaptively performing nonlinear recalibration on the classification and centerness responses. In addition, PRMTrack achieves the best performance under ROT, with 73.5% Precision and 50.0% Success Rate, suggesting that the multi-scale contextual modeling and sparse feature interaction in MSPN improve representation stability under geometric structural variations. For ARC, which more directly reflects aspect-ratio variations, PRMTrack does not achieve the best ranking, but still shows clear improvements over the Baseline, indicating that the proposed multi-scale structural modeling contributes positively to handling non-uniform object appearance changes.

It is worth noting that PRMTrack still exhibits a certain performance gap compared with the best-performing methods under ARC, DEF, and IV. In particular, when objects undergo severe deformation, substantial aspect-ratio variations, or abrupt imaging-condition changes, the current multi-scale structural modeling may still be insufficient to fully characterize complex geometric and appearance

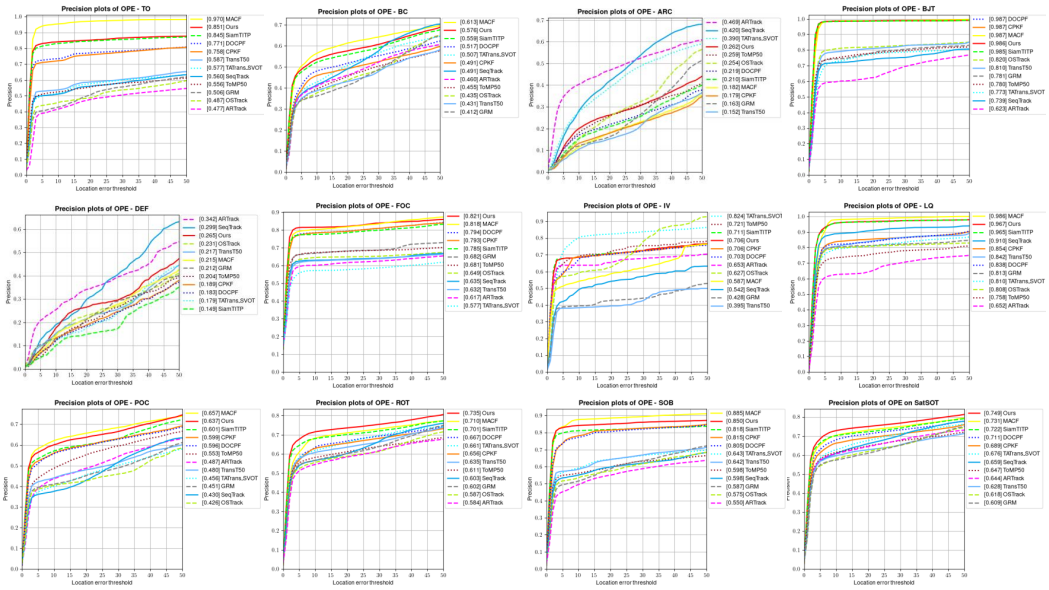


Figure 7: Precision comparison of SatSOT across all attributes

variations. Future work will explore more direction- and deformation-aware structural representations and incorporate long-term temporal information to model abrupt appearance changes, further improving tracking robustness in extreme satellite video scenarios.It should be noted that the results of methods without publicly available source code in Table 1 are not shown in Figure 7 and Figure 8.

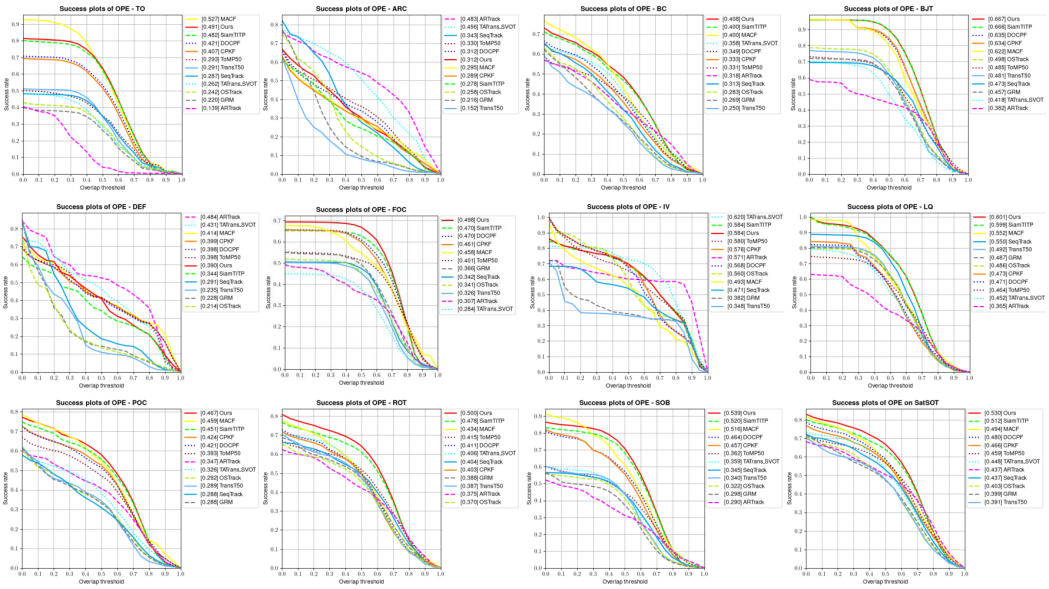


Figure 8: Success rate comparison of SatSOT across all attributes.

E VISUALIZATION AND DISCUSSION OF TRACKING RESULTS

To further provide an intuitive comparison of different trackers in complex satellite video scenarios, we select three representative sequences from SatSOT, namely Car_58, Ship_01, and Train_04, and

visualize the tracking results of PRMTrack together with three representative trackers, SiamTITP, DOCPF, and GRM, as shown in Figure 9. In the Car_58 sequence, multiple objects with similar appearances exist around the tracked object, while persistent occlusion occurs during motion. As the interference becomes more severe, the compared methods gradually exhibit localization drift or even lose the object, whereas PRMTrack is still able to maintain stable tracking of the true object across consecutive frames. This result indicates that, when the current observation is simultaneously affected by occlusion and similar-object interference, more reliable object representation, dynamic template renewal, and response discrimination can help reduce the accumulation of erroneous information during continuous tracking.

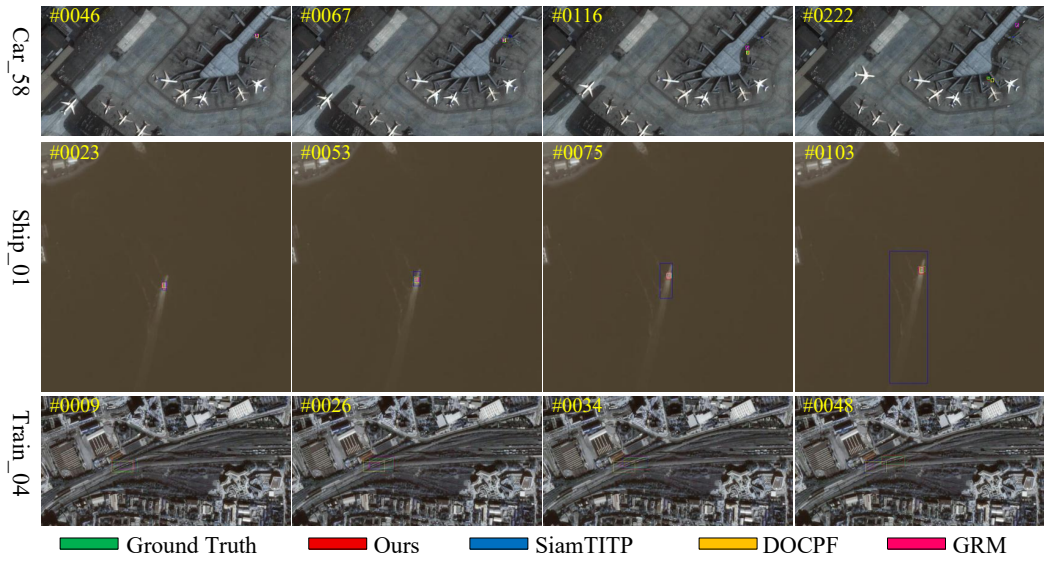


Figure 9: Visual comparison of tracking results.

In the Ship_01 sequence, the object is relatively small and shows weak appearance contrast against the surrounding water surface, while the wake generated by the moving ship introduces additional background interference. It can be observed that the predicted boxes of some compared methods are gradually affected by background clutter and deviate from the true object, whereas the predictions of PRMTrack remain highly consistent with the ground-truth location. This result demonstrates that PRMTrack can maintain stronger object discrimination under low-contrast and complex-background conditions, thereby reducing the influence of background interference on object localization.

The Train_04 sequence illustrates another typical challenge faced by slender objects in satellite videos. As the train enters a curved track, its motion direction, scale, and aspect ratio change noticeably, which can easily cause structural representation mismatch between the template and the current object. Compared with the other methods, the predicted boxes of PRMTrack adapt better to the continuous geometric changes of the train and maintain higher consistency with the ground-truth bounding boxes across several key frames, demonstrating stronger adaptability to geometric appearance variations.

F THE USE OF LARGE LANGUAGE MODELS

We use Large Language Models only to a minimal extent to assist with or polish the writing.