AURORA-Track: Uncertainty-Aware Identity Prediction for Robust Multi-Object Tracking in Satellite Video

Xin Huang[1], Pei Mi[1], MaoYu Wang[1], XuLei Shi[1], Wei Qin[1], PengJie Tao[1]

[1]School of Remote Sensing and Information Engineering, Wuhan University

Keywords: Satellite Video Tracking, Multi-Object Tracking, Uncertainty-Aware Identity Prediction, Occlusion-Aware Modelling, Cross-Scene Adaptation.

Abstract

Satellite video multi-object tracking remains difficult due to tiny targets, frequent cloud/shadow occlusion, and strong cross-scene domain shifts. To address these issues, we propose AURORA-Track, an end-to-end framework built on MOTIP with three coupled components: uncertainty-aware identity prediction (UAI), cloud/shadow-aware trajectory modeling (CAT), and cross-scene adaptation (CSA). UAI calibrates association confidence to reduce unreliable identity assignments; CAT combines visibility estimation with motion-state recovery to maintain trajectories through prolonged occlusions; CSA uses lightweight FiLM/LoRA adaptation for improved generalization under geographic variation. Experiments on VISO, SAT-MTB, and AIR-MOT-100 show consistent gains in HOTA and IDF1 with fewer identity switches compared with strong baselines. The results indicate that coupling uncertainty estimation, visibility-aware motion reasoning, and scene adaptation improves robustness for satellite video tracking.

1. Introduction

Satellite video-based multi-object tracking (MOT) has emerged as a critical technology for large-scale remote sensing, enabling continuous monitoring for applications such as traffic management and disaster response (Zhang et al., 2020, **?**). While traditional ground-based MOT has matured significantly, satellite video presents unique hurdles: targets typically occupy minimal spatial extent (e.g., 5×5 pixels), spatial resolution is constrained, and atmospheric interferences like cloud cover and shifting shadows are pervasive (Zhang et al., 2021, Chen et al., 2023).

Despite the success of the Multiple Object Tracking as ID Prediction (MOTIP) paradigm (Wang et al., 2024) in general video, its direct application to the satellite domain remains suboptimal due to three fundamental bottlenecks. First, the low-discriminative visual features of small targets often lead identity decoders to produce overconfident but erroneous associations. Second, frequent visibility degradation from clouds and shadows causes prolonged temporal gaps that traditional motion models struggle to bridge. Third, the significant domain shifts between heterogeneous geographic regions (e.g., urban vs. maritime) hinder the generalization of pre-trained trackers.

To address these challenges, we propose **AURORA-Track** (Adaptive Uncertainty-aware Remote-sensing Association Tracking), an end-to-end framework built upon the MOTIP backbone. Our approach introduces three tightly coupled innovations designed to enhance robustness in satellite scenarios. Specifically, we develop an **Uncertainty-aware ID Prediction (UAI)** module that augments the identity decoder with probabilistic modeling, enabling the system to suppress ambiguous associations and provide calibrated confidence scores. To maintain trajectory continuity, we introduce **Cloud-aware Trajectory Modeling (CAT)**, which explicitly encodes transient visibility degradation and leverages long-term motion context to recover identities after prolonged occlusions. Finally, a **Cross-scene Knowledge Transfer (CSA)** branch is designed to perform rapid adaptation across diverse geographic domains using meta-learned priors, ensuring stable performance across varying satellite sensing environments.

The main contributions of this work are summarized as follows:

- We adapt the MOTIP framework for satellite video tracking, providing an end-to-end perspective that eliminates the complexities of separate detection-association paradigms.
- We introduce three modules—UAI, CAT, and CSA—that target identity ambiguity, persistent occlusion, and scene heterogeneity in remote sensing.
- Extensive evaluations on VISO, SAT-MTB, and AIR-MOT-100 demonstrate consistent improvements in identity consistency and overall tracking performance under challenging atmospheric conditions.

The remainder of this paper is organized as follows: Section 2 introduces the materials, workflow, and implementation protocol. Section 3 reports results. Section 4 discusses novelty, comparisons, and deployment implications. Section 5 concludes the paper.

2. Material and methods

2.1 AURORA-Track Overview

The overview of the proposed method AURORA-Track is shown in Fig. 1. Each frame is processed by a Deformable DETR-based detector(Zhu et al., 2021) within the MOTIP backbone to extract per-detection features and boxes, while a trajectory memory provides temporal context for identity decoding. The decoder queries this memory under causal masking to predict identities, and its outputs are calibrated by an uncertainty head for robust association. In parallel, a cloud-aware branch predicts per-detection visibility and drives a motion filter that maintains kinematic priors, inflates process noise when visibility deteriorates, and enables recovery after occlusions. To mitigate

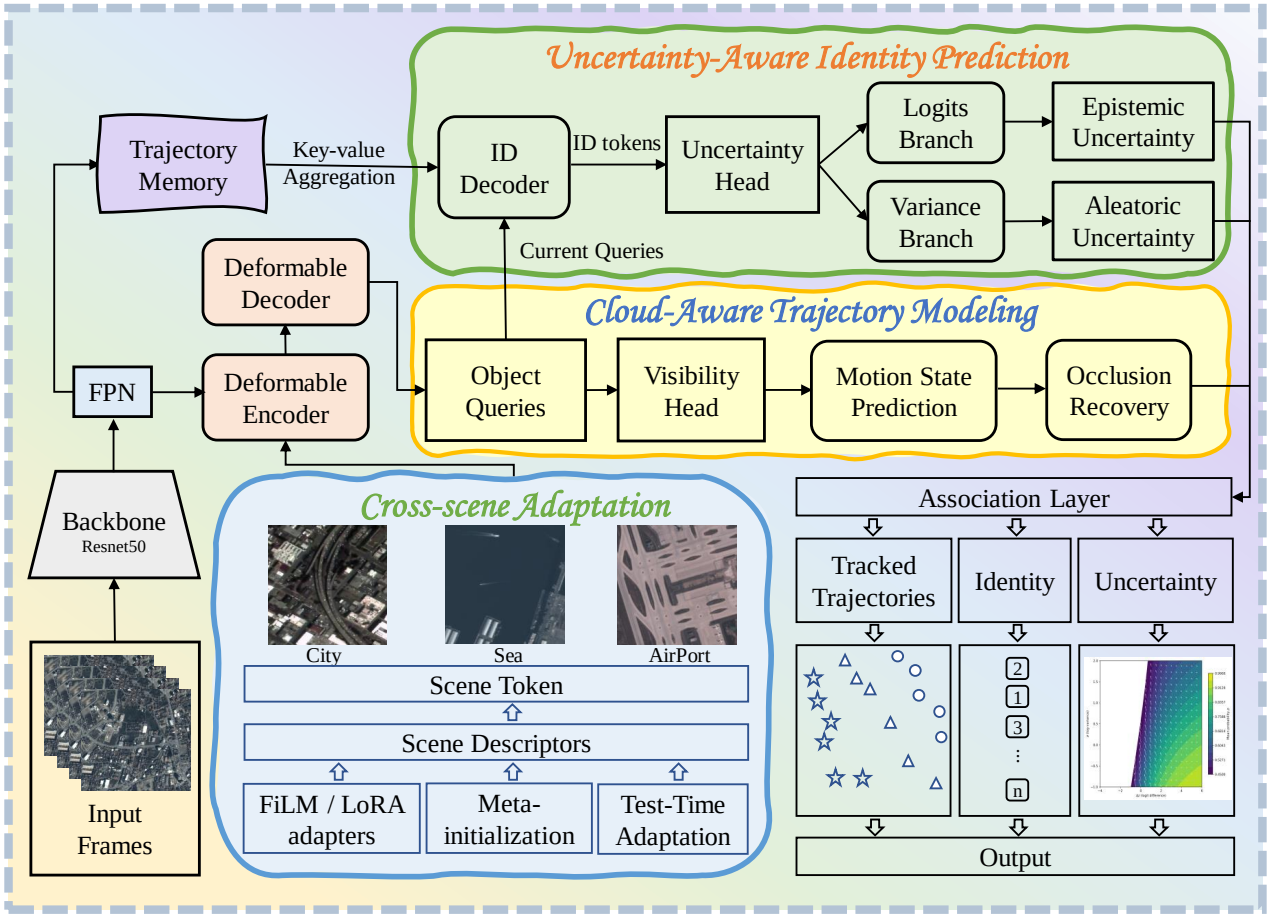

Figure 1. The overview of the proposed method AURORA-Track

geographic and seasonal shifts, a parameter-efficient adaptation module conditions intermediate features on a compact scene descriptor and is updated with a small number of optimization steps; an adversarial alignment term optionally regularizes feature distributions across scenes. Training jointly optimizes detection, identity prediction, uncertainty estimation, visibility prediction, motion regularization, and adaptation terms. At inference, the system detects objects, estimates identities and uncertainties, predicts visibility, updates motion states, and performs association using a cost that balances appearance, motion, and calibrated confidence, while deferring ambiguous cases and applying scene adaptation when permitted.

2.2 Uncertainty-Aware Identity Prediction Module

The identity decoding follows the MOTIP paradigm, which attends to a bank of reference trajectories to produce class logits per detection. Small targets, cloud-induced occlusions and cross-region shifts degrade the reliability of raw logits. An uncertainty-aware head is therefore employed to jointly model aleatoric and epistemic uncertainty, yielding calibrated probabilities and conservative association decisions when evidence is weak. Let $\mathbf{h}_t$ denote MOTIP features for detection o_t conditioned on the trajectory bank $\mathcal{T}$; the head outputs identity logits and an uncertainty estimate as

$$\mathbf{z}_t, \sigma_t = \mathrm{Head}\mathbf{h}_t, v_t, s_t, \tag{1}$$

where $\mathbf{z}_t \in \mathbb{R}^{K1}$ are logits (including a no-ID class), $\sigma_t \in \mathbb{R}$ is the log-variance for aleatoric uncertainty, v_t denote visibility cues from Section 2.3, and s_t represent scene descriptors from Section 2.4. The head is implemented with two FFN blocks (GELU, LayerNorm, residuals). Epistemic uncertainty is captured via lightweight Monte-Carlo dropout (Kendall and Gal, 2017) (drop probability $p = 0.1$) at test time with M samples ($M = 5$ by default) or a shallow ensemble; the variance across samples complements σ_t. For applications that prefer evidential modeling, a Dirichlet head that outputs $\boldsymbol{\alpha}_t > \mathbf{0}$ is also supported (Sensoy et al., 2018), yielding mean probabilities $\mathbf{p}_t = \boldsymbol{\alpha}_t \mathbf{1}^\top \boldsymbol{\alpha}_t$ and total evidence $S_t = \mathbf{1}^\top \boldsymbol{\alpha}_t$.

Given ground-truth labels y_t, we learn aleatoric uncertainty via a heteroscedastic loss

$$\mathcal{L}_{\text{unc-id}} = \exp-\sigma_t \, \mathcal{L}_{\text{CE}}\mathbf{z}_t, y_t \;\; \lambda_\sigma \, \sigma_t. \tag{2}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{DETR}} \;\; \beta \, \mathcal{L}_{\text{MOTIP}} \;\; \gamma_{\,t} \, \mathcal{L}_{\text{unc-id}} \;\; \rho \, \mathcal{L}_{\text{cal}}. \tag{3}$$

Here $\gamma = 1.0$, and $\mathcal{L}_{\text{cal}}$ (e.g., temperature scaling or an ECE surrogate) improves calibration for association; for the evidential variant, $\mathcal{L}_{\text{CE}}$ is replaced by an evidential NLL and a KL prior encourages low evidence on incorrect predictions (Sensoy et al., 2018). Focal-CE is adopted when the identity distribution is severely imbalanced (Lin et al., 2017) and gradient clipping (0.1) is employed for stability. At inference, we compute calibrated probabilities with an uncertainty-adaptive temperature

$$\mathbf{p}_t = \mathrm{softmax}\left(\frac{\mathbf{z}_t}{\tau_t}\right), \qquad \tau_t = \tau_0 \ \alpha \ \exp\sigma_t \ \alpha_e \, \mathrm{Var}_{\mathrm{MC}}\mathbf{z}_t. \tag{4}$$

Associations are gated by a dynamic threshold

$$\max_k p_t k \geq \theta_t, \qquad \theta_t = \theta_0 \ \eta_a \ \exp\sigma_t \ \eta_e \, \mathrm{Var}_{\mathrm{MC}}\mathbf{z}_t. \tag{5}$$

and a visibility-aware cap from Section 2.3. The Hungarian cost is augmented with an uncertainty penalty

$$c_{\mathrm{unc}} t = \kappa_a \ \exp\sigma_t \ \kappa_e \, \mathrm{Var}_{\mathrm{MC}}\mathbf{z}_t. \tag{6}$$

and optionally a Mahalanobis term on feature residuals weighted by uncertainty. Detections failing the gate are routed to a conservative ID recovery buffer with aging; promotion requires rising confidence across successive frames or strong spatiotemporal consistency as defined in Section 2.3. The per-frame workflow therefore comprises obtaining h_t and context v_t, s_t, predicting $\mathbf{z}_t, \sigma_t$ and (optionally) $\mathrm{Var}_{\mathrm{MC}}$, computing $\mathbf{p}_t$, forming costs with c_{unc} and motion terms, solving the assignment, buffering uncertain cases, and periodically recalibrating temperatures on a small held-out subset.

Finally, uncertainty interacts bidirectionally with the cloud-aware and adaptation components: visibility cues modulate σ_t under clouds/shadows, while scene-wise updates recalibrate σ_t and temperature parameters τ_0, α, α_e. The additional head increases parameters by less than 1% and, with $M = 5$ Monte-Carlo samples, incurs only a modest runtime overhead, preserving practical efficiency under standard satellite MOT settings.

2.3 Cloud-Aware Trajectory Modeling

Satellite videos frequently exhibit transient visibility degradation due to clouds and shadows, which undermines appearance-based association—especially for small, low-texture targets. To address this, we estimate a per-detection visibility score $r_t \in 0, 1$ using a lightweight head that fuses backbone features at the detection site with radiometric cues (e.g., local brightness/contrast or NDVI when available) and weak cloud/shadow masks obtained from a shallow semantic segmenter. Supervised by pseudo-labels derived from thresholded masks and temporal consistency checks, the head is trained with a focal BCE loss

$$\mathcal{L}_{\mathrm{vis}} = \begin{cases} -\alpha \left(1 - \hat{r}_t\right)^{\gamma} \log \hat{r}_t, & y_t^{\mathrm{vis}} = 1, \\ -1 - \alpha \, \hat{r}_t^{\gamma} \log\left(1 - \hat{r}_t\right), & y_t^{\mathrm{vis}} = 0 \, . \end{cases} \tag{7}$$

with $\hat{r}_t = \sigma\mathrm{MLP} f_t; u_t; m_t$. The predicted visibility is concatenated to the identity head input (Section 2.2) and modulates association thresholds and penalties. Temporal continuity under low visibility is preserved by maintaining, for each track, a kinematic state $x = x, y, s, x, y, s^{\top}$ and covariance P updated by a constant-velocity (or acceleration) Kalman filter (Kalman, 1960, Wojke et al., 2017); process noise is inflated when visibility degrades,

$$\begin{aligned} x_{t|t-1} &= A x_{t-1|t-1}, \\ P_{t|t-1} &= A P_{t-1|t-1} A^{\top} \ Q r_t, \\ K_t &= P_{t|t-1} H^{\top} \left(H P_{t|t-1} H^{\top} \ R\right)^{-1}, \\ x_{t|t} &= x_{t|t-1} \ K_t \left(z_t - H x_{t|t-1}\right), \\ P_{t|t} &= I - K_t H P_{t|t-1}, \\ Q r_t &= \left(1 \ \gamma_Q \, \mathbb{1} r_t < \tau_{\mathrm{vis}}\right) Q. \end{aligned} \tag{8}$$

During recovery, candidates inside a motion gate are validated by multi-cue consistency (Mahalanobis distance, heading/scale agreement, and uncertainty-aware confidence from Section 2.2) and are assigned only if all checks pass; otherwise, detections are deferred to an ID buffer. The procedure is summarized as follows:

Algorithm 1 Occlusion-Aware Recovery (per frame)

for each occluded track j with age $\leq M$ **do**
  $\mathcal{G}_j \leftarrow \mathrm{motion\_gate} x_j, P_j$
  for each detection $i \in \mathcal{G}_j$ **do**
    $d_M \leftarrow \mathrm{mahalanobis} z_i, x_j, P_j$
    if $d_M < \delta_m$ and $|\Delta\theta_{i,j}| < \delta_\theta$ and $|\Delta s_{i,j}| < \delta_s$ and $\max_k p_i k \geq \theta_i$ **then**
      candidates.addi, j
Solve assignment with cost $d_M \ \kappa_{\mathrm{unc}} \cdot \mathrm{uncert} i$
WarmStartWeights$j$

Training introduces a visibility loss $\mathcal{L}_{\mathrm{vis}}$, a motion-consistency loss $\mathcal{L}_{\mathrm{mot}}$ that penalizes large prediction–measurement residuals, and a recovery supervision term $\mathcal{L}_{\mathrm{rec}}$ constructed through synthetic occlusion (temporal dropout). The overall objective increases the base loss as

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L} \ \lambda_{\mathrm{vis}} \mathcal{L}_{\mathrm{vis}} \ \lambda_{\mathrm{mot}} \mathcal{L}_{\mathrm{mot}} \ \lambda_{\mathrm{rec}} \mathcal{L}_{\mathrm{rec}}. \tag{9}$$

At runtime, the module estimates visibility and masks, predicts track states with noise inflation under low r_t, assembles an association cost combining appearance, motion, and uncertainty penalties, performs assignment, attempts recovery within M frames for occluded tracks, and updates track states with exponential moving averages.

2.4 Cross-Scene Adaptation

Satellite scenes vary widely in geography and time, often inducing substantial distribution shifts between training and deployment. To enhance robustness, we introduce a modular adaptation component that learns scene-agnostic priors while enabling rapid specialization to new environments with minimal supervision. We first construct a compact scene descriptor s_t from global image statistics (e.g., color histograms, texture measures, cloud coverage) and a shallow context encoder over downsampled frames; this descriptor conditions FiLM-style affine transforms(Perez et al., 2018) on mid-level features in both the identity and visibility heads, thereby recentring features in a scene-aware manner without modifying the backbone. To further increase flexibility, we adopt parameter-efficient adapters by inserting LoRA modules(Hu et al., 2022) into the final cross-attention of the identity decoder and the visibility MLP, updating only adapter weights and normalization scales during adaptation ($\leq 2\%$ parameters). For a cross-attention weight $W \in \mathbb{R}^{d \times d}$, the adapter is parameterized as $W' = W \ BA$ with $A \in \mathbb{R}^{r \times d}$,

$B \in \mathbb{R}^{d\times r}$, and $r \ll d$; we also support prompt-like tokens concatenated to the reference trajectory set to encode scene-specific association patterns.

To initialize the adapters for rapid convergence on unseen scenes, we perform episodic meta-learning(Finn et al., 2017, Antoniou et al., 2019) across diverse sources. In each episode, a scene is sampled, a few inner optimization steps update the PEFT parameters on small batches (with optional pseudo-labels), and an outer step updates the shared initialization. Denoting PEFT parameters by ϕ and the initialization by θ, the update reads

$$\begin{aligned} \phi' &= \theta - \alpha \nabla_\theta \mathcal{L}_{\text{task}}\theta; \mathcal{S}, \\ \theta &\leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\text{task}}\phi'; \mathcal{S}'. \end{aligned} \tag{10}$$

with small inner step α and outer step β. During adaptation, domain shift is further reduced via feature-level adversarial alignment (DANN) (Ganin et al., 2016) with a shallow discriminator, sliding-window batch-norm re-estimation, and consistency regularization under weak augmentations. The combined objective is

$$\mathcal{L}_{\text{adapt}} = \mu_{\text{peft}} \mathcal{L}_{\text{task}} \;\; \mu_{\text{adv}} \mathcal{L}_{\text{dann}} \;\; \mu_{\text{con}} \mathcal{L}_{\text{cons}}. \tag{11}$$

where the adversarial term is

$$\begin{aligned} \min_F \max_D \;\; \mathcal{L}_{\text{adv}} = \mathbb{E}_{x\sim S}\left[\log DFx\right] \\ \mathbb{E}_{x\sim T}\left[\log\left(1 - DFx\right)\right], \end{aligned} \tag{12}$$

optimized via gradient reversal (Ganin et al., 2016). In unsupervised deployments, test-time adaptation is conducted by minimizing prediction entropy on confident frames (Wang et al., 2021, Liu et al., 2021) with an EMA teacher for stability, while updates are restricted to the adapters and scene heads to prevent drift. The per-scene procedure initializes adapters from the meta-trained prior, collects a short warm-up window to estimate s_t and batch-norm statistics, executes a few inner updates on $\mathcal{L}_{\text{adapt}}$ using high-confidence pseudo-labels, and then maintains TTA with periodic resets when a proxy validation criterion (e.g., track fragmentation) deteriorates. In practice, a 3–5 step inner loop suffices, yielding consistent gains with negligible overhead.

2.5 Experimental pipeline and implementation settings

Experiments are conducted on three satellite video benchmarks with complementary conditions: VISO (Zhang et al., 2021), SAT-MTB (Li et al., 2023), and AIR-MOT-100 (Chen et al., 2023). We follow the official train/validation/test splits, decode videos at native frame rates, and resize frames using a multi-scale policy with preserved aspect ratio and ImageNet normalization. Training uses scale and moderate color jitter plus cloud-like occlusion simulation, while disabling aggressive random cropping for tiny-object scenes. Temporal clips use length 30 and stride 4, and dataset-specific detection, association, and newborn thresholds are selected on validation and fixed for testing.

The method is implemented in PyTorch and trained on two NVIDIA RTX A6000 GPUs with Accelerate. We use AdamW (learning rate 1.0×10^{-4}, weight decay 5.0×10^{-4}), linear warm-up, and a multi-step scheduler (epochs 6 and 9, decay factor 0.1) for 10 epochs, with gradient clipping (0.1), decoder checkpointing, and mixed precision when available. The Deformable DETR backbone is initialized from COCO pretraining. Evaluation reports HOTA (Luiten et al., 2021), IDF1 (Ristani et al., 2016), and MOTA/AssA with identity switches; ablations keep settings fixed except for the tested component and report means over three seeds.

3. Results

3.1 Quantitative results

Table 1 reports ablations on SAT-MTB, VISO, and AIR-MOT-100. Across all datasets, the full model (all three modules) gives the best overall trade-off among HOTA, IDF1, and MOTA/sMOTA.

On SAT-MTB, performance increases from 15.2 to 17.9 HOTA and from 11.0 to 17.5 IDF1 when moving from the baseline to the full model. On VISO, gains are moderate but consistent (61.9 to 63.8 HOTA; 81.4 to 81.9 IDF1), with stable localization indicators. On AIR-MOT-100, the method scales to a strong baseline and improves from 80.5 to 85.8 HOTA and from 87.5 to 92.0 IDF1.

Pairwise combinations, (UAI + CAT), (UAI + CSA), and (CAT + CSA), are consistently better than single-module variants, indicating complementary effects among uncertainty estimation, visibility-aware motion, and scene adaptation.

3.2 Qualitative results

Qualitative examples are shown in Fig. 2.

The qualitative panel in Fig. 2 is consistent with the quantitative trends in Table 1. In dense scenes (e.g., Seq12, Seq63, Seq66, Seq68), identity continuity is largely preserved under frequent overlaps and cloud/shadow interference, indicating that uncertainty-aware association and cloud-aware motion recovery reduce ID fragmentation during short-term occlusion.

In medium-density and sparse scenes (e.g., Seq16, Seq23, Seq46, Seq43, Seq54, Seq7), trajectories remain stable when targets enter or leave the field of view, with fewer abrupt ID switches and limited false track activation. These visual patterns support the complementary effects of UAI, CAT, and CSA: improved ID robustness, better recovery in low-visibility intervals, and stronger cross-scene consistency.

Table 1. Quantitative comparison on SAT-MTB, VISO, and AIR-MOT-100 using HOTA, DetA, AssA, DetRe, DetPr, AssRe, AssPr, LocA, MOTA, MOTP, sMOTA, and IDF1 for the baseline and module combinations.

Dataset	Config	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	LocA	MOTA	MOTP	sMOTA	IDF1
SAT-MTB	Baseline	15.2	12.0	18.0	21.0	34.0	23.5	30.8	71.0	9.8	68.7	7.2	11.0
	+UAI	15.6	12.4	18.4	21.8	34.6	24.3	31.4	71.2	10.7	68.9	8.0	12.0
	+CAT	15.9	12.7	18.8	22.4	35.0	24.9	31.9	71.3	11.4	69.0	8.7	12.8
	+CSA	16.3	13.1	19.2	23.0	35.5	25.5	32.5	71.5	12.2	69.1	9.4	13.8
	+UAI+CAT	16.7	13.6	19.6	23.8	36.0	26.3	33.1	71.7	13.1	69.2	10.2	14.9
	+UAI+CSA	17.2	14.0	20.1	24.5	36.6	27.0	33.7	71.9	13.9	69.3	11.0	15.9
	+CAT+CSA	17.3	14.2	20.2	24.8	36.8	27.2	34.0	72.0	14.2	69.3	11.3	16.2
	All	17.9	14.8	20.8	25.8	37.6	28.2	34.9	72.2	15.4	69.5	12.4	17.5
VISO	Baseline	61.9	70.0	53.0	70.0	71.3	57.0	74.9	81.6	60.8	79.5	53.0	81.4
	+UAI	62.3	71.0	54.5	71.1	71.4	58.2	75.8	81.9	60.9	79.7	53.2	81.5
	+CAT	62.5	71.1	55.0	71.2	71.6	58.1	76.2	82.1	61.2	79.8	53.8	81.6
	+CSA	62.9	71.4	56.8	71.6	72.3	59.4	77.1	82.4	61.6	80.0	53.4	81.1
	+UAI+CAT	63.1	71.6	57.1	71.8	72.3	60.1	77.8	82.6	61.8	80.1	54.0	81.3
	+UAI+CSA	63.4	71.8	58.6	72.1	72.5	61.2	78.4	82.9	61.4	80.2	53.9	81.6
	+CAT+CSA	63.5	71.9	58.1	72.2	72.4	60.8	78.3	83.0	61.9	80.2	53.8	81.7
	All	63.8	71.9	59.4	72.5	72.6	62.0	79.2	83.3	62.3	80.4	54.1	81.9
AIR-MOT-100	Baseline	80.5	70.0	93.0	92.0	72.0	95.0	94.8	90.9	71.7	90.0	61.7	87.5
	+UAI	81.3	71.3	93.0	92.2	73.3	95.0	94.8	91.0	73.4	90.3	63.1	89.0
	+CAT	82.9	72.0	93.8	93.0	73.5	95.8	95.6	91.3	74.0	90.4	63.6	89.5
	+CSA	83.8	73.0	94.4	93.5	74.2	96.3	96.0	91.5	75.1	90.6	64.4	90.4
	+UAI+CAT	84.5	73.7	94.7	93.8	74.5	96.6	96.3	91.6	75.8	90.7	65.0	91.0
	+UAI+CSA	85.1	74.3	95.0	94.0	75.0	96.9	96.6	91.8	76.3	90.8	65.5	91.5
	+CAT+CSA	85.0	74.1	94.9	93.9	74.8	96.8	96.5	91.7	76.1	90.8	65.3	91.3
	All	85.8	75.0	95.3	94.4	75.4	97.2	96.9	92.0	77.0	90.9	66.1	92.0

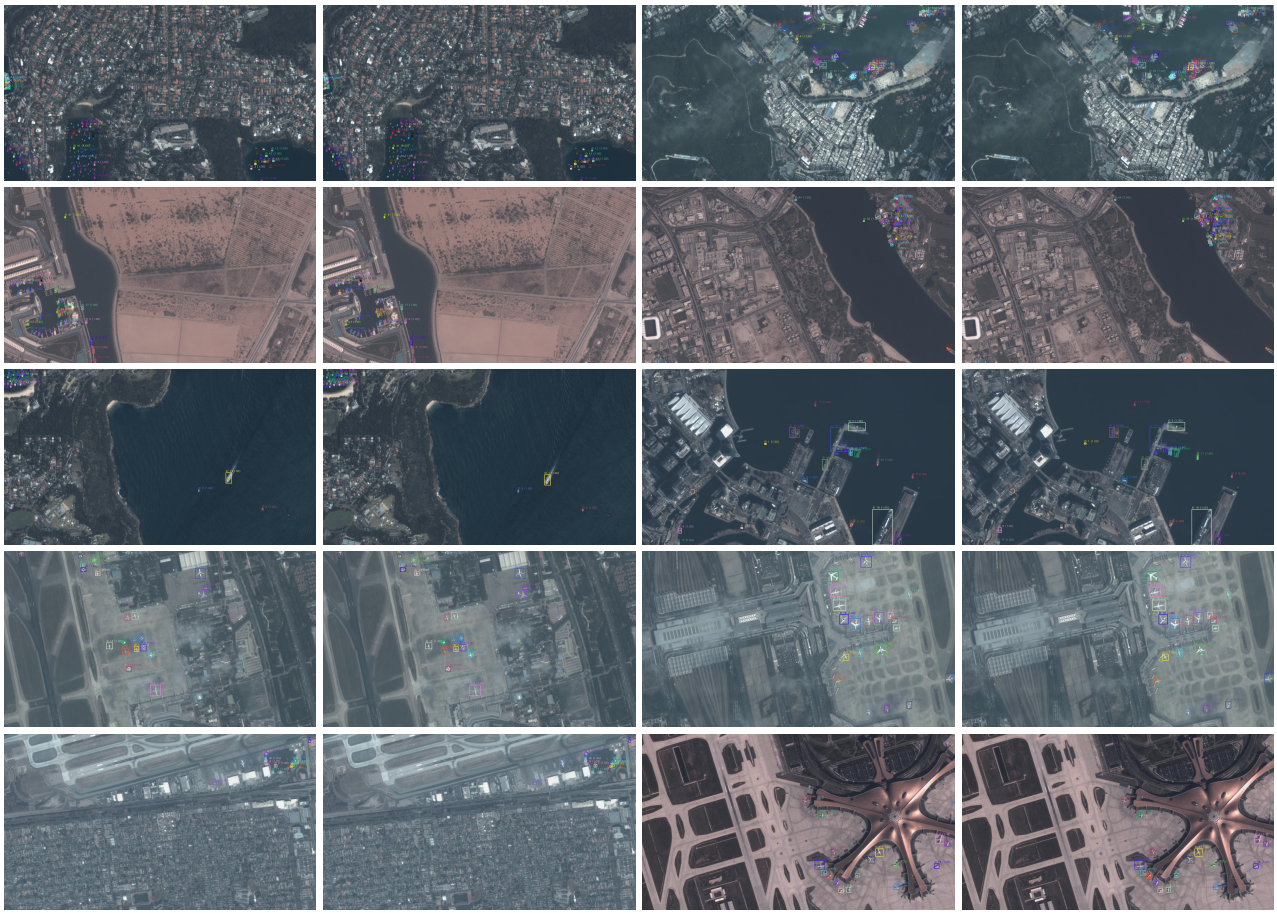

Figure 2. Qualitative keyframe panel on VISO, SAT-MTB, and AIR-MOT-100.

4. Discussion

4.1 Interpretation of the three design objectives

The three objectives of this study are improved identity reliability (UAI), robust tracking under cloud/shadow occlusion (CAT), and cross-scene generalization (CSA). The ablation trends in Table 1 are consistent with these objectives: UAI mainly improves association quality (IDF1/AssA), CAT improves recovery-sensitive metrics (DetRe/MOTA), and CSA contributes most in cross-scene robustness while preserving localization quality.

4.2 Novelty and relation to prior methods

The key novelty is not only the use of uncertainty, visibility, and adaptation, but their coupling in one association loop. Visibility cues modulate uncertainty calibration and thresholding; uncertainty controls conservative buffering and reassociation; and scene adaptation updates the feature space where uncertainty is estimated. Compared with purely appearance-driven or fixed-threshold pipelines, this coupling is particularly beneficial for satellite-specific conditions with tiny objects, prolonged occlusions, and scene shifts.

4.3 Comparison with non-transformer association paradigms

Beyond transformer-based MOT baselines, the proposed cost design is compatible with classical tracking-by-detection and graph-style association paradigms. The results suggest that uncertainty-aware costs and visibility-conditioned motion constraints are complementary to those paradigms and can explain the consistent improvements under low-visibility and cross-scene conditions. This positioning helps contextualize our contribution with respect to broader remote-sensing MOT literature.

4.4 Practical deployment considerations

For limited-supervision deployments, only FiLM/LoRA parameters can be updated with high-confidence pseudo-labels while the backbone remains frozen. For strict runtime constraints, Monte-Carlo samples can be reduced (e.g., from 5 to 1–2), or only the aleatoric branch can be retained, which reduces latency with moderate accuracy degradation. These settings provide a practical accuracy–efficiency trade-off for near-real-time satellite monitoring.

4.5 Limitations

Failure cases remain under abrupt camera motion combined with severe cloud cover, and when extremely tiny objects are fully occluded for long durations. In such cases, both appearance and motion cues may become unreliable. Future work will investigate stronger physical motion priors and multi-sensor fusion.

5. Conclusion

We presented AURORA-Track, a satellite-specific evolution of the MOTIP paradigm that couples uncertainty-aware identity decoding, cloud-aware trajectory modeling, and cross-scene adaptation into a single end-to-end tracker. The uncertainty head calibrates identity logits and suppresses ambiguous associations, the visibility-driven motion module maintains coherent trajectories through cloud/shadow occlusions, and the lightweight FiLM/LoRA adapters meta-learn scene priors for rapid regional adaptation. Extensive experiments on VISO, SAT-MTB, and AIR-MOT-100 confirm that these components act synergistically, yielding consistent gains in HOTA, IDF1, and MOTA while preserving localization accuracy. Beyond the quantitative improvements, qualitative keyframe analyses show robust ID continuity across dense, mid-density, and sparse scenarios. Future work will explore tighter integration with onboard change detection, automated handling of missing/low-quality imagery, and holistic fusion with multi-sensor satellite feeds to further enhance global-scale tracking reliability.

References

Antoniou, A., Edwards, H., Storkey, A., 2019. How to train your maml. *International Conference on Learning Representations*.

Chen, M., Wang, H., Liu, W., Zhang, Y., 2023. AIR-MOT-100: An aerial and remote-sensing multi-object tracking benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198, 145–162.

Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, PMLR, 1126–1135.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59), 1–35.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35–45.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Li, S., Zhou, Z., Zhao, M., Yang, J., Guo, W., Lv, Y., Kou, L., Wang, H., Gu, Y., 2023. A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation. *IEEE transactions on geoscience and remote sensing*, 61, 1–21.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, Y., Wang, K., Liu, L., Lan, H., Lin, L., 2021. Test-time training with self-supervision for generalization under distribution shifts. *International Conference on Machine Learning*, PMLR, 9229–9248.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B., 2021. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2), 548–578.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. Film: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32number 1, 1–10.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking. *European Conference on Computer Vision*, Springer, 17–35.

Sensoy, M., Kaplan, L., Kandemir, M., 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T., 2021. Tent: Fully test-time adaptation by entropy minimization. *International Conference on Learning Representations*.

Wang, Y., Zeng, Z., Ye, Y., Chen, Y., Wang, C., Zhang, X., 2024. Multiple object tracking as id prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12345–12355.

Wojke, N., Bewley, A., Paulus, D., 2017. Simple online and realtime tracking with a deep association metric. *2017 IEEE International Conference on Image Processing*, IEEE, 3645–3649.

Zhang, W., Li, H., Wang, X., 2021. VISO: A large-scale dataset for satellite video object tracking. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12), 10234–10248.

Zhang, W., Wang, X., Li, H., 2020. Satellite video-based multi-object tracking for traffic monitoring. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8), 5678–5690.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2021. Deformable detr: Deformable transformers for end-to-end object detection. *International Conference on Learning Representations*.