ADATRACK: ANISOTROPIC DIFFUSION AWARE MULTI-OBJECT TRACKER FOR SATELLITE VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-object tracking in satellite video requires preserving identities when limited appearance cues and uncertain localization make nearby targets difficult to distinguish. This motivates a central question: how can historical evidence, current observations, and motion direction jointly shape the spatial support used for association? We propose ADATrack, an anisotropic diffusion aware tracking framework that makes this support a recurrent state for each track. Its central design couples cross-frame evidence accumulation directly to association scoring through an explicit field evolution mechanism. A learned encoder predicts an anchor and evolution parameters from track history. The retained field is warped to the predicted anchor, then updated through anisotropic diffusion, decay, and an anchor-centered source modulated by current detector responses. Warping accounts for displacement, directional diffusion spreads support preferentially along the historical motion axis, and the source injects current visual evidence. Normalized field responses sampled at detection centers define the costs for one-to-one association. After matching, the field is aligned with the observation when available and carried into the next frame. Anchor and association supervision jointly train the encoder through this recurrence. On VISO and SAT-MTB, ADATrack attains 67.1 and 60.5 HOTA, respectively, while achieving the highest IDF1 and fewest identity switches among the compared trackers.

Keywords. Remote sensing video, multi-object tracking, anisotropic diffusion field, track association, temporal propagation.

1 INTRODUCTION

Multi-object tracking in satellite video requires preserving target identities when objects occupy few pixels and provide limited appearance cues (Yin et al., 2022). In crowded scenes, localization errors can make neighboring targets difficult to distinguish, placing greater demands on spatial and temporal evidence for association. The central challenge is to maintain a useful spatial account of each track as both its position and the available observations change. This motivates studying how historical evidence and current detections jointly determine where a track can continue.

Trackers have improved association through confidence-aware detections, corrected motion, uncertainty-aware costs, and temporal propagation (Zhang et al., 2022; Cao et al., 2023; Solano-Carrillo et al., 2024; Mancusi et al., 2023; Zhao et al., 2024b). The gap we address is how to retain a track's spatial responses as a state, update that state with current detector evidence, and use the updated responses themselves to score candidate detections.

This representation matters when nearby small targets have similar appearance and uncertain locations: a single predicted position does not express how support extends around a track or how a new response changes it. A recurrent spatial field can preserve that support across frames while incorporating current evidence before assignment. Its spread can follow the historical motion direction without forcing the track to a single candidate location.

ADATrack implements this state with a learned anchor and evolution parameters. It warps the retained field to the predicted anchor, then applies anisotropic diffusion, decay, and a source modulated by detector objectness. Responses sampled at detection centers define one-to-one association costs; the resulting field is carried to the next frame.

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.

Experiments on VISO (Yin et al., 2022) and SAT-MTB (Li et al., 2023) evaluate ADATrack against general and satellite-video trackers. ADATrack achieves HOTA scores of 67.1 and 60.5, respectively, with the highest IDF1 and lowest identity-switch counts among the compared methods. Component ablations examine temporal field warping, response modulation, directional propagation, and association decisions. The results demonstrate improvements in both overall tracking quality and identity preservation.

Our contributions are threefold:

- We formulate satellite-video association using a recurrent track-specific spatial field, connecting cross-frame evidence retention with candidate scoring through an explicit field update.
- We develop an anisotropic diffusion mechanism that combines aligned historical fields, observation-modulated sources, and directional propagation, providing a structured way to generate spatial association support.
- We show improvements in tracking and identity metrics on VISO and SAT-MTB; VISO ablations test the contribution of the recurrent field and motion-conditioned diffusion.

2 Related Work

2.1 Motion and Association Modeling

Online trackers score candidate associations from different cues. SORT uses Kalman prediction and geometric matching (Bewley et al., 2016); DeepSORT adds appearance embeddings (Wojke et al., 2017), and FairMOT jointly learns detection and re-identification features (Zhang et al., 2021). ByteTrack incorporates low-confidence detections (Zhang et al., 2022), while OC-SORT corrects motion estimates from observations (Cao et al., 2023). UTrack accounts for detection uncertainty (Solano-Carrillo et al., 2024); TrackFlow models conditional densities of association cues and uses negative log-likelihoods as costs (Mancusi et al., 2023). These works chiefly differ in which evidence enters matching and how candidate costs are formed.

DiffMOT uses a diffusion probabilistic model to predict nonlinear object motion conditioned on trajectory history (Lv et al., 2024). Its denoising process generates motion predictions. In ADATrack, anisotropic diffusion instead acts as a spatial propagation operator on a retained track field, combined with decay and an observation-modulated source. The relevant distinction is the quantity being evolved and its role in association: predicted motion versus spatial support from which candidate scores are read.

2.2 Temporal Evidence Propagation

Temporal information can be carried through explicit displacements, image features, or instance representations. CenterTrack estimates offsets linking current object centers to previous locations (Zhou et al., 2020). TraDeS derives offsets from a cost volume and warps historical features for detection and segmentation (Wu et al., 2021). TrackFormer and MOTR propagate track queries across frames (Meinhardt et al., 2022; Zeng et al., 2022), while CO-MOT studies cooperative label assignment and shadow sets for end-to-end tracking (yan et al., 2025).

In our formulation, the retained state is location-resolved: its scalar value at each candidate center provides the association response after the current observation update. Matching remains a discrete one-to-one assignment.

2.3 Satellite Video Tracking

Satellite-video trackers exploit temporal evidence to compensate for limited target appearance. MP^{2}Net combines mask propagation with motion prediction (Zhao et al., 2024b); MFACNet aggregates consecutive-frame features and their correlations (Zhao et al., 2024a); and HRTracker combines high-resolution feature fusion with adaptive association (Wu et al., 2024). PS2PAT-MOT propagates location and appearance information through a spatio-temporal state module before progressive trajectory association (Zhu et al., 2026). ADATrack instead uses a track-specific field as

its association state: the field update localizes detector objectness around the predicted anchor and spreads support along motion before candidate scoring.

3 Proposed Methodology

3.1 Overview and Track State

ADATrack maintains a spatial support field for each active track and derives association scores from its recurrent update. A frozen detector supplies detections $\mathcal{D}_t = \{(B_t^j, \mathbf{c}_t^j, p_t^j, y_t^j)\}_{j=1}^{M_t}$ and an objectness map $R_t \in [0,1]^{H' \times W'}$, where y_t^j is the predicted class. The response layer and detector configuration are fixed for each experiment. Image coordinates are mapped to the field grid by $\tilde{\mathbf{c}} = \mathbf{c}/s$, with stride s.

Definition 1. ***Track support field.*** For track i, $u_t^i \in \mathbb{R}_{\geq 0}^{H' \times W'}$ is a spatial association response stored after frame t. Values at noninteger coordinates are obtained by bilinear interpolation. The field encodes support for candidate locations; it is not a spatially normalized probability density.

For each of the N_{t-1} active tracks, the recurrent state is $\mathcal{T}_{t-1}^i = (B_{t-2}^i, B_{t-1}^i, u_{t-1}^i, p_{t-1}^i, a_{t-1}^i, y^i)$: two successive boxes, the stored field, confidence, missed-frame count, and class. The update predicts an anchor, aligns the historical field, and incorporates current observations before matching (Figs. 1 and 2).

3.2 Recurrent Anisotropic Field Update

State encoding. A shared encoder E_θ takes $\mathbf{z}_{t-1}^i = (\bar{\mathbf{c}}_{t-2}^i, \bar{\boldsymbol{\ell}}_{t-2}^i, \bar{\mathbf{c}}_{t-1}^i, \bar{\boldsymbol{\ell}}_{t-1}^i, p_{t-1}^i, a_{t-1}^i/A_{\max})$ and outputs

$$E_\theta(\mathbf{z}_{t-1}^i) = (\Delta\mathbf{c}_t^i, \Delta \log \boldsymbol{\ell}_t^i, a_{\perp,t}^i, a_{\Delta,t}^i, a_{\alpha,t}^i, a_{\lambda,t}^i), \tag{1}$$

where $\boldsymbol{\ell} = (w, h)$ and bars denote elementwise normalization by image dimensions (W, H). Predicted center offsets are converted back to image coordinates. The encoder is an MLP with two 128-dimensional hidden layers, SiLU activations, and a linear eight-parameter output.

$$\begin{aligned} \mathbf{c}_{t,\mathrm{anc}}^i &= \mathbf{c}_{t-1}^i + \Delta\mathbf{c}_t^i, & \boldsymbol{\ell}_{t,\mathrm{anc}}^i &= \boldsymbol{\ell}_{t-1}^i \odot \exp(\Delta \log \boldsymbol{\ell}_t^i), \\ d_\perp &= \mathrm{softplus}(a_\perp) + \varepsilon_d, & d_\parallel &= d_\perp + \mathrm{softplus}(a_\Delta), \\ \alpha &= \mathrm{softplus}(a_\alpha), & \lambda &= \mathrm{softplus}(a_\lambda). \end{aligned} \tag{2}$$

Track and frame indices are suppressed below where unambiguous.

Historical alignment and source. The prior field is translated by the predicted displacement:

$$u^{(0)}(\mathbf{x}) = \mathcal{W}\left(u_{t-1}, \frac{\mathbf{c}_{t,\mathrm{anc}} - \mathbf{c}_{t-1}}{s}\right)(\mathbf{x}), \qquad \mathcal{W}(u, \boldsymbol{\delta})(\mathbf{x}) = u(\mathbf{x} - \boldsymbol{\delta}). \tag{3}$$

Warping uses bilinear sampling with zero padding. The source uses an axis-aligned Gaussian whose scales follow the predicted box:

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathrm{diag}\left[(\kappa w_{t,\mathrm{anc}}/s)^2, (\kappa h_{t,\mathrm{anc}}/s)^2\right], \\ G(\mathbf{x}) &= \exp\left[-\tfrac{1}{2}(\mathbf{x} - \tilde{\mathbf{c}}_{t,\mathrm{anc}})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \tilde{\mathbf{c}}_{t,\mathrm{anc}})\right], \\ f_t^i(\mathbf{x}) &= \alpha G(\mathbf{x})\left[1 + \beta R_t(\mathbf{x})\right], \qquad \beta \geq 0. \end{aligned} \tag{4}$$

The box scales set the source support rather than a calibrated motion covariance. Objectness strengthens this track-local source; even when $R_t = 0$, the baseline αG remains. The source does not explicitly shift the anchor.

Directional propagation. Let $\mathbf{v} = (\mathbf{c}_{t-1} - \mathbf{c}_{t-2})/s$. When two historical boxes exist and $\|\mathbf{v}\| > \varepsilon_v$, set $\mathbf{e} = \mathbf{v}/\|\mathbf{v}\|$ and $\mathbf{e}_\perp = (-e_y, e_x)^\top$; otherwise use isotropic diffusion:

$$\mathbf{D} = \begin{cases} d_\parallel \mathbf{e}\mathbf{e}^\top + d_\perp \mathbf{e}_\perp \mathbf{e}_\perp^\top, & \text{reliable displacement,} \\ d_\perp \mathbf{I}, & \text{otherwise.} \end{cases} \tag{5}$$

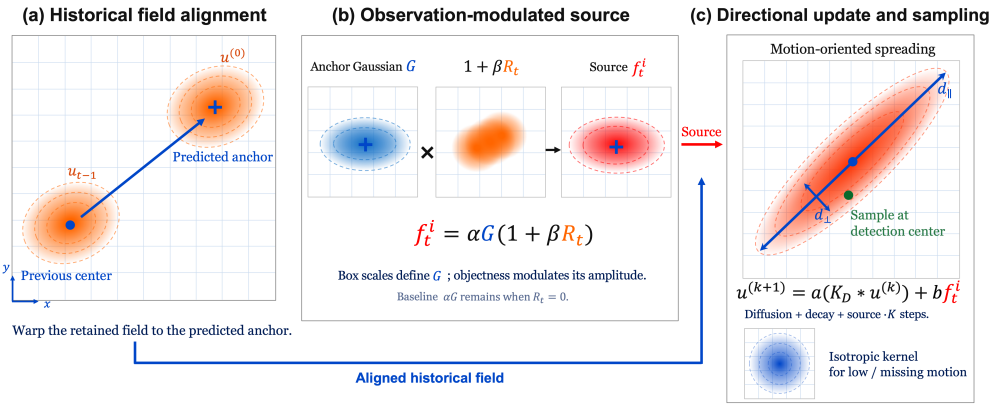

Figure 1: Schematic of the recurrent field mechanism. (a) The retained field is translated to the predicted anchor. (b) An axis-aligned Gaussian defines the source support, modulated by current objectness. (c) The aligned field and source jointly drive diffusion and decay; candidate responses are sampled at detection centers. The inset shows the isotropic diffusion kernel used for low or missing motion.

This tensor is positive definite and has the stated directional eigenvalues (Appendix B). Its preference is axial, since reversing $\mathbf{e}$ leaves $\mathbf{D}$ unchanged. Translation is supplied by warping; diffusion spreads spatial support.

The continuous model is

$$\partial_\rho u = \nabla \cdot (\mathbf{D}\nabla u) - \lambda u + f_t^i, \tag{6}$$

where ρ is within-frame pseudo-time. For spatially constant $\mathbf{D}$, Gaussian convolution gives the diffusion operator on $\mathbb{R}^2$. We approximate it on the field grid with a sampled, finitely truncated non-negative kernel $\mathcal{K}_{\Delta\rho,\mathbf{D}}$, normalized to sum to one, with covariance parameter $2\Delta\rho\mathbf{D}$. Diffusion is followed by the exact decay/source update for each substep:

$$u^{(k+1)} = a\big(\mathcal{K}_{\Delta\rho,\mathbf{D}} * u^{(k)}\big) + b\, f_t^i, \quad a = e^{-\lambda\Delta\rho}, \quad b = \frac{1-a}{\lambda}, \quad \tilde{u}_t^i = u^{(K)}. \tag{7}$$

The limit at $\lambda = 0$ is $b = \Delta\rho$. This is a split approximation, not an exact solution of the complete forced equation. Kernel construction and boundary handling are specified in Appendix C; K and $\Delta\rho$ control the within-frame propagation interval.

Proposition 1. ***Non-negative, bounded updates.*** *For finite non-negative initialization and source, a non-negative unit-sum kernel, zero padding, and parameters fixed within a frame, equation 7 preserves non-negativity. For $\lambda > 0$,*

$$\|u^{(K)}\|_\infty \le a^K \|u^{(0)}\|_\infty + \frac{1-a^K}{\lambda}\|f_t^i\|_\infty. \tag{8}$$

The proof is given in Appendix C.

3.3 Association and Track Management

For each detection center, bilinear sampling gives a response normalized by the maximum over the entire track field:

$$r_{ij} = \tilde{u}_t^i(\mathbf{c}_t^j/s), \qquad \bar{r}_{ij} = \frac{r_{ij}}{\max_{\mathbf{x}} \tilde{u}_t^i(\mathbf{x}) + \varepsilon_r}, \qquad C_{ij} = 1 - \bar{r}_{ij}. \tag{9}$$

Here $\varepsilon_r > 0$ gives zero scores for a zero field. Normalization is independent of the candidate set. A pair is eligible only if its classes agree, its center satisfies the loose gate $\|(\mathbf{c}_t^j - \mathbf{c}_{t,\mathrm{anc}}^i) \oslash \boldsymbol{\ell}_{t,\mathrm{anc}}^i\|_\infty \le g$, and $\bar{r}_{ij} > \tau$, with $g > 0$ and $0 < \tau < 1$. Ineligible costs are infinite. We append N_{t-1} dummy columns, each with cost $1-\tau$, and solve the rectangular assignment by Hungarian matching. Dummy assignments leave tracks unmatched; unused detections remain unmatched. Each eligible

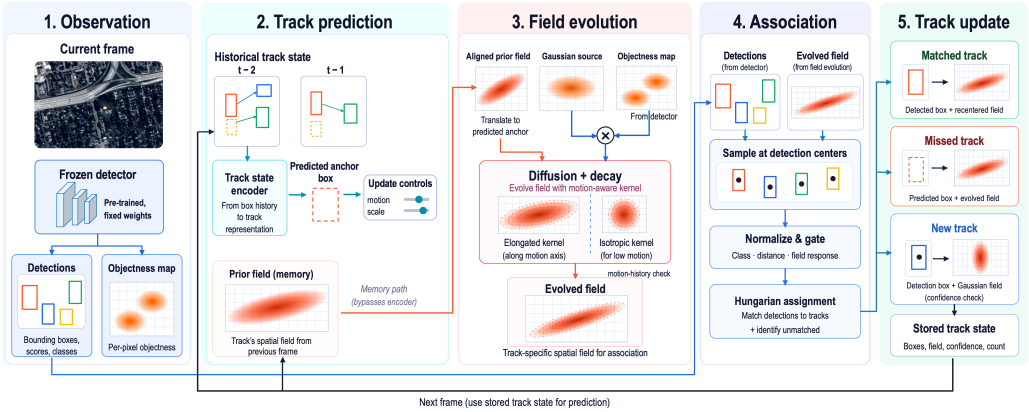

Figure 2: Overview of ADATrack. Historical track states and current detector responses drive field construction and evolution. Field responses determine association costs, followed by matching and track-state updates.

real pair costs less than a dummy assignment, but global one-to-one competition can favor leaving a track unmatched to preserve a stronger match. Tracks with no eligible detection are assigned to dummies. With no detections or tracks, only the corresponding state-management rules apply.

A matched track takes the detected box and confidence and resets its missed count. The field stored for the next frame is

$$u_t^i = \begin{cases} \mathcal{W}\left(\tilde{u}_t^i, \dfrac{\mathbf{c}_t^j - \mathbf{c}_{t,\mathrm{anc}}^i}{s}\right), & \text{matched to } j, \\ \tilde{u}_t^i, & \text{unmatched.} \end{cases} \tag{10}$$

An unmatched track uses its predicted anchor box, retains its confidence, and increments its missed count; it is removed when that count exceeds $A_{\max}$. The two-box history advances in both cases. An unmatched detection above δ_{new} starts a track with its box, class, confidence, and field $u_{\mathrm{init}}^i(\mathbf{x}) = G(\mathbf{x}; B_{\mathrm{init}}^i)$, where $G(\cdot; B)$ is the axis-aligned Gaussian in equation 4 evaluated using box B, without source scaling or objectness modulation. Both history slots duplicate this initial box, giving isotropic diffusion on the first update.

3.4 Learning Objectives

We train the encoder on clips of L consecutive frames with the detector frozen, using teacher forcing with ground-truth (GT) box histories and identities. At each frame, a class-consistent, one-to-one IoU assignment between detector outputs and GT boxes defines positive detections; pairs below threshold η_{det} are excluded. This label assignment is fixed for differentiation. Historical confidence is the assigned detection's score, or zero if absent, and the missed count records successive frames without an assigned detection.

Let $\mathcal{V}$ contain track-frame pairs with a preceding GT state and a current GT box, and let $\mathcal{P} \subseteq \mathcal{V}$ additionally require an assigned detection inside the class/spatial gate. The anchor loss uses all pairs in $\mathcal{V}$; association supervision uses only $\mathcal{P}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{anchor}} &= \sum_{(i,t)\in\mathcal{V}} \left[\|\mathbf{c}_{t,\mathrm{anc}}^i - \mathbf{c}_t^{i,\mathrm{gt}}\|_1 + \eta_\ell \|\log \boldsymbol{\ell}_{t,\mathrm{anc}}^i - \log \boldsymbol{\ell}_t^{i,\mathrm{gt}}\|_1 \right], \\ P_{ij} &= \frac{\exp(\gamma \bar{r}_{ij})}{\sum_{j'\in\mathcal{C}_{it}} \exp(\gamma \bar{r}_{ij'})}, \qquad \mathcal{L}_{\mathrm{assoc}} = -\sum_{(i,t)\in\mathcal{P}} \log P_{ij_i^+}, \\ \mathcal{L} &= \lambda_{\mathrm{anc}} \mathcal{L}_{\mathrm{anchor}} + \lambda_a \mathcal{L}_{\mathrm{assoc}}. \end{aligned} \tag{11}$$

Here $\mathcal{C}_{it}$ contains detections passing the class/spatial gate, before score rejection; j_i^+ is the assigned positive, and $\gamma > 0$ is an inverse temperature. Association supervision ranks candidates only when

the positive passes this gate. Missing or out-of-gate positives contribute no association loss; empty sums are zero, and singleton candidate sets yield no discriminative association gradient. Only anchor and association supervision are used. Unmatched decisions follow the assignment rule with threshold τ selected on the validation split; they are not directly supervised by this loss.

At clip entry or object birth, the first GT box initializes $u^i_{\text{init}} = G(\cdot; B^{i,\text{gt}}_{\text{init}})$ and is copied into both history slots. This frame initializes the state; losses start from the next eligible frame. After each training update, the evolved field is aligned to the current GT center using the matched branch of equation 10, with the detection center replaced by the GT center. Gradients propagate through field recurrence within a clip and stop at clip boundaries; no gradient is taken through discrete label assignment or detector outputs. At inference, predicted/detected histories and the state transitions above replace GT forcing.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on VISO (Yin et al., 2022), a satellite-video benchmark containing small moving vehicles, aircraft, ships, and trains, and SAT-MTB (Li et al., 2023), a multitask satellite-video benchmark with object-tracking annotations. We use the multi-object tracking annotations of both datasets.

Training and inference protocol. Training follows the GT-forced recurrence in Section 3, with predicted and detected track histories used at inference.

Data splits and training configuration. For both datasets, we partitioned whole sequences into training, validation, and test sets at a 7:1:2 ratio with a fixed seed of 42. We initialized YOLOX-S from COCO weights, fine-tuned it on each dataset's training split, and froze it for tracker training. Its stride-8 objectness output served as R_t. We trained the state encoder with AdamW, an initial learning rate of 10^{-4}, weight decay of 10^{-4}, batches of eight clips of eight frames, and 50 epochs. Checkpoints and tracking thresholds were selected on the validation split.

Metrics and baselines. We report HOTA and its DetA and AssA components (Luiten et al., 2021), IDF1 (Ristani et al., 2016), MOTA, and identity switches (IDS) (Bernardin & Stiefelhagen, 2008). DetA and AssA distinguish detection from association quality; IDF1 and IDS assess identity preservation, while MOTA is also sensitive to detector quality. General MOT baselines are ByteTrack (Zhang et al., 2022), OC-SORT (Cao et al., 2023), DiffMOT (Lv et al., 2024), DeconfuseTrack (Huang et al., 2024), and CO-MOT (yan et al., 2025). Satellite- or aerial-video baselines are MP^{2}Net (Zhao et al., 2024b), TLSH-MOT (Yuan et al., 2025), and MGTrack (Ren et al., 2024). All methods were evaluated on the same test sequences using identical detector outputs.

4.2 Main Results

Table 1 compares ADATrack with the listed trackers on VISO and SAT-MTB using the six metrics defined above.

On VISO, ADATrack leads the listed baselines on all six metrics. Its 4.4-point HOTA gain over the strongest baseline on that metric, MP^{2}Net, is accompanied by higher AssA and fewer identity switches. The gains span overall tracking quality and identity preservation.

On SAT-MTB, ADATrack again leads on all six metrics, showing that the identity-oriented gains extend beyond VISO. The following VISO ablations examine which design choices contribute to these results.

Category-level IDF1 comparisons are reported in Appendix A (Table A.1).

4.3 Mechanism Analysis

Table 2 tests the field recurrence, observation modulation, state correction, and training losses on VISO. The source-only variant loses the most HOTA and AssA, showing the strongest sensitivity

Table 1: Overall quantitative comparison on the VISO and SAT-MTB datasets.

Method	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑	IDS↓
(a) VISO						
ByteTrack (Zhang et al., 2022)	56.2	59.5	53.1	67.8	60.7	1012
OC-SORT (Cao et al., 2023)	57.0	58.9	55.2	69.1	59.8	853
DiffMOT (Lv et al., 2024)	60.3	62.4	58.3	72.5	64.2	487
DeconfuseTrack (Huang et al., 2024)	61.1	62.7	59.6	73.3	65.1	448
CO-MOT (yan et al., 2025)	59.2	61.3	57.2	70.7	62.1	521
MP^{2}Net (Zhao et al., 2024b)	62.7	64.1	61.4	75.1	68.3	509
TLSH-MOT (Yuan et al., 2025)	58.0	60.4	55.7	69.4	61.6	437
MGTrack (Ren et al., 2024)	59.9	61.2	58.6	72.0	65.4	463
ADATrack	67.1	68.3	65.8	77.2	72.6	255
(b) SAT-MTB						
ByteTrack (Zhang et al., 2022)	48.1	51.9	44.5	59.1	49.1	2341
OC-SORT (Cao et al., 2023)	48.7	51.4	46.1	60.2	48.3	1873
DiffMOT (Lv et al., 2024)	51.2	53.9	48.6	62.7	52.2	1507
DeconfuseTrack (Huang et al., 2024)	52.0	54.5	49.7	63.6	53.1	1590
CO-MOT (yan et al., 2025)	49.6	52.2	47.0	60.8	49.9	2016
MP^{2}Net (Zhao et al., 2024b)	56.0	58.3	53.7	67.5	58.5	1125
TLSH-MOT (Yuan et al., 2025)	49.7	52.3	47.3	61.5	51.1	1982
MGTrack (Ren et al., 2024)	52.5	54.8	50.2	64.9	54.3	1457
ADATrack	60.5	62.4	58.7	69.4	61.6	843

among these comparisons to removing retained spatial state. The weaker results without temporal warping, objectness modulation, or post-assignment correction show that retaining a field alone is insufficient; its alignment and updates also matter. The isotropic and single-loss variants further favor directional propagation and joint anchor/association supervision in this setting. The association loss specifically ranks eligible detection candidates, as defined in Section 3.

Table 2: Component ablations on VISO.

Variant	HOTA↑	AssA↑	IDF1↑	IDS↓
Full ADATrack	67.1	65.8	77.2	255
Source only (no recurrent field)	61.6	57.7	70.4	514
Without temporal warping	64.1	60.2	73.3	416
Without objectness modulation	65.6	63.2	75.6	319
Learned isotropic diffusion ($d_{\parallel} = d_{\perp}$)	65.0	62.1	74.9	354
Without post-assignment correction	65.2	62.4	75.1	343
Without association loss	64.1	60.9	73.6	402
Without anchor loss	63.3	59.6	72.5	447

Table 3: Propagation and association comparisons on VISO.

Variant	HOTA↑	AssA↑	IDF1↑	IDS↓
Field propagation				
No diffusion	64.1	60.6	73.5	405
Fixed isotropic diffusion	64.5	61.2	74.1	381
Learned isotropic diffusion	65.0	62.1	74.9	354
Motion-conditioned anisotropic diffusion	67.1	65.8	77.2	255
Perpendicular-direction anisotropic diffusion	63.7	60.0	72.8	433
Association				
Candidate-set maximum normalization	65.8	63.4	75.7	314
Without unmatched option	62.9	59.1	71.9	482
Without spatial gate	64.9	61.9	74.6	366
Full association	67.1	65.8	77.2	255

Ablation definitions. In Table 3, "No diffusion" replaces each diffusion convolution with the identity operator while retaining temporal warping, decay, source injection, and all other components. "Fixed isotropic diffusion" uses one scalar diffusion coefficient selected on the validation split and held constant for all tracks and frames; "Learned isotropic diffusion" instead predicts that coefficient per track while constraining $d_{\parallel} = d_{\perp}$. "Without unmatched option" removes the dummy columns but retains class, spatial, and response-threshold eligibility: Hungarian matching uses only eligible real pairs, and tracks with no feasible assignment remain unmatched. The full association uses dummy assignments to resolve missing candidates and allow unmatched tracks under global one-to-one competition.

The propagation controls in Table 3 distinguish retained-field memory from the effect of diffusion. Isotropic diffusion improves on no diffusion, but the full motion-conditioned update performs better; spreading along the perpendicular direction instead performs worse than no diffusion. These comparisons support using motion direction to orient the spread, rather than treating any smoothing as equivalent. The learned isotropic row is the same variant as in Table 2.

The association controls favor whole-field normalization, a spatial gate, and dummy unmatched assignments; removing the dummies has the largest effect in this block.

4.4 Qualitative Analysis

Fig. 3 shows three frames from VISO car sequence 013, with many small vehicles around a road junction. ADATrack shows fewer missed and false detections than the displayed baselines at frames 100 and 200. At frame 300, it ties MP^{2}Net in true positives and false negatives but has fewer false positives. These local detection gains complement the sequence-level identity improvements in Table 1(a).

5 Conclusion

ADATrack makes a recurrent spatial field the association state of each track. Historical support is aligned with a learned anchor, updated with detector evidence, and propagated preferentially along motion before candidate detections are scored. This connects temporal memory and current observations through the same representation used for matching. On VISO and SAT-MTB, ADATrack achieves the best HOTA, IDF1, and MOTA and the fewest identity switches among the compared methods using identical detection inputs. VISO ablations show that retained fields, temporal alignment, motion-conditioned anisotropic diffusion, and explicit unmatched assignments each contribute to these gains. Together, these results establish recurrent spatial support as an effective association representation for small targets in satellite video.

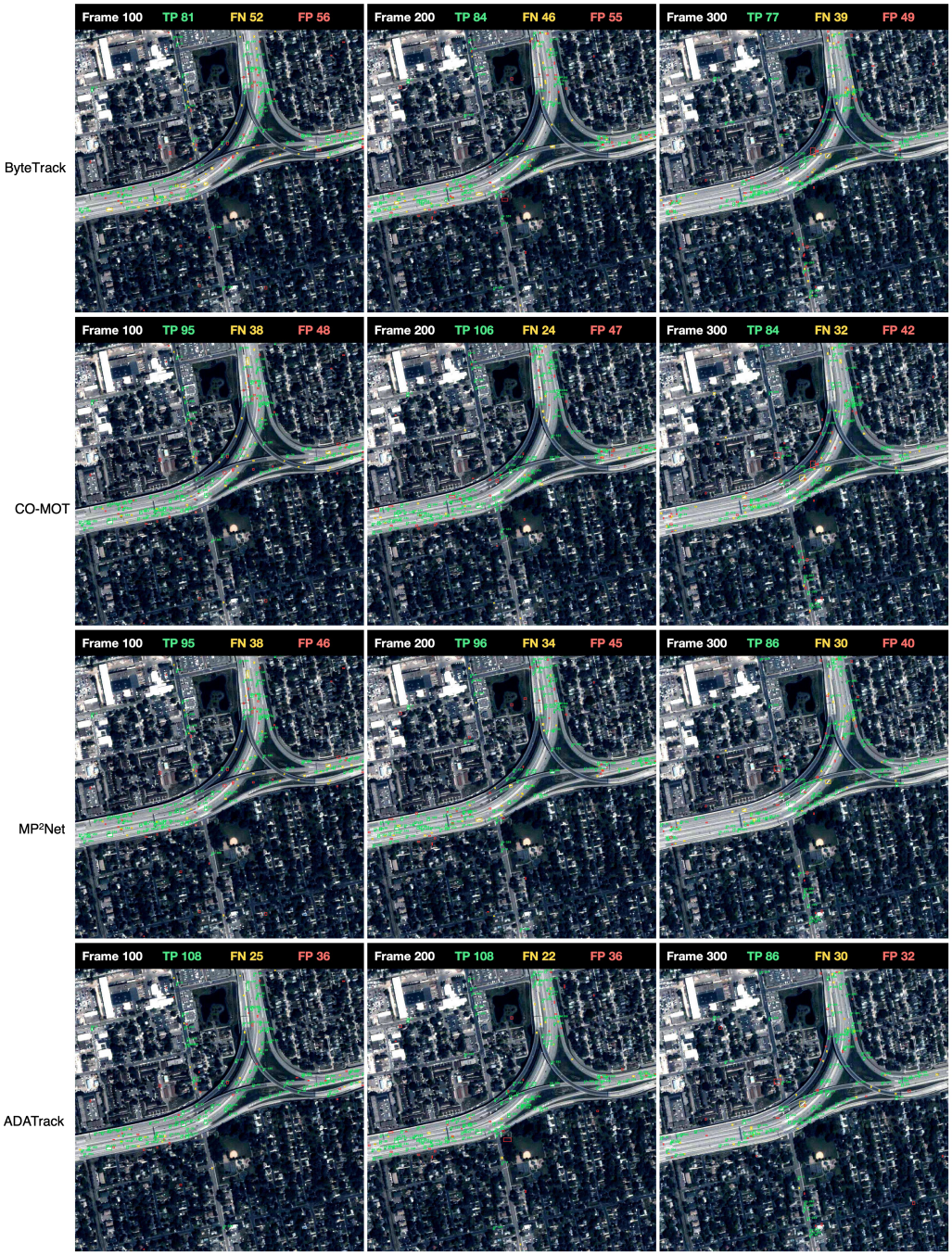

Figure 3: VISO car sequence 013 at frames 100, 200, and 300. Rows: ByteTrack, CO-MOT, MP^{2}Net, and ADATrack. Green, yellow, and red boxes mark true positives, false negatives, and false positives, respectively.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and identify related work, and support research ideation and execution through discussion and technical assistance. We did not use these tools to draft paper sections, generate synthetic datasets, or prove mathematical claims. The authors reviewed all AI-assisted material and take responsibility for the final text, references, claims, and research outputs.

REFERENCES

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, 2008.

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

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 (CVPR)*, pp. 9686–9696, June 2023.

Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao Wei. Deconfusetrack: Dealing with confusion for multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19290–19299, June 2024.

Shengyang Li, Zhuang Zhou, Manqi Zhao, Jian Yang, Weilong Guo, Yixuan Lv, Longxuan Kou, Han Wang, and Yanfeng Gu. A multitask benchmark dataset for satellite video: Object detection, tracking, and segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023. doi: 10.1109/TGRS.2023.3278075.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19321–19330, June 2024.

Gianluca Mancusi, Aniello Panariello, Angelo Porrello, Matteo Fabbri, Simone Calderara, and Rita Cucchiara. TrackFlow: Multi-object tracking with normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9531–9543, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Mancusi_TrackFlow_Multi-Object_tracking_with_Normalizing_Flows_ICCV_2023_paper.html.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8844–8854, June 2022.

Libo Ren, Wenxin Yin, Wenhui Diao, Kun Fu, and Xian Sun. Motion-guided multiobject tracking model for high-speed aerial objects in satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024. doi: 10.1109/TGRS.2024.3401573.

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 17–35, Cham, 2016. Springer International Publishing. ISBN 978-3-319-48881-3.

Edgardo Solano-Carrillo, Felix Sattler, Antje Alex, Alexander Klein, Bruno Pereira Costa, Angel Bueno Rodriguez, and Jannis Stoppe. UTrack: Multi-object tracking with uncertain detections. *arXiv preprint arXiv:2408.17098*, 2024. doi: 10.48550/arXiv.2408.17098. URL https://arxiv.org/abs/2408.17098.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017. doi: 10.1109/ICIP.2017.8296962.

Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12352–12361, June 2021.

Yuqi Wu, Qiaoyuan Liu, Haijiang Sun, and Donglin Xue. Hrtracker: Multi-object tracking in satellite video enhanced by high-resolution feature fusion and an adaptive data association. *Remote Sensing*, 16(17), 2024. ISSN 2072-4292. doi: 10.3390/rs16173347. URL https://www.mdpi.com/2072-4292/16/17/3347.

Feng yan, Weixin Luo, Yujie Zhong, Yiyang Gan, and Lin Ma. CO-MOT: Boosting end-to-end transformer-based multi-object tracking via coopetition label assignment and shadow sets. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0ov0dMQ3mN.

Qian Yin, Qingyong Hu, Hao Liu, Feng Zhang, Yingqian Wang, Zaiping Lin, Wei An, and Yulan Guo. Detecting and tracking small and dense moving objects in satellite videos: A benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. doi: 10.1109/TGRS.2021.3130436.

Yubin Yuan, Yiquan Wu, Langyue Zhao, Yuqi Liu, and Yaxuan Pang. Tlsh-mot: Drone-view video multiple object tracking via transformer-based locally sensitive hash. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. doi: 10.1109/TGRS.2025.3545081.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 659–675, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19812-0.

Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. doi: 10.1007/s11263-021-01513-4. URL https://doi.org/10.1007/s11263-021-01513-4.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 1–21, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20047-2.

Hu Zhao, Yanyun Shen, Zhipan Wang, and Qingling Zhang. Mfacnet: A multi-frame feature aggregating and inter-feature correlation framework for multi-object tracking in satellite videos. *Remote Sensing*, 16(9), 2024a. ISSN 2072-4292. doi: 10.3390/rs16091604. URL https://www.mdpi.com/2072-4292/16/9/1604.

Manqi Zhao, Shengyang Li, Han Wang, Jian Yang, Yuhan Sun, and Yanfeng Gu. Mp2net: Mask propagation and motion prediction network for multiobject tracking in satellite videos. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024b. doi: 10.1109/TGRS.2024.3385406.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 474–490, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8.

Kun Zhu, Haitao Guo, Guanzhou Chen, Xiaodong Zhang, Lei Ding, Xiangyun Liu, and Jun Lu. Propagating spatio-temporal state and progressively associating trajectory for satellite video multi-object tracking. *Expert Systems with Applications*, 301:130507, 2026. doi: 10.1016/j.eswa.2025.130507.

APPENDIX

A CATEGORY-LEVEL IDF1 RESULTS

Table A.1 reports IDF1 by object category on VISO. ADATrack achieves 77.17, 85.2, 89.3, and 81.2 IDF1 on cars, trains, airplanes, and ships, respectively, leading the listed methods in all four categories. Cars have the lowest IDF1 among these categories for every listed method. On cars, ADATrack exceeds MGTrack by 5.20 points and the strongest listed baseline, MP^{2}Net, by 2.11 points.

Table A.1: IDF1 comparison for different object categories on VISO.

Method	Car	Train	Airplane	Ship
ByteTrack (Zhang et al., 2022)	67.71	75.8	79.8	71.8
OC-SORT (Cao et al., 2023)	69.05	77.1	81.1	73.1
DiffMOT (Lv et al., 2024)	72.48	80.5	84.5	76.5
DeconfuseTrack (Huang et al., 2024)	73.27	81.3	85.3	77.3
CO-MOT (yan et al., 2025)	70.63	78.7	82.9	74.7
MP^{2}Net (Zhao et al., 2024b)	75.06	83.1	87.1	79.1
TLSH-MOT (Yuan et al., 2025)	69.32	77.4	81.4	73.4
MGTrack (Ren et al., 2024)	71.97	80.0	84.0	76.0
ADATrack	**77.17**	**85.2**	**89.3**	**81.2**

B PROPERTIES OF THE DIFFUSION OPERATOR

This section establishes the diffusion tensor properties and energy dissipation of the continuous source-free system. Track and frame indices are suppressed throughout.

B.1 POSITIVE DEFINITENESS AND DIRECTIONALITY

Proposition B.1. ***Diffusion tensor.*** *Let* $\mathbf{e} \in \mathbb{R}^2$ *be a unit vector and* $\mathbf{e}_\perp = (-e_y, e_x)^\top$. *For* $d_\parallel \geq d_\perp > 0$, *define*

$$\mathbf{D} = d_\parallel \mathbf{e}\mathbf{e}^\top + d_\perp \mathbf{e}_\perp \mathbf{e}_\perp^\top. \tag{B.1}$$

Then $\mathbf{D}$ *is symmetric positive definite, with eigenvalues* $d_\parallel$ *and* $d_\perp$ *along* $\mathbf{e}$ *and* $\mathbf{e}_\perp$, *respectively. Equality of the coefficients gives the isotropic tensor* $d_\perp \mathbf{I}$.

Proof. Both outer products are symmetric. Since the two directions form an orthonormal basis, any $\mathbf{z} \neq 0$ satisfies

$$\mathbf{z} = (\mathbf{z}^\top \mathbf{e})\mathbf{e} + (\mathbf{z}^\top \mathbf{e}_\perp)\mathbf{e}_\perp, \qquad \mathbf{z}^\top \mathbf{D}\mathbf{z} = d_\parallel (\mathbf{z}^\top \mathbf{e})^2 + d_\perp (\mathbf{z}^\top \mathbf{e}_\perp)^2 > 0. \tag{B.2}$$

Multiplication by each basis vector gives $\mathbf{D}\mathbf{e} = d_\parallel \mathbf{e}$ and $\mathbf{D}\mathbf{e}_\perp = d_\perp \mathbf{e}_\perp$. Finally, $\mathbf{e}\mathbf{e}^\top + \mathbf{e}_\perp \mathbf{e}_\perp^\top = \mathbf{I}$ proves the isotropic case. □

The tensor is unchanged by $\mathbf{e} \mapsto -\mathbf{e}$. Its directional preference is therefore axial; displacement of the field is provided separately by temporal warping.

B.2 ENERGY DISSIPATION WITHOUT A SOURCE

Theorem B.1. ***Continuous source-free energy dissipation.*** *Let* Ω' *be a bounded domain with a sufficiently regular boundary. Assume a sufficiently smooth solution of*

$$\partial_\rho u = \nabla \cdot (\mathbf{D}\nabla u) - \lambda u, \qquad \mathbf{n}^\top \mathbf{D}\nabla u = 0 \quad \textit{on } \partial\Omega', \tag{B.3}$$

where $\mathbf{n}$ *is the outward unit normal,* $\mathbf{D}$ *is symmetric positive semidefinite, and* $\lambda \geq 0$*. Then*

$$\frac{d}{d\rho}\frac{1}{2}\int_{\Omega'} u^2 \, d\mathbf{x} \leq 0. \tag{B.4}$$

Proof. With $E(\rho) = \frac{1}{2}\int_{\Omega'} u^2 \, d\mathbf{x}$, substitution and integration by parts yield

$$\begin{aligned} \frac{dE}{d\rho} &= \int_{\Omega'} u\big[\nabla \cdot (\mathbf{D}\nabla u) - \lambda u\big] \, d\mathbf{x} \\ &= \int_{\partial\Omega'} u \, \mathbf{n}^\top \mathbf{D}\nabla u \, dS - \int_{\Omega'} (\nabla u)^\top \mathbf{D}\nabla u \, d\mathbf{x} - \int_{\Omega'} \lambda u^2 \, d\mathbf{x} \\ &\leq 0. \end{aligned} \tag{B.5}$$

The boundary term vanishes by the zero-flux condition, and the remaining integrands are non-negative before their minus signs. □

For a nonzero source f, the energy balance includes the additional term $\int_{\Omega'} uf \, d\mathbf{x}$.

C Discrete Field Update

C.1 Kernel, Boundaries, and Splitting

For each track and frame, hold $\mathbf{D}$, λ, and the source f fixed during K within-frame updates of length $\Delta\rho > 0$. On a unit-spaced field grid, choose a truncation multiplier $r_K > 0$ and define

$$\begin{aligned} m &= \left\lceil r_K \sqrt{2\Delta\rho \, \lambda_{\max}(\mathbf{D})} \right\rceil, \qquad \mathcal{Q} = \{-m, \ldots, m\}^2, \\ \mathcal{K}(\mathbf{q}) &= \frac{\exp[-\mathbf{q}^\top \mathbf{D}^{-1}\mathbf{q}/(4\Delta\rho)]}{\sum_{\mathbf{p}\in\mathcal{Q}} \exp[-\mathbf{p}^\top \mathbf{D}^{-1}\mathbf{p}/(4\Delta\rho)]}. \end{aligned} \tag{C.1}$$

Here $\mathcal{K}$ abbreviates $\mathcal{K}_{\Delta\rho,\mathbf{D}}$ and is zero outside $\mathcal{Q}$. Its continuous covariance parameter is $2\Delta\rho\mathbf{D}$; truncation and sampling mean that its discrete covariance need not equal that parameter exactly. The discrete support is held fixed during each backward pass; kernel weights remain differentiable with respect to $\mathbf{D}$ on that support.

Convolution uses zero extension outside the image and retains the original field dimensions. No location-dependent kernel renormalization is applied at image boundaries. Bilinear warping and candidate sampling use the same zero extension, so support leaving the image is lost. The zero-flux boundary in Section B is an assumption of that separate continuous theorem, not the boundary used in this implementation.

Gaussian convolution represents constant-tensor diffusion on $\mathbb{R}^2$. The sampled convolution approximates that operation. It is followed by the exact pointwise decay/source solution over one substep:

$$u^{(k+1)} = a(\mathcal{K} * u^{(k)}) + bf, \qquad a = e^{-\lambda\Delta\rho}, \qquad b = \begin{cases} (1-a)/\lambda, & \lambda > 0, \\ \Delta\rho, & \lambda = 0. \end{cases} \tag{C.2}$$

This composition is a splitting approximation to the forced PDE. For $\lambda > 0$, the equivalent expression $b = -\operatorname{expm1}(-\lambda\Delta\rho)/\lambda$ avoids cancellation when $\lambda\Delta\rho$ is small; its zero-rate limit is $\Delta\rho$.

C.2 Proof of the Discrete Bound

Proposition C.1. ***Non-negative, bounded field evolution.*** *For finite non-negative* $u^{(0)}$, f*, a non-negative unit-sum kernel, zero extension, and parameters fixed within a frame, each iterate is non-negative. If* $\lambda > 0$*, then*

$$\|u^{(K)}\|_\infty \leq a^K \|u^{(0)}\|_\infty + \frac{1-a^K}{\lambda}\|f\|_\infty. \tag{C.3}$$

At $\lambda = 0$*, the bound becomes* $\|u^{(K)}\|_\infty \leq \|u^{(0)}\|_\infty + K\Delta\rho\|f\|_\infty$.

Proof. Zero extension preserves non-negativity, and convolution forms weighted sums with non-negative weights. Since $a, b \geq 0$, induction gives non-negative iterates. The kernel weights contributing at an image location sum to at most one; hence $\|\mathcal{K} * u\|_\infty \leq \|u\|_\infty$. Therefore

$$\|u^{(k+1)}\|_\infty \leq a\|u^{(k)}\|_\infty + b\|f\|_\infty. \tag{C.4}$$

Unrolling this recurrence gives $a^K\|u^{(0)}\|_\infty + b\sum_{k=0}^{K-1} a^k\|f\|_\infty$. For $\lambda > 0$, substitute $b = (1-a)/\lambda$ and evaluate the geometric sum. At $\lambda = 0$, use $a = 1$ and $b = \Delta\rho$. □

Bilinear zero-padded warping is also a non-negative weighted interpolation with weight sum at most one. Thus historical alignment and post-assignment correction of an existing field preserve non-negativity and do not increase its maximum response. A newly initialized Gaussian field is non-negative and bounded by one. The source obeys $\|f\|_\infty \leq \alpha(1 + \beta)$ when $G, R \in [0, 1]$.