NOT ALL PIXELS FLOW ALIKE: SPATIAL-CLOCK FLOW MATCHING FOR REMOTE SENSING CHANGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing change detection (RSCD) aims to identify genuine scene changes from bitemporal imagery, yet visual discrepancies caused by illumination, blur, noise, and residual misregistration are often entangled with true changes. Existing methods mainly learn discriminative mappings from bitemporal observations to change masks. However, implicitly coupling nuisance suppression and change structure recovery makes it difficult to distinguish appearance variations from genuine changes, especially under weak changes and severe nuisance interference. To explicitly model this correction process, we formulate RSCD as continuous evidence-to-geometry transport using Flow Matching (FM). Nevertheless, applying FM to RSCD requires explicit modeling of nuisance-consistent target coupling and geometry-guided spatial conditioning. We therefore propose TempoFlow-CD, which reformulates FM through target coupling and temporal parameterization. Nuisance-Invariant Geometry Flow (NIGF) aligns nuisance-varied observations with shared mask–distance geometry endpoints, while Change-Adaptive Spatial Clock (CASC) enables content-aware transport progression through adaptive local phases and velocity refinement. Extensive experiments on six RSCD benchmarks demonstrate that TempoFlow-CD achieves superior F1 and IoU performance with improved boundary localization and robustness to radiometric perturbations.

1 INTRODUCTION

Remote sensing change detection (RSCD) aims to identify meaningful changes from multitemporal remote sensing imagery, supporting applications such as urban monitoring, environmental assessment, and disaster response Cheng et al. (2024). Its fundamental challenge lies in that visual discrepancy does not necessarily correspond to physical change. Illumination variation, seasonal appearance, blur, noise, and residual misregistration may introduce significant differences in unchanged regions, whereas small or weakly contrasted changes can leave only subtle evidence Chen & Shi (2020). Reliable RSCD therefore requires suppressing nuisance-induced variations while recovering stable change structures from ambiguous bitemporal observations.

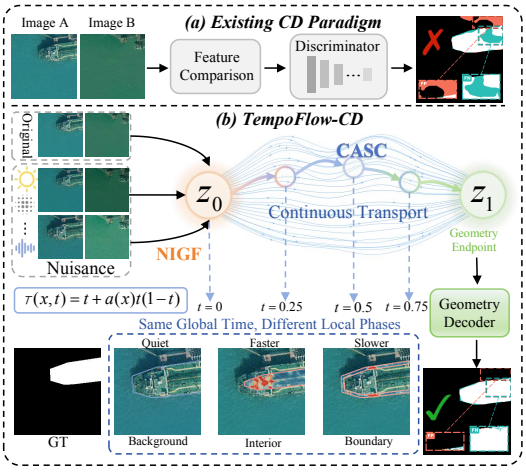

Figure 1: From discriminative prediction to continuous evidence-to-geometry transport. TempoFlow-CD reformulates Flow Matching for RSCD by aligning nuisance-varied observations toward shared change geometry and enabling spatially adaptive evolution.

Existing methods primarily address this challenge through improved bitemporal representation learning. Siamese networks compare paired observations through shared feature extractors Daudt et al. (2018), attention and Transformer architectures enhance contextual

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.

reasoning and multiscale interaction Chen & Shi (2020); Chen et al. (2021); Bandara & Patel (2022), and state-space models further extend spatiotemporal dependency modeling Chen et al. (2024). Despite expressive representations, most methods ultimately decode bitemporal evidence directly into binary change masks. As illustrated in Figure 1(a), nuisance-induced discrepancies and genuine change cues are compressed into a single discriminative prediction, requiring the same mapping to suppress irrelevant variations and recover fine-grained change structures. Consequently, the transition from nuisance-entangled evidence to stable change geometry remains implicit, limiting explicit modeling of how ambiguous observations should evolve toward reliable change representations.

This perspective motivates formulating RSCD as a continuous evidence-to-geometry transport problem. Flow Matching is well suited to this formulation: by learning vector fields that govern continuous state evolution Lipman et al. (2023); Liu et al. (2023), it provides an explicit mechanism for progressively transforming ambiguous bitemporal evidence into structured change geometry. While Flow Matching offers a natural foundation for this task, adapting it to RSCD still presents two key modeling challenges. First, nuisance-varied observations of the same underlying physical change should converge toward a shared and stable geometric representation, whereas standard source-target coupling does not by itself impose an explicit consistency constraint across these observations. Second, boundaries, interiors, and background differ in local ambiguity and structural complexity, yet their geometry-dependent local phases remain implicit under globally shared temporal conditioning. These challenges motivate a task-specific formulation of Flow Matching that incorporates nuisance-consistent transport coupling and explicit spatial phase conditioning.

Based on this reformulation principle, we introduce TempoFlow-CD, a continuous evidence-to-geometry transport framework tailored for RSCD. As illustrated in Figure 1(b), TempoFlow-CD reformulates Flow Matching from two fundamental aspects. At the coupling level, Nuisance-Invariant Geometry Flow (NIGF) establishes nuisance-consistent source-target alignment by mapping multiple nuisance-varied observations of the same image pair toward a shared mask–distance geometry endpoint. Rather than treating nuisance variations as independent perturbations, NIGF couples label-preserving appearance variations to a shared geometry target and penalizes disagreement among their endpoint estimates. At the temporal level, Change-Adaptive Spatial Clock (CASC) extends globally shared temporal conditioning by introducing spatially adaptive local phases while retaining a shared integration grid. Through selective velocity refinement, different regions evolve according to their local ambiguity and structural complexity. Extensive experiments on six RSCD benchmarks demonstrate that TempoFlow-CD achieves superior F1 and IoU performance, with consistent improvements in boundary localization and robustness to radiometric perturbations. Our main contributions are summarized as follows:

- We reformulate RSCD as continuous evidence-to-geometry transport and address two task-specific design requirements: nuisance-invariant source-target coupling and geometry-supervised spatial phase conditioning.
- We propose NIGF to reformulate source-target coupling by aligning nuisance-varied observations toward a shared geometry endpoint, enforcing nuisance-consistent transport.
- We introduce CASC as a temporal reformulation that enables spatially adaptive transport on a shared integration grid through local phase adjustment and velocity refinement.

2 Related Work

Remote Sensing Change Detection. Deep learning has substantially advanced remote sensing change detection by enabling more effective bitemporal representation learning and temporal interaction modeling. Early convolutional methods, such as FC-Siamese Daudt et al. (2018), employ shared-weight encoders for temporal comparison, while STANet Chen & Shi (2020) introduces spatial–temporal attention to enhance change discrimination capability under appearance variation and spatial misalignment. Transformer-based approaches further improve contextual reasoning and long-range interaction: BIT Chen et al. (2021) captures semantic relationships through compact token representations, and ChangeFormer Bandara & Patel (2022) adopts hierarchical Transformer features for multiscale change modeling. More recently, ChangeMamba Chen et al. (2024) explores state-space modeling for efficient spatiotemporal dependency learning. Despite increasingly powerful representations, existing methods mainly learn discriminative mappings from bitemporal fea-

tures to binary change masks, leaving the continuous transition from nuisance-entangled evidence to structured change geometry largely implicit.

Generative and Flow-Based Change Detection. Generative models have been explored in RSCD to exploit continuous representations and iterative prediction. DDPM-CD Bandara et al. (2025) leverages intermediate diffusion features learned from unlabeled remote sensing imagery, while GCD-DDPM Wen et al. (2024) generates change maps through conditional denoising trajectories. Beyond diffusion models, Flow Matching Lipman et al. (2023) and Rectified Flow Liu et al. (2023) provide continuous transport formulations by learning vector fields between source and target distributions. ChangeFlow Rolih et al. (2026) investigates latent rectified flow for bitemporal change-mask generation. However, existing generative and flow-based approaches typically define transport processes in latent or stochastic spaces, with image observations mainly serving as conditioning signals rather than transport states. Consequently, the structured evidence in bitemporal observations and its progressive transformation toward stable change geometry remain underexplored.

Geometry-Aware Continuous Dynamics. Continuous geometric representations provide richer structural information beyond binary change labels by encoding object shape, interior geometry, and boundary characteristics. Distance transforms and signed distance functions have been widely used to encode shape and boundary structures, while FlowSDF (Bogensperger et al., 2025) integrates distance-based representations into Flow Matching for medical image segmentation. Meanwhile, adaptive evolution mechanisms have been investigated in conditional generative modeling. Feature-wise linear modulation (FiLM) (Perez et al., 2018) enables condition-dependent feature modulation through learned affine transformations, and Patch Forcing (Schusterbauer et al., 2026) introduces spatially varying noise scales and difficulty-aware sampling to support heterogeneous denoising. Nevertheless, most continuous flow formulations do not explicitly incorporate geometry-dependent local phase conditioning, motivating task-specific parameterization for heterogeneous transport. These limitations motivate a Flow Matching formulation that models nuisance-consistent evidence coupling and spatially adaptive transport dynamics for dense change understanding.

3 METHOD

3.1 OVERVIEW

Appearance perturbations can change bitemporal evidence without altering the underlying geometry, calling for a consistent geometric interpretation across views. To address this challenge, we propose TempoFlow-CD, a flow-matching framework for continuous evidence-to-geometry transport (Figure 2). NIGF couples nuisance-varied sources to a shared mask–distance target and constrains their endpoint estimates to agree. A common destination, however, leaves geometry-dependent phase conditioning unspecified. We therefore introduce CASC to condition velocity prediction through supervised local phases and boundary-guided refinement, while retaining a shared integration grid. Given co-registered images $I_A, I_B \in \mathbb{R}^{3\times H\times W}$, we predict a change probability map M_p and a distance map D_p using a shared decoder, with ground-truth masks $M \in \{0,1\}^{H\times W}$ providing geometric supervision during training. For nuisance views $k \in \{0,\dots,K-1\}$ and scales $s \in \{1,\dots,4\}$, NIGF constructs source states $Z_0^k = \{z_{0,s}^k\}_{s=1}^4$ and contexts $C^k = \{c_s^k\}_{s=1}^4$, with state dynamics:

$$\frac{\mathrm{d}Z^k(t)}{\mathrm{d}t} = V_\theta\big(Z^k(t), t, C^k\big), \qquad Z^k(0) = Z_0^k, \tag{1}$$

where $t \in [0,1]$ is the global integration time. The transported endpoint $\hat{z}_{G,s}^k = z_s^k(1)$ is distinct from the label-derived target $z_{G,s}$, which is available only during training.

3.2 NUISANCE-INVARIANT GEOMETRY FLOW

Appearance perturbations can change the source evidence while preserving the underlying change geometry. The corresponding views should therefore be coupled to a common destination, yet independent training does not directly penalize disagreement among their endpoint estimates. To encode this relationship, we propose NIGF, which combines a shared mask–distance target with cross-view endpoint consistency. The target specifies region membership and interior geometry, while the consistency constraint discourages appearance-dependent disagreement in the predicted geometry.

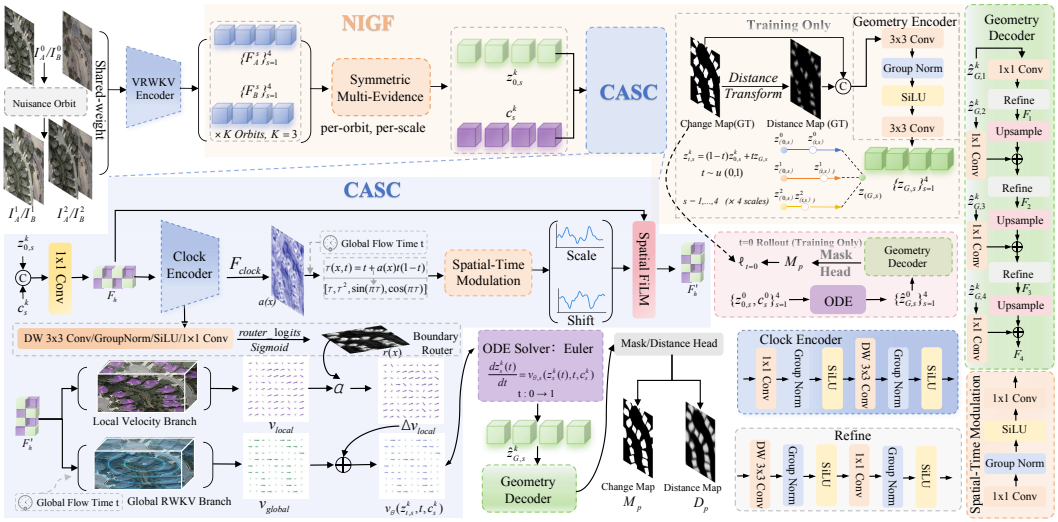

Figure 2: Overview of TempoFlow-CD. NIGF aligns nuisance-varied observations with shared mask–distance geometry, while CASC provides local phase conditioning and boundary-guided velocity refinement. Transported states decode into M_p and D_p. Training uses nuisance augmentation and label-derived geometry, whereas inference requires only the original bitemporal pair.

Symmetric evidence and shared geometry. For each training pair, we construct $K = 3$ aligned views: the original pair ($k = 0$) and two appearance-perturbed pairs. Perturbations are sampled independently across temporal images while preserving change labels. A shared-weight Vision-RWKV (VRWKV) encoder (Duan et al., 2025) extracts features $F_A^{k,s}$ and $F_B^{k,s}$. Suppressing indices and denoting $d = F_B - F_A$, we construct:

$$e = \mathrm{Cat}\big(|d|,\ P_\Sigma(F_A + F_B),\ P_\times(F_A \odot F_B),\quad 1 - \cos(F_A, F_B),\ |\phi(d) + \phi(-d)|\big). \tag{2}$$

Here, Cat denotes channel concatenation, $\odot$ elementwise multiplication, and cosine similarity is computed across channels. The projections $P_\Sigma, P_\times$ and shared nonlinear projection ϕ preserve complementary cues with exchange invariance. Fusion yields $h_s^k = \psi_s(e_s^k)$, from which $z_{0,s}^k = P_{z,s}(h_s^k)$ and $c_s^k = P_{c,s}(h_s^k)$ are derived. Only $z_{0,s}^k$ is transported. To encode interior geometry beyond binary membership, we pair M with its normalized interior distance transform D. Let d_{in} denote the nearest-background distance within changed regions and zero elsewhere. We define:

$$D(x) = \frac{M(x) d_{\mathrm{in}}(x)}{\max(1, \max_y d_{\mathrm{in}}(y))}, \qquad z_{G,s} = E_{G,s}(\mathrm{Resize}_s([M, D])), \tag{3}$$

where $[\cdot,\cdot]$ denotes channel concatenation and $E_{G,s}$ is a learnable two-layer convolutional geometry encoder. The geometry encoder and transport network are jointly trained without stopping gradients through the shared target $z_{G,s}$. Geometry reconstruction $\mathcal{L}_{\mathrm{geo}}$ supervises the decoded mask–distance representation. Gradient paths are provided in Appendix A.

Consistent geometry-directed transport. Sharing a target does not require identical transport from different sources. Since nuisance-varied views start from different states, their target velocities can differ even when the endpoint is the same. We therefore impose consistency on their implied endpoint estimates while allowing view-specific intermediate states and velocities. Following flow matching (Lipman et al., 2023), we sample $t \sim \mathcal{U}(0, 1)$, shared across views and scales, and form:

$$z_{t,s}^k = (1 - t) z_{0,s}^k + t z_{G,s}, \qquad v_{\mathrm{tar},s}^k = z_{G,s} - z_{0,s}^k. \tag{4}$$

Writing $v_s^k = [V_\theta(Z_t^k, t, C^k)]_s$ with $Z_t^k = \{z_{t,s}^k\}_{s=1}^4$, an intermediate prediction implies the endpoint estimate $\tilde{z}_{G,s}^k = z_{t,s}^k + (1-t) v_s^k$. Its orbit mean is $\bar{z}_{G,s} = K^{-1} \sum_k \tilde{z}_{G,s}^k$. We combine velocity regression with cross-view endpoint consistency:

$$\mathcal{L}_{\mathrm{trans}} = \mathbb{E}_t \sum_{s=1}^{4} \frac{\omega_s}{K} \sum_{k=0}^{K-1} \Big[\|v_s^k - v_{\mathrm{tar},s}^k\|_{2,\mathrm{av}}^2 + \lambda_{\mathrm{cons}} \|\tilde{z}_{G,s}^k - \bar{z}_{G,s}\|_{1,\mathrm{av}} \Big]. \tag{5}$$

Here, $\|\cdot\|^2_{2,\mathrm{av}}$ and $\|\cdot\|_{1,\mathrm{av}}$ denote the mean squared and mean absolute values, respectively, with ω_s weighting the contribution of scale s. Expectations are taken over training pairs and nuisance perturbations. No separate endpoint regression is required, as $\tilde{z}^k_{G,s} - z_{G,s} = (1-t)(v^k_s - v^k_{\mathrm{tar},s})$ captures endpoint errors through velocity residuals.

3.3 Change-Adaptive Spatial Clock

NIGF encourages endpoint agreement across nuisance views, but leaves geometry-dependent local phase conditioning unspecified. Boundaries, change interiors, and background have distinct geometric roles that a shared global time does not explicitly encode. To capture these distinctions in the velocity field, we propose CASC, combining geometry-supervised local phase conditioning with boundary-routed velocity refinement on a shared integration grid. Scales are processed coarse-to-fine with state–context fusion $F^k_{h,s} = P_{h,s}([z^k_s(t) + g^k_s, c^k_s])$, where g^k_s is the resized and projected velocity from the preceding scale ($g^k_1 = 0$). During flow-matching training, $z^k_{t,s}$ replaces $z^k_s(t)$. View and scale indices are omitted below.

Spatial phase conditioning and velocity refinement. A pointwise–depthwise clock encoder maps F_h to F_{clock}. A prediction head A produces the bounded offset and local phase:

$$a(x) = \rho \tanh(A(F_{\mathrm{clock}})(x)), \qquad \tau(x,t) = t + a(x)t(1-t), \tag{6}$$

where $\rho = 0.75$. This construction preserves both temporal endpoints and keeps $\tau \in [0,1]$. Positive offsets advance phase conditioning and negative offsets delay it. We embed τ as $b(\tau) = [\tau, \tau^2, \sin(\pi\tau), \cos(\pi\tau)]$ and apply spatial FiLM (Perez et al., 2018):

$$(\gamma, \beta) = \mathcal{M}(b(\tau)), \qquad F'_h = F_h \odot (1 + \epsilon \tanh\gamma) + \epsilon\beta, \quad \epsilon = 0.1. \tag{7}$$

The spatial clock conditions velocity on local phase τ, recomputing offsets at each field evaluation while retaining t for interpolation and ordinary differential equation (ODE) integration. Both branches use F'_h: RWKV predicts v_{global} via residual spatial and channel mixing, with chunked spatial mixing and t-modulated WKV decay and self-bias. Depthwise–pointwise convolutions predict v_{local} for boundary-gated refinement:

$$r(x) = \sigma(\mathcal{R}(F_{\mathrm{clock}})(x)), \quad \Delta v_{\mathrm{local}} = \alpha r \odot v_{\mathrm{local}}, \quad v^k_s = v_{\mathrm{global}} + \Delta v_{\mathrm{local}}, \quad \alpha = 0.1. \tag{8}$$

Geometry-guided clock learning. Spatially varying offsets and routing weights provide flexibility, but do not by themselves specify a geometric interpretation. We therefore supervise the offsets with distance-derived targets and the router with boundary targets, linking their predictions explicitly to the change mask. Let M_s, D_s be resized targets. We derive the offset target $a^*_s = \rho M_s \odot (2D_s - 1)$ and boundary target $B_s = \mathrm{Resize}^{\mathrm{NN}}_s(\mathrm{Dilate}(M) - \mathrm{Erode}(M))$. The former favors delayed phases where interior distance is below $1/2$, advanced phases above $1/2$, and zero offset in background. To discourage abrupt variations of clock offsets, we define:

$$\ell_{\mathrm{sm}}(a^k_s) = \mathrm{mean}\,|\Delta_x a^k_s| + \mathrm{mean}\,|\Delta_y a^k_s|. \tag{9}$$

Here, Δ_x and Δ_y are first-order differences between adjacent columns and rows, respectively, on the native feature grid. Each is averaged over valid neighbors and batch samples without padding. We combine smooth-L_1 offset regression, boundary supervision using balanced binary cross-entropy plus Dice loss, and offset smoothness:

$$\mathcal{L}_{\mathrm{clock}} = \mathbb{E}\left[\frac{1}{K\Omega}\sum_{s=1}^{4}\omega_s \sum_{k=0}^{K-1}(\lambda_{\mathrm{off}}\ell_{\mathrm{off}} + \lambda_{\mathrm{bnd}}\ell_{\mathrm{bnd}} + \lambda_{\mathrm{sm}}\ell_{\mathrm{sm}})\right]. \tag{10}$$

The batch-averaged terms are $\ell_{\mathrm{off}}(a^k_s, a^*_s)$, $\ell_{\mathrm{bnd}}(r^k_s, B_s)$, and $\ell_{\mathrm{sm}}(a^k_s)$, with arguments omitted in the sum and $\Omega = \sum_{s=1}^{4}\omega_s = 3.25$. The expectation covers mini-batches, nuisance perturbations, and interpolation time. Smoothness applies only to offsets a^k_s. All geometric targets derive from existing change masks, requiring no additional annotations.

3.4 Geometry Decoding and Optimization

NIGF and CASC define target coupling and the geometry-conditioned velocity field. However, flow matching uses interpolated states containing label-derived geometry, whereas inference integrates from image-derived sources. We therefore use a shared decoder for target reconstruction and intermediate endpoint supervision, with an additional mask loss on the original-view rollout.

Integration and shared decoding. Starting from Z_0^k, we use N uniform Euler steps:

$$Z_{n+1}^k = Z_n^k + \frac{1}{N} V_\theta(Z_n^k, t_n, C^k), \qquad t_n = \frac{n}{N}, \quad n = 0, \ldots, N-1. \tag{11}$$

The decoder progressively fuses the terminal transport states $\hat{z}_{G,s}^k = [Z_N^k]_s$ through a coarse-to-fine pathway. Suppressing k, its multiscale features $F_1, \ldots, F_4$ follow:

$$F_1 = \mathrm{Refine}_1(P_1 \hat{z}_{G,1}), \quad F_s = \mathrm{Refine}_s(\mathrm{Up}_s(F_{s-1}) + P_s \hat{z}_{G,s}), \quad s = 2, 3, 4. \tag{12}$$

Here, P_s is a lateral projection and Refine_s a residual depthwise–pointwise convolutional block. After resizing and refining F_4 at input resolution, two heads output sigmoid change probabilities M_p and raw distances D_p. The decoder and heads are shared across training branches.

Unified output supervision. Decoding $\{z_{G,s}\}$, $\{\tilde{z}_{G,s}^k\}$, and the original-view rollout endpoint yields (M_G, D_G), $(\tilde{M}^k, \tilde{D}^k)$, and (M_p^0, D_p^0), respectively. Let ℓ_m denote weighted binary cross-entropy plus soft Dice loss and ℓ_d denote smooth-L_1 distance regression. Writing the geometry reconstruction error as $\mathcal{L}_{\mathrm{geo}} = \ell_m(M_G, M) + \ell_d(D_G, D)$, we obtain:

$$\mathcal{L}_{\mathrm{sup}} = \mathbb{E}_t \sum_{k=0}^{K-1} \left[w_k \ell_m(\tilde{M}^k, M) + \frac{\lambda_d}{K} \ell_d(\tilde{D}^k, D) \right] + \lambda_{\mathrm{geo}} \mathcal{L}_{\mathrm{geo}} + \lambda_0 \ell_m(M_p^0, M). \tag{13}$$

Weights $w_0 = \lambda_m(1 + \mu/K)$ and $w_k = \mu\lambda_m/K$ for $k > 0$ emphasize the original pair while retaining supervision across views. The final term supervises its source-initialized rollout without label-derived latent inputs ($t = 0$ rollout in Figure 2).

Quiet-region velocity regularization. Zero background offsets restore $\tau = t$ but do not imply zero velocity. We therefore penalize the complete velocity $v = v_{\mathrm{global}} + \alpha r \odot v_{\mathrm{local}}$ in unchanged regions at randomly sampled interpolated states. With $U_{b,s} = 1 - \mathrm{Resize}_s^{\mathrm{NN}}(M_b)$, we use:

$$\mathcal{L}_{\mathrm{quiet}} = \mathbb{E}\left[\sum_{s=1}^{4} \omega_s \frac{\sum_{b,k,c,x} U_{b,s}(x) |v_{b,s,c}^k(x,t)|}{\max\left(1, KC_s \sum_{b,x} U_{b,s}(x)\right)} \right]. \tag{14}$$

Here, b, k, c, x index samples, views, channels, and spatial positions, with C_s channels at scale s. The expectation covers mini-batches, perturbations, and time. The ratio pools background statistics across the batch and views, with zero contribution for empty backgrounds. No additional averaging, division by Ω, or accumulation along Euler steps is applied.

Overall objective. The complete objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{sup}} + \lambda_{\mathrm{trans}} \mathcal{L}_{\mathrm{trans}} + \lambda_{\mathrm{clock}} \mathcal{L}_{\mathrm{clock}} + \lambda_{\mathrm{quiet}} \mathcal{L}_{\mathrm{quiet}}. \tag{15}$$

At inference, we integrate original-view source states with CASC and decode them into $M_p = M_p^0$ and $D_p = D_p^0$, without nuisance augmentation or label-derived geometry. Appendix B analyzes shared-target consistency, residual-dependent rollout error, and spatial phase bounds. Algorithm 1 in Appendix C details training, inference, and optimization.

4 Experiments

Dataset descriptions, evaluation metrics, implementation details, and radiometric perturbation settings are provided in Appendix D.

4.1 Comparison with State-of-the-Art Methods

Quantitative comparison. Table 1 compares TempoFlow-CD with representative methods across six RSCD benchmarks. It achieves the best F1 and IoU scores on all datasets, improving F1 by 0.83, 0.55, and 3.56 percentage points on SYSU-CD, CLCD, and HRCUS-CD, respectively. The results verify the effectiveness of evidence-to-geometry transport for suppressing nuisance ambiguity and recovering fine-grained structures. Figure 3 demonstrates favorable accuracy-efficiency trade-offs.

Table 1: Comparison with state-of-the-art methods on six benchmarks. Results are reported as F1 / IoU (%), with the best and second-best results highlighted in **bold** and underline, respectively.

Method	LEVIR-CD	SYSU-CD	WHU-CD	CLCD	HRCUS-CD	LEVIR-CD+
FC-EF Daudt et al. (2018)	80.35 / 67.16	72.88 / 57.34	75.36 / 60.46	58.96 / 41.80	38.74 / 24.02	65.31 / 48.49
BIT Chen et al. (2021)	90.62 / 82.84	78.96 / 65.24	89.02 / 80.22	73.13 / 57.64	52.99 / 36.04	81.92 / 69.37
ChangerEx Fang et al. (2023)	90.68 / 82.95	81.21 / 68.36	89.22 / 80.53	69.74 / 53.54	54.42 / 37.38	79.66 / 66.19
BAN Li et al. (2024)	91.42 / 84.20	81.59 / 68.90	94.44 / 89.46	80.30 / 67.09	74.23 / 59.02	86.23 / 75.79
CDMamba Zhang et al. (2025)	89.37 / 80.78	79.60 / 66.11	91.97 / 85.13	68.41 / 51.98	65.36 / 48.55	79.95 / 66.60
Change3D Zhu et al. (2025)	91.74 / 84.73	83.01 / 70.95	94.24 / 89.11	76.99 / 62.58	75.99 / 61.28	79.87 / 66.49
Dual UNet (FSC-180k) Benidir et al. (2025)	90.07 / 81.93	83.06 / 71.03	92.97 / 86.87	75.74 / 60.95	72.52 / 56.88	80.42 / 67.26
ChangeRWKV Yang et al. (2026)	92.11 / 85.38	82.05 / 69.56	91.38 / 84.12	74.98 / 59.97	77.65 / 63.47	84.68 / 73.43
FSMamba Guo & Xu (2026)	88.66 / 79.63	79.68 / 66.23	82.74 / 70.57	66.49 / 49.80	65.50 / 48.70	79.21 / 65.58
UniChange Zhang et al. (2026)	91.70 / 84.68	82.19 / 69.76	95.37 / 91.15	72.40 / 56.75	76.90 / 62.47	86.58 / 76.34
ChangeFlow Rolih et al. (2026)	92.10 / –	85.60 / –	– / –	84.50 / –	– / –	– / –
Ours	**92.68 / 86.36**	**86.43 / 76.11**	**96.79 / 93.78**	**85.05 / 73.99**	**81.21 / 68.36**	**87.39 / 77.61**

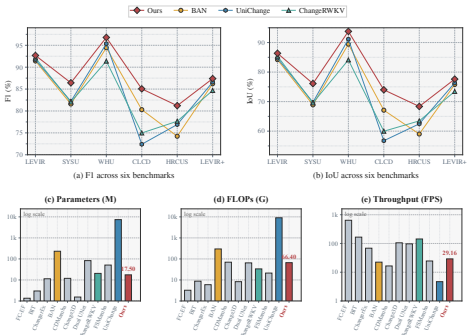

Figure 3: Accuracy-efficiency comparison.

Table 2: Performance comparison on SECOND and XiongAn (%).

Dataset	Method	mIoU ↑
SECOND	UniChange Zhang et al. (2026)	72.85
	FILFSCD Huang et al. (2026)	73.39
	ChangeFlow Rolih et al. (2026)	73.90
	Ours	**74.10**
Dataset	**Method**	**F1 ↑ IoU ↑**
XiongAn	HeteCD Jing et al. (2025)	80.60 67.50
	FreqHete Jing et al. (2026)	83.25 71.30
	Ours	**84.26 72.80**

Performance comparison on SECOND and XiongAn. Table 2 extends the evaluation to SECOND Yang et al. (2021) and the optical–SAR XiongAn dataset Jing et al. (2025). TempoFlow-CD achieves 74.10% mIoU on SECOND, exceeding ChangeFlow by 0.20 percentage points. On XiongAn, it achieves 84.26% F1 and 72.80% IoU, outperforming the strongest compared baseline, FreqHete, by 1.01 and 1.50 percentage points, respectively. These results support the applicability of the proposed framework to diverse land-cover changes and heterogeneous optical–SAR observations.

Qualitative comparison. Figure 4 compares TempoFlow-CD with existing methods across six RSCD benchmarks. TempoFlow-CD achieves cleaner and more structurally consistent change maps, reducing false responses while preserving irregular boundaries and small-scale changes. The results validate that evidence-to-geometry transport mitigates nuisance ambiguity and recovers fine-grained structures, aligning with quantitative improvements.

Boundary quality. Figure 5 evaluates boundary quality using Boundary F1 and Boundary IoU (full results in Appendix E). TempoFlow-CD achieves the best scores across all six benchmarks, with substantial gains (in percentage points) in Boundary F1 on SYSU-CD (+7.60) and CLCD (+6.28), and Boundary IoU on WHU-CD (+5.32) and LEVIR-CD+ (+3.70), demonstrating superior boundary localization and fine-grained structure recovery.

4.2 Ablation Studies

Transport formulation ablation. Table 3 evaluates the proposed Flow Matching reformulation. Compared with the standard transport setting (90.01% F1 / 81.83% IoU), nuisance-consistent coupling and spatially adaptive temporal evolution improve performance to 91.71% / 84.69% and 91.33% / 84.04%, respectively. Their combination achieves 92.68% F1 and 86.36% IoU, validating the complementary effects of both components. Additional supervision ablations are in Appendix F.

Nuisance-consistent coupling analysis. Table 4 evaluates transport coupling strategies. Absolute feature differences, mask-only geometry, and removing distance constraints reduce F1 by 1.72,

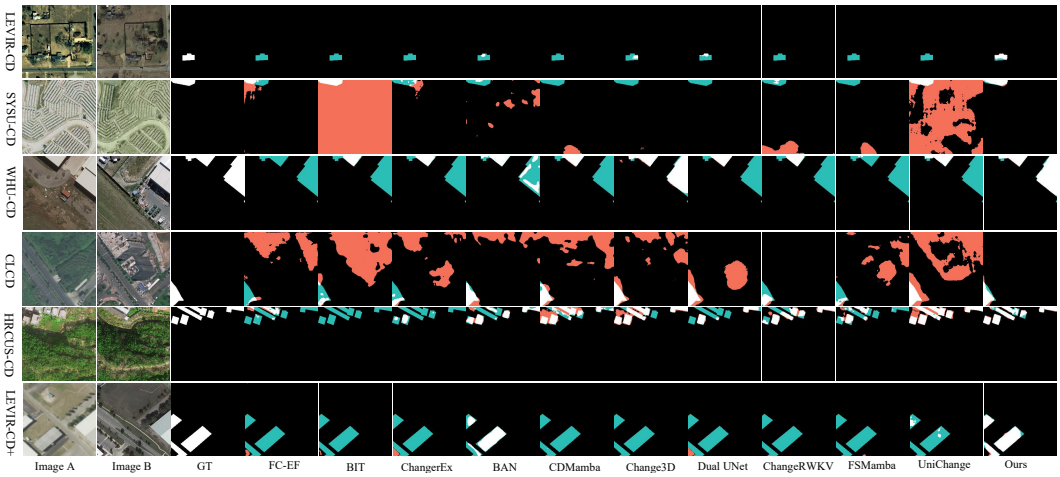

Figure 4: Qualitative comparison on LEVIR-CD, SYSU-CD, WHU-CD, CLCD, HRCUS-CD, and LEVIR-CD+. White, black, coral, and cyan denote TP, TN, FP, and FN, respectively.

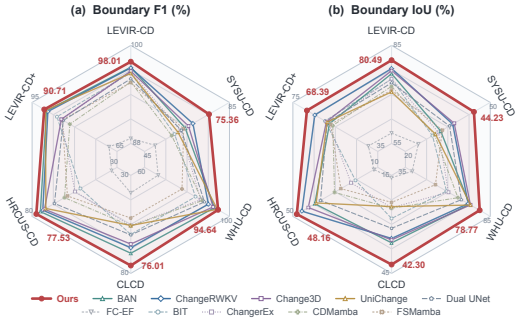

Figure 5: Boundary quality comparison.

Table 3: Ablation study of the proposed Flow Matching reformulation on LEVIR-CD.

NIGF	CASC	F1	IoU	P	R	ΔF1
×	×	90.01	81.83	90.90	89.13	−2.67
✓	×	91.71	84.69	90.47	92.98	−0.97
×	✓	91.33	84.04	89.29	93.46	−1.35
✓	✓	**92.68**	**86.36**	92.78	92.59	–

Table 4: NIGF design ablation on LEVIR-CD.

Variant	F1	IoU	P	R	ΔF1
AbsDiff Source	90.96	83.41	88.57	93.47	−1.72
Mask-only Geometry	91.12	83.69	88.51	93.89	−1.56
w/o Distance-field Loss	91.12	83.70	88.23	94.21	−1.56
w/o Nuisance Orbit	91.39	84.14	89.54	93.31	−1.29
Orbit Augmentation Only	91.09	83.64	88.34	94.02	−1.59
Ours	**92.68**	**86.36**	92.78	92.59	–

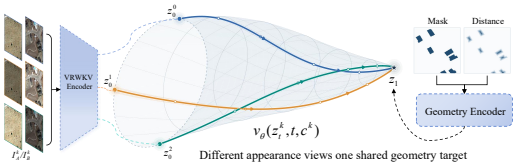

Figure 6: Nuisance-consistent transport coupling. Different appearance views converge toward a shared geometry endpoint.

1.56, and 1.56 percentage points, respectively. Removing nuisance-orbit training or using it only as augmentation further reduces F1 by 1.29 and 1.59 percentage points. Partial or alternative couplings provide no clear gains over the CASC-only setting in Table 3, highlighting the necessity of the full NIGF formulation and its coordinated design. Figure 6 further shows appearance-varied views converging to a shared geometry endpoint.

Spatially adaptive temporal evolution analysis. Table 5 reports F1 drops of 2.32, 2.28, and 1.62 percentage points without the spatial clock, local velocity branch, and clock supervision, respectively. Removing either the clock or local branch leaves their coupled pathway incomplete. The offset, boundary, and smoothness objectives each improve performance. In additional comparisons (Appendix F), auxiliary geometry supervision alone yields 91.57%/84.45% F1/IoU, while alternative spatial conditioning also underperforms full CASC, supporting its coupled design.

4.3 Analysis of Learned Transport Dynamics

Spatial response visualization. Figure 7 visualizes learned transport responses under different formulations. The baseline exhibits diffuse activations in unchanged regions, while incomplete for-

Table 5: CASC design ablation on LEVIR-CD.

Variant	F1	IoU	P	R	ΔF1
w/o Spatial Clock	90.36	82.42	86.00	95.19	−2.32
w/o Clock Supervision	91.06	83.58	87.92	94.43	−1.62
w/o Local Velocity Branch	90.40	82.49	85.98	95.30	−2.28
w/o Offset Supervision	91.75	84.75	90.46	93.08	−0.93
w/o Boundary Supervision	91.50	84.33	89.32	93.78	−1.18
w/o Smoothness Supervision	91.47	84.29	89.23	93.84	−1.21
Ours	**92.68**	**86.36**	92.78	92.59	–

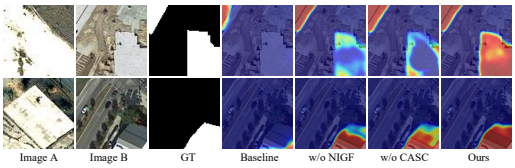

Figure 7: Spatial transport response visualization. TempoFlow-CD produces cleaner and more geometry-aligned transport responses.

mulations retain stronger nuisance-sensitive or less localized responses. In contrast, the complete formulation produces cleaner and more geometry-aligned responses, validating the effectiveness of the proposed Flow Matching reformulation for dense change understanding.

Regional and pixel-wise dynamics. Figure 8 shows that the global-clock setting yields similar boundary and change-interior trajectories with increasing background responses. TempoFlow-CD maintains weak background dynamics, learns distinct regional trajectories, and progressively aligns pixel-wise responses with local change evidence. For a fixed $N = 4$ checkpoint, increasing inference steps from one to eight improves F1/IoU from 92.08%/85.32% to 93.04%/87.00% (Appendix F), showing gains from additional inference steps for the same checkpoint.

4.4 Robustness to Radiometric Disturbances

We apply severe radiometric perturbations only to the second-temporal image on LEVIR-CD (Table 6, settings in Appendix D.4). TempoFlow-CD shows the smallest F1/IoU drops among compared methods: 3.89/6.53 and 9.74/15.51 percentage points under linear and nonlinear disturbances, respectively. Nuisance-orbit training further improves F1/IoU from 85.88%/75.26% to 88.79%/79.83% under linear disturbances and from 81.11%/68.23% to 82.94%/70.85% under nonlinear disturbances, indicating additional robustness gains.

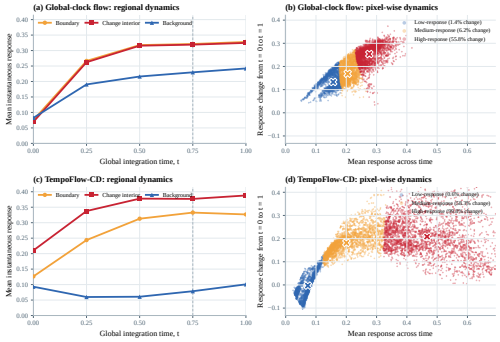

Figure 8: Spatially adaptive transport dynamics. TempoFlow-CD captures heterogeneous evolution across regions and pixels.

Table 6: Robustness on LEVIR-CD under radiometric perturbations.

Method	Linear	Nonlinear
Change3D	46.06 / 29.92 (↓45.68 / 54.81)	44.55 / 28.66 (↓47.19 / 56.07)
Dual UNet	42.35 / 26.86 (↓47.72 / 55.07)	40.03 / 25.02 (↓50.04 / 56.91)
FSMamba	26.17 / 15.06 (↓62.49 / 64.57)	24.70 / 14.09 (↓63.96 / 65.54)
UniChange	73.80 / 58.48 (↓17.90 / 26.20)	58.23 / 41.08 (↓33.47 / 43.60)
w/o Nuisance Orbit	85.88 / 75.26 (↓5.51 / 8.88)	81.11 / 68.23 (↓10.28 / 15.91)
Ours	**88.79 / 79.83** (↓**3.89 / 6.53**)	**82.94 / 70.85** (↓**9.74 / 15.51**)

5 Conclusion

We present TempoFlow-CD, a flow-matching framework that formulates remote sensing change detection as continuous evidence-to-geometry transport. NIGF couples nuisance-varied sources through a shared mask–distance target and endpoint consistency, while CASC combines geometry-supervised local phase conditioning with boundary-guided velocity refinement on a shared integration grid. Experiments on six RSCD benchmarks demonstrate improved detection accuracy and boundary localization, with additional evaluations supporting robustness to radiometric disturbances. These findings support continuous geometry transport with nuisance-consistent coupling and spatially adaptive dynamics as an effective formulation for remote sensing change detection.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and to provide feedback on the clarity and presentation of the manuscript. The tools were not used to generate experimental results, fabricate or modify data, or replace the authors' scientific analysis. All AI-assisted suggestions were reviewed, verified, and revised by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Sections 3.2–3.4. Dataset descriptions, evaluation metrics, implementation details, and radiometric perturbation settings are provided in Appendix D. Complete quantitative results and additional ablation studies are reported in Appendices E and F, respectively.

REFERENCES

Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE international geoscience and remote sensing symposium*, pp. 207–210. IEEE, 2022.

Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for remote sensing change detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5250–5262, 2025.

Yanis Benidir, Nicolas Gonthier, and Clément Mallet. The change you want to detect: Semantic change detection in earth observation with hybrid data generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2204–2214. IEEE, 2025.

Lea Bogensperger, Dominik Narnhofer, Alexander Falk, Konrad Schindler, and Thomas Pock. Flowsdf: Flow matching for medical image segmentation using distance transforms. *International Journal of Computer Vision*, 133(7):4864–4876, 2025.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15329–15337. IEEE, 2021.

Guangliang Cheng, Yunmeng Huang, Xiangtai Li, Shuchang Lyu, Zhaoyang Xu, Hongbo Zhao, Qi Zhao, and Shiming Xiang. Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing*, 16(13):2355, 2024.

Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE international conference on image processing (ICIP)*, pp. 4063–4067. IEEE, 2018.

Yuchen Duan, Weiyun Wang, Zhe Chen, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Hongsheng Li, Jifeng Dai, and Wenhai Wang. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. In *International Conference on Learning Representations*, volume 2025, pp. 83166–83182, 2025.

Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–11, 2023.

Yingting Guo and Yan Xu. Fsmamba: A frequency-enhanced spatial-temporal mamba network for local-global remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Junqing Huang, Shucheng Ji, Yapeng Wang, Min Xia, and Xiaochen Yuan. A sam fine-tuning framework with frequency-domain interactive lora for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.

Wei Jing, Haichen Bai, Binbin Song, Weiping Ni, Junzheng Wu, and Qi Wang. Hetecd: Feature consistency alignment and difference mining for heterogeneous remote sensing image change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 223:317–327, 2025.

Wei Jing, Qiang Li, and Qi Wang. Heterogeneous change detection via frequency domain interaction and statistical style embedding. *Pattern Recognition*, pp. 114393, 2026.

Kaiyu Li, Xiangyong Cao, and Deyu Meng. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–12, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Mengxi Liu, Zhuoqun Chai, Haojun Deng, and Rong Liu. A cnn-transformer network with multi-scale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4297–4306, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Blaž Rolih, Matic Fučka, Filip Wolf, and Luka Čehovin Zajc. Changeflow–latent rectified flow for change detection in remote sensing. *arXiv preprint arXiv:2605.15375*, 2026.

Johannes Schusterbauer, Ming Gui, Yusong Li, Pingchuan Ma, Felix Krause, and Björn Ommer. Denoising, fast and slow: Difficulty-aware adaptive sampling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43260–43270, June 2026.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE transactions on geoscience and remote sensing*, 60:1–16, 2021.

Yihan Wen, Xianping Ma, Xiaokang Zhang, and Man-On Pun. Gcd-ddpm: A generative change detection model based on difference-feature-guided ddpm. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

Zhenyu Yang, Gensheng Pei, Tao Chen, Xia Yuan, Haofeng Zhang, Xiangbo Shu, and Yazhou Yao. Beyond quadratic: Linear-time change detection with rwkv. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11811–11819, 2026.

Haotian Zhang, Keyan Chen, Chenyang Liu, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025.

Jindou Zhang, Zhenfeng Shao, Qing Ding, Xiao Huang, Yu Wang, Xuechao Zhou, and Deren Li. Aernet: An attention-guided edge refinement network and a dataset for remote sensing building change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Xu Zhang, Danyang Li, Xiaohang Dong, Tianhao Wu, Hualong Yu, Jianye Wang, Qicheng Li, and Xiang Li. Unichange: Unifying change detection with multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42169–42179, 2026.

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3d: Revisiting change detection and captioning from a video modeling perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24011–24022. IEEE, 2025.

APPENDIX

A JOINT OPTIMIZATION OF GEOMETRY AND TRANSPORT

The labels M and D are fixed, whereas their latent representation is learned. Let ϑ denote the geometry-encoder parameters, and write Z_G, Z_0^k, and Z_t^k as flattened, concatenated multiscale states. The image-derived source states and contexts do not depend on ϑ. For the velocity residual $\Delta^k = V_\theta(Z_t^k, t, C^k) - (Z_G - Z_0^k)$, the chain rule gives

$$\frac{\partial \Delta^k}{\partial \vartheta} = \left(tJ_V^k - I\right) \frac{\partial Z_G}{\partial \vartheta}, \qquad J_V^k = \frac{\partial V_\theta(Z_t^k, t, C^k)}{\partial Z_t^k}. \tag{16}$$

Here, I is the identity matrix. The tJ_V^k term corresponds to the interpolated-state route, and $-I$ to the target-velocity route, neither is detached. With network parameters held fixed, Eq. (4) remains a linear path in the global time t. Its endpoints are updated across optimization steps rather than along that interpolation path.

Geometry reconstruction supervises the learned endpoints through the shared mask–distance decoder. An identical latent representation for all samples cannot accurately reconstruct distinct mask–distance pairs through this decoder. However, reconstruction does not fix the latent scale or provide a general guarantee against representation drift or collapse. The rollout objective separately supervises predictions from image-derived initial states without injecting label-derived latents.

B THEORETICAL ANALYSIS OF NIGF AND CASC

We analyze how shared-target coupling constrains nuisance-dependent endpoint estimates and how spatial clocks modify conditioning without changing the integration variable. The results distinguish exact algebraic properties from conditional bounds on the implemented Euler rollout. They do not assume that training reaches a global optimum or that the model is invariant to arbitrary unseen perturbations.

Scope and notation. Fix the network parameters and one label-preserving nuisance orbit. This is an analysis at a given parameter setting, not a frozen-encoder training scheme: the geometry encoder is jointly optimized without stop-gradient, as described in Appendix A. Write $Z_G = \{z_{G,s}\}_{s=1}^4$, $\tilde{Z}_G^k = \{\tilde{z}_{G,s}^k\}_{s=1}^4$, and $\bar{Z}_G = K^{-1}\sum_k \tilde{Z}_G^k$. For a multiscale tensor $U = \{U_s\}_{s=1}^4$, define the norms

$$\|U\|_\omega^2 = \sum_{s=1}^4 \omega_s \|U_s\|_{2,\mathrm{av}}^2, \qquad \|U\|_{\omega,1} = \sum_{s=1}^4 \omega_s \|U_s\|_{1,\mathrm{av}}, \tag{17}$$

where $\omega_s > 0$ and the elementwise reductions match the training objective. We use Z_t^k for the interpolated training state and Z_n^k for the state after n Euler updates.

B.1 SHARED-TARGET COUPLING AND ENDPOINT CONSISTENCY

The reference paths satisfy $Z_t^k = (1-t)Z_0^k + tZ_G$. Because the endpoint is shared, any two nuisance views obey

$$Z_t^k - Z_t^\ell = (1-t)(Z_0^k - Z_0^\ell). \tag{18}$$

Thus, nuisance-dependent differences contract along the prescribed interpolation. This identity concerns reference paths, not automatically the trajectories produced by the learned solver. To characterize prediction errors, define $\Delta^k(t) = V_\theta(Z_t^k, t, C^k) - (Z_G - Z_0^k)$ and $\bar{\Delta}(t) = K^{-1}\sum_k \Delta^k(t)$.

Proposition B.1 (Endpoint error and within-orbit dispersion)**.** *For every $t \in [0,1]$, the one-step endpoint estimates satisfy*

$$\tilde{Z}_G^k - Z_G = (1-t)\Delta^k(t), \qquad \tilde{Z}_G^k - \bar{Z}_G = (1-t)(\Delta^k(t) - \bar{\Delta}(t)). \tag{19}$$

Consequently, their mean squared dispersion is

$$\frac{1}{K}\sum_k \|\tilde{Z}_G^k - \bar{Z}_G\|_\omega^2 = (1-t)^2 \left[\frac{1}{K}\sum_k \|\Delta^k(t)\|_\omega^2 - \|\bar{\Delta}(t)\|_\omega^2\right]. \tag{20}$$

In particular, this dispersion is bounded by $(1-t)^2 K^{-1}\sum_k \|\Delta^k(t)\|_\omega^2$.

Proof. Substituting the linear path and $V_\theta(Z_t^k, t, C^k) = Z_G - Z_0^k + \Delta^k(t)$ into $\tilde{Z}_G^k = Z_t^k + (1-t)V_\theta(Z_t^k, t, C^k)$ gives $\tilde{Z}_G^k = Z_G + (1-t)\Delta^k(t)$. Averaging over views and subtracting proves the two identities. Expanding the squared norm around $\bar{\Delta}(t)$ gives the dispersion formula, since $\sum_k(\Delta^k(t) - \bar{\Delta}(t)) = 0$. □

The implemented consistency term uses an L_1 norm rather than the squared norm above. Let $\mathcal{C}(t) = K^{-1}\sum_k \|\tilde{Z}_G^k - \bar{Z}_G\|_{\omega,1}$ denote its contribution at time t before multiplication by λ_{cons}. Its relation to pairwise endpoint disagreement is

$$\mathcal{C}(t) \le \frac{1}{K^2}\sum_{k,\ell}\|\tilde{Z}_G^k - \tilde{Z}_G^\ell\|_{\omega,1} \le 2\mathcal{C}(t). \tag{21}$$

The first inequality follows by expressing each deviation from the mean as an average of pairwise differences. The second follows from the triangle inequality through $\bar{Z}_G$. These relations clarify the joint objective: velocity regression controls error relative to the shared target, while consistency penalizes cross-view disagreement. Agreement can coexist with a shared prediction error, which requires target-directed supervision. The constraints overlap at zero residual. These identities characterize the training constraints, with their practical benefit evaluated empirically.

B.2 From Velocity Residuals to Euler Endpoint Error

The distinction between $\tilde{Z}_G^k$ and the integrated endpoint is essential because only the latter is used at inference. The next result connects residuals on the reference path to the actual N-step Euler prediction.

Proposition B.2 (Conditional Euler endpoint bound)**.** *Let $h = 1/N$, $t_n = nh$, and $Z_{n+1}^k = Z_n^k + hV_\theta(Z_n^k, t_n, C^k)$. Suppose that, for each t_n and fixed C^k, the full multiscale field is L-Lipschitz in its state under $\|\cdot\|_\omega$ on a region containing the reference and Euler states. With the same initial state Z_0^k, the terminal error satisfies*

$$\|Z_N^k - Z_G\|_\omega \le h\sum_{n=0}^{N-1}(1+hL)^{N-1-n}\|\Delta^k(t_n)\|_\omega =: E_N^k. \tag{22}$$

Proof. The reference path is linear, so $Z_{t_{n+1}}^k = Z_{t_n}^k + h(Z_G - Z_0^k)$ exactly. For $q_n^k = Z_n^k - Z_{t_n}^k$, subtracting this update from the Euler step and adding and subtracting $V_\theta(Z_{t_n}^k, t_n, C^k)$ gives

$$\|q_{n+1}^k\|_\omega \le (1+hL)\|q_n^k\|_\omega + h\|\Delta^k(t_n)\|_\omega. \tag{23}$$

Unrolling the recursion from $q_0^k = 0$ and using $Z_{t_N}^k = Z_G$ proves the result. □

To relate the discrete bound to velocity regression, additionally assume $\|\Delta^k(t) - \Delta^k(u)\|_\omega \le T_k|t-u|$ on $[0,1]$, and define $\mathcal{E}_{\text{fm}}^k = \int_0^1 \|\Delta^k(t)\|_\omega^2\,\mathrm{d}t$. Then

$$E_N^k \le e^L\left(\sqrt{\mathcal{E}_{\text{fm}}^k} + \frac{T_k}{2N}\right). \tag{24}$$

Indeed, integration over each grid interval gives $h\sum_n\|\Delta^k(t_n)\|_\omega \le \int_0^1\|\Delta^k(t)\|_\omega\,\mathrm{d}t + T_k h/2$. Applying Cauchy–Schwarz and $(1+hL)^{N-1} \le e^L$ yields the bound. Averaging $\mathcal{E}_{\text{fm}}^k$ over views and training pairs recovers the velocity-regression component of the transport objective.

For two views sharing Z_G, the triangle inequality gives $\|Z_N^k - Z_N^\ell\|_\omega \le E_N^k + E_N^\ell$. If the shared mask-probability decoder $\mathcal{H}_m$ is L_m-Lipschitz from the latent norm to the output root-mean-square norm, then

$$\|M_p^k - M_p^\ell\|_{2,\text{av}} \le L_m(E_N^k + E_N^\ell), \qquad M_p^k = \mathcal{H}_m(Z_N^k). \tag{25}$$

For the implemented solver, $N = 4$ and $h = 0.25$. Under the stated regularity assumptions, these bounds relate velocity residuals to latent endpoint error and probability-map disagreement. Their application requires bounds on L, T_k, and L_m beyond the phase-amplitude constraint. Four-step

performance and thresholded-mask F1/IoU are assessed empirically, while rollout supervision directly evaluates the source-initialized prediction path. For fixed endpoint representations, exact velocity matching along the reference segment makes a single Euler step sufficient. With finite approximation errors, predictions can depend on the evolving state and the number of field evaluations. The step comparisons therefore characterize the learned model, without establishing that multistep integration is necessary for the linear reference path.

B.3 Bounded Spatial Phase Conditioning

The offset is recomputed from the evolving state. We therefore write it as $a(x,t)$ here, making explicit the dependence suppressed in the main-text notation $a(x)$. CASC retains the global solver time and uses $\tau(x,t) = t + a(x,t)t(1-t)$ only as a conditioning signal.

Proposition B.3 (Phase range and offset sensitivity)**.** *If $|a(x,t)| \le \rho < 1$, then for every x and $t \in [0,1]$,*

$$\tau(x,0) = 0, \quad \tau(x,1) = 1, \quad 0 \le \tau(x,t) \le 1, \quad |\tau(x,t) - t| \le \rho t(1-t) \le \rho/4. \tag{26}$$

For two offset maps a and a' evaluated at the same t, their corresponding phases satisfy

$$\|\tau_a - \tau_{a'}\|_\infty = t(1-t)\|a - a'\|_\infty \le \tfrac{1}{4}\|a - a'\|_\infty. \tag{27}$$

Proof. The temporal endpoints follow directly from $t(1-t) = 0$. Since $\tau = t[1 + a(1-t)]$ and $1 - \tau = (1-t)(1-at)$, both quantities are nonnegative for $|a| \le \rho < 1$. The displacement and sensitivity bounds follow from $t(1-t) \le 1/4$. □

With $\rho = 0.75$, the phase displacement is at most 0.1875. The phase embedding b also has bounded sensitivity: $\|b'(\tau)\|_2^2 = 1 + 4\tau^2 + \pi^2 \le 5 + \pi^2$ on $[0,1]$. Thus, the clock changes the conditioning in a controlled range and attenuates offset differences near the temporal endpoints. These are properties of the conditioning signal, not guarantees of unchanged latent endpoints or reduced velocity-field sensitivity.

Interpretation of the spatial clock. If the offset were held constant in time, the phase derivative would be at least $1 - \rho > 0$. For the actual state-dependent offset, however, the total derivative, wherever it exists, is

$$\frac{\mathrm{d}\tau(x,t)}{\mathrm{d}t} = 1 + (1-2t)a(x,t) + t(1-t)\frac{\mathrm{d}a(x,t)}{\mathrm{d}t}. \tag{28}$$

Here, the spatial clock denotes the local phase supplied to the velocity field, while integration retains the shared time t. Since the offset depends on the evolving state, its amplitude bound alone does not ensure trajectory-wise monotonicity of τ. Positive offsets advance the phase input, while the resulting velocity is determined by the learned field.

Routed velocity refinement. The sigmoid router satisfies $0 \le r(x) \le 1$, giving the pointwise bound $\|\Delta v_{\mathrm{local}}(x)\|_2 \le \alpha \|v_{\mathrm{local}}(x)\|_2$, with $\alpha = 0.1$. This bounds the routed correction relative to its local branch, not relative to the global velocity or the complete field's Lipschitz constant. Setting the spatial FiLM output and local velocity output to zero recovers the corresponding global-time branch, so CASC admits the unmodified field as a special case. Boundary supervision encourages the router to concentrate local corrections on geometrically difficult regions.

Why an explicit spatial phase? CASC introduces a geometry-supervised inductive bias for finite-step transport: it makes the local phase an explicit intermediate variable that can be supervised independently of the complete velocity prediction. The target $a_s^*(x) = \rho M_s(x)(2D_s(x) - 1)$ assigns zero offset to background, delayed conditioning to low-distance change regions, and advanced conditioning to high-distance interiors. These geometric roles are communicated to both velocity branches through spatial FiLM at every field evaluation. The offset objective therefore specifies how local geometry should organize phase conditioning, while flow matching continues to supervise the resulting transport velocity.

This distinction can be stated directly in terms of the supervised quantity. Define $\tau_s^*(x,t) = t + a_s^*(x)t(1-t)$. For any predicted offset $a_s^k(x,t)$,

$$\begin{aligned} |\tau_s^k(x,t) - \tau_s^*(x,t)| &= t(1-t)|a_s^k(x,t) - a_s^*(x)| \\ &\le \tfrac{1}{4}|a_s^k(x,t) - a_s^*(x)|. \end{aligned} \tag{29}$$

For two aligned nuisance views sharing the same mask–distance targets, the triangle inequality similarly gives the following bound on cross-view phase disagreement:

$$\begin{aligned} |\tau_s^k(x,t) - \tau_s^\ell(x,t)| \le t(1-t)\big(&|a_s^k(x,t) - a_s^*(x)| \\ &+ |a_s^\ell(x,t) - a_s^*(x)|\big). \end{aligned} \tag{30}$$

Thus, offset prediction error controls both geometry-aligned phase error and cross-view phase disagreement. Together with offset smoothness and boundary supervision, this gives CASC a directly supervised organization of local conditioning within the shared integration grid. These relations characterize the added constraints, their benefit for change detection is assessed empirically. Table 5 reports F1 reductions from 92.68% to 90.36% without the spatial clock, to 91.06% without clock supervision, and to 91.75% without offset supervision. Together, the phase-error bounds and component ablations identify a concrete role for the spatial clock: promoting geometry-aligned local conditioning during finite-step transport.

C Training and Inference Procedures

Algorithm 1 summarizes the training and inference procedures. Nuisance perturbations preserve spatial alignment and change labels, with their parameters sampled independently for the two temporal images. Every evaluation of V_θ processes the four scales from coarse to fine, recomputes the spatial offsets and boundary routers, and applies the CASC operations in Eqs. (6)–(8). In particular, $\tau(x,t) = t + a(x)t(1-t)$ conditions the velocity field, whereas interpolation and Euler integration use the global time t. The contexts C^k remain fixed during integration. All decoding branches share the geometry decoder and output heads, the mask head applies a sigmoid, while the distance head returns a raw prediction. Geometry targets and nuisance views are used only during training.

Algorithm 1: Training and inference of TempoFlow-CD

Require: Training image pairs and change masks (I_A, I_B, M), nuisance views $K = 3$, Euler steps $N = 4$, loss weights.

Ensure: Trained model, change probability map M_p and distance prediction D_p.

Training

for each optimization step **do**
  Sample a mini-batch of (I_A, I_B, M).
  Construct aligned nuisance views $\{(I_A^k, I_B^k)\}_{k=0}^{K-1}$, with $(I_A^0, I_B^0) = (I_A, I_B)$.
  Extract $\{F_A^{k,s}, F_B^{k,s}\}_{s=1}^4$ with the shared VRWKV encoder.
  for each view k and scale s **do**
    Construct symmetric evidence e_s^k using Eq. (2).
    $h_s^k \leftarrow \psi_s(e_s^k)$, $z_{0,s}^k \leftarrow P_{z,s}(h_s^k)$, $c_s^k \leftarrow P_{c,s}(h_s^k)$.
  end for
  Form $Z_0^k = \{z_{0,s}^k\}_{s=1}^4$ and $C^k = \{c_s^k\}_{s=1}^4$ for each k.
  Compute normalized interior distance D and shared targets $z_{G,s} = E_{G,s}(\mathrm{Resize}_s([M, D]))$ by Eq. (3).
  Sample $t \sim \mathcal{U}(0,1)$, shared across the mini-batch, views, and scales.
  for $k = 0, \dots, K-1$ **do**
    $z_{t,s}^k \leftarrow (1-t)z_{0,s}^k + t z_{G,s}$, $v_{\mathrm{tar},s}^k \leftarrow z_{G,s} - z_{0,s}^k$, for all s.
    $Z_t^k \leftarrow \{z_{t,s}^k\}_{s=1}^4$, $\{v_s^k\}_{s=1}^4 \leftarrow V_\theta(Z_t^k, t, C^k)$.
    Retain $\{a_s^k\}$ and $\{r_s^k\}$ from the CASC evaluation in Eqs. (6)–(8).
    $\tilde{z}_{G,s}^k \leftarrow z_{t,s}^k + (1-t)v_s^k$, for all s.
  end for
  $\bar{z}_{G,s} \leftarrow K^{-1}\sum_{k=0}^{K-1} \tilde{z}_{G,s}^k$, for all s.
  Decode $\{z_{G,s}\}$ into (M_G, D_G) and $\{\tilde{z}_{G,s}^k\}$ into $(\tilde{M}^k, \tilde{D}^k)$ for each k.
  Integrate from Z_0^0 with fixed C^0 using N Euler steps as in Eq. (11), and decode the terminal states into (M_p^0, D_p^0).

Derive offset targets $a_s^* = \rho M_s \odot (2D_s - 1)$ and boundary targets B_s as in Section 3.3.
Compute $\mathcal{L}_{\text{trans}}$, $\mathcal{L}_{\text{clock}}$, $\mathcal{L}_{\text{sup}}$, and $\mathcal{L}_{\text{quiet}}$ using Eqs. (5), (10), (13), and (14), respectively.
$\mathcal{L} \leftarrow \mathcal{L}_{\text{sup}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \lambda_{\text{clock}}\mathcal{L}_{\text{clock}} + \lambda_{\text{quiet}}\mathcal{L}_{\text{quiet}}$.
Jointly update all trainable parameters using $\nabla\mathcal{L}$, retaining gradients through $z_{G,s}$ in both $z_{t,s}^k$ and $v_{\text{tar},s}^k$.
end for

Inference

Require: A bitemporal image pair (I_A, I_B) and the trained model.
Construct Z_0^0 and C^0 from (I_A, I_B) using the shared encoder and symmetric evidence construction above.
for $n = 0, \ldots, N-1$ **do**
$t_n \leftarrow n/N$, $Z_{n+1}^0 \leftarrow Z_n^0 + N^{-1}V_\theta(Z_n^0, t_n, C^0)$.
end for
$\hat{z}_{G,s}^0 \leftarrow [Z_N^0]_s$, for $s = 1, \ldots, 4$.
Decode $\{\hat{z}_{G,s}^0\}_{s=1}^4$ into (M_p^0, D_p^0) using the shared geometry decoder and mask–distance heads.
return $M_p = M_p^0$ and $D_p = D_p^0$.

D Experimental Setup

D.1 Datasets

We conduct experiments on six widely used remote sensing change detection benchmarks: LEVIR-CD (Chen & Shi, 2020), SYSU-CD (Shi et al., 2021), WHU-CD (Ji et al., 2018), CLCD (Liu et al., 2022), HRCUS-CD (Zhang et al., 2023), and LEVIR-CD+ (Shen et al., 2021). These benchmarks cover diverse geographic regions, scene types, and change patterns, providing a comprehensive evaluation of model performance under varying remote sensing conditions. We follow the official data splits for all datasets except WHU-CD, where the commonly adopted 6,096/762/762 train/validation/test split is used.

D.2 Evaluation Metrics

We evaluate binary change detection using precision (P), recall (R), F1 score, and intersection over union (IoU), with F1 and IoU as the primary metrics. Let $\widehat{M}$ and M denote the final binary prediction and ground-truth mask, respectively. Taking changed pixels as the positive class, TP, FP, FN, and TN denote the numbers of true-positive, false-positive, false-negative, and true-negative pixels. Precision and recall are defined as

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{31}$$

F1 and IoU are computed as

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{32}$$

To assess contour localization, we report tolerance-based Boundary F1 (BF1). Let $\mathcal{C}_p = \partial\widehat{M}$ and $\mathcal{C}_g = \partial M$ denote the extracted prediction and ground-truth contours. We use a fixed matching tolerance of $\delta = 8$ pixels under the L_∞ distance, where $\|x - y\|_\infty = \max(|x_1 - y_1|, |x_2 - y_2|)$. Let $\mathcal{N}_\delta(S)$ denote the pixels whose minimum L_∞ distance to a pixel set S does not exceed δ. Boundary precision, recall, and F1 are then given by

$$\mathrm{P_b} = \frac{|\mathcal{C}_p \cap \mathcal{N}_\delta(\mathcal{C}_g)|}{|\mathcal{C}_p|}, \qquad \mathrm{R_b} = \frac{|\mathcal{C}_g \cap \mathcal{N}_\delta(\mathcal{C}_p)|}{|\mathcal{C}_g|}, \qquad \mathrm{BF1} = \frac{2\mathrm{P_b}\mathrm{R_b}}{\mathrm{P_b} + \mathrm{R_b}}. \tag{33}$$

Here, $|\cdot|$ denotes pixel cardinality. The first two terms measure the fractions of predicted and ground-truth contour pixels matched within the specified tolerance.

We additionally report Boundary IoU (BIoU) (Cheng et al., 2021), which evaluates the overlap of the foreground-side boundary bands. For a binary mask S, identified with its foreground pixel set, let $\mathcal{B}_d(S) \subseteq S$ denote its inner boundary band of width d pixels. We set $d = 8$ pixels and compute

$$\mathrm{BIoU} = \frac{|\mathcal{B}_d(\widehat{M}) \cap \mathcal{B}_d(M)|}{|\mathcal{B}_d(\widehat{M}) \cup \mathcal{B}_d(M)|}. \tag{34}$$

The two parameters have distinct roles: δ controls contour matching, whereas d defines the foreground band used for overlap evaluation. Both are measured in pixels on the evaluation grid.

For the binary change detection evaluation on SECOND, mIoU averages the IoUs of the changed and unchanged classes:

$$\mathrm{mIoU} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\right). \tag{35}$$

All metrics are reported as percentages, with higher values indicating better performance.

D.3 Implementation Details

We train the model for 200 epochs using AdamW with a learning rate of 1×10^{-5} and a batch size of 16. Input image pairs are cropped to 256×256 patches and normalized with ImageNet statistics. We apply standard geometric augmentations jointly to both temporal images, while photometric and blur perturbations are sampled independently across time. For nuisance-orbit training, each pair yields $K = 3$ geometrically aligned views: the original pair and two perturbed pairs generated using randomly sampled photometric, noise, blur, frequency-amplitude, and JPEG perturbations. The output-supervision coefficients are $\lambda_m = 1.0$, $\lambda_d = 0.2$, $\lambda_0 = 1.0$, and $\lambda_{\mathrm{geo}} = 0.5$, with $\mu = 0.5$ controlling view weighting. In the overall objective, we use $\lambda_{\mathrm{trans}} = 1.0$, $\lambda_{\mathrm{clock}} = 0.1$, and $\lambda_{\mathrm{quiet}} = 0.05$. The cross-view consistency coefficient within $\mathcal{L}_{\mathrm{trans}}$ is $\lambda_{\mathrm{cons}} = 0.03$. Within $\mathcal{L}_{\mathrm{clock}}$, the offset, boundary, and clock-offset smoothness coefficients are $\lambda_{\mathrm{off}} = 1.0$, $\lambda_{\mathrm{bnd}} = 0.5$, and $\lambda_{\mathrm{sm}} = 0.05$, respectively. The scale weights are $(\omega_1, \omega_2, \omega_3, \omega_4) = (1.0, 1.0, 0.75, 0.5)$, ordered from coarse to fine. The clock objective averages over views and batch samples and is normalized by $\Omega = \sum_s \omega_s = 3.25$. In contrast, quiet-region regularization pools unchanged-pixel statistics across the entire mini-batch at each scale and sums the scale-weighted penalties without division by Ω. The overall smoothness multiplier is $\lambda_{\mathrm{clock}}\lambda_{\mathrm{sm}} = 0.005$, applied to the normalized scale-and-view average. Quiet-region regularization is weighted separately by $\lambda_{\mathrm{quiet}} = 0.05$. We integrate from $t = 0$ to $t = 1$ using $N = 4$ uniform Euler steps with $\Delta t = 0.25$. All baselines with publicly available code are faithfully reproduced using their official implementations and recommended configurations. For ChangeFlow (Rolih et al., 2026), whose code is not publicly available, we report the results from the original paper.

D.4 Radiometric Perturbation Settings

Separation from nuisance-orbit training. The robustness evaluation is deliberately distribution-shifted from nuisance-orbit training. Training perturbs both temporal images independently and randomly composes 2–4 moderate operations from nine nuisance types, whereas robustness testing perturbs only I_B using fixed high-intensity radiometric compositions. As shown in Table 7, the two settings differ in temporal application, perturbation composition, operator family, and severity. Shared photometric operators use substantially broader test ranges, and the nonlinear setting further introduces several operators not explicitly used during training. The detailed parameter ranges of the two high-intensity robustness conditions are reported in Table 8. Therefore, the robustness results measure transfer beyond the training augmentation distribution rather than performance under the same perturbation scheme.

E Additional Quantitative Results

Table 9 reports Boundary F1 and Boundary IoU across all six benchmarks.

F Additional Ablation Studies

Analysis of transport supervision. Table 10 analyzes the supervision strategies for learning the proposed transport formulation. Removing the flow, geometry, quiet-region, or distance-field objective degrades F1 by 1.37–1.56 percentage points, demonstrating their complementary roles in constraining transport dynamics and geometric reconstruction. The largest F1 drop occurs without the distance-field objective, reducing F1/IoU from 92.68%/86.36% to 91.12%/83.70%. Removing

Table 7: Nuisance-orbit training versus high-intensity robustness evaluation.

Property	**Nuisance-orbit training (medium)**	**High-intensity robustness evaluation**
Temporal application	Both I_A and I_B, parameters are sampled independently	Only I_B is perturbed, I_A remains unchanged
View construction	$K = 3$: original pair plus two perturbed pairs	One fixed corrupted test set per evaluation condition
Perturbation composition	Randomly sample 2–4 operations from 9 nuisance types	Linear High applies all 3 operations jointly, Nonlinear High applies all 5 operations jointly
Exposure / multiplicative gain	$[0.75, 1.25]$	$[0.30, 0.52]$ or $[1.65, 2.15]$
Channel / white-balance gain	$[0.85, 1.15]$	$[0.70, 1.30]$
Gamma correction	$[0.75, 1.35]$	$[0.25, 0.45]$ or $[2.50, 3.80]$
Camera response / illumination gradient / vignetting / haze	Not explicitly used	All included in Nonlinear High
Noise / blur / frequency amplitude / JPEG-like degradation	Included	Not used in either High setting

Table 8: Parameter ranges for high-intensity perturbations on LEVIR-CD.

Linear High		**Nonlinear High**	
Exposure gain	$[0.30, 0.52]$ or $[1.65, 2.15]$	Gamma correction	$[0.25, 0.45]$ or $[2.50, 3.80]$
White-balance gain	$[0.70, 1.30]$	Camera-response slope	$[3, 6]$
Brightness offset	$[-0.18, 0.20]$	Illumination-gradient strength	$[0.25, 0.55]$
		Vignetting strength	$[0.25, 0.55]$
		Haze coefficient	$[0.03, 0.12]$

Table 9: Boundary F1 / Boundary IoU (%) on six benchmarks. Best and second-best results are bold and underlined, respectively.

Method	**LEVIR-CD**	**SYSU-CD**	**WHU-CD**	**CLCD**	**HRCUS-CD**	**LEVIR-CD+**
FC-EF Daudt et al. (2018)	88.60 / 58.17	49.98 / 24.58	66.38 / 38.98	38.93 / 15.74	33.14 / 11.80	67.81 / 39.44
BIT Chen et al. (2021)	96.64 / 74.70	59.13 / 29.58	83.96 / 59.23	60.15 / 28.19	51.54 / 20.52	86.22 / 59.87
ChangerEx Fang et al. (2023)	96.77 / 74.79	62.00 / 31.68	86.42 / 60.34	55.65 / 24.90	54.74 / 22.60	84.73 / 58.22
BAN Li et al. (2024)	<u>97.33</u> / 76.35	64.11 / 32.48	92.93 / 72.73	<u>69.73</u> / <u>35.66</u>	<u>75.11</u> / 39.67	89.92 / 57.72
CDMamba Zhang et al. (2025)	95.54 / 72.49	57.81 / 30.27	87.16 / 65.94	55.84 / 25.06	60.90 / 30.39	81.68 / 56.19
Change3D Zhu et al. (2025)	96.86 / 77.27	65.68 / <u>37.25</u>	92.30 / 73.31	65.04 / 34.76	72.70 / 42.74	84.66 / 59.21
Dual UNet (FSC-180k) Benidir et al. (2025)	95.91 / 73.55	62.32 / 35.59	88.20 / 67.96	60.39 / 31.21	66.83 / 36.98	83.98 / 58.87
ChangeRWKV Yang et al. (2026)	97.25 / <u>77.93</u>	<u>67.76</u> / 36.72	90.04 / 71.97	66.91 / 34.32	73.70 / <u>45.75</u>	89.43 / <u>64.69</u>
FSMamba Guo & Xu (2026)	95.43 / 71.35	63.41 / 33.06	77.62 / 52.70	51.79 / 23.26	59.18 / 27.37	84.34 / 57.17
UniChange Zhang et al. (2026)	96.59 / 70.79	60.72 / 30.61	<u>94.26</u> / <u>73.45</u>	55.57 / 24.57	71.91 / 39.04	<u>90.25</u> / 58.88
Ours	**98.01 / 80.49**	**75.36 / 44.23**	**94.64 / 78.77**	**76.01 / 42.30**	**77.53 / 48.16**	**90.71 / 68.39**

the $t = 0$ rollout objective also degrades performance to 91.48%/84.30%, confirming the importance of directly supervising the integrated transport outcome from the image-derived source state. These results support the joint supervision of transport dynamics and geometric predictions.

Prediction and vector-field designs. Table 11 compares TempoFlow-CD with a discriminative endpoint predictor and alternative vector-field designs. The full model achieves 92.68%/86.36% F1/IoU, compared with 91.63%/84.56% for direct endpoint prediction, a gain of 1.05/1.80 percentage points under the evaluated configurations. Replacing the RWKV field with convolutions or removing time conditioning lowers F1 by 1.28 and 1.49 percentage points, respectively. Replacing global-time decay modulation with FiLM lowers F1 by 1.71 percentage points. This substitution concerns the RWKV branch rather than CASC's spatial FiLM. These comparisons favor the proposed transport implementation and time-conditioned RWKV field among the tested designs.

Table 10: Supervision ablation for evidence-to-geometry transport on LEVIR-CD (%).

Variant	F1	IoU	P	R
w/o FM Loss	91.23	83.87	88.75	93.84
w/o Geometry Loss	91.31	84.00	89.06	93.68
w/o Quiet-region Loss	91.24	83.89	89.18	93.39
w/o Distance-field Loss	91.12	83.70	88.23	94.21
w/o $t = 0$ Rollout	91.48	84.30	91.82	91.14
Ours	**92.68**	**86.36**	92.78	92.59

Table 11: Comparison of prediction and vector-field designs on LEVIR-CD (%).

Variant	F1	IoU	P	R
Discriminative (w/o FM)	91.63	84.56	90.50	92.79
Conv. Vector Field	91.40	84.16	88.76	94.20
w/o Time Conditioning	91.19	83.80	88.76	93.75
FiLM instead of Decay	90.97	83.44	88.28	93.83
Ours	**92.68**	**86.36**	92.78	92.59

Table 12: Comparison of CASC and alternative spatial-conditioning designs on LEVIR-CD (%).

Variant	F1	IoU	P	R
Auxiliary Supervision Only	91.57	84.45	91.40	91.74
Positional-Time Conditioning	91.60	84.50	90.20	93.04
Direct Offset FiLM	91.34	84.06	88.93	93.87
Velocity Scaling	91.58	84.46	89.38	93.89
Ours	**92.68**	**86.36**	92.78	92.59

Table 13: Effect of Euler transport steps on LEVIR-CD. (a) Separately trained models. (b) A fixed $N = 4$ checkpoint evaluated with different inference steps. F1/IoU are in %. Time is ms/image.

N	F1	IoU	FPS	Time
(a) Separately trained models				
1	91.86	84.94	67.59	14.80
2	92.25	85.62	45.80	21.84
4	92.68	86.36	29.16	34.30
6	92.72	86.42	19.21	52.05
8	**92.81**	**86.59**	16.12	62.03
(b) Fixed $N = 4$ checkpoint				
1	92.08	85.32	69.36	14.42
2	92.30	85.70	47.33	21.13
4	92.68	86.36	29.16	34.30
6	92.74	86.47	20.80	48.08
8	**93.04**	**87.00**	16.22	61.65

Alternative spatial-conditioning designs. Table 12 compares CASC with auxiliary supervision and alternative spatial-conditioning designs. *Auxiliary Supervision Only* retains the same offset, boundary, and smoothness objectives while disabling their modulation of transport velocity, yielding 91.57%/84.45% F1/IoU, comparable to the NIGF-only setting of 91.71%/84.69% in Table 3. The evaluated *Positional-Time Conditioning*, *Direct Offset FiLM*, and *Velocity Scaling* achieve 91.60%/84.50%, 91.34%/84.06%, and 91.58%/84.46%, respectively, whereas CASC reaches 92.68%/86.36%. These results favor the proposed combination of geometry-supervised phase conditioning and routed velocity refinement over the evaluated alternatives.

Effect of the number of transport steps. Table 13(a) compares separately trained configurations. Increasing N from 1 to 4 improves F1/IoU from 91.86%/84.94% to 92.68%/86.36%, while increasing inference time from 14.80 to 34.30 ms/image. Additional steps yield smaller gains at further computational cost, motivating $N = 4$ as the default trade-off. Table 13(b) evaluates a fixed $N = 4$-trained checkpoint with different inference steps. F1/IoU increases from 92.08%/85.32% at $N = 1$ to 93.04%/87.00% at $N = 8$, showing that this checkpoint can benefit from additional field evaluations beyond its four-step training rollout.