RIFT-CD: PRESERVING SPATIAL DETAIL AND RELATIONAL INVARIANCE FOR ROBUST REMOTE SENSING CHANGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing change detection requires simultaneously preserving fine-grained spatial evidence and learning temporal representations that generalize across diverse scenes. However, modern pretrained vision backbones often compress spatial information aggressively, while conventional temporal fusion can still exploit scene-specific appearance cues, limiting robustness under small structural changes and distribution shifts. We introduce RIFT-CD, a relationally normalized, resolution-preserving backbone for change detection. Rather than treating temporal reasoning as an auxiliary prediction component, RIFT-CD integrates change-sensitive representation learning directly into the feature trunk. The architecture maintains high-resolution semantic representations throughout the hierarchy, adopts information-preserving feature transitions, and performs temporal interaction in a content-normalized relational space. We further regularize backbone adaptation to retain transferable pretrained representations while preventing late-stage specialization to training scenes. This design provides a unified way to balance spatial fidelity, temporal discrimination, and cross-scene generalization within a single end-to-end model. Experiments across multiple change-detection benchmarks demonstrate consistent improvements in segmentation quality, boundary localization, and robustness to appearance variation, suggesting that preserving information and constraining representation drift are central to building more generalizable change detectors.

1 INTRODUCTION

Remote sensing change detection identifies structural changes between two observations while rejecting differences caused by illumination, seasonal appearance, and viewing conditions. Building updates often occupy only a few pixels, yet appearance variation can affect an entire scene. A useful representation must therefore preserve fine spatial evidence and distinguish meaningful temporal differences from nuisance variation, as illustrated by LEVIR-CD and S2Looking (Chen & Shi, 2020; Shen et al., 2021).

Siamese networks compare multiscale features through concatenation or differencing (Daudt et al., 2018; Fang et al., 2022). Attention and transformer models enrich these comparisons with spatial and temporal context (Chen et al., 2022; Bandara & Patel, 2022); feature-interaction methods such as Changer let one date influence the encoding of the other (Fang et al., 2023). Strong pretrained representations (Siméoni et al., 2026) make this interaction attractive, but also sharpen a central question: *what temporal evidence should be fed back into a pretrained encoder?* A cross-date route can carry both structural change and irrelevant appearance differences into later features.

Our controlled experiments reveal that the answer depends on the evidence being exchanged. With raw differences and pair-mean appearance, cross-date feedback reduces F1 by 0.09 points. Normalization alone stabilizes the feedback message but produces no clean F1 gain; replacing the pair mean alone increases the relative update magnitude in changed regions but remains sensitive to appearance perturbations. Combining these changes yields both stable messages and more selective updates: feedback then adds 0.11 F1 and 2.56 boundary-band F1 points. The resulting Full-versus-Raw feedback interaction is positive, with 95% paired-scene intervals of [0.09, 0.31] for F1 and [1.50,

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.

3.54] for boundary-band F1. These findings identify **nuisance stability and change selectivity as complementary properties of useful feedback**.

RIFT-CD builds this principle into a resolution-preserving backbone. A stride-two stem exposes four times as many semantic positions per level while retaining pretrained convolutional weights. Shared cross-window attention gathers context from the other date; normalized discrepancies and channel-wise agreement represent temporal relations without a direct pair-mean input. Directed residual feedback passes this evidence into both branches before the next encoder stage, and native image differences refine localization during decoding. Dense sampling preserves the structures to be compared, while relational conditioning determines what the feedback route carries.

Our contributions are threefold:

- **An empirical account of feedback utility.** An eight-model normalization–pair-mean–feedback study shows how evidence construction reverses the effect of cross-date feedback. Matched perturbation tests and internal measurements connect this reversal to reduced nuisance propagation and greater change selectivity.
- **A backbone that conditions encoding on normalized relations.** Dense pretrained semantics, discrepancy-and-agreement evidence, and directed residual updates integrate spatial preservation with temporal reasoning. We establish the descriptor's positive-scale invariance and the feedback path's temporal symmetry.
- **Evidence across detection, localization, and mechanism.** Results on seven benchmarks are complemented by matched spatial and compute controls, scene-paired interaction intervals, tiny-component evaluation, and learned-feature probes. Together they connect the design to both accuracy and the behavior of the exchanged representation.

RIFT-CD achieves test F1 scores of 92.92%, 87.43%, 69.23%, 96.96%, and 99.95% on LEVIR-CD, LEVIR-CD+, S2Looking, WHU-CD, and CDD, respectively. Under matched LEVIR-CD validation settings, the full model exceeds a stride-two SimpleDiff baseline by 0.21 F1 points (95% CI: [0.09, 0.32]). The mechanism study explains why relational conditioning makes the same feedback route beneficial.

2 Related Work

Siamese representations and temporal context. FC-Siam uses shared convolutional encoders with concatenation or feature differencing (Daudt et al., 2018); SNUNet-CD strengthens multiscale fusion through densely connected Siamese decoding (Fang et al., 2022). STANet introduces spatial-temporal attention (Chen & Shi, 2020), BIT models context through compact semantic tokens (Chen et al., 2022), and ChangeFormer uses hierarchical transformer features (Bandara & Patel, 2022). More recently, RS-Mamba and SPMNet employ state-space modeling for remote sensing dense prediction and change detection (Zhao et al., 2024; Wang et al., 2025), while ConvFormer-CD combines convolutional and transformer processing with temporal attention (Yang et al., 2025). These approaches enrich context and feature fusion. Our focus is complementary: preserving the spatial density of pretrained semantics and controlling the evidence used to compare them.

Feature interaction and nuisance variation. Changer introduces interaction during feature extraction through aggregation-distribution and feature exchange (Fang et al., 2023). RIFT-CD develops this direction with shared windowed attention, normalized relational states, and residual feedback across successive stages. We study which evidence makes this interaction useful. The normalization–pair-mean–feedback comparison connects feedback gains to message stability and change selectivity; normalized discrepancy and agreement provide the corresponding evidence design.

Adapting pretrained semantics. ConvNeXt supplies a multiscale convolutional hierarchy (Liu et al., 2022), and DINOv3 learns transferable dense representations through self-supervised pretraining (Siméoni et al., 2026). ChangeDINO combines frozen DINOv3 features with a lightweight Siamese stream, a spatial–spectral differential decoder, and learnable morphology (Cheng & Hsu, 2026). RIFT-CD adapts the DINOv3 ConvNeXt hierarchy through three linked choices: a stride-two stem preserves denser semantics, normalized relations constrain temporal comparison, and feedback conditions subsequent encoder stages on both dates.

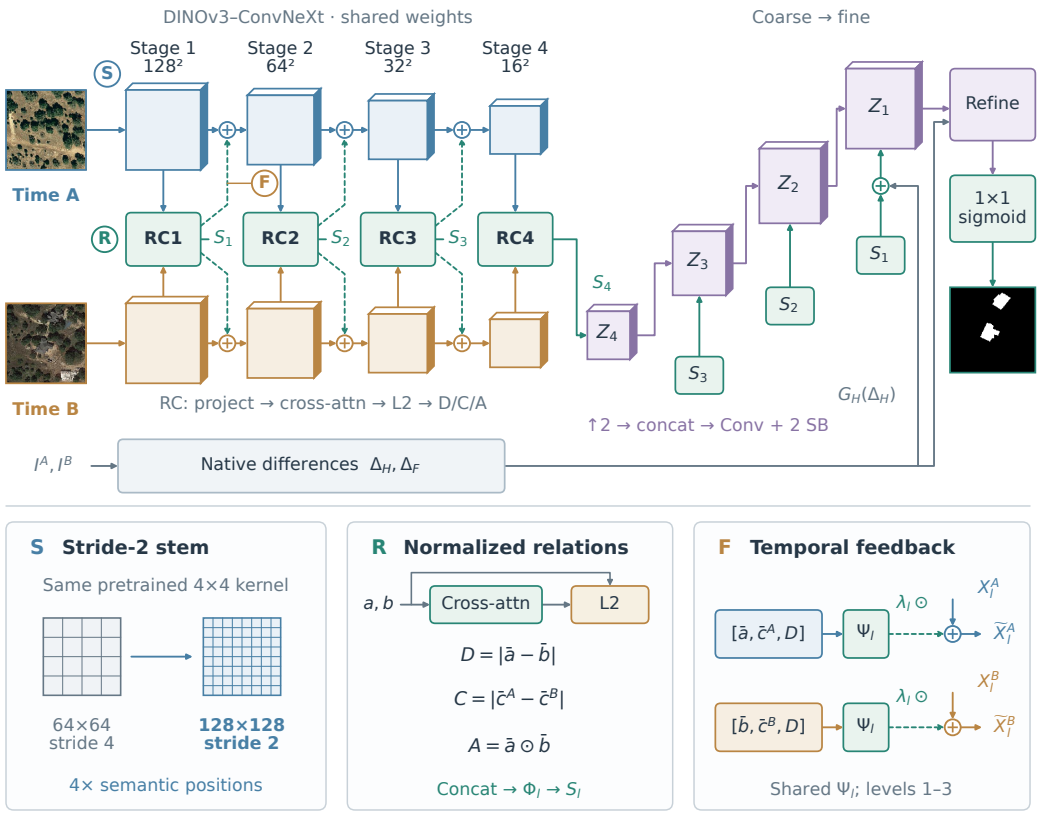

Figure 1: RIFT-CD overview. The two dates share encoder weights. Tags S/R/F identify the details below; matched S_l labels indicate decoder skips. Dashed arrows update both dates before stages 2–4. Native differences refine S_1 and the output. RC: relational coupler; SB: spatial block. The mask is ground truth; auxiliary supervision is omitted.

3 METHOD

Given co-registered images $I^A, I^B \in \mathbb{R}^{3\times H\times W}$, RIFT-CD predicts a binary change map through a shared four-stage encoder and a change-state decoder (Figure 1). The design first preserves dense semantic evidence, then constrains its temporal comparison, and finally feeds directed relational evidence into subsequent encoding. At level l, temporal features X_l^A, X_l^B enter a relational coupler that produces a symmetric change state S_l. The first three couplers also update both temporal branches before the next encoder stage. We denote these factors by S (semantic resolution), R (relational evidence), and F (feedback).

3.1 HIGH-RESOLUTION FOUNDATION SEMANTICS (S)

We initialize the encoder with DINOv3 ConvNeXt-Large and fully fine-tune its retained stages. The 4×4 stem retains its pretrained kernel, but uses stride two and padding one. The hierarchy therefore operates at $1/2$, $1/4$, $1/8$, and $1/16$ input resolution, with stage widths $(192, 384, 768, 1536)$ and depths $(3, 3, 27, 3)$. Every level contains four times as many positions as its stride-four counterpart, allowing finer structures to undergo semantic processing before fusion.

The spatial benefit follows a simple sampling argument. Under an idealized nearest-grid localization model with uniform offset in $[-s/2, s/2]$, the per-axis quantization error has variance $s^2/12$. Halving the stride reduces this sampling variance by 75%, motivating denser semantic processing for small structures and boundaries. Section 4.2 evaluates the resulting stride adaptation in terms of segmentation quality and computational cost.

Shared native paths complement the semantic hierarchy: $H^t = G_{1/2}(\mathrm{Unshuffle}_2(I^t))$ and $F^t = G_1(I^t)$ for $t \in \{A, B\}$. Pixel unshuffle rearranges each 2×2 RGB neighborhood into channels before learned projection. The decoder receives normalized differences from these paths at half and full resolution.

3.2 Normalized Relational Evidence (R)

Figure 2(a) details the relational construction. Shared convolutional projections with channel normalization and GELU map each stage to width 128, giving $a_l = P_l(X_l^A)$ and $b_l = P_l(X_l^B)$. A shared channel-dropout mask preserves temporal correspondence during training. Shared cross-window attention then constructs context from the other date:

$$c_l^A = \mathcal{A}(a_l, b_l), \quad c_l^B = \mathcal{A}(b_l, a_l), \qquad \mathcal{A}(x, y) = W_o\left[\mathrm{softmax}\left(\frac{Q(x)K(y)^\top}{\sqrt{d_h}}\right)V(y)\right]. \tag{1}$$

The expression applies per head and window, with head concatenation before W_o. A shared depthwise positional convolution and channel normalization precede the projections. We use four heads and window sides $(8, 8, 16, 8)$ across the four levels.

For each channel vector, let $\mathcal{N}(x) = x/\max(\|x\|_2, \epsilon)$ with $\epsilon = 10^{-6}$, and use a bar for normalized features. The relational descriptor is

$$D_l = |\bar{a}_l - \bar{b}_l|, \qquad C_l = |\bar{c}_l^A - \bar{c}_l^B|, \qquad A_l = \bar{a}_l \odot \bar{b}_l, \tag{2}$$

$$R_l = D_l \parallel C_l \parallel A_l, \qquad S_l = \Phi_l(R_l), \tag{3}$$

where $\parallel$ denotes concatenation and $\odot$ elementwise multiplication. Φ_l applies a 3×3 projection followed by kernel-three and kernel-five spatial blocks. Each block uses depthwise convolution, channel normalization, and pointwise expansion/projection (ratio two, with GELU) within a scaled residual path. Direct discrepancy, contextual discrepancy, and channel-wise agreement provide complementary evidence without a direct absolute pair-mean input. For unit vectors u, v, $\|u - v\|_2^2 = 2 - 2u^\top v$ and $\sum_c u_c v_c = u^\top v$, connecting the descriptor to angular separation while retaining channel structure.

Geometric property. For $\alpha > 0$, normalization satisfies $\mathcal{N}(\alpha x) = \mathcal{N}(x)$ whenever both norms exceed ϵ. Away from that floor, its Jacobian is

$$J_{\mathcal{N}}(x) = \frac{I - \bar{x}\bar{x}^\top}{\|x\|_2}, \qquad J_{\mathcal{N}}(x)x = 0. \tag{4}$$

Thus radial magnitude perturbations vanish to first order, while tangential changes remain, as illustrated in Figure 2(b). Under the same norm condition, independent positive rescaling of the four descriptor inputs leaves the relational state unchanged. This property applies to the projected features and attention outputs entering normalization. Appendix C gives the derivations; Appendix B.1 tests learned representations and the norm-floor transition.

3.3 Cross-Temporal Feedback Inside the Backbone (F)

The symmetric state serves change decoding, while directed residual feedback conditions subsequent semantic extraction (Figure 2(c)). For $l < 4$,

$$\widetilde{X}_l^A = X_l^A + \lambda_l \odot \Psi_l(\bar{a}_l \parallel \bar{c}_l^A \parallel D_l),$$

$$\widetilde{X}_l^B = X_l^B + \lambda_l \odot \Psi_l(\bar{b}_l \parallel \bar{c}_l^B \parallel D_l), \qquad X_{l+1}^t = E_{l+1}(\widetilde{X}_l^t). \tag{5}$$

The two dates share Ψ_l, a width-128 projection followed by a 1×1 convolution to the stage width, and the learned per-channel scale λ_l, initialized to 10^{-3}. Feedback therefore starts with a small residual multiplier and lets deeper features depend on the other observation.

Information flow and symmetry. With fixed weights and stochastic masks, independent temporal encoding, including self-temporal feedback, satisfies $\partial X_{l+1}^A / \partial I^B = 0$. Cross-temporal feedback introduces a path for this derivative to be nonzero, allowing later stages to extract features conditioned on the image pair. Swapping the dates exchanges the two attention outputs and feedback updates while preserving R_l. Induction across stages, together with symmetric native differences, yields $f_\theta(I^A, I^B) = f_\theta(I^B, I^A)$ at deterministic inference, up to numerical precision.

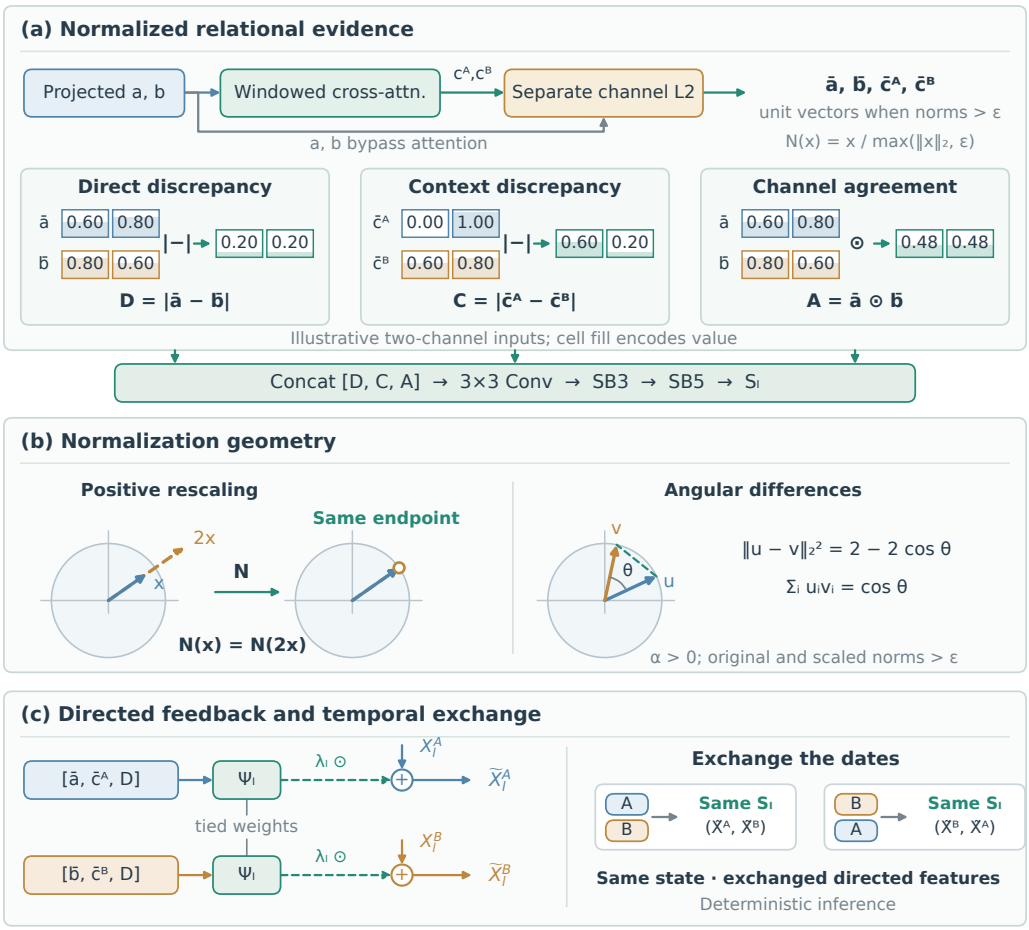

Figure 2: Relational evidence and feedback. (a) Two-channel mathematical examples illustrate discrepancy and agreement using independently specified contextual vectors. SB3/SB5 denote spatial blocks with kernel sizes three/five. (b) Positive rescaling preserves direction when the original and scaled norms exceed ϵ. (c) Temporal exchange preserves the change state and swaps directed updates.

3.4 Decoding and Supervision

Native differences $\Delta_H = |\mathcal{N}(H^A) - \mathcal{N}(H^B)|$ and $\Delta_F = |\mathcal{N}(F^A) - \mathcal{N}(F^B)|$ refine localization. We first set $S_1' = S_1 + G_H(\Delta_H)$. Starting from $Z_4 = S_4$, a top-down decoder concatenates bilinearly upsampled coarse states with finer states, using S_1' at the finest level. Each fusion applies a 3×3 projection and two kernel-five spatial blocks at width 128. Half/full-resolution refinements incorporate Δ_H and Δ_F at widths 64/32 before the logit z (Figure 1); the auxiliary head reads Z_1 before refinement. Training minimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}}(z, Y) + \mathcal{L}_{\mathrm{Dice}}(z, Y) + 0.20\, \mathcal{L}_{\mathrm{BCE}}(z_{1/2}, \mathrm{AreaDown}_{1/2}(Y)). \tag{6}$$

The auxiliary target represents fractional changed-area occupancy. Dice statistics are aggregated within each distributed microbatch, with its loss disabled for all-background targets. Inference uses one EMA model, threshold 0.50, and no test-time augmentation.

4 Experiments

4.1 Setup and Quantitative Comparison

Table 1 compares ten representative methods on seven datasets: LEVIR-CD and its expanded LEVIR-CD+ release (Chen & Shi, 2020), S2Looking (Shen et al., 2021), SYSU-CD (Shi et al.,

2022), WHU-CD (Ji et al., 2019), CDD (Lebedev et al., 2018), and MSRSCD (Liu et al., 2025). Segmentation scores are percentages.

Benchmark implementation. The model processes 256×256 image pairs with semantic/native widths 128/32. On LEVIR-CD, AdamW trains for 16,000 updates with head/stem/backbone learning rates $(2 \times 10^{-4}, 2 \times 10^{-5}, 6 \times 10^{-6})$, 600 warmup updates, cosine decay, weight decay 0.05, and EMA decay $1 - 10^{-3}$. The lower backbone learning rate limits updates to pretrained features, while the head adapts more rapidly to change detection. Two devices use four pairs each and two accumulation steps. Synchronized geometry, temporal swaps, photometric changes, channel dropout (0.05), and CutMix (0.15) regularize training. Evaluation uses threshold 0.50 without test-time augmentation. Appendix A details the comparison and mechanism-probe protocols.

Table 1: F1 (%) across seven datasets for ten representative comparison methods.

Method	LEVIR	LEVIR+	S2L	SYSU	WHU	CDD	MSRSCD
FC-Siam-Di	86.31	59.14	32.09	64.08	59.88	69.63	65.07
STANet	88.71	62.58	37.72	75.06	78.94	87.50	70.57
SNUNet-CD	88.16	69.86	44.91	77.43	86.73	96.77	69.63
BIT	89.31	71.08	55.94	79.39	89.87	95.52	73.18
ChangeFormer	90.50	80.46	63.39	80.71	92.25	96.31	72.58
Changer	92.06	86.93	67.08	78.02	91.51	95.89	72.90
RS-Mamba	89.77	80.91	62.56	80.46	92.79	94.41	74.40
SPMNet	91.34	83.09	63.78	77.53	91.80	96.04	71.10
ConvFormer-CD	91.44	85.30	68.00	79.35	92.13	93.96	74.60
ChangeDINO	92.31	84.89	67.13	84.70	94.18	99.76	78.22
RIFT-CD	**92.92**	**87.43**	**69.23**	**85.87**	**96.96**	**99.95**	**78.35**

Detection performance. RIFT-CD achieves the highest F1 in Table 1 on all seven datasets. It exceeds the strongest comparison method by 0.61 points on LEVIR-CD, 0.50 on LEVIR-CD+, 1.23 on S2Looking, 1.17 on SYSU-CD, 2.78 on WHU-CD, and 0.13 on MSRSCD. On CDD, F1 reaches 99.95 compared with ChangeDINO's 99.76. These improvements span building-change, urban, and multiscale change-detection settings.

Comparison with DINOv3-based methods. ChangeDINO uses a frozen DINOv3 feature stream (Cheng & Hsu, 2026), providing a comparison within the same pretraining family. RIFT-CD improves its F1/IoU by 2.54/3.93 points on LEVIR-CD+ and 0.19/0.39 on CDD. On MSRSCD, RIFT-CD reaches 78.35/64.41 F1/IoU, compared with 78.22/64.23 for ChangeDINO, a gain of 0.13/0.18 points.

Localization quality. We evaluate boundary localization with *Band F1*, the pixel F1 within a one-pixel ground-truth boundary band. LEVIR-CD reaches 86.78 IoU and 71.96 Band F1, and LEVIR-CD+ reaches 77.67 and 61.59. On the more challenging S2Looking dataset, IoU is 52.94, Band F1 is 50.47, and recall is 67.32.

4.2 Spatial resolution and computational cost

We separate the spatial and compute controls from the evidence-by-feedback study. Table 2 compares RIFT-S4, SimpleDiff-S2, and complete RIFT-S2 under shared DINOv3 initialization, seed policy, 16,000 updates, decoder, augmentation, and threshold. SimpleDiff-S2 keeps the dense backbone and native-detail paths, using feature differencing in place of relational coupling and feedback.

RIFT-S2 improves on SimpleDiff-S2 by 0.21 F1 points (95% paired-scene CI: [0.09, 0.32]) and 2.68 Band F1 points, with 0.89% more parameters and 3.56% more counted operations. Relative to RIFT-S4, denser semantic processing adds 0.32 F1 and 2.96 Band F1 points at $3.75\times$ the counted operations. Appendix D specifies the counting convention. These controls establish the value and cost of spatial density before we examine the evidence carried by feedback.

Table 2: Spatial and compute controls on LEVIR-CD validation. Detection metrics are percentages. Parameters (M) and counted operations (G) refer to one 256×256 image pair. RIFT-S2 is the full cross-feedback model.

Model	F1	Precision	Recall	Band F1	Params	Ops
RIFT-S4	92.82	92.98	92.66	69.90	201.17	106.32
SimpleDiff-S2	92.93	93.00	92.86	70.18	199.40	385.32
RIFT-S2	**93.14**	**93.19**	**93.10**	**72.86**	**201.17**	**399.03**

Table 3: Evidence changes the benefit of feedback. Eight matched LEVIR-CD validation models cross normalization (N), pair-mean replacement by agreement (M), and feedback route (F). Each row reports Self $\rightarrow$ Cross; differences are percentage points with 95% paired-scene bootstrap intervals. The final row is the Full-versus-Raw difference of feedback gains, $\mathcal{I}_F$.

Evidence (N, M)	F1	Δ_F [95% CI]	Band F1	Δ_F [95% CI]
Raw (0,0)	92.98→92.89	−0.09 [-0.18, -0.02]	70.26→70.29	+0.03 [-0.31, +0.34]
Norm only (1,0)	93.08→93.08	+0.00 [-0.06, +0.06]	70.62→70.74	+0.12 [-0.15, +0.39]
No mean only (0,1)	93.07→93.08	+0.01 [-0.05, +0.07]	70.58→70.81	+0.23 [-0.07, +0.53]
Full (1,1)	93.03→93.14	+0.11 [+0.03, +0.19]	70.30→72.86	+2.56 [+1.45, +3.57]
Full − Raw	—	+0.20 [+0.09, +0.31]	—	+2.53 [+1.50, +3.54]

4.3 Evidence determines the benefit of feedback

We train eight models crossing three binary factors: relational normalization (N), replacement of the pair mean by channel-wise agreement (M), and self- versus cross-temporal feedback (F). All use DINOv3 ConvNeXt-Large foundation-only initialization, stride two, seed 20260919, the same decoder and native-detail paths, and the same 16,000-update training and checkpoint-selection policy. No change-detection checkpoint is used for initialization. Appendix A details the shared protocol.

For a metric Q, define the feedback gain and the Full-versus-Raw interaction as

$$\Delta_F^Q(N, M) = Q_{N,M,\text{Cross}} - Q_{N,M,\text{Self}}, \qquad \mathcal{I}_F^Q = \Delta_F^Q(1, 1) - \Delta_F^Q(0, 0). \tag{7}$$

The interaction asks whether replacing raw evidence with the full relational design changes the benefit of the same feedback route. Table 3 reports 95% paired-bootstrap intervals over 64 original scenes, keeping all compared models paired within each resample.

From harmful to useful feedback. Raw evidence yields a feedback gain of -0.09 F1 points, with an interval entirely below zero. Normalization alone yields 0.00, and pair-mean replacement alone yields $+0.01$. Full evidence yields $+0.11$ [0.03, 0.19], together with $+2.56$ [1.45, 3.57] Band F1. The Full-versus-Raw interaction reaches $+0.20$ [0.09, 0.31] F1 and $+2.53$ [1.50, 3.54] Band F1. Both interaction intervals lie above zero: the value of feedback changes with the evidence it carries.

Evidence suited to feedback. Full-Self has lower F1 than the two partial variants (93.03 versus 93.08 and 93.07), yet Full-Cross reaches the highest F1, 93.14. Stand-alone self-feedback performance therefore does not predict which evidence will benefit most from cross-date exchange. The larger boundary gain further localizes the benefit to structural detail. We next examine the messages and updates that accompany this reversal.

4.4 Stable messages and selective updates

We probe the cross-feedback models with two complementary measurements. *Message drift* measures the relative change in the A-directed feedback message when only date B receives a radiometric perturbation. *Update selectivity* is the ratio of changed-region to unchanged-region mean norms of the actual injected update $u_l^A = \lambda_l \odot m_l^A$. Ground-truth area occupancy weights the two regions at each feature resolution. Appendix A.4 gives their definitions.

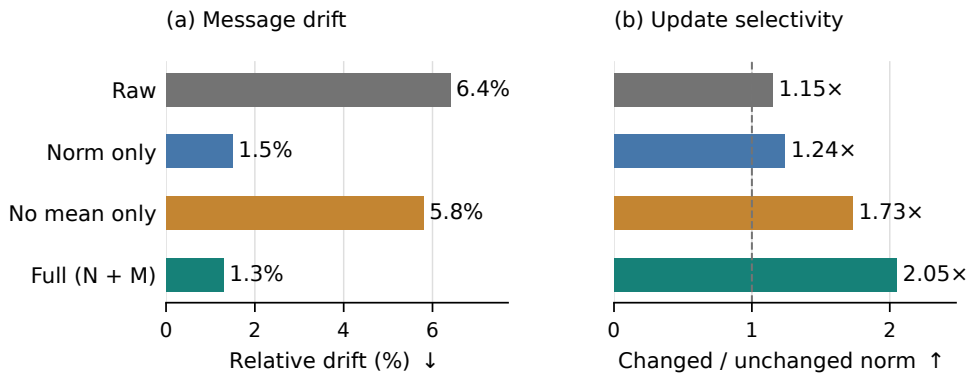

Figure 3: Complementary effects of evidence construction in the cross-feedback models. (a) Relative B-to-A message drift under radiometric stress ($s = 2$) on LEVIR-CD validation. (b) Changed/unchanged norm ratio of the injected feedback update; the dashed line marks equal mean magnitude. Normalization is associated with lower drift, and pair-mean replacement with greater change selectivity. Full evidence combines both properties. Bars are reported point estimates; Appendix A.4 defines the quantities.

Table 4: Feedback under appearance stress. Radiometric shifts (Rad) and Gaussian noise perturb date B at $s = 2$ on the same 64 scenes. Δ_F is Cross minus Self F1 under each perturbation. ΔU-FPR is the Cross-minus-Self difference in the perturbation-induced unchanged-pixel false-positive increase; negative values indicate reduced nuisance amplification. All entries are percentage points. Appendix A.4 defines the metrics.

	Δ_F F1 ↑		ΔU-FPR ↓	
Evidence	Rad	Noise	Rad	Noise
Raw	-0.17	-0.13	+0.055	+0.041
Norm only	-0.02	-0.01	+0.006	+0.009
No mean only	-0.05	-0.04	+0.044	+0.032
Full	+0.09	+0.07	-0.012	-0.008
Full − Raw	+0.26	+0.20	-0.067	-0.049

The two evidence changes have distinct effects. Normalization alone reduces message drift from 6.4% to 1.5%, while its update ratio remains 1.24. Pair-mean replacement alone raises the ratio from 1.15 to 1.73, while drift remains 5.8%. Full evidence combines low drift (1.3%) with the largest ratio (2.05). This joint pattern explains the feedback comparison: stable messages alone yield little benefit, while change-focused updates alone still carry appearance variation. The ratio also distinguishes greater spatial selectivity from a uniform reduction in update magnitude.

Reduced propagation of appearance nuisance. We perturb date B before tiling using the same radiometric and Gaussian-noise protocol as the full-model stress tests, at strength $s = 2$: RGB gain $(1.16, 0.88, 1.04)$ with offset 0.04, or Gaussian noise with standard deviation 0.04. We compare the Cross-minus-Self F1 gain and the difference in perturbation-induced unchanged-pixel false positives, ΔU-FPR (Table 4). With raw evidence, feedback reduces stressed F1 by 0.17/0.13 points and adds 0.055/0.041 points of nuisance-induced false positives under radiometric/noise perturbations. Full evidence reverses both patterns: feedback adds 0.09/0.07 F1 and reduces nuisance amplification by 0.012/0.008 points. The corresponding Full-versus-Raw contrasts are $+0.26/+0.20$ F1 and $-0.067/-0.049$ false-positive points. Thus the clean interaction is accompanied by improved feedback behavior under appearance variation.

4.5 Spatial and learned-feature checks

Separate fixed-endpoint diagnostics connect the remaining architectural choices to their intended behavior (Appendix B). On 49 ground-truth components of area 1–64 pixels, stride two improves Recall50 over stride four from 20.41% to 32.65% with raw evidence and to 34.69% with normalized evidence; paired-scene intervals for both gains lie above zero. Positive rescaling changes normalized descriptors by only 1.94–5.05×10^{-7} across stages, compared with 1.85–3.50 for raw descriptors. With identical weights, finite-difference and gradient probes detect a B-to-A information path under cross-feedback and exactly zero response under self-feedback.

The full-model perturbation sweep also distinguishes appearance changes from lost spatial information. At maximum strength, radiometric shifts and noise reduce F1 by 0.08 and 0.28 points, whereas Gaussian blur reduces it by 9.24 points and lowers recall to 81.60. This sensitivity to strong smoothing reinforces the role of fine spatial evidence. Appendix B reports the complete sweep and percentile-selected predictions, including missed and overextended regions.

5 Conclusion

RIFT-CD preserves dense pretrained semantics and conditions subsequent encoding on normalized temporal relations. Its central finding is that the benefit of cross-date feedback depends on the evidence being exchanged. The factorial study separates two complementary effects: normalization stabilizes messages under appearance variation, while replacing the pair mean increases the relative update magnitude in changed regions. Their combination reverses the feedback effect, improves boundary localization, and reduces nuisance-induced false positives. Together with seven-benchmark results and spatial controls, these findings provide a concrete design principle for change detection: preserve fine structural evidence and make temporal feedback both stable to nuisance variation and selective for change.

AI use. Generative AI tools were used only to provide assistance with method implementation and result interpretation. Experimental design and mathematical derivations and proofs were conducted independently by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

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. doi: 10.1109/igarss46834.2022.9883686.

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. doi: 10.3390/rs12101662.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. doi: 10.1109/tgrs.2021.3095166.

Ching-Heng Cheng and Chih-Chung Hsu. ChangeDINO: DINOv3-Driven Building Change Detection in Optical Remote Sensing Imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XI-3-2026:523–530, 2026. doi: 10.5194/isprs-annals-XI-3-2026-523-2026.

Rodrigo Caye Daudt, Bertrand 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. doi: 10.1109/icip.2018.8451652.

Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. doi: 10.1109/lgrs.2021.3056416.

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. doi: 10.1109/tgrs.2023.3277496.

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, 2019. doi: 10.1109/tgrs.2018.2858817.

M. A. Lebedev, Y. V. Vizilter, O. V. Vygolov, V. A. Knyaz, and A. Y. Rubis. Change Detection in Remote Sensing Images Using Conditional Adversarial Networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2:565–571, 2018. doi: 10.5194/isprs-archives-xlii-2-565-2018.

Shenbo Liu, Dongxue Zhao, Yuheng Zhou, Ying Tan, Huang He, Zhao Zhang, and Lijun Tang. Network and Dataset for Multiscale Remote Sensing Image Change Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:2851–2866, 2025. doi: 10.1109/jstars.2024.3522135.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01167.

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. doi: 10.3390/rs13245094.

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, 2022. doi: 10.1109/tgrs.2021.3085870.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=2NlGyqNjns.

Jiashu Wang, Jinze Song, Hao Zhang, Zekai Zhang, Yunlong Ji, Wenyin Zhang, Jinglin Zhang, and Xing Wang. SPMNet: A Siamese Pyramid Mamba Network for Very-High-Resolution Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. doi: 10.1109/tgrs.2025.3565801.

Feng Yang, Mengtao Li, Wenqiang Shu, Anyong Qin, Tiecheng Song, Chenqiang Gao, and Gui-Song Xia. ConvFormer-CD: Hybrid CNN–Transformer With Temporal Attention for Detecting Changes in Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/tgrs.2025.3544651.

Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. RS-Mamba for Large Remote Sensing Image Dense Prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/tgrs.2024.3425540.

A Experimental Protocol

A.1 Spatial and compute controls

Table 2 evaluates complete LEVIR-CD validation with shared DINOv3 initialization, seed policy, 16,000 updates, decoder, augmentation, and threshold. RIFT-S4 changes semantic stride; SimpleDiff-S2 retains the dense backbone and native-detail paths while replacing relational coupling and feedback with feature differencing. RIFT-S2 implements the complete model in Section 3. The RIFT-S2-minus-SimpleDiff comparison yields $+0.21$ F1 points, with 95% paired-scene CI [0.09, 0.32].

A.2 Eight-model evidence-by-feedback study

Table 3 crosses normalization ($N \in \{0, 1\}$), pair-mean replacement ($M \in \{0, 1\}$), and feedback route ($F \in \{\mathrm{Self}, \mathrm{Cross}\}$). The pair-mean input is retained when $M = 0$ and replaced by channel-wise agreement when $M = 1$. Thus M denotes a descriptor replacement. The full $(N, M) = (1, 1)$ descriptor is given by Eqs. 2–3, and Eq. 5 gives cross-temporal feedback. The self route uses own-date feedback with the same projection architecture.

All eight models start from DINOv3 ConvNeXt-Large foundation-only weights, with seed 20260919 and no initialization from a change-detection checkpoint. They share a stride-two stem, decoder, native-detail paths, widths 128/32, loss, augmentation, 256×256 crops, and effective batch size 16. Training uses 16,000 updates, 600 warmup updates, head/stem/backbone learning rates $(2 \times 10^{-4}, 2 \times 10^{-5}, 6 \times 10^{-6})$, AdamW with weight decay 0.05, cosine decay, and EMA decay 0.999. Validation interval and checkpoint-selection rule are shared across all arms. Evaluation uses threshold 0.50, one model, and no test-time augmentation. The primary study uses one common training seed.

A.3 Paired-scene evaluation and interaction

The original scene is the resampling unit for the 95% paired-bootstrap intervals over 64 validation scenes. The same resampled scenes are used for all models entering a comparison, retaining the grouping of tiles within each scene. For the interaction, the Full-Cross, Full-Self, Raw-Cross, and Raw-Self results stay paired when computing Eq. 7. This evaluates the difference of feedback gains directly. F1, precision, recall, and Band F1 are percentages; differences and interval endpoints are percentage points.

A.4 Feedback behavior metrics

At feedback site $l \in \{1, 2, 3\}$, write the message and injected update as

$$m_l^A = \Psi_l(\bar{a}_l \,\|\, \bar{c}_l^A \,\|\, D_l), \qquad u_l^A = \lambda_l \odot m_l^A, \qquad \widetilde{X}_l^A = X_l^A + u_l^A. \tag{8}$$

The input normalization follows the evidence arm. Figure 3(a) reports the drift of the message m, and panel (b) uses the magnitude of the injected update u.

Message drift. Date A is fixed and only date B is perturbed, $B' = T(B)$. For an original scene j, the relative message drift at site l is

$$d_{m,l}^{(j)}(T) = \frac{\|m_l^A(A, B') - m_l^A(A, B)\|_F}{\|m_l^A(A, B)\|_F + \epsilon_d}. \tag{9}$$

Here $\epsilon_d > 0$ stabilizes the denominator, and the Frobenius norm covers channel and spatial entries. The displayed drift values are percentages. The corresponding injected-update drift is obtained by substituting u for m; the two quantities distinguish the message representation from its learned channel-wise gating.

Change selectivity. Area-downsampling the binary ground truth yields fractional occupancy $q_l(p) = \mathrm{AreaDown}_l(Y)(p) \in [0,1]$. Let $r_l(p) = \|u_l^A(p)\|_2$. Define

$$\mu_l^{\mathrm{chg}} = \frac{\sum_p q_l(p) r_l(p)}{\sum_p q_l(p) + \epsilon_d}, \qquad \mu_l^{\mathrm{unch}} = \frac{\sum_p [1 - q_l(p)] r_l(p)}{\sum_p [1 - q_l(p)] + \epsilon_d}, \tag{10}$$

$$\rho_l = \frac{\mu_l^{\mathrm{chg}}}{\mu_l^{\mathrm{unch}} + \epsilon_d}. \tag{11}$$

Fractional occupancy retains small changes within coarse feature cells. A ratio near one indicates similar mean update magnitude in changed and unchanged regions; a larger ratio indicates greater relative emphasis on changed regions. The reported Raw and Full ratios are 1.15 and 2.05, respectively, a difference of 0.90 ratio units.

Appearance perturbations. Images have RGB values in $[0,1]$. At strength s, radiometric and noise perturbations are

$$T_{\mathrm{rad}}(I_B) = \mathrm{clip}\left(I_B \odot (1 + 0.08s, 1 - 0.06s, 1 + 0.02s) + 0.02s, 0, 1\right), \tag{12}$$

$$T_{\mathrm{noise}}(I_B) = \mathrm{clip}(I_B + 0.02s\,\eta, 0, 1), \qquad \eta \sim \mathcal{N}(0, I). \tag{13}$$

Each perturbation is applied before scene tiling, with fixed weights and labels. The matched mechanism table uses $s = 2$. A fixed noise field is shared across compared arms and nested strengths. The additional full-model sweep in Appendix B uses $s \in \{0, 0.5, 1, 2\}$ and also includes Gaussian blur.

Nuisance amplification. For model f and condition T, let $\mathrm{FPR}_U(f,T)$ be the false-positive rate on ground-truth unchanged pixels. The perturbation-induced increase and its feedback contrast are

$$\delta_U(f,T) = \mathrm{FPR}_U(f,T) - \mathrm{FPR}_U(f,\mathrm{clean}), \tag{14}$$

$$\Delta U\text{-FPR}(T) = \delta_U(\mathrm{Cross},T) - \delta_U(\mathrm{Self},T). \tag{15}$$

Positive values indicate that cross-feedback adds nuisance-induced false positives relative to self-feedback; negative values indicate suppression. Table 4 expresses these differences in percentage points. Its final row subtracts the Raw feedback contrast from the Full contrast.

A.5 Mechanism-probe protocol

The tiny-component and learned-feature probes in Section B.1 use a dedicated set of fixed-endpoint diagnostic runs. These runs use 445 training scenes, 256×256 crops, and AdamW head/stem/backbone learning rates $(1.80 \times 10^{-4}, 1.50 \times 10^{-5}, 4.00 \times 10^{-6})$, 800 warmup updates, cosine decay, and weight decay 0.05. Each update accumulates two microbatches of eight pairs; Dice is aggregated per microbatch. Crops mix grid and arbitrary locations equally. The augmentation policy includes synchronized geometry, temporal swaps, photometric changes, channel dropout (0.05), and CutMix (0.15). Evaluation uses the EMA $(1 - 10^{-3})$ at exactly 16,000 updates, threshold 0.50, and no TTA.

The diagnostic wrapper preserves parameter keys, shapes, and initial tensors. Stride-four controls area-resize native evidence and restore half-resolution features for the auxiliary head. The relational switch jointly changes normalization in semantic, native, and feedback evidence and replaces the pair mean with agreement. Self-feedback uses own-date attention and discrepancies with shared projection weights. Both routes are evaluated, with the inactive route multiplied by zero.

Validation contains 1,024 tiles from 64 original scenes. Tiny-component intervals use stitched-scene counts and 20,000 paired scene resamples within one training repeat. Learned-feature and routing probes use frozen fixed-step EMA weights in FP32: 32 consecutive crops from two scenes for scale and finite differences, and four crops from the first scene for VJP. Stress tests use a separate fixed checkpoint and clean reference.

B Additional Diagnostics and Predictions

B.1 Understanding the mechanisms

We examine the three design choices through tiny-component recall, feature rescaling, and cross-date sensitivity. Table 5 summarizes the learned-feature probes; Appendix A details the fixed-endpoint diagnostic runs.

Preserving tiny changes. We evaluate 49 ground-truth components of area 1–64 pixels using Recall50, which counts a component as recovered when at least half its area is predicted changed. Switching from stride four to stride two under self-feedback raises Recall50 from 20.41% to 32.65% with raw evidence and from 20.41% to 34.69% with normalized relational evidence. The paired-scene 95% intervals for these gains are [3.63, 24.45] and [3.45, 30.43] percentage points. Both comparisons show that denser semantic sampling recovers more tiny changes.

Table 5: Direct probes on learned features. Descriptor columns report mean relative change under positive rescaling; feedback columns report mean date-A feature change (FD) and VJP gradient RMS induced by date B. Both feedback probes are exactly zero under self-feedback with identical weights. Feedback is probed at the first three locations only.

	Descriptor rescaling		Cross-temporal feedback	
Level	Normalized (10^{-7})	Raw	FD (10^{-5})	VJP (10^{-6})
1	1.94	1.85	10.66	1.61
2	3.00	3.37	7.34	1.32
3	4.16	3.50	1.16	11.02
4	5.05	2.45	—	—

Scale invariance on learned representations. On 32 validation crops, we independently rescale descriptor inputs a, b, c^A, c^B with spatially varying positive factors drawn log-uniformly from $[0.25, 4.00]$. Mean relative changes range from 1.94×10^{-7} to 5.05×10^{-7} for the normalized descriptor, versus 1.85–3.50 for raw evidence across four stages. The raw comparator uses unnormalized discrepancies and the pair mean; this tests complete descriptor designs on learned features.

The same probe reveals the transition at the normalization floor. At stage one, relative change is 1.70×10^{-7} at target norm ϵ and rises to 0.53 and 0.91 at 0.50ϵ and 0.10ϵ. This behavior agrees with the normalization formula: above the floor, positive rescaling preserves the descriptor; below it, the fixed denominator makes the output depend on magnitude.

Feedback routing with identical weights. We switch between cross-temporal and equal-capacity self-feedback while retaining identical weights. Gaussian RGB noise of standard deviation $1/255$ on date B changes date-A features after couplers 1–3 on 32 crops. Vector–Jacobian product (VJP) input-gradient RMS is also nonzero on four crops. Both responses are exactly zero under self-feedback. Together, these probes show how cross-date feedback makes deeper features depend directly on the other observation.

B.2 Robustness under input perturbations

We test sensitivity to appearance and spatial perturbations using a fixed LEVIR-CD EMA checkpoint on all 64 validation scenes. Each perturbation is applied to date B before scene tiling at strengths $s \in \{0, 0.50, 1, 2\}$, preserving date A and the labels. Radiometric changes use RGB gains $1 + s(0.08, -0.06, 0.02)$ and offset $0.02s$ for values in $[0, 1]$. Gaussian noise has standard deviation $0.02s$, with the same noise field across strengths, and Gaussian blur has spatial standard deviation s pixels. RGB values are clipped to $[0, 1]$. Each family is evaluated separately at threshold 0.50 with fixed weights and no test-time augmentation or ensembling; F1 drops are measured against the clean reference from the same evaluation.

Table 6: Fixed-model stress tests on the complete LEVIR-CD validation split. Clean F1 is 93.12. Δ_2 is the F1 drop at $s = 2$ (percentage points), with pointwise 95% paired-scene intervals. Recall and unchanged-pixel false-positive rate (FPR) also refer to $s = 2$.

Perturbation	$F1_{0.50}$	$F1_1$	$F1_2$	Δ_2 [95% CI]	Recall	FPR
Radiometric	93.12	93.10	93.05	0.08 [-0.03, 0.23]	93.05	0.30
Gaussian noise	93.10	93.07	92.84	0.28 [0.16, 0.44]	92.59	0.30
Gaussian blur	93.03	92.07	83.89	9.24 [6.69, 12.66]	81.60	0.57

RIFT-CD remains stable under the tested radiometric shifts and Gaussian noise: at $s = 2$, F1 decreases by 0.08 and 0.28 points, respectively (Table 6). Strong blur produces a 9.24-point drop, with recall falling to 81.60 and unchanged-pixel false-positive rate increasing from 0.30 to 0.57. This contrast highlights the importance of spatial detail: appearance perturbations preserve much of the structural evidence, whereas strong smoothing removes fine boundaries and small changes. Paired-scene intervals summarize variation across the 64 scenes.

B.3 Additional benchmark metrics

The selected WHU-CD checkpoint obtains validation F1 of 95.94. Its test precision and recall are 96.97 and 96.96, with IoU 94.11. On MSRSCD, validation/test F1 is 78.36/78.35, with test precision 77.41, recall 79.33, IoU 64.41, and Band F1 53.56.

For CDD, validation/test results are 99.94/99.95 F1, 99.93/99.95 precision, 99.95/99.95 recall, 99.89/99.90 IoU, 99.25/99.28 Band F1, and 99.99/99.99 accuracy.

B.4 Qualitative results

Figure 4 shows predictions from the selected LEVIR-CD model. Rows (a) and (b) recover the annotated building footprints despite visible background appearance changes, with errors concentrated near their contours. Row (c) exposes overextended regions and a missed small component.

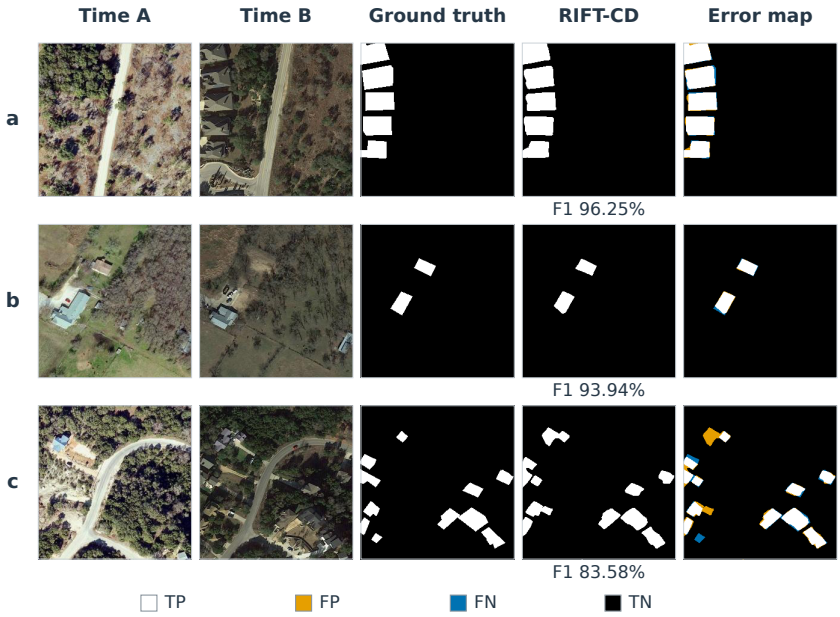

Figure 4: LEVIR-CD validation predictions at threshold 0.50. Rows (a)–(c) are distinct-scene 256×256 crops nearest the 90th, 50th, and 10th F1 percentiles among tiles with at least 256 changed pixels. All columns share crop coordinates. Error maps show correct change (TP, white), false positives (FP, orange), false negatives (FN, blue), and true negatives (TN, black). F1 scores refer to individual crops.

C Relational Invariance and Feedback

The following properties explain how the evidence construction treats feature magnitude and temporal order. They apply at the normalization inputs in Eqs. 2–5.

Positive-scale invariance. Let $x \neq 0$ and $\alpha > 0$, with $\|x\|_2 \geq \epsilon$ and $\alpha\|x\|_2 \geq \epsilon$. Then

$$\mathcal{N}(\alpha x) = \frac{\alpha x}{\alpha \|x\|_2} = \mathcal{N}(x). \tag{16}$$

Apply this identity independently to a_l, b_l, c_l^A, c_l^B, allowing positive scale factors to vary spatially. Every normalized input remains unchanged, so the direct discrepancy D_l, contextual discrepancy C_l, and agreement A_l remain unchanged. Their concatenation and its image under Φ_l are therefore unchanged as well. At fixed weights, the same argument preserves each feedback message m_l^t and its gated update $\lambda_l \odot m_l^t$. The normalization thus removes positive scalar magnitude before it enters either the decoded relation or the feedback update.

First-order geometry. For $\|x\|_2 > \epsilon$, differentiating $x/\|x\|_2$ gives Eq. 4. Decompose a perturbation as $\delta x = \beta \bar{x} + v$, where $v^\top \bar{x} = 0$. Then

$$J_{\mathcal{N}}(x)\, \delta x = \frac{v}{\|x\|_2}. \tag{17}$$

The radial component disappears to first order, while the tangential component remains. Below the floor, $\mathcal{N}(x) = x/\epsilon$, so magnitude again affects the output. The learned-feature sweep in Appendix B evaluates this transition explicitly. Image-level appearance changes can also rotate feature directions; their behavior is measured directly by the matched perturbation experiments in Section 4.4.

Temporal exchange symmetry. Assume shared weights and deterministic inference. Exchanging the dates swaps a_l with b_l and c_l^A with c_l^B. Absolute discrepancies and elementwise agreement are symmetric, so R_l and S_l are preserved. The two directed feedback formulas exchange their outputs under the same swap. The shared encoder therefore preserves this exchange relation at the next level; induction establishes it throughout the hierarchy. Symmetric native differences and the decoder then yield the same final change prediction.

The cross-date information path. In independent encoding with self-temporal feedback, date-A features are functions of date A alone. Their derivative with respect to date B is zero. Cross-feedback inserts the other date through contextual attention and discrepancy, introducing a route for a nonzero derivative. The identical-weight finite-difference and VJP experiments in Appendix B directly test this route. The N×M×F study then evaluates which evidence makes that route beneficial.

D Compute and Parameter Accounting

D.1 Models and relative cost

SimpleDiff-S2 retains the encoder, native paths, and decoder, replacing relational couplers with projected differences. All graphs are evaluated with dropout inactive. The RIFT-S4 graph used for counting has stride four, bilinear native resizing, and a quarter-resolution auxiliary head; the diagnostic training controls in Appendix A use area resizing and a half-resolution head. Table 7 reports the evaluation-graph counts, including shared parameters once and excluding inactive diagnostic routes.

Table 7: Parameters and counted operations per image pair. M/G denote $10^6/10^9$. The total combines fvcore-counted operations with the analytical SDPA contribution under the convention below.

Model	Params (M)	fvcore (G)	SDPA (G)	Counted + SDPA (G)
SimpleDiff-S2	199.40	385.32	0.00	385.32
RIFT-S2	201.17	398.21	0.81	399.03
RIFT-S4	201.17	106.12	0.20	106.32

Using unrounded counts, RIFT-S2 adds 0.89% parameters and 3.56% counted compute over SimpleDiff-S2. Stride-two processing uses $3.75\times$ the counted operations of RIFT-S4 with identical parameter count.

D.2 Operation-count convention

We use `fvcore` v0.1.5.post20221221 for one paired 256×256 RGB input in evaluation mode, including auxiliary output and disabling block checkpointing. Its MAC-style convention counts each multiply–accumulate position in convolution, linear, and matrix-product operators once. The convolution handler omits accumulation and bias additions.

Affine LayerNorm is included at approximately five operations per input element (four without affine parameters). GELU, unsupported additions, multiplications, subtractions, absolute values, and descriptor L_2 normalization are omitted. Bilinear interpolation and other supported operators follow their explicit handlers; reshaping and memory traffic receive no arithmetic cost.

We account for fused `scaled_dot_product_attention` analytically by adding its $QK^\top$ and AV products under the same MAC-style convention. At level l, let $H_l \times W_l$ be the spatial size, s_l the window side, d the total feature width across heads, and $n_l = s_l^2$ the tokens per window. With $N_l = H_l W_l / s_l^2$ windows per temporal direction, both $A \to B$ and $B \to A$ contribute:

$$C_l^{\mathrm{SDPA}} = 2N_l \left(n_l^2 d + n_l^2 d\right) = 4H_l W_l s_l^2 d. \tag{18}$$

The two terms represent $QK^\top$ and AV. Query scaling, softmax, and other elementwise work inside SDPA are excluded. The Q/K/V and output projections are already counted as convolutions and are not added again.

For RIFT-S2, $d = 128$, spatial sizes $(128^2, 64^2, 32^2, 16^2)$, and windows $(8, 8, 16, 8)$ yield a 0.81 G supplement; RIFT-S4's $(64^2, 32^2, 16^2, 8^2)$ yield 0.20 G. SimpleDiff-S2 contains no SDPA. Adding these contributions gives the totals in Table 7.