# Modeling Temporal Differences for Spatial Reconstruction in Remote Sensing Change Detection

**Anonymous authors**
Paper under double-blind review

## Abstract

Binary remote sensing change detection combines image-content features with explicit temporal descriptors. We ask whether the benefit of those descriptors depends on how the two feature groups are updated during multiscale decoding. JOINT forms a 192-channel state from two stems that encode temporal concatenation and absolute-difference/product descriptors. Joint residual updates precede all-scale aggregation and progressive mask reconstruction, with supervision from binary change masks alone. Two-factor comparisons cross feature construction with joint or independent refinement on LEVIR-CD and WHU-CD, using five common training seeds. The mean gain from explicit relations is larger under joint refinement, with F1 interaction contrasts of 0.270 and 0.246 percentage points. Average precision and boundary F1 show the same direction. Across five common training seeds, mean test F1 exceeds BASE by 0.626, 0.660, 0.500, and 0.610 percentage points on LEVIR-CD, WHU-CD, GZ-CD, and CDD, respectively. The findings associate descriptor gains with update organization within this decoder family. JOINT requires more inference time and memory than BASE, and its empty-target false positives increase on LEVIR-CD.

**Keywords:** binary change detection; remote sensing; temporal feature comparison; multiscale decoding; spatial reconstruction.

## 1 Introduction

Binary remote sensing change detection predicts a change mask from two registered images of the same scene. Shared encoders extract features for each date; the decoder must compare those features and recover the spatial extent of change. Temporal concatenation exposes both observations to a learned comparison, whereas differences and products supply explicit pairwise descriptors. Both approaches appear in existing detectors (Daudt et al., 2018; Shao et al., 2026; Zhang et al., 2026).

A gain from an explicit relation branch can depend on both its descriptors and the operations that update them. A direct comparison with a content-only baseline does not separate these effects. We therefore ask whether relation descriptors provide the same mean gain under joint and independent pyramid refinement. We cross these two update organizations with content-only and content–relation inputs while retaining the same pyramid-state width.

JOINT retains a 192-channel state formed from two 96-channel stems, one for content and one for absolute-difference/product descriptors. Residual blocks jointly refine the state at each scale; finer levels first add the resized state from the next coarser level. All four refined states then contribute to a shared reconstruction head. The architecture retains the PVTv2-B2 backbone and uses the original binary masks as its sole supervision.

The central evidence is a two-factor comparison on LEVIR-CD and WHU-CD, which measures the relation-feature gain within each update organization in detection F1, score-ranking AP, and boundary F1. Four-dataset, five-seed comparisons establish the complete model's empirical benefit relative to a shared-encoder baseline. An early-compression alternative provides a secondary comparison of pyramid-state width arrangements. Pixel errors, region coverage, and inference costs characterize the gains and their practical trade-offs.



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.

## 2 Related Work

**Siamese and hierarchical change detection.** Siamese detectors compare shared-encoder features through concatenation or differencing (Daudt et al., 2018). SNUNet-CD uses densely connected decoding (Fang et al., 2022); BIT refines pixel features through compact temporal-context tokens (Chen et al., 2022). ChangeFormer uses hierarchical transformer features (Bandara & Patel, 2022a). Its V6 decoder reduces projected temporal concatenations before top-down addition, all-scale aggregation, and progressive reconstruction (Bandara & Patel, 2022b). JOINT follows this hierarchy with PVTv2 features (Wang et al., 2022) and residual pyramid updates (Lin et al., 2017; He et al., 2016). We study how content and relation features are organized within these residual pyramid updates.

**Content and explicit temporal relations.** ICIF-Net, DMINet, and SEIFNet study temporal interaction and multilevel fusion (Feng et al., 2022; 2023; Huang et al., 2024). DIPFNet separately processes differences, sums, and concatenations, then merges and reduces them before hierarchical fusion (Shao et al., 2026; Zhang, 2026). AdaDINO combines both date features, their absolute difference, and their product to adapt a frozen encoder (Zhang et al., 2026). JOINT instead updates a shared 192-channel content–relation state after coarse-to-fine addition, reducing it to 96 channels only at all-scale aggregation. EMDNet couples multiscale difference reasoning with edge-aware encoding and progressive spatial refinement (Liu et al., 2026). Table 13 compares these architectures with JOINT.

## 3 Content–Relation Pyramid

### 3.1 Shared temporal encoding and paired stems

Let $I_1, I_2 \in \mathbb{R}^{3\times H\times W}$ be a registered image pair and $Y \in \{0,1\}^{H\times W}$ its ground-truth (GT) binary change mask. A shared encoder $E$ produces four pairs of features $F_i^1, F_i^2$, with spatial strides $(4, 8, 16, 32)$ and channel counts $(64, 128, 320, 512)$. A scale-specific projection $P_i$, shared between dates, maps both features to width $w = 96$:

$$a_i = P_i(F_i^1), \qquad b_i = P_i(F_i^2). \tag{1}$$

Each projection uses a bias-free $1 \times 1$ convolution, eight-group normalization (Wu & He, 2018), and a Gaussian error linear unit (GELU) (Hendrycks & Gimpel, 2016). Projection parameters are shared between dates, but not across the four scales.

Two independent convolutional branches, called stems, encode the paired features in complementary forms:

$$e_i = C_i(a_i \,\|\, b_i), \qquad q_i = Q_i\big(|a_i - b_i| \,\|\, (a_i \odot b_i)\big), \qquad z_i = e_i \,\|\, q_i. \tag{2}$$

Here $\|$ is channel concatenation and $\odot$ is the elementwise product. We call $e_i$ the content features and $q_i$ the relation features. Both stems use a $2w \to w$ pointwise convolution, group normalization, and GELU, giving $z_i$ a width of $2w = 192$. Difference and product arithmetic use 32-bit floating point (FP32), including during mixed-precision training. The relation descriptors are symmetric under date exchange, whereas ordered content concatenation is not. The complete architecture therefore does not guarantee invariance to exchanging the input dates.

### 3.2 Joint top-down refinement

Starting at the coarsest level, we form

$$v_4 = z_4, \qquad h_4 = J_4(v_4), \tag{3}$$

$$v_i = z_i + \mathcal{U}_i(h_{i+1}), \qquad h_i = J_i(v_i), \quad i = 3, 2, 1, \tag{4}$$

where $\mathcal{U}_i$ is bilinear interpolation to the spatial size of level $i$, with `align_corners=False`. Every refiner operates on the entire 192-channel state:

$$J_i(v) = \mathrm{GELU}\big(v + R_i(v)\big), \qquad R_i = \mathrm{GN} \circ K_{3\times3}^{(2)} \circ \mathrm{GELU} \circ \mathrm{GN} \circ K_{3\times3}^{(1)}. \tag{5}$$

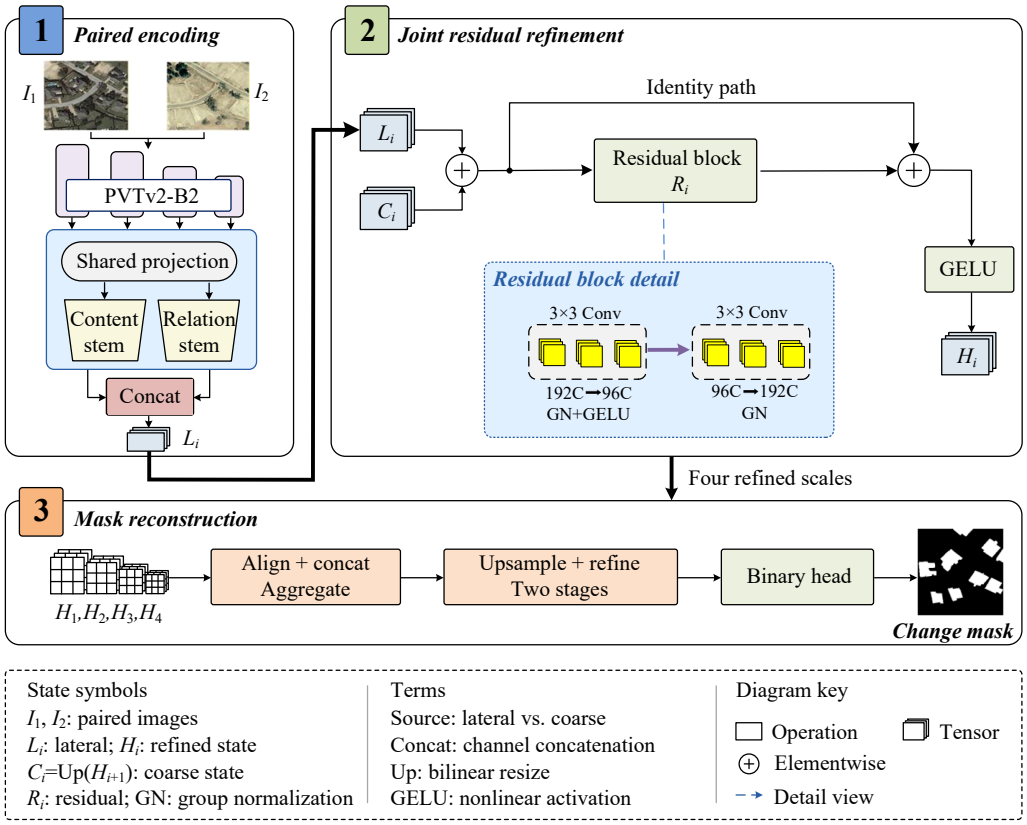


Figure 1: JOINT architecture. (1) Shared temporal encoding and paired stems. (2) Joint residual refinement with a 192-channel identity path and a 96-channel residual bottleneck. (3) All-scale aggregation and mask reconstruction. In the schematic, $L_i = z_i$ and $H_i = h_i$ in Equation 4; $C_i$ is the resized coarser state, with $C_4 = 0$. The dashed connector indicates an expanded view of the residual block. Only the final mask is supervised. Image tiles illustrate input and output formats.

The channel sequence inside $R_i$ is $192 \to 96 \to 192$, while its identity branch retains the full 192-channel state. At each level, the two stem outputs occupy separate channel blocks in $z_i$. They are concatenated without an additional projection to fewer channels. The residual transform mixes these channels, so the resulting state no longer contains isolated content and relation streams. The delayed reduction therefore concerns the pyramid state, rather than every intermediate transform inside the decoder.

### 3.3 Aggregation and Spatial Reconstruction

All four states are aligned to the finest feature scale and concatenated:

$$t = A\big(h_1 \,\|\, \mathcal{U}_1(h_2) \,\|\, \mathcal{U}_1(h_3) \,\|\, \mathcal{U}_1(h_4)\big), \tag{6}$$

where $A$ is a $768 \to 96$ pointwise convolution with group normalization and GELU. Reconstruction uses two stages,

$$t_{1/2} = D_{1/2}(\mathcal{U}_{H/2}(t)), \qquad t_1 = D_1(\mathcal{U}_H(t_{1/2})), \qquad \ell = K_{1\times1}(t_1). \tag{7}$$

Here $\mathcal{U}_{H/2}$ and $\mathcal{U}_H$ resize to $(H/2, W/2)$ and $(H, W)$, respectively. $D_{1/2}$ and $D_1$ use $3 \times 3$ convolutions with channel transitions $96 \to 48$ and $48 \to 24$, followed by group normalization and GELU. The final classifier maps 24 channels to one logit, with change probability $p = \sigma(\ell)$. No boundary head or prediction-time post-processing is used. Appendix A gives the tensor shapes and forward computation.

### 3.4 SUPERVISION

JOINT and BASE combine full-resolution binary cross-entropy with a Dice-overlap loss (Milletari et al., 2016), using the following empty-target-masked formulation:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \frac{1}{B}\sum_{n=1}^{B} \mathbf{1}\left[\sum_{x} Y_n(x) > 0\right]\left(1 - \frac{2\sum_x p_n(x)Y_n(x) + \epsilon}{\sum_x p_n(x) + \sum_x Y_n(x) + \epsilon}\right), \tag{8}$$

with $\epsilon = 10^{-6}$. Binary cross-entropy (BCE) is averaged over every pixel in the batch. Empty-target images contribute to BCE but contribute zero to the Dice term. Dice is divided by the original batch size $B$, not the number of positive images. Supervision applies only to the final mask, without intermediate feature or component labels.

## 4 EXPERIMENTAL SETUP

**Datasets and evaluation.** We evaluate LEVIR-CD (Chen & Shi, 2020), WHU-CD (Ji et al., 2019; GPCV Laboratory, Wuhan University, n.d.), GZ-CD (Peng et al., 2021), and the season-varying CDD (Lebedev et al., 2018). CDD contains general binary changes, not only building changes. Models are trained separately on each dataset, so the comparison measures within-dataset performance rather than zero-shot transfer. Within each dataset, all methods use the same training, validation, and test partitions. Validation F1 selects one checkpoint per training run, with the earliest epoch breaking ties. All accuracy results below evaluate those selected checkpoints on the test split, using a fixed probability threshold of 0.5. There is no per-metric checkpoint selection, test-time augmentation, or test-fitted decision threshold.

**Training protocol.** JOINT and its controls initialize PVTv2-B2 from ImageNet pretraining (Deng et al., 2009), with newly initialized fusion and decoder layers. Training updates all parameters for 100 epochs with batch size eight. AdamW (Loshchilov & Hutter, 2019) uses an initial learning rate of $10^{-4}$ and weight decay $10^{-2}$. Cosine warm restarts (Loshchilov & Hutter, 2017) use $T_0 = 15$ and $T_{\mathrm{mult}} = 2$; the gradient norm is clipped to 0.5. Paired horizontal and vertical flips, quarter-turn rotations, and date exchange preserve image–mask correspondence. RGB normalization uses mean and standard deviation 0.5, without color jitter or Mosaic. Supervision remains Equation 8. The experiments use NVIDIA GeForce RTX 4090 GPUs on Ubuntu. Appendix B details preprocessing and model initialization.

**Training repetitions and metrics.** All in-family models use the same five training seeds. F1 is reported as mean ± sample standard deviation (SD, $n = 5$), which describes variation across training runs rather than a confidence interval. Precision (P), recall (R), intersection over union (IoU), overall accuracy (OA), average precision (AP), and boundary F1 (BF1) accompany F1. All metrics use the same validation-selected checkpoint for each run. Differences are computed from the three-decimal means and expressed in percentage points (pp). Appendix D defines the reported means, SDs, and paired contrasts.

**Reference models.** BASE is the shared-PVT reference. Its temporal concatenations are mapped to the backbone widths before top-down decoding. It has 31,687,745 parameters, compared with 26,553,257 for JOINT. Several decoder details differ, so this comparison evaluates the complete redesign.

WIDE-CONTENT retains JOINT's shapes and connections, feeding temporal concatenation to both stems instead of using explicit relations in one. EARLY-COMPRESS projects the two-stem state from 192 to 96 channels before the pyramid. SEPARATE keeps both 96-channel streams but refines them independently until aggregation. The crossed WIDE-CONTENT + SEPARATE control completes a two-factor comparison of feature construction and update organization. We abbreviate that label as WIDE+SEPARATE in tables and figures.

Table 1: Test performance across five common training seeds. F1 is mean ± sample SD (%); differences are JOINT minus BASE in pp. Each run uses its validation-F1-selected checkpoint. Appendix D reports all seven metrics.

| Dataset | BASE F1 | JOINT F1 | ΔF1 | ΔIoU |
|---|---|---|---|---|
| LEVIR-CD | 91.136 ± 0.078 | 91.762 ± 0.072 | +0.626 | +1.062 |
| WHU-CD | 94.670 ± 0.078 | 95.330 ± 0.083 | +0.660 | +1.197 |
| GZ-CD | 89.350 ± 0.170 | 89.850 ± 0.140 | +0.500 | +0.821 |
| CDD | 97.510 ± 0.067 | 98.120 ± 0.053 | +0.610 | +1.168 |

Table 2: Structural comparisons on the test splits. F1 is mean ± sample SD over five common training seeds (%); IoU is in percent. WIDE+SEPARATE abbreviates WIDE-CONTENT + SEPARATE.

| | LEVIR-CD | | WHU-CD | |
|---|---|---|---|---|
| Model | F1 | IoU | F1 | IoU |
| BASE | 91.136 ± 0.078 | 83.716 | 94.670 ± 0.078 | 89.880 |
| WIDE-CONTENT | 91.436 ± 0.115 | 84.223 | 95.046 ± 0.109 | 90.560 |
| EARLY-COMPRESS | 91.448 ± 0.076 | 84.244 | 95.038 ± 0.097 | 90.545 |
| SEPARATE | 91.450 ± 0.104 | 84.247 | 95.044 ± 0.075 | 90.556 |
| WIDE+SEPARATE | 91.394 ± 0.083 | 84.152 | 95.006 ± 0.092 | 90.487 |
| JOINT | 91.762 ± 0.072 | 84.778 | 95.330 ± 0.083 | 91.077 |

# 5 Results

## 5.1 Five-seed test performance

JOINT achieved a higher mean test F1 than BASE on all four datasets (Table 1). The improvements were 0.626, 0.660, 0.500, and 0.610 pp on LEVIR-CD, WHU-CD, GZ-CD, and CDD, respectively. IoU also increased on all four datasets. The improvement across all four datasets establishes the benefit of the complete decoder redesign relative to BASE.

## 5.2 Relation-feature gains under joint and independent updates

Table 2 compares the structural alternatives on LEVIR-CD and WHU-CD, using five repetitions for every reported row. JOINT had the highest mean F1 and IoU among the listed controls on both datasets.

**Does the relation-feature gain depend on update organization?** Within each update organization, we replace the second stem's temporal concatenation with absolute differences and products. Layer shapes, learned parameter budgets, and the 192-channel pyramid-state width remain fixed within each pair. Under joint refinement, this substitution increased mean F1 by 0.326 pp on LEVIR-CD and 0.284 pp on WHU-CD. Under independent refinement, the corresponding gains were 0.056 and 0.038 pp (Table 3). For a metric $M$, we summarize this difference using

$$\begin{aligned}\Delta_M = &(\bar{M}^{\mathrm{JOINT}} - \bar{M}^{\mathrm{WIDE}}) \\ &- (\bar{M}^{\mathrm{SEPARATE}} - \bar{M}^{\mathrm{WIDE+SEPARATE}}),\end{aligned} \tag{9}$$

where $\bar{M}$ is the mean across the five common training seeds. The F1 interaction means were 0.270 and 0.246 pp, respectively.

**Does the pattern extend beyond fixed-threshold F1?** Table 3 extends the F1 comparison in Figure 2 to AP and BF1. The AP interaction was 0.280 pp on LEVIR-CD and 0.270 pp on WHU-CD; the BF1 interaction was 1.510 and 1.970 pp, respectively. Thus, the larger mean descriptor gain under joint refinement also appears in score ranking and boundary agreement, not only F1 at threshold 0.5. These metrics summarize the same predictions and are not independent replications. The contrasts describe means, not run-by-run stability. Appendix D reports all seven metric contrasts and the statistical scope of the paired comparison.

Table 3: Mean gain from explicit relations under each update organization (pp). Joint: JOINT minus WIDE-CONTENT; independent: SEPARATE minus WIDE+SEPARATE. The interaction $\Delta_M$ is joint minus independent. All values are differences of five-seed means.

| | LEVIR-CD | | | WHU-CD | | |
|---|---|---|---|---|---|---|
| Metric | Joint | Independent | $\Delta_M$ | Joint | Independent | $\Delta_M$ |
| F1 | +0.326 | +0.056 | +0.270 | +0.284 | +0.038 | +0.246 |
| AP | +0.300 | +0.020 | +0.280 | +0.290 | +0.020 | +0.270 |
| BF1 | +1.770 | +0.260 | +1.510 | +2.200 | +0.230 | +1.970 |

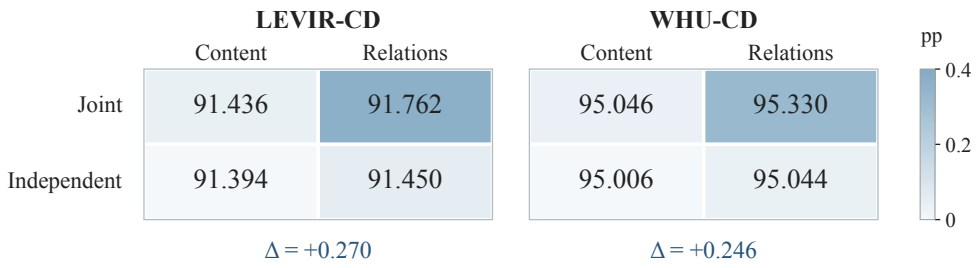


Figure 2: Five-seed mean test F1 (%). Columns specify second-stem inputs; color shows F1 minus WIDE+SEPARATE (pp). $\Delta$ denotes the mean interaction (pp).

**Secondary comparison of state width.** Against EARLY-COMPRESS, JOINT gained 0.314 pp on LEVIR-CD and 0.292 pp on WHU-CD. This comparison favors retaining the 192-channel state in this implementation. Residual widths and normalization grouping also change, so the result compares complete width arrangements (Appendix C).

### 5.3 Pixel errors and region coverage

The complete JOINT model improved both precision and recall relative to BASE on LEVIR-CD. Mean FN and FP counts decreased by 29,930.8 and 55,539.0 pixels, while completely missed GT regions decreased from 451.4 to 377.6. Empty-target FP nevertheless increased from 37,599.8 to 38,527.8 pixels. These spatial summaries characterize the complete-model gain; they do not isolate the feature-by-update interaction. Appendix E provides the full six-model counts and Figure A1 compares BASE and JOINT.

**Where predictions differ.** Figure 3 shows matched image examples: JOINT covers buildings missed by BASE in (a) and removes a background false alarm in (b). Panel (c) shows fewer missed change pixels and fewer false-positive pixels. These single-checkpoint examples locate individual prediction errors and are separate from the five-seed test summaries. Appendix J adds aligned predictions from external methods.

### 5.4 Inference cost

Table 4 compares RTX 4090 inference costs under FP32 and automatic mixed precision (AMP), at batch sizes one and eight. JOINT has 16.20% fewer parameters than BASE, but its measured time and peak memory are higher in every listed setting. FP32 B1 time increased by 10.53%, from 60.80 to 67.20 ms per pair. AMP B8 amortized time increased by 38.46%, from 8.84 to 12.24 ms per pair. Fewer parameters therefore do not imply a lower inference cost here.

## 6 Discussion and Limitations

The central finding is a larger mean benefit from explicit relations under joint refinement than under independent refinement on both ablation datasets. Within each update organization, the content-only and content–relation variants have matching layer shapes and learned parameter budgets. The contrast

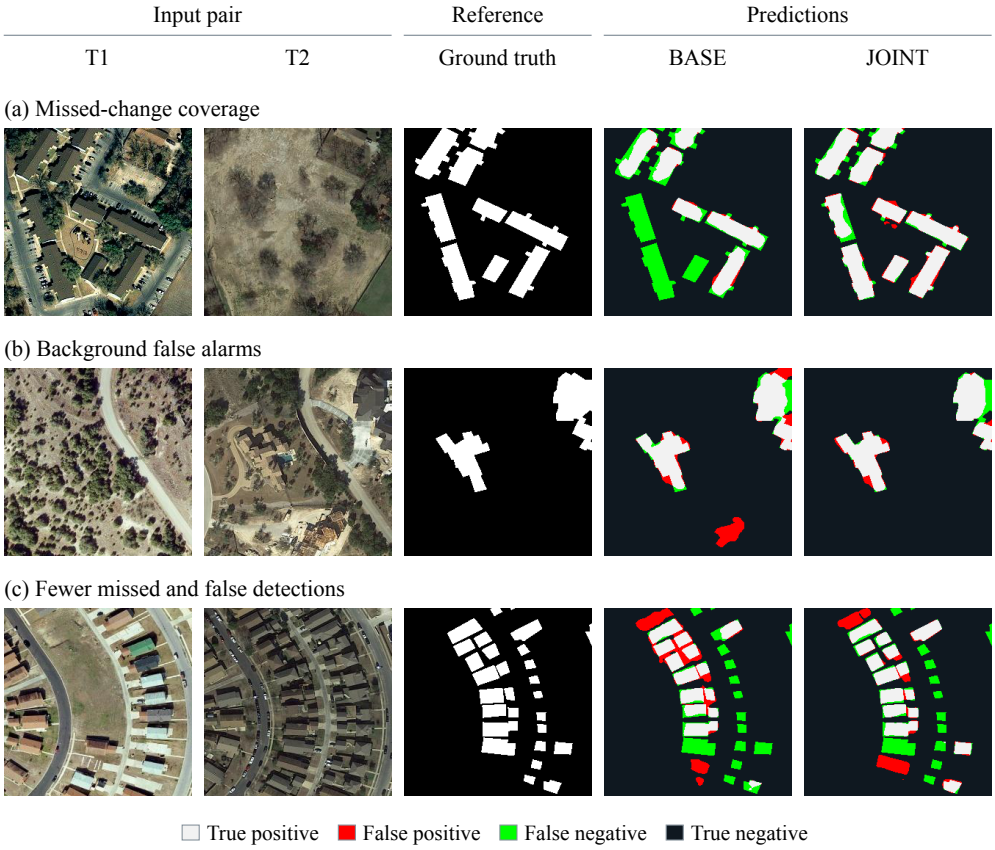


Figure 3: Selected LEVIR-CD validation examples from single BASE and JOINT checkpoints. Each row shows both dates, GT, BASE, and JOINT over the full $256 \times 256$ patch. Red marks false positives, green false negatives, white true positives, and dark pixels true negatives. (a) Improved changed-building coverage; (b) fewer background false alarms; (c) simultaneous reductions in missed and false detections.

Table 4: Inference time / peak memory on RTX 4090. Units are ms per image pair / MiB. B1 and B8 denote batches of one and eight pairs; B8 time is amortized per pair, not single-pair latency. Appendix G reports all six models under these settings.

| Model | FP32 B1 | FP32 B8 | AMP B1 | AMP B8 |
|---|---|---|---|---|
| BASE | 60.80 / 185.0 | 7.95 / 345.0 | 66.10 / 235.0 | 8.84 / 340.0 |
| JOINT | 67.20 / 236.0 | 8.95 / 950.0 | 77.50 / 287.0 | 12.24 / 1,003.0 |

therefore links the descriptor gain to update organization, beyond the complete model's improvement over BASE. The same direction in F1, AP, and BF1 connects this observation to fixed-threshold detection, score ranking, and boundary agreement. For this decoder, descriptor choice and refinement organization should be evaluated together.

ChangeFormer supplies a related reconstruction hierarchy, while DIPFNet and AdaDINO provide precedents for combining temporal descriptors. JOINT places content and relation features in a shared residual state before all-scale aggregation. The crossed comparison evaluates how this organization changes the mean benefit of established descriptors. Its evidence concerns two datasets within the present decoder family. Joint and independent updates also differ in residual parameterization and normalization, so the contrast describes their combined organization. It does not isolate a unique cross-channel mechanism or establish transfer to other decoder architectures.

The EARLY-COMPRESS comparison is secondary evidence about state-width arrangements, with residual widths and normalization grouping changing together. It favors the retained 192-channel state in this implementation without measuring information loss in isolation.

Empty-target false positives increase despite lower total FP and FN, which matters when unchanged scenes dominate deployment. The wider intermediate states also accompany higher measured inference time and memory despite a smaller parameter count than BASE.

## 7 Conclusion

The two-dataset crossed comparison associates larger mean gains from explicit temporal relations with joint refinement in the evaluated decoder, across F1, AP, and boundary F1. JOINT implements this arrangement through a shared content–relation pyramid and exceeds BASE in five-seed mean test F1 on four datasets. Pixel and region statistics locate part of the improvement, while empty-target false positives and inference cost define practical limitations.

## Reproducibility Statement

Section 3 and Appendix A specify the forward computation, tensor shapes, normalization, and reconstruction operations. Section 4 and Appendices B–C describe initialization, optimization, augmentation, checkpoint selection, and the structural controls. Appendices D–G provide the seven-metric results, metric definitions, region statistics, and hardware-specific costs.

## AI Use Statement

AI was used only for language polishing.

## Ethics Statement

We use existing remote sensing datasets and collect no new human-subject data. Unverified change maps should not be sole evidence for safety-critical planning, property decisions, or individual monitoring. Redistribution of data and artifacts remains subject to the relevant licenses.

## References


Wele Gedara Chaminda Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. In *IEEE International Geoscience and Remote Sensing Symposium*, 2022a. URL https://arxiv.org/abs/2201.01293.

Wele Gedara Chaminda Bandara and Vishal M. Patel. ChangeFormer: Author implementation, V6 decoder. Source code, 2022b. URL https://github.com/wgcban/ChangeFormer/blob/afd1b7ed640aa265a2c730de958416ae7356a2f9/models/ChangeFormer.py.

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.

Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *IEEE International Conference on Image Processing*, 2018. URL https://arxiv.org/abs/1810.08462.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024. doi: 10.1016/j.isprsjprs.2024.01.004.

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.

Yuchao Feng, Honghui Xu, Jiawei Jiang, Hao Liu, and Jianwei Zheng. ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2022.3168331.

Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2023.3241257.

GPCV Laboratory, Wuhan University. WHU building dataset: Building change detection data. Dataset publisher website, n.d. URL https://gpcv.whu.edu.cn/data/building_dataset.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Yanyuan Huang, Xinghua Li, Zhengshun Du, and Huanfeng Shen. Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3360516.

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.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1612.03144.

Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2021. doi: 10.1109/LGRS.2020.2988032.

Ying Liu, Haoyu Jiang, Xinying Zhong, Cong Lin, and Yuanyuan Wu. EMDNet: Edge-aware multi-scale difference network for remote sensing image change detection. *Journal of King Saud University Computer and Information Sciences*, 38:733, 2026. doi: 10.1007/s44443-026-01166-8.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.03983.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Jianhao Miao, Shuang Li, Xuechen Bai, Wenxia Gan, Jianwei Wu, and Xinghua Li. RS-NormGAN: Enhancing change detection of multi-temporal optical remote sensing images through effective radiometric normalization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:324–346, 2025. doi: 10.1016/j.isprsjprs.2025.02.005.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Daifeng Peng, Lorenzo Bruzzone, Yongjun Zhang, Haiyan Guan, Haiyong Ding, and Xu Huang. SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 5891–5906, 2021. doi: 10.1109/TGRS.2020.3011913.

Zhenfeng Shao, Jindou Zhang, Tao Cheng, Deren Li, Yaofeng Hu, Jingyi Guo, Ashraf Dewan, Orhan Altan, Tsehaie Woldai, and Neema Simon Sumari. A novel framework and benchmark dataset for fine-grained urban building change detection in ultra-high-resolution UAV imagery. *Geo-spatial Information Science*, pp. 1–23, 2026. doi: 10.1080/10095020.2026.2695542.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. doi: 10.1007/s41095-022-0274-8.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1803.08494.

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.

Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166: 183–200, 2020. doi: 10.1016/j.isprsjprs.2020.06.003.

Jindou Zhang. DIPFNet: Author implementation. Source code, 2026. URL https://github.com/zjd1836/DIPFNet.

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. doi: 10.1109/TGRS.2023.3300533.

Xu Zhang, Xinqing Li, Jianpeng Xie, Zeshuai Zhu, Xin He, and Yun Liu. AdaDINO: Pair-aware in-backbone adaptation of frozen DINO for efficient remote sensing change detection. *arXiv preprint arXiv:2608.07982*, 2026. URL https://arxiv.org/abs/2608.07982.

## A Tensor Specification and Forward Computation

Table 5 gives the tensor shapes for Figure 1.

Table 5: JOINT tensor specification for a $256 \times 256$ pair; batch dimension omitted.

| Stage | Channels and spatial size | Operation |
|---|---|---|
| Input, each date | $3 \times 256 \times 256$ | Registered RGB pair |
| Encoder level 1 | $64 \times 64 \times 64$ | Shared PVTv2-B2 |
| Encoder level 2 | $128 \times 32 \times 32$ | Shared PVTv2-B2 |
| Encoder level 3 | $320 \times 16 \times 16$ | Shared PVTv2-B2 |
| Encoder level 4 | $512 \times 8 \times 8$ | Shared PVTv2-B2 |
| $a_i, b_i$ | $96 \times H_i \times W_i$ each | Date-shared scale projection |
| $e_i, q_i$ | $96 \times H_i \times W_i$ each | Two independent pair stems |
| $z_i, v_i, h_i$ | $192 \times H_i \times W_i$ | Concat, top-down add, refiner |
| Aligned pyramid | $768 \times 64 \times 64$ | Resize four $h_i$ and concatenate |
| $t$ | $96 \times 64 \times 64$ | $1 \times 1$ aggregation |
| $t_{1/2}$ | $48 \times 128 \times 128$ | Resize and $3 \times 3$ refinement |
| $t_1$ | $24 \times 256 \times 256$ | Resize and $3 \times 3$ refinement |
| Logit $\ell$ | $1 \times 256 \times 256$ | $1 \times 1$ classifier |

All decoder convolutions are bias-free except the final classifier, and every group normalization uses eight groups. The residual transform ends with normalization; the identity is added before applying GELU. Pyramid blocks have scale-specific weights, while encoder and projection weights are shared between dates.

```
F1, F2 = encoder(I1), encoder(I2)
for i in 1..4:
  a, b = project[i](F1[i]), project[i](F2[i])
  e = content[i](concat(a, b))
  q = relation[i](concat(abs(a-b), a*b))
  z[i] = concat(e, q)
for i in 4..1:
  v = z[i] if i == 4 else z[i] + resize(h[i+1])
  h[i] = GELU(v + residual[i](v))
t = aggregate(concat(resize(h[1]), ..., resize(h[4])))
t = refine_half(resize_half(t))
t = refine_full(resize_full(t))
logits = classifier(t)
```

## B Implementation and Training Details

**Hardware and repeats.** The five-seed experiments used NVIDIA GeForce RTX 4090 GPUs under Ubuntu. Every in-family model was trained with the same set of five random seeds. For each run, validation F1 determined the checkpoint evaluated on the test set.

**Training recipe and data processing.** The training recipe uses four data-loader workers, batch size eight, and no dropped final batch. Inputs are registered $256 \times 256$ RGB pairs. Horizontal and vertical flips each have probability 0.5; rotations are sampled from multiples of 90 degrees. The two dates and target mask receive the same spatial transformation. Date exchange uses probability 0.5. Grayscale masks are binarized above 127, and RGB values are normalized from $[0, 1]$ to $[-1, 1]$. All in-family comparisons use this preprocessing and augmentation recipe.

**Model initialization and comparison scope.** The encoder uses ImageNet pretraining, with no trained change-detection checkpoint used for initialization. Common modules use seeded initialization before the differing pyramid blocks. BASE also uses backbone-aligned fusion widths and `align_corners=True` for bilinear resizing; JOINT uses `align_corners=False`. BASE versus JOINT evaluates the complete decoder redesign; the structural controls instead retain the JOINT encoder and reconstruction head.

## C Structural Control Definitions

The primary structural comparisons in Table 2 cross explicit relations with the organization of residual updates. EARLY-COMPRESS separately examines a different state-width arrangement. They share the JOINT encoder, reconstruction head, objective, and training protocol. In SEPARATE, each 192-channel state is split into two 96-channel streams. Independent $96 \to 96 \to 96$ residual refiners process them before concatenation. Aggregation can still mix the streams. Both variants contain $4 \cdot 9w^2$ convolution weights per pyramid block, but their group-normalization parameters differ.

Table 6: Structural controls and parameter budgets, including encoder and decoder. State width is measured before pyramid refinement.

| Model | Second-stem input | Width | Update | Params (M) |
|---|---|---|---|---|
| JOINT | Abs. diff. + product | 192 | Joint | 26.553 |
| WIDE-CONTENT | Concatenation | 192 | Joint | 26.553 |
| EARLY-COMPRESS | Abs. diff. + product | 96 | Joint | 26.590 |
| SEPARATE | Abs. diff. + product | $2 \times 96$ | Independent | 26.554 |

**Early compression.** Early compression uses the stems in Equation 2, followed by an independent, bias-free linear projection $L_i : 2w \to w$ at each scale. No activation or normalization follows this projection before the top-down computation:

$$\widetilde{z}_i = L_i(z_i), \tag{10}$$

$$\widetilde{v}_i = \widetilde{z}_i + \mathcal{U}_i(\widetilde{h}_{i+1}), \qquad \widetilde{h}_i = \mathrm{GELU}\big(\widetilde{v}_i + \widetilde{R}_i(\widetilde{v}_i)\big), \tag{11}$$

omitting the top-down term at the coarsest level, where no higher state exists. The state width is $w$, and the residual convolutions use $w \to 2w \to w$, following Equation 5's normalization and activation sequence. All-scale aggregation maps $4w$ to $w$, rather than $8w$ to $w$, before the unchanged reconstruction head. Both stems remain present; compression is learned rather than implemented by averaging or deleting a branch.

Each residual block retains $4 \cdot 9w^2$ convolution weights at its corresponding feature scale. Four added projections and reduced aggregation increase the total parameter count by $4w^2 = 36{,}864$, or about 0.139%. The control therefore has a similar parameter budget but different computation. State width, update width, and channels per normalization group change together. These coupled changes prevent assigning a score difference solely to irreversible information loss.

**Wide content.** The second control retains every JOINT layer shape and connection, replacing only the relation-stem input:

$$q_i^{\mathrm{content}} = Q_i(a_i \,\|\, b_i), \qquad z_i^{\mathrm{content}} = C_i(a_i \,\|\, b_i) \,\|\, q_i^{\mathrm{content}}. \tag{12}$$

The stems remain independently parameterized, with unchanged state width, residual transforms, aggregation, and reconstruction. The model also retains the same total parameter count of 26,553,257. Only the deterministic inputs to the second stem change; the image pair and learned parameter count remain fixed. Descriptor arithmetic can still affect runtime despite identical convolution shapes.

**Crossed content-only control.** WIDE-CONTENT + SEPARATE combines the second-stem input of WIDE-CONTENT with the independent refinement of SEPARATE. Both stems receive $a_i \,\|\, b_i$ and each produces 96 channels. After top-down addition, split the 192-channel state into $v_i^{(1)}$ and $v_i^{(2)}$ and update them as

$$h_i = J_i^{(1)}\big(v_i^{(1)}\big) \,\|\, J_i^{(2)}\big(v_i^{(2)}\big), \tag{13}$$

where each $J_i^{(k)}$ is the independent $96 \to 96 \to 96$ residual refiner defined for SEPARATE. The encoder, top-down additions, all-scale aggregation, and reconstruction head are unchanged. The resulting two-factor comparison varies the second stem's input (temporal concatenation or difference/product) and refinement organization (joint or separate).

## D Complete Five-Seed Test Results

Tables 7–9 report all seven metrics for the evaluated models, using the same five training seeds. Each run is selected by validation F1 before test evaluation. F1 dispersion is the sample SD:

$$\bar{F}_1 = \frac{1}{5}\sum_{s=1}^{5} F_{1,s}, \qquad s_{F_1} = \sqrt{\frac{1}{4}\sum_{s=1}^{5}(F_{1,s} - \bar{F}_1)^2}. \tag{14}$$

The $\pm$ term describes variation across training runs, not uncertainty over images, pixels, or paired model differences. F1 is reported with its sample SD; the other columns report means. Means of nonlinear metrics need not satisfy the identities of a single pooled confusion matrix exactly.

Table 7: Complete LEVIR-CD test summaries (%). F1 is mean $\pm$ sample SD across five training seeds. WIDE+SEPARATE abbreviates WIDE-CONTENT + SEPARATE.

| Model | P | R | F1 | IoU | OA | AP | BF1 |
|---|---|---|---|---|---|---|---|
| BASE | 91.868 | 90.416 | 91.136 $\pm$ 0.078 | 83.716 | 99.121 | 96.200 | 60.100 |
| WIDE-CONTENT | 92.292 | 90.596 | 91.436 $\pm$ 0.115 | 84.223 | 99.151 | 96.500 | 61.650 |
| EARLY-COMPRESS | 92.066 | 90.838 | 91.448 $\pm$ 0.076 | 84.244 | 99.151 | 96.460 | 61.910 |
| SEPARATE | 92.171 | 90.740 | 91.450 $\pm$ 0.104 | 84.247 | 99.152 | 96.490 | 61.740 |
| WIDE+SEPARATE | 92.064 | 90.734 | 91.394 $\pm$ 0.083 | 84.152 | 99.146 | 96.470 | 61.480 |
| JOINT | 92.680 | 90.862 | 91.762 $\pm$ 0.072 | 84.778 | 99.184 | 96.800 | 63.420 |

Table 8: Complete WHU-CD test summaries (%); five-seed reporting as in Table 7.

| Model | P | R | F1 | IoU | OA | AP | BF1 |
|---|---|---|---|---|---|---|---|
| BASE | 95.206 | 94.140 | 94.670 $\pm$ 0.078 | 89.880 | 99.491 | 97.910 | 56.180 |
| WIDE-CONTENT | 95.623 | 94.476 | 95.046 $\pm$ 0.109 | 90.560 | 99.527 | 98.200 | 58.340 |
| EARLY-COMPRESS | 95.533 | 94.548 | 95.038 $\pm$ 0.097 | 90.545 | 99.526 | 98.160 | 58.620 |
| SEPARATE | 95.560 | 94.534 | 95.044 $\pm$ 0.075 | 90.556 | 99.527 | 98.180 | 58.510 |
| WIDE+SEPARATE | 95.562 | 94.456 | 95.006 $\pm$ 0.092 | 90.487 | 99.523 | 98.160 | 58.280 |
| JOINT | 95.958 | 94.710 | 95.330 $\pm$ 0.083 | 91.077 | 99.555 | 98.490 | 60.540 |

Table 9: GZ-CD and CDD test summaries (%); F1 is mean $\pm$ sample SD over the same five training seeds.

| Model | P | R | F1 | IoU | OA | AP | BF1 |
|---|---|---|---|---|---|---|---|
| **GZ-CD** | | | | | | | |
| BASE | 91.389 | 87.400 | 89.350 $\pm$ 0.170 | 80.750 | 97.916 | 92.640 | 36.910 |
| JOINT | 91.965 | 87.830 | 89.850 $\pm$ 0.140 | 81.571 | 98.016 | 93.110 | 38.790 |
| **CDD** | | | | | | | |
| BASE | 97.620 | 97.400 | 97.510 $\pm$ 0.067 | 95.141 | 99.383 | 99.570 | 75.230 |
| JOINT | 98.220 | 98.020 | 98.120 $\pm$ 0.053 | 96.309 | 99.534 | 99.790 | 79.420 |

**Mean interaction contrasts.** For common training seeds, a difference of model means equals the mean of seed-paired differences. The same linearity applies to Equation 9. The equivalent update-gain contrast is $(\bar{F}_1^{\mathrm{JOINT}} - \bar{F}_1^{\mathrm{SEPARATE}}) - (\bar{F}_1^{\mathrm{WIDE}} - \bar{F}_1^{\mathrm{WIDE+SEPARATE}})$. Joint-update gains with relation inputs are 0.312 and 0.286 pp on LEVIR-CD and WHU-CD; content-only gains are 0.042 and 0.040 pp. Their differences equal the 0.270 and 0.246 pp contrasts in Table 3. Table 10 extends the same contrast to every reported metric. All seven mean interactions are positive on both datasets. This agreement across metrics concerns the same model predictions and does not constitute additional independent repetitions. The contrasts are not paired significance tests or a claim that every seed has the same ordering. Individual model SDs describe between-run dispersion. They do not determine the uncertainty of the paired interaction, which also depends on cross-model covariances within seeds.

Table 10: Conditional descriptor gains and mean interactions across all seven metrics (pp). Joint and independent gains use the same model pairs as Table 3; $\Delta_M$ follows Equation 9. All entries are differences of five-seed means.

| | LEVIR-CD | | | WHU-CD | | |
|---|---|---|---|---|---|---|
| Metric | Joint | Independent | $\Delta_M$ | Joint | Independent | $\Delta_M$ |
| P | +0.388 | +0.107 | +0.281 | +0.335 | -0.002 | +0.337 |
| R | +0.266 | +0.006 | +0.260 | +0.234 | +0.078 | +0.156 |
| F1 | +0.326 | +0.056 | +0.270 | +0.284 | +0.038 | +0.246 |
| IoU | +0.555 | +0.095 | +0.460 | +0.517 | +0.069 | +0.448 |
| OA | +0.033 | +0.006 | +0.027 | +0.028 | +0.004 | +0.024 |
| AP | +0.300 | +0.020 | +0.280 | +0.290 | +0.020 | +0.270 |
| BF1 | +1.770 | +0.260 | +1.510 | +2.200 | +0.230 | +1.970 |

## E Spatial Error and Region Summaries

Table 11 compares six models using mean counts across five runs and a fixed total of 3,371 GT connected regions. Completely missed and at-least-half-covered regions are disjoint; the remaining regions have positive coverage below 50%.

Table 11: LEVIR-CD spatial summaries, averaged over five runs. FN, FP, near-GT FP, and empty-target FP are pixel counts. Missed and covered are GT region counts; covered requires at least 50% of the region's pixels to be predicted positive. The evaluation set contains 3,371 GT connected regions.

| Model | FN | FP | Near-GT FP | Empty FP | Missed | Covered |
|---|---|---|---|---|---|---|
| BASE | 643,171.4 | 537,137.8 | 424,338.8 | 37,599.8 | 451.4 | 2,810.2 |
| WIDE-CONTENT | 631,091.8 | 507,793.6 | 401,157.0 | 35,545.2 | 409.4 | 2,854.6 |
| EARLY-COMPRESS | 614,851.4 | 525,321.2 | 415,003.8 | 36,772.6 | 410.2 | 2,856.2 |
| SEPARATE | 621,428.0 | 517,222.2 | 408,605.4 | 36,205.8 | 408.2 | 2,860.2 |
| WIDE+SEPARATE | 621,830.8 | 524,902.8 | 414,673.0 | 36,743.4 | 416.0 | 2,851.0 |
| JOINT | 613,240.6 | 481,598.8 | 366,015.2 | 38,527.8 | 377.6 | 2,901.0 |

Figure A1 summarizes the complete-model comparison. Mean FN and FP decrease by 29,930.8 and 55,539.0 pixels, respectively. Near-GT FP decreases by 58,323.6 pixels, more than the net total reduction, so false positives outside that category increase overall. JOINT has 928.0 more empty-target FP pixels than BASE. Among the 3,371 GT regions, the number with at least half their pixels detected increases from 2,810.2 to 2,901.0. Their share increases from 83.364% to 86.058%, a gain of 2.694 pp. Coverage does not penalize spillover outside GT, so it must be interpreted alongside pixel-level false positives. These are connected-mask regions, not separately annotated building instances. Net count differences do not identify which individual regions were recovered or newly missed between models.

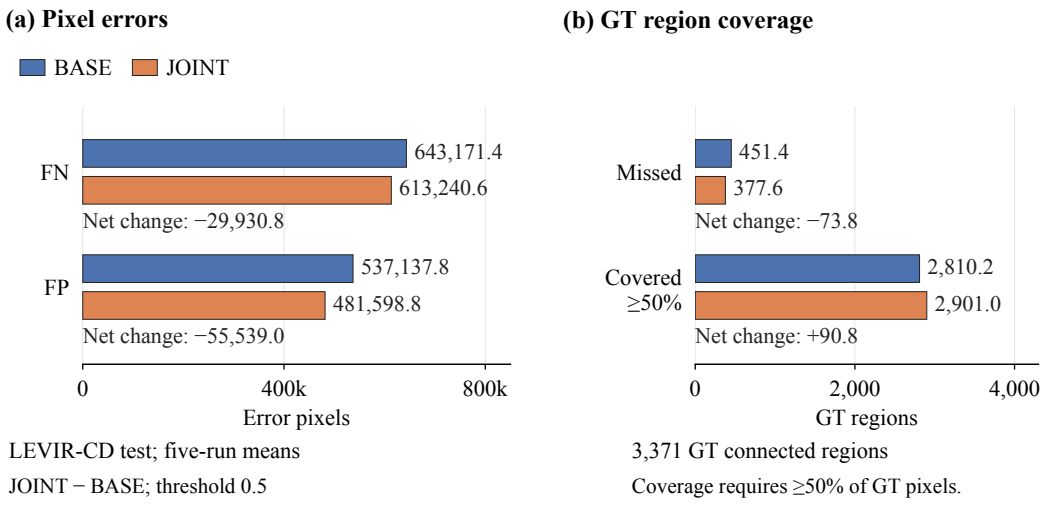


Figure A1: BASE and JOINT on LEVIR-CD, using five-run test means. (a) FN and FP pixel counts; (b) completely missed and at-least-half-covered GT regions. Blue and orange identify models, and both axes start at zero. Bars show mean counts; net changes are JOINT minus BASE. Table 11 reports all six models.

## F Metric Definitions and Scope

Within one run, pooled-pixel metrics use counts summed over the evaluation split:

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2TP}{2TP + FP + FN}, \quad \text{IoU} = \frac{TP}{TP + FP + FN}. \tag{15}$$

$$\text{OA} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{16}$$

Metrics are computed from pooled counts within each run and then averaged across the five runs. AP evaluates the ordering of continuous change scores. BF1 evaluates boundary agreement: boundaries are extracted by subtracting $3 \times 3$ mask erosions from dilations. Boundary precision and recall use exact overlap between the resulting boundary pixels, and BF1 is their harmonic mean.

GT components use eight-neighbor connectivity, with zero and at least 50% pixel coverage defining the missed and covered categories. Near-GT FP occupies the exterior three-pixel band around GT; empty-target FP counts predicted positives in images whose GT masks contain no change. The region statistics use the same 3,371 GT connected regions for all six LEVIR-CD models. Figure 3 provides selected validation examples, whereas the quantitative tables report five-run test means.

## G Complete Runtime Comparison

Table 12 reports every runtime setting on RTX 4090. The four columns correspond to two numerical precisions and two batch sizes.

Table 12: Inference time in ms per pair / peak GPU memory in MiB. B8 times are amortized across eight image pairs.

| Model | FP32 B1 | FP32 B8 | AMP B1 | AMP B8 |
|---|---|---|---|---|
| BASE | 60.80 / 185.0 | 7.95 / 345.0 | 66.10 / 235.0 | 8.84 / 340.0 |
| WIDE-CONTENT | 65.90 / 233.0 | 8.73 / 930.0 | 74.70 / 280.0 | 11.64 / 978.0 |
| EARLY-COMPRESS | 63.30 / 208.0 | 8.21 / 598.0 | 71.20 / 254.0 | 10.32 / 661.0 |
| SEPARATE | 70.40 / 240.0 | 9.26 / 970.0 | 80.00 / 293.0 | 12.58 / 1,032.0 |
| WIDE+SEPARATE | 69.10 / 237.0 | 9.04 / 954.0 | 78.30 / 289.0 | 12.25 / 1,009.0 |
| JOINT | 67.20 / 236.0 | 8.95 / 950.0 | 77.50 / 287.0 | 12.24 / 1,003.0 |

The runtime comparison is specific to the listed RTX 4090 settings. Batch-eight time is the batch inference time divided by eight. JOINT trades higher measured time and memory for improved detection, despite having fewer learned parameters than BASE.

## H Comparison With Related Architectures

Table 13 compares how three related architectures construct temporal representations and incorporate them into decoding.

Table 13: Construction and use of temporal features in nearby architectures. Descriptions follow the cited author implementations or method equations. Accuracy comparisons are reported separately in Appendix I.

| Method | Temporal feature construction | Subsequent processing |
|---|---|---|
| ChangeFormer V6 (Bandara & Patel, 2022b) | Projects each date and maps their concatenation from $2d$ to $d$ channels. | Adds the coarser comparison map, aggregates all scales, and reconstructs the mask progressively. |
| DIPFNet (Zhang, 2026) | Encodes absolute difference, sum, and concatenation in three branches; merges and reduces their outputs. | Sends one reduced map per scale into a hierarchical fusion decoder. |
| AdaDINO (Zhang et al., 2026) | Combines both date features with their absolute difference and product inside a frozen DINO encoder. | Uses pair-conditioned residual adaptation and carries its responses into prior-guided coarse-to-fine decoding. |
| JOINT | Encodes temporal concatenation and difference/product descriptors in two independent 96-channel stems. | Jointly updates their 192-channel state after top-down addition; compresses all four states during aggregation before reconstruction. |

# I External-Method Comparisons

Table 14 compares reproduced external methods with JOINT on the four test sets. All methods use the same dataset-specific partitions, validation-F1 checkpoint selection, threshold of 0.5, and pooled-pixel metric definitions. Architectures, pretraining, and training recipes remain method-specific.

The table includes 43 external method–dataset records after a uniform arithmetic check of P, R, F1, and IoU. Six inconsistent records and three unreported combinations are omitted.

Table 14: Test-set comparisons with external methods (%). JOINT rows are five-seed means, with SD reported in Table 1. External rows retain their original aggregation and two-decimal precision.

**(a) LEVIR-CD**

| Method | P | R | F1 | IoU |
|---|---|---|---|---|
| FC-Siam-Conc | 84.17 | 79.49 | 81.77 | 69.16 |
| IFNet | 94.02 | 82.93 | 88.13 | 78.77 |
| SNUNet | 89.18 | 87.17 | 88.16 | 78.83 |
| BIT | 89.24 | 89.37 | 89.31 | 80.68 |
| DTCDSCN | 88.53 | 86.83 | 87.67 | 78.05 |
| ChangeFormer | 92.05 | 88.80 | 90.40 | 82.48 |
| AERNet | 89.97 | 91.59 | 90.78 | 83.11 |
| SEIFNet | 94.29 | 87.67 | 90.86 | 83.25 |
| ChangeCLIP | 92.93 | 90.89 | 91.89 | 85.01 |
| RS-NormGAN | 91.72 | 91.03 | 91.37 | 84.11 |
| JOINT | 92.680 | 90.862 | 91.762 | 84.778 |

**(b) WHU-CD**

| Method | P | R | F1 | IoU |
|---|---|---|---|---|
| IFNet | 96.91 | 73.19 | 83.40 | 71.52 |
| SNUNet | 91.34 | 85.53 | 88.34 | 79.11 |
| BIT | 88.71 | 86.27 | 87.47 | 77.73 |
| DTCDSCN | 91.84 | 89.16 | 90.48 | 82.62 |
| ChangeFormer | 88.50 | 85.33 | 86.88 | 76.81 |
| ICIF-Net | 92.93 | 88.70 | 90.77 | 83.09 |
| DMINet | 92.65 | 90.35 | 91.49 | 84.31 |
| AERNet | 92.47 | 91.89 | 92.18 | 85.49 |
| SEIFNet | 93.99 | 92.61 | 93.29 | 87.43 |
| ChangeCLIP | 95.18 | 93.64 | 94.40 | 89.40 |
| RS-NormGAN | 94.21 | 93.48 | 93.85 | 88.41 |
| JOINT | 95.958 | 94.710 | 95.330 | 91.077 |

**(c) GZ-CD**

| Method | P | R | F1 | IoU |
|---|---|---|---|---|
| FC-Siam-Conc | 80.37 | 68.97 | 74.23 | 59.03 |
| IFNet | 92.19 | 74.08 | 82.15 | 69.71 |
| BIT | 82.40 | 78.18 | 80.23 | 66.99 |
| DTCDSCN | 88.19 | 78.38 | 83.00 | 70.93 |
| ChangeFormer | 84.59 | 65.23 | 73.66 | 58.30 |
| ICIF-Net | 89.90 | 80.76 | 85.09 | 74.05 |
| DMINet | 87.92 | 76.79 | 81.98 | 69.46 |
| AERNet | 88.06 | 81.07 | 84.42 | 73.03 |
| SEIFNet | 89.64 | 85.43 | 87.48 | 77.75 |
| ChangeCLIP | 91.86 | 85.15 | 88.38 | 79.18 |
| RS-NormGAN | 89.73 | 84.68 | 87.13 | 77.20 |
| ConvFormer | 90.32 | 88.52 | 89.41 | 80.85 |
| JOINT | 91.965 | 87.830 | 89.850 | 81.571 |

**(d) CDD**

| Method | P | R | F1 | IoU |
|---|---|---|---|---|
| FC-Siam-Conc | 93.58 | 63.13 | 75.40 | 60.51 |
| IFNet | 99.07 | 85.27 | 91.65 | 84.59 |
| SNUNet | 98.67 | 89.06 | 93.62 | 88.00 |
| BIT | 97.52 | 88.12 | 92.59 | 86.19 |
| DTCDSCN | 96.71 | 86.85 | 91.52 | 84.36 |
| ChangeFormer | 97.35 | 84.95 | 90.73 | 83.03 |
| ICIF-Net | 97.21 | 90.50 | 93.73 | 88.20 |
| DMINet | 97.30 | 88.94 | 92.93 | 86.79 |
| AERNet | 99.05 | 89.31 | 93.93 | 88.56 |
| SEIFNet | 99.18 | 89.08 | 93.86 | 88.43 |
| JOINT | 98.220 | 98.020 | 98.120 | 96.309 |

**Method references.** FC-Siam-Conc (Daudt et al., 2018); IFNet (DSIFN) (Zhang et al., 2020); SNUNet (SNUNet-CD) (Fang et al., 2022); BIT (Chen et al., 2022); DTCDSCN (Liu et al., 2021); ChangeFormer (Bandara & Patel, 2022a); ICIF-Net (Feng et al., 2022); DMINet (Feng et al., 2023); AERNet (Zhang et al., 2023); SEIFNet (Huang et al., 2024); ChangeCLIP (Dong et al., 2024); RS-NormGAN (Miao et al., 2025); ConvFormer (ConvFormer-CD) (Yang et al., 2025).

## J Aligned Qualitative Comparisons

Figure 4 compares single-checkpoint test predictions from BIT (Chen et al., 2022), DMINet (Feng et al., 2023), and SEIFNet (Huang et al., 2024) with JOINT. Each dataset contributes a nonempty patch selected by median GT area from the available matched prediction pool. All methods are shown on identical image patches and crops. Each shared crop maximizes disagreement among all four predictions.

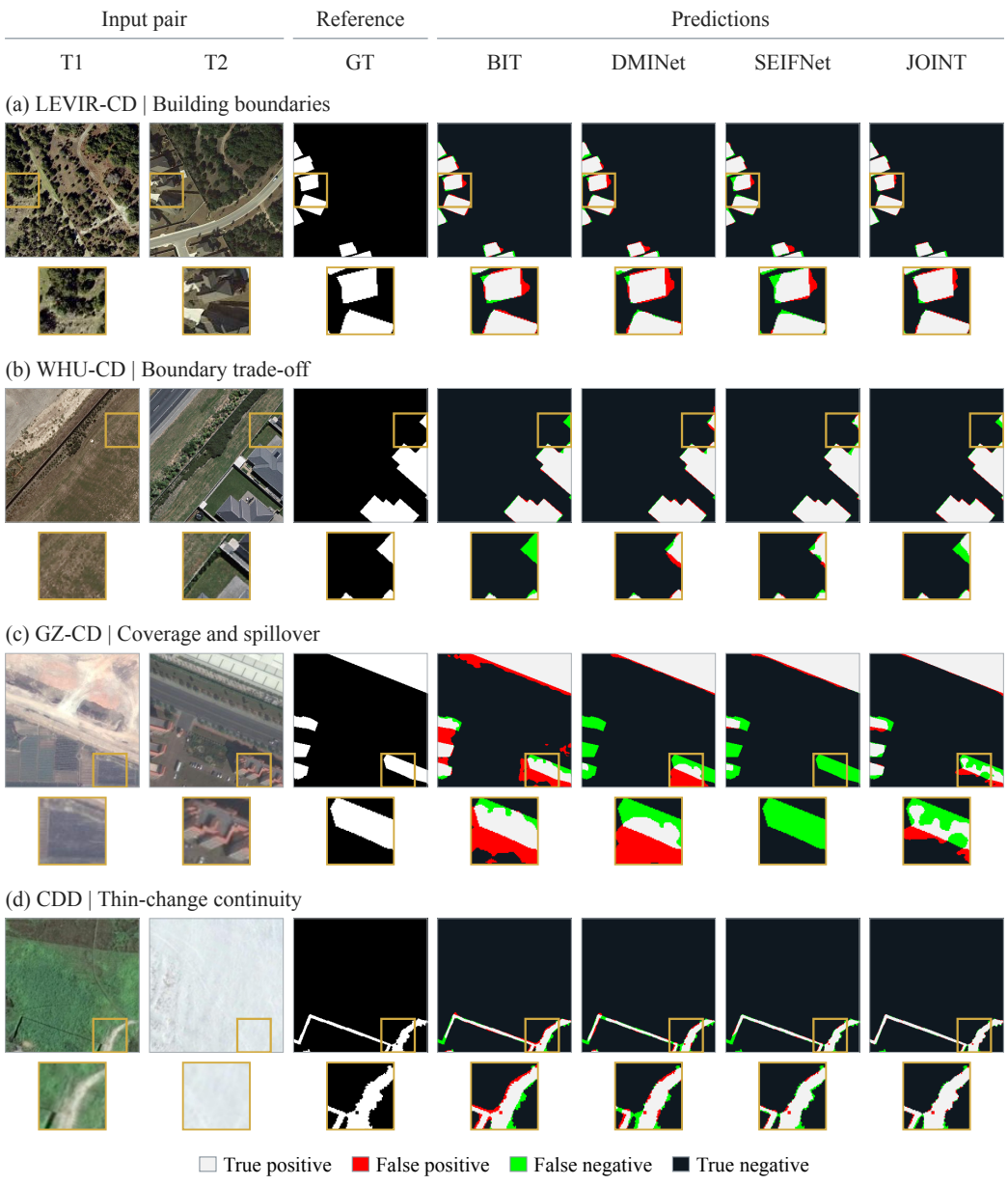


Figure 4: Aligned single-checkpoint predictions on four test examples. Gold $64 \times 64$ crops are enlarged at twice the relative scale; FP is red and FN green, as in Figure 3. LEVIR-CD and CDD show improved local coverage; GZ-CD shows a coverage–spillover trade-off. In the WHU-CD example, other methods recover more boundary pixels than JOINT. These single-checkpoint illustrations do not depict the five-seed means reported in the quantitative tables.