PAIRTRANS: REAL-PAIR-GROUNDED SEMANTIC-VISUAL TRANSFORMATION FOR SEMANTIC CHANGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic change detection (SCD) learns land-cover semantics and temporal transitions from densely annotated bi-temporal image-mask pairs. Such paired annotations are expensive to acquire. Existing generative augmentation methods construct additional change data from prompts, simulation rules, noise manipulation, or single-temporal states. Although effective, they do not explicitly construct new supervision by conditioning each synthesized temporal side on the opposite observation of the same annotated pair. We formulate SCD augmentation as bidirectional source-conditioned target-side generation from real annotated pairs and propose PairTrans. For target semantics, PairTrans derives a class-to-class transition matrix from the paired masks and perturbs it to propose new states with structured class transitions. For target appearance, it realizes each generated state from both its mask and the opposite-time real image. Applying the two directed generators symmetrically produces a fully synthetic pair whose semantic transitions are pair-guided and whose visual realizations are conditioned on observed scene context. Across three datasets and five SCD architectures, PairTrans achieves strong gains on SECOND and Landsat in low-data regimes, improves change-region metrics on HiUCD, and remains competitive as real supervision increases.

1 INTRODUCTION

Semantic change detection (SCD) (Caye Daudt et al., 2019; Yang et al., 2022; Tang et al., 2025) identifies land-cover categories and their transitions from bi-temporal remote sensing images. Unlike binary change detection (Lu et al., 2004), it predicts the semantic categories before and after change. Training SCD models (Zheng et al., 2022; Ding et al., 2022; 2024; Shen et al., 2026) requires spatially aligned image pairs with dense labels for both temporal states (Yang et al., 2022; Yuan et al., 2022; Tian et al., 2020), making annotation costly. The scarcity is also structural: even a modest set of land-cover categories gives rise to numerous directed transitions, many with few annotated examples. Observing a category in the training set does not imply coverage of the different transitions in which it can participate. Standard geometric or photometric augmentation preserves observed transition labels and cannot substantially enrich this transition space.

Generative augmentation reduces the dependence on expensive SCD annotations. Existing methods (Zheng et al., 2025; Zang et al., 2025; Liu et al., 2026) render images from generated layouts, transform a single temporal observation through simulation, or coordinate image pairs through noise-domain manipulation. Their conditioning structures expose different evidence to semantic-state generation and target-image realization.

For SCD, a target histogram does not identify a directed transition: different source-to-target mappings can yield the same composition but different change labels. Likewise, a mask specifies category and geometry but leaves appearance and source-scene context underdetermined. Illumination, acquisition conditions, and material texture are not encoded by the mask. A generated target may therefore be realistic in isolation and aligned with its mask, yet remain weakly related to the corresponding source observation. A complete annotated pair supplies both class-to-class relations and visual context. What remains underexplored is using its two endpoints as reciprocal source condi-

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.

Table 1: Conditioning structures of generative augmentation methods for SCD. I, M, and S denote image, semantic mask, and semantic state, respectively. Reciprocal pair conditioning means that both observations of one real annotated pair condition generation in opposite temporal directions.

Method	Mask generation	Image realization	Reciprocal pair conditioning
ChangeDiff	prompt $\rightarrow (\tilde{M}_1, \tilde{M}_2)$	$\tilde{M}_1 \rightarrow \tilde{I}_1,\ \tilde{M}_2 \rightarrow \tilde{I}_2$	$\times$
Changen2	$S_t +$ event $\rightarrow S_{t+1}$	$(I_t, S_t, S_{t+1}) \rightarrow I_{t+1}$	$\times$
Noise2Change	$M_1 +$ noise $\rightarrow (\tilde{M}_1, \tilde{M}_2)$	$(\tilde{M}_1, \tilde{M}_2) \rightarrow (\tilde{I}_1, \tilde{M}_1, \tilde{I}_2, \tilde{M}_2)$	$\times$
PairTrans (Ours)	$(M_1, M_2) \rightarrow \mathbf{T}_{1\rightarrow 2} \rightarrow \tilde{\mathbf{T}}_{1\rightarrow 2} \rightarrow \tilde{M}_2$ $(M_2, M_1) \rightarrow \mathbf{T}_{2\rightarrow 1} \rightarrow \tilde{\mathbf{T}}_{2\rightarrow 1} \rightarrow \tilde{M}_1$	$I_1 + \tilde{M}_2 \rightarrow \tilde{I}_2$ $I_2 + \tilde{M}_1 \rightarrow \tilde{I}_1$	✓

tions, so that each pseudo side is generated relative to the opposite real observation. Table 1 contrasts these conditioning structures.

The central challenge is to create new semantic states while retaining the transition information and scene context available in a real pair. This raises two linked questions: which target state should be proposed relative to a source observation, and how should that state be visually realized using the source scene as context?

We formulate SCD augmentation as *bidirectional source-conditioned target-side generation from real annotated pairs*. Given $\mathcal{X} = (I_1, M_1, I_2, M_2)$, each observed side conditions a new opposite side, specifying both its semantic transition and its appearance relative to the source scene. Since SCD jointly models both temporal observations, we use each observation to condition generation of the opposite side.

PairTrans addresses these questions with two coupled generators. The semantic generator perturbs a class-to-class transition matrix from the real mask pair to propose new transitions, subject to constraints on the proposal's deviation from the observed matrix. Category-constrained denoising then guides target-mask generation. The generated mask connects the two generators: it serves as both the synthetic target's annotation and the semantic condition for image realization. The visual generator renders it using the full opposite-time real image as appearance context. The same generators operate in both temporal directions, and their generated endpoints form a new pair with the same image-mask interface as real SCD supervision, without copying either real endpoint. Generator adaptation and downstream training use the same limited real subset.

Our main contributions are summarized as follows:

- We identify reciprocal source conditioning as an underexplored dimension of generative SCD augmentation and formulate the task as bidirectional target-side generation from complete real annotated pairs.
- We develop PairTrans, which combines transition-matrix perturbation and category-constrained semantic generation with target-mask-guided visual realization conditioned on the full source image, then assembles the two directed outputs into synthetic bi-temporal supervision.
- We evaluate PairTrans across three datasets, five downstream architectures, and multiple supervision ratios, with targeted ablations of its conditioning design and comparisons of augmentation benefits and trade-offs under limited supervision.

2 Related Work

2.1 Generative Augmentation for Change Detection

Synthetic augmentation addresses the high annotation cost of multi-temporal remote-sensing data. Early pipelines use procedural simulation, object insertion or removal, copy-paste editing, or rule-based scene modification (Song et al., 2024; Chen et al., 2022), with change patterns constrained

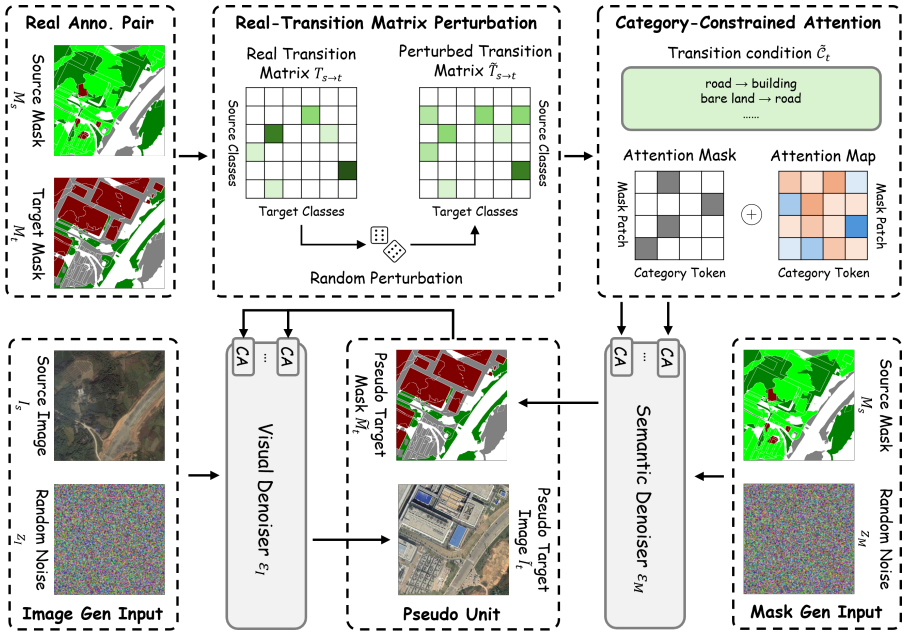

Figure 1: PairTrans in one temporal direction. The semantic branch perturbs the real pair-transition matrix and combines its condition with the source-mask latent to generate $\tilde{M}_t$. The visual branch then denoises a target latent jointly with the full source-image latent under $\tilde{M}_t$. Applying the same construction in reverse produces the opposite pseudo side, and the two directional outputs are assembled into a complete synthetic SCD pair.

by predefined rules or templates. Recent methods use learned generators with different conditioning sources. ChangeDiff generates semantic layouts from text or ratio conditions and then synthesizes image-mask pairs (Zang et al., 2025). Changen2 models temporal variation through stochastic change processes from single-temporal images or semantic states (Zheng et al., 2025). Noise2Change manipulates or optimizes noise-derived semantic states and jointly generates temporally aligned image-mask pairs (Liu et al., 2026).

PairTrans studies a complementary question: how both endpoints of a complete annotated pair can jointly guide the semantic transitions and visual realization of another pair.

2.2 Conditional Visual Generation

Image-to-image translation learns mappings from conditional inputs to target images (Isola et al., 2017), while diffusion models support controllable generation and editing (Rombach et al., 2022; Kawar et al., 2023). Spatial controls such as masks, layouts, and sketches further guide pretrained models (Zhang et al., 2023; Wang et al., 2026). In remote-sensing generation (Sastry et al., 2024; Toker et al., 2024), these controls help align plausible imagery with spatial structures and dense annotations.

PairTrans uses conditional generation at the level of bi-temporal supervision rather than individual-image synthesis: a generated mask specifies the target semantics, while the opposite-time real image supplies scene context. This realization process is coupled with pair-derived semantic transition in both temporal directions.

3 Method

Figure 1 summarizes PairTrans, which augments real bi-temporal data through semantic target-state generation and source-conditioned image synthesis. For each temporal direction, we perturb transitions observed in the real mask pair and generate a target mask conditioned on the source

mask. This generated mask serves both as the synthetic annotation and as the semantic condition for image synthesis, while the real source image provides appearance context. The two directional image–mask outputs are then assembled into a synthetic pair for downstream SCD training.

3.1 Problem Formulation and Decomposition

Let $\mathcal{D}_r = \{\mathcal{X}_n\}_{n=1}^{N}$ be a real SCD training set, where $\mathcal{X}_n = (I_1, M_1, I_2, M_2)$, $I_1, I_2 \in \mathbb{R}^{H\times W\times 3}$, and $M_1, M_2 \in \{1, \dots, K\}^{H\times W}$. A downstream model consumes (I_1, I_2) and predicts the two semantic states. We augment the real training set with a pseudo set $\mathcal{D}_p$ sharing the same image–mask interface, yielding $\mathcal{D}_a = \mathcal{D}_r \cup \mathcal{D}_p$.

For each direction $(s, t) \in \{(1, 2), (2, 1)\}$, let $\mathcal{G}_{s\to t}(\mathcal{X}; \xi_t)$ generate a target image–mask pair, where ξ_t denotes the stochastic generation state. Applying both directions gives

$$(\tilde{I}_1, \tilde{M}_1) = \mathcal{G}_{2\to 1}(\mathcal{X}; \xi_1), \quad (\tilde{I}_2, \tilde{M}_2) = \mathcal{G}_{1\to 2}(\mathcal{X}; \xi_2), \quad \tilde{\mathcal{X}} = (\tilde{I}_1, \tilde{M}_1, \tilde{I}_2, \tilde{M}_2). \tag{1}$$

Directional mask candidates are matched by joint changed-area ratio relative to the real anchor, then rendered using the corresponding real source images. Both endpoints are synthetic; matching does not enforce either proposal's classwise transition matrix. We perturb the real-pair transition as $\tilde{\mathbf{T}}_{s\to t} = \mathcal{P}(\mathbf{T}_{s\to t}; \pi_s, \xi)$, where π_s is the source-class composition and ξ specifies the perturbation proposal. With transition summary $\tilde{\mathbf{r}}_{s\to t} = \Gamma(\tilde{\mathbf{T}}_{s\to t})$, the directed conditional factorizes as

$$\hat{q}_\Theta(\tilde{I}_t, \tilde{M}_t \mid I_s, M_s, \tilde{\mathbf{r}}_{s\to t}) = q_{\theta_s}(\tilde{M}_t \mid M_s, \tilde{\mathbf{r}}_{s\to t}) q_{\theta_v}(\tilde{I}_t \mid I_s, \tilde{M}_t). \tag{2}$$

The semantic factor uses the source mask and transition summary without source radiometry. Its generated mask $\tilde{M}_t$ serves as both the pseudo label and the visual factor's semantic condition, while I_s provides appearance context. The two factors learn semantic transitions and source-conditioned appearance from observed pairs.

We train both directions on observed pairs: the semantic factor reconstructs M_t from M_s and $\Gamma(\mathbf{T}_{s\to t})$, and the visual factor reconstructs I_t from (I_s, M_t). Generated masks are never adaptation targets. At generation, the feasible proposal $\tilde{\mathbf{T}}_{s\to t}$ replaces the observed transition to sample $\tilde{M}_t$; this same mask labels and conditions the pseudo image.

3.2 Real-Transition-Guided Target-State Generation

The semantic factor generates $\tilde{M}_t$ from the source mask and a transition summary derived from the same real mask pair. It comprises pair-transition perturbation, transition-aware text conditioning, and a category constraint in semantic denoiser.

3.2.1 Pair-Transition Perturbation

A target state should reflect plausible land-cover transitions rather than an arbitrary category mixture. Let $\mathcal{A}_s = \{a : \pi_s^a > 0\}$ denote the source classes present in M_s. For each $a \in \mathcal{A}_s$, we summarize the directed real mask pair by a row-normalized matrix

$$\mathbf{T}_{s\to t}^{a,b} = \frac{|\{p \in \Omega \mid M_s(p) = a,\ M_t(p) = b\}|}{|\{p \in \Omega \mid M_s(p) = a\}|}, \tag{3}$$

where Ω is the pixel lattice. Each row is the empirical target distribution conditioned on source class a, and absent source classes are excluded.

Let π_s^a denote the fraction of source pixels assigned to class a. The target histogram induced by a transition matrix is

$$h_t^b = \sum_{a\in\mathcal{A}_s} \pi_s^a \mathbf{T}_{s\to t}^{a,b}. \tag{4}$$

Because h_t discards which source classes contribute to each target, PairTrans uses both target composition and active transition support.

For example, exchanging the destinations of two source classes can preserve the target histogram while changing transition labels. PairTrans therefore perturbs the full matrix and uses h_t as only one conditioning summary.

We construct a candidate with a stochastic class-swap operator

$$\bar{\mathbf{T}}_{s\to t} = \mathcal{C}_{\text{swap}}(\mathbf{T}_{s\to t}; \xi). \tag{5}$$

The operator selects two active source rows when available and proposes transferring probability mass between their target-class entries, and unselected rows remain unchanged. At least one selected row must introduce new target support. The state ξ determines the selected rows, source and destination entries, transfer magnitude, and random seed.

A raw class swap can overproduce change or move too far from the observed pair. We therefore constrain it using the global changed-area ratio induced by a transition matrix,

$$\chi(\mathbf{U}; \pi_s) = \sum_{a\in\mathcal{A}_s} \pi_s^a \left(1 - \mathbf{U}^{a,a}\right), \tag{6}$$

and define $\chi_{\mathbf{T}} = \chi(\mathbf{T}; \pi_s)$ for $\mathbf{T} = \mathbf{T}_{s\to t}$. Let $\ell_\chi = \max(\chi_{\min}, \gamma_{\min}\chi_{\mathbf{T}})$ and $u_\chi = \min(\chi_{\max}, \gamma_{\max}\chi_{\mathbf{T}})$. The feasible set contains nonnegative matrices satisfying

$$\begin{aligned} \mathbf{U}\mathbf{1} = \mathbf{1}, \quad \mathbf{U}_{\bar{\mathcal{A}}_\xi} = \mathbf{T}_{\bar{\mathcal{A}}_\xi}, \\ \max_{a\in\mathcal{A}_\xi} \|\mathbf{U}^{a,:} - \mathbf{T}^{a,:}\|_1 \le \tau_{\text{row}}, \quad \ell_\chi \le \chi(\mathbf{U}; \pi_s) \le u_\chi. \end{aligned} \tag{7}$$

where $\mathcal{A}_\xi \subseteq \mathcal{A}_s$ is the selected row set and $\bar{\mathcal{A}}_\xi$ its complement. The perturbed transition is

$$\tilde{\mathbf{T}}_{s\to t} = \mathcal{P}(\mathbf{T}_{s\to t}; \pi_s, \xi) = \Pi_{\mathcal{F}}\left(\bar{\mathbf{T}}_{s\to t}\right), \tag{8}$$

where $\Pi_{\mathcal{F}}$ denotes the feasibility projection used after proposal. This construction preserves valid row distributions and bounds departure from the observed pair while permitting both transition destinations and overall changed area to vary. It therefore introduces structured diversity without treating the matrix entries as independent noise.

The proposal is summarized as

$$\begin{aligned} \tilde{\mathcal{C}}_t &= \{b \mid \exists a \in \mathcal{A}_s : \tilde{T}^{a,b}_{s\to t} > 0\}, \\ \tilde{\Delta}_{s\to t} &= \{(a,b) \mid a \in \mathcal{A}_s, a \neq b, \tilde{T}^{a,b}_{s\to t} > 0\}, \\ \tilde{h}^b_t &= \sum_{a\in\mathcal{A}_s} \pi^a_s \tilde{T}^{a,b}_{s\to t}, \quad c_t = \phi(\tilde{h}_t, \tilde{\Delta}_{s\to t}). \end{aligned} \tag{9}$$

Here ϕ verbalizes proportions and active transitions, while $\tilde{\mathcal{C}}_t$ supplies the gate below. Thus $\tilde{\mathbf{r}}_{s\to t} = \Gamma(\tilde{\mathbf{T}}_{s\to t}) = (\tilde{h}_t, \tilde{\Delta}_{s\to t}, \tilde{\mathcal{C}}_t)$ conditions denoising without imposing an exact pixel quota.

Text communicates composition and transition intent, while the category gate suppresses unrelated classes; neither replaces the source mask used to locate target semantics.

3.2.2 Category-Constrained Text Attention

Text conditioning alone may activate categories outside the specified target set. We therefore define a category gate

$$g^k_t = \mathbb{I}[k \in \tilde{\mathcal{C}}_t], \qquad k = 1, \ldots, K. \tag{10}$$

We apply this gate as a signed bias to category-token responses in the text-conditioning layers. For mask-latent queries Q and category-token keys and values K_c, V_c, the constrained attention is

$$\text{Attn}_c(Q, K_c, V_c; g_t) = \text{softmax}\left(\frac{QK_c^\top}{\sqrt{d}} + \mathcal{B}(g_t)\right) V_c, \tag{11}$$

where $\mathcal{B}(g_t)_k = \alpha_+$ for allowed categories and $-\alpha_-$ otherwise. We use them to enhance allowed responses and suppress excluded ones. After decoding, an allowed-class projection repairs only illegal labels. It constrains category support but does not impose exact occupancy or class proportions.

We adapt a pretrained latent diffusion model to masks and convert it into a source-mask-conditioned editor. Adaptation uses the observed matrix and summary $\mathbf{r}_{s\to t} = \Gamma(\mathbf{T}_{s\to t})$, aligning c_t and g_t with M_t. Proposal sampling is used only for pseudo-target generation. Let $z^m_s = \mathcal{E}_m(M_s)$ and $z^m_t = \mathcal{E}_m(M_t)$ be four-channel mask latents. At step τ,

$$z^m_{t,\tau} = \sqrt{\bar{\alpha}_\tau} z^m_t + \sqrt{1-\bar{\alpha}_\tau}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{12}$$

Table 2: Main augmentation results on SECOND under different ratios. We report representative SCD metrics in the main text.

Backbone	Training data	5% real					20% real				
		Score	mIoU	SeK	F_{scd}	F1	Score	mIoU	SeK	F_{scd}	F1
SSCDL	Real only	28.68	66.51	12.46	53.09	64.30	30.31	68.37	14.00	54.16	67.21
	+ ChangeDiff	29.71	67.29	13.61	53.93	64.90	33.57	70.19	17.88	58.30	68.98
	+ Noise2Change	29.35	67.25	13.10	53.68	65.79	31.27	68.77	15.19	55.03	68.64
	+ Ours	**30.78**	**68.43**	**14.65**	**55.46**	**66.89**	**34.05**	**70.65**	**18.37**	**58.88**	**69.74**
BiSRNet	Real only	28.71	66.53	12.50	53.45	64.67	29.88	68.05	13.52	53.84	66.92
	+ ChangeDiff	29.59	67.42	13.38	53.35	65.28	**34.48**	**70.85**	**18.89**	**59.56**	69.97
	+ Noise2Change	29.43	67.25	13.21	53.90	65.52	31.25	69.32	14.93	54.69	68.92
	+ Ours	**31.28**	**68.67**	**15.26**	**56.32**	**67.69**	34.19	70.80	18.50	58.79	**70.14**
TED	Real only	28.72	66.56	12.51	53.30	64.49	30.03	68.15	13.69	53.87	67.16
	+ ChangeDiff	29.88	67.60	13.72	54.07	65.71	34.05	70.51	18.42	58.86	69.42
	+ Noise2Change	29.79	67.59	13.59	54.65	66.02	31.59	68.96	15.58	54.98	69.15
	+ Ours	**30.93**	**68.53**	**14.81**	**55.30**	**66.97**	**34.50**	**70.90**	**18.90**	**59.52**	**70.33**
A2Net	Real only	30.09	67.85	13.91	54.40	67.08	34.10	70.88	18.33	58.40	**70.63**
	+ ChangeDiff	30.03	67.88	13.81	53.87	66.18	34.18	70.83	18.48	58.70	70.21
	+ Noise2Change	30.37	68.03	14.23	54.83	66.81	33.16	70.28	17.26	57.40	69.53
	+ Ours	**31.32**	**68.70**	**15.30**	**55.86**	**67.72**	**34.66**	**71.09**	**19.04**	**59.73**	70.38
SCanNet	Real only	29.57	67.04	13.51	54.51	65.35	34.21	70.39	18.70	59.68	69.69
	+ ChangeDiff	30.38	67.69	14.38	55.22	65.56	34.81	70.82	19.37	**60.24**	69.89
	+ Noise2Change	31.04	68.05	15.17	56.36	66.95	33.10	69.74	17.40	57.78	69.71
	+ Ours	**32.02**	**68.82**	**16.25**	**57.38**	**67.54**	**34.99**	**71.06**	**19.53**	60.16	**70.74**

We expand the denoiser input from four to eight channels and feed the concatenation $[z^m_{t,\tau}; z^m_s]$. The transition text c_t enters cross-attention under gate g_t, yielding the training objective

$$\mathcal{L}_s = \mathbb{E}\left[\left\|\epsilon - \epsilon_{\theta_s}([z^m_{t,\tau}; z^m_s], \tau, c_t; g_t)\right\|_2^2\right]. \tag{13}$$

At generation, denoising from M_s and the sampled condition produces $\tilde{M}_t$. Decoding, allowed-class repair, and spatial cleanup are detailed in the supplement.

3.3 Source-Conditioned Target Image Realization

We adapt PLACE, a layout-semantic fusion model (Lv et al., 2024), to reconstruct I_t from (I_s, M_t) during training and synthesize $\tilde{I}_t$ from $(I_s, \tilde{M}_t)$ during generation. We first adapt the four-channel PLACE denoiser to mask-conditioned remote-sensing image generation, then use its weights to initialize an eight-channel source-conditioned editor. The editor is trained on (I_1, M_2, I_2) and (I_2, M_1, I_1), ordered as source image, target mask, and reconstruction target.

Let $z^I_s = \mathcal{E}_I(I_s)$ and $z^I_t = \mathcal{E}_I(I_t)$ be four-channel image latents, and let $z^I_{t,\tau}$ be the noised target latent. Rather than conditioning on a mask alone, the denoiser receives

$$u^I_\tau = [z^I_{t,\tau}; z^I_s] \in \mathbb{R}^{8\times h\times w}, \tag{14}$$

while PLACE uses M_t as a spatial hint and injects its class identities through class-token attention; we denote this mask condition by $\psi(M_t)$. Its diffusion objective is

$$\mathcal{L}_v = \mathbb{E}\left[\left\|\epsilon - \epsilon_{\theta_v}(u^I_\tau, \tau, \psi(M_t))\right\|_2^2\right]. \tag{15}$$

The source latent and target mask provide separate appearance and semantic conditions. Source appearance is incorporated through latent conditioning and the visual diffusion objective.

For each directed transition proposal, we sample $J = 8$ target-mask candidates, denoted by $\mathcal{Q}_t = \{\tilde{M}^{(j)}_t\}_{j=1}^J$. For $(A, B) \in \mathcal{Q}_1 \times \mathcal{Q}_2$, let $\chi(A, B) = |\Omega|^{-1} \sum_{p\in\Omega} \mathbb{I}[A(p) \neq B(p)]$ and define the anchor-compatibility score

$$d_\chi(A, B; \mathcal{X}) = |\chi(A, B) - \chi(M_1, M_2)|. \tag{16}$$

After allowed-class repair and spatial cleanup, pairs outside the predefined valid range of $\chi(A, B)$ are discarded and the remainder are ranked by d_χ. For the selected masks $A = \tilde{M}_1$ and $B = \tilde{M}_2$, the visual factor samples $\tilde{I}_1 \sim q_{\theta_v}(\cdot \mid I_2, A)$ and $\tilde{I}_2 \sim q_{\theta_v}(\cdot \mid I_1, B)$ to complete Eq. (1).

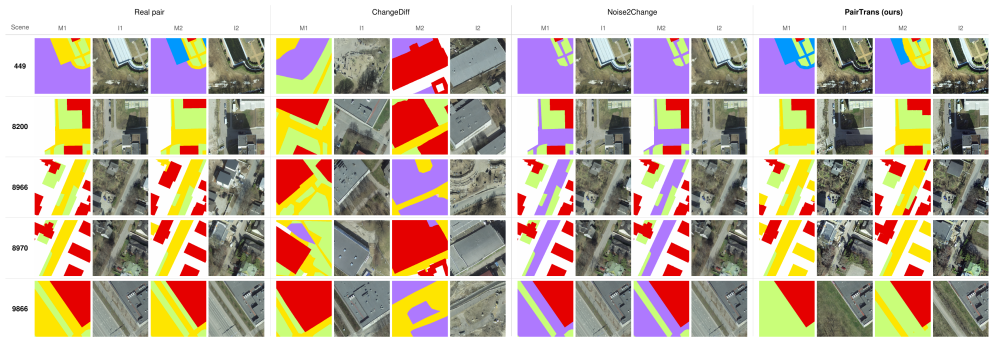

Figure 2: Qualitative comparison on HiUCD under the 5% setting. Each row denotes one scene, and each block displays (M_1, I_1, M_2, I_2). Complete pairs are shown because the methods do not share an identical synthesized source endpoint. All panels use the original 512×512 outputs.

4 Experiments

4.1 Experimental Setup

Datasets and Training Protocol. We use three complementary benchmarks. SECOND (Yang et al., 2022) has 4,662 aligned 512×512 pairs in a 2,968/1,694 train/test split and six land-cover classes. The Landsat-SCD release (Yuan et al., 2022) contains 8,468 416×416 pairs at 30-m resolution, including pre-generated augmentations. We exclude these augmentations and use only the 2,385 original pairs. HiUCD (Tian et al., 2020) has 40,800 urban pairs at 0.1-m resolution, nine classes, 48 change types, and strong unchanged-pixel dominance. Following ChangeDiff (Zang et al., 2025), the same fixed real subset adapts the generators and trains downstream models. We test SECOND at 5% and 20%, and Landsat-SCD and HiUCD at 5%. Generated pairs augment only these subsets, while evaluation remains entirely real. Dataset-specific subset construction, seeds, counts, and training schedules are reported in the supplement. The 5% subsets contain 148, 71, and 87 real pairs, respectively; HiUCD samples only training pairs containing change.

Compared methods. We reproduce ChangeDiff (Zang et al., 2025) and Noise2Change (Liu et al., 2026) on the selected subsets. Augmentation methods share real subsets, five synthetic pairs per real pair, downstream hyperparameters, and epoch counts. Generator adaptation uses method-specific configurations. Changen2 (Zheng et al., 2025) appears in Table 1 but lacks official training code for reproduction.

Downstream SCD models. Following ChangeDiff (Zang et al., 2025), we use five downstream architectures: SSCDL (Ding et al., 2022), BiSRNet (Ding et al., 2022), TED (Ding et al., 2024), A2Net (Li et al., 2023), and SCanNet (Ding et al., 2024).

Metrics. Following ChangeDiff (Zang et al., 2025) and the standard SECOND-style SCD evaluation protocol, we report OA, Score, mIoU, SeK, F_{scd}, Kappa, IoU, F1, Recall, and Precision. OA, Score, mIoU, SeK, F_{scd}, and Kappa mainly evaluate semantic change detection quality, while IoU, F1, Recall, and Precision summarize binary change-region quality. Higher values indicate better performance for all metrics.

4.2 Main Results

Qualitative Visualization. Figure 2 illustrates semantic changes realized within a largely stable spatial layout. In scene 9866, PairTrans changes the broad diagonal region between its masks, and the corresponding images show a grassy strip in one endpoint and a paved strip in the other, while the adjacent roof remains recognizable. Each method retains its own (M_1, I_1, M_2, I_2) endpoints, so this comparison concerns within-pair coherence rather than reconstruction from a shared source input.

Results on SECOND. Table 2 shows that PairTrans's relative advantage depends on the amount of real supervision. With 5% real data, gains span semantic-change and change-region metrics

Table 3: Main augmentation results on Landsat and HiUCD with 5% real training data, evaluated on real test images. Full metrics are provided in the supplementary material.

Backbone	Training data	Landsat					HiUCD				
		Score	mIoU	SeK	F_{scd}	F1	Score	mIoU	SeK	F_{scd}	F1
SSCDL	Real only	33.15	69.80	17.45	62.01	69.86	**14.77**	49.22	**0.00**	0.00	0.00
	+ ChangeDiff	33.71	70.37	18.00	63.15	70.01	14.49	49.49	-0.51	9.61	17.06
	+ Noise2Change	32.62	68.82	17.11	59.58	70.03	13.42	51.36	-2.83	9.50	18.24
	+ Ours	**37.35**	**72.91**	**22.11**	**66.94**	**72.86**	12.21	**54.55**	-5.94	**12.85**	**23.23**
BiSRNet	Real only	33.01	69.66	17.30	62.18	69.54	**14.76**	49.22	**-0.00**	0.00	0.00
	+ ChangeDiff	33.30	69.96	17.59	62.60	69.74	14.09	50.46	-1.49	9.59	17.25
	+ Noise2Change	31.39	68.08	15.66	57.45	69.36	13.80	49.75	-1.61	8.12	16.57
	+ Ours	**37.68**	**72.87**	**22.59**	**67.56**	**72.89**	12.08	**53.46**	-5.66	**10.96**	**20.57**
TED	Real only	37.08	72.68	21.83	66.14	72.96	13.19	55.26	-4.84	14.46	25.53
	+ ChangeDiff	37.39	73.02	22.12	66.32	73.46	13.93	48.12	-0.72	7.04	14.93
	+ Noise2Change	35.19	71.16	19.78	62.38	72.44	**14.52**	49.29	**-0.38**	0.20	0.31
	+ Ours	**40.68**	**75.01**	**25.97**	**69.96**	**75.41**	13.01	**55.90**	-5.37	**14.69**	**26.85**
A2Net	Real only	35.74	71.53	20.40	65.55	71.54	15.37	53.56	-1.00	13.77	24.05
	+ ChangeDiff	36.02	72.04	20.58	65.39	72.12	**15.59**	53.77	**-0.78**	14.08	24.24
	+ Noise2Change	37.37	72.84	22.17	64.51	74.23	14.49	53.51	-2.24	13.34	22.86
	+ Ours	**39.56**	**74.01**	**24.79**	**69.32**	**74.31**	14.54	**57.19**	-3.73	**19.22**	**30.51**
SCanNet	Real only	39.24	73.64	24.50	68.40	74.31	13.01	54.91	-4.94	11.61	25.13
	+ ChangeDiff	39.09	73.70	24.25	68.75	74.12	13.50	50.73	-2.46	7.85	16.83
	+ Noise2Change	36.55	71.20	21.71	64.51	72.61	**14.77**	49.22	**-0.00**	0.00	0.00
	+ Ours	**42.80**	**76.02**	**28.57**	**72.16**	**76.73**	11.57	**57.81**	-8.24	**13.65**	**30.88**

across all five backbones, indicating a benefit shared by different architectures. With 20%, the competing augmentation methods become more closely matched, and ChangeDiff leads Score for BiSRNet. The strongest evidence for PairTrans is therefore in the scarce-label regime: additional real pairs narrow its relative advantage, although its Score remains above real-only training for all tested backbones. Table 2 also reveals a cross-budget pattern: with 5% real data, PairTrans exceeds 20% real-only training in Score, mIoU, SeK, and F_{scd} for SSCDL, BiSRNet, and TED. This ordering does not hold for A2Net or SCanNet, nor consistently for binary F1, indicating architecture-dependent benefits under a reduced annotation budget.

Results on Landsat. Table 3 shows consistent gains in semantic-change and change-region quality on medium-resolution imagery. The full metrics in the supplementary material reveal how the error balance differs: Noise2Change favors Recall, whereas PairTrans achieves higher Precision and F1 across all backbones. PairTrans therefore improves the reliability of predicted changes without requiring higher change recall.

Results on HiUCD. On HiUCD, Table 3 reveals a trade-off between change-region quality and semantic agreement. Relative to real-only training, PairTrans improves binary F1 and F_{scd} across backbones. For SSCDL and BiSRNet, it recovers changes that real-only models miss; for TED, A2Net, and SCanNet, higher Precision accompanies lower Recall. Meanwhile, Score and SeK decrease across all five models.

4.3 Ablation Studies

We ablate PairTrans on SECOND with 5% real training data using SCanNet. The supplementary material reports full results for all backbones.

Core components. With both semantic controls enabled, source-image conditioning further improves the main SCD metrics in Table 4; the complete model also outperforms image conditioning alone. Across backbones, the full combination leads Score, mIoU, and SeK. However, the ranking of partial configurations varies. For SCanNet in Table 4, adding mask attention without transition perturbation lowers Score, mIoU, and SeK, whereas adding it with perturbation improves all reported metrics, with image conditioning enabled throughout. This contrast suggests that the benefit of category control depends on the transition proposals it accompanies, supporting their joint use.

Table 4: Component ablation on SECOND with 5% real training data using SCanNet. Img. Cond., Matrix Pert., and Mask Attn. denote source-image conditioning, real-transition matrix perturbation, and category-constrained mask attention, respectively.

Img. Cond.	Matrix Pert.	Mask Attn.	Score	mIoU	SeK	F_{scd}	F1
	Real only		29.57	67.04	13.51	54.51	65.35
×	✓	✓	31.22	68.33	15.31	56.35	67.18
✓	×	×	31.10	68.18	15.21	56.20	66.68
✓	✓	×	31.43	68.46	15.56	56.60	67.44
✓	×	✓	30.99	68.04	15.10	56.27	66.87
✓	✓	✓	**32.02**	**68.82**	**16.25**	**57.38**	**67.54**

Table 5: Direction, data-ratio, and grounding ablations on SECOND with 5% real training data using SCanNet. Full PairTrans uses bidirectional generation, a 1:5 real-to-synthetic ratio, and matched references. Uni. pairs a generated target with its real source; Mismatch-ref breaks source–transition correspondence. Real only uses no synthetic data.

Factor	Setting	Score	mIoU	SeK	F_{scd}	F1
Direction	Uni. $1 \rightarrow 2$	26.67	64.14	10.61	50.99	59.94
	Uni. $2 \rightarrow 1$	27.87	65.49	11.74	52.30	62.88
Data ratio	1:1	30.48	67.57	14.58	55.54	66.32
	1:10	31.66	68.46	15.88	56.87	66.73
Pair grounding	Mismatch-ref	31.26	68.32	15.38	56.22	66.87
Reference	Real only	29.57	67.04	13.51	54.51	65.35
	Full PairTrans	**32.02**	**68.82**	**16.25**	**57.38**	**67.54**

Direction-symmetric generation. Table 5 and the supplementary results show that both unidirectional variants fall below real-only training in Score across all tested backbones. At equal synthetic-data volume, the full pair construction outperforms both variants. These results support the full pair construction as an effective design for synthetic SCD supervision.

Generated-data amount. Table 5 shows that increasing synthetic-data volume is not uniformly beneficial. For SCanNet, moving from 1:5 to 1:10 raises Precision but lowers Recall, F1, and the main SCD metrics. The ratio favored by Score also depends on the backbone. The generated-data amount thus changes the balance of detection errors and should be selected for the downstream model and target metric. Under the fixed-epoch protocol, larger augmented datasets entail more optimization steps.

Real-pair grounding. Table 5 compares matched and mismatched references at equal synthetic-data volume. Mismatch-ref improves Score over real-only training, and matched references provide additional gains in Score and semantic-change metrics across backbones. Source–transition correspondence therefore contributes beyond the presence of generated samples, while its effects on OA and Precision vary across backbones.

5 Conclusion

PairTrans constructs synthetic SCD supervision by using real-pair transition structure to guide target semantics and the opposite-time image to condition target appearance. Its bidirectional construction improves downstream performance on SECOND and Landsat in limited-data settings, while HiUCD exposes a trade-off between change-region quality and semantic agreement. The component, direction, and grounding ablations support the value of the complete conditioning design. Future work can investigate explicit local consistency objectives for fine-grained temporal correspondence.

AI USE STATEMENT

We used generative AI tools primarily for language polishing and improving the readability of the manuscript. The authors take full responsibility for the final manuscript, including its text, claims, and results.

REPRODUCIBILITY STATEMENT

Section 3 describes the semantic–visual factorization, transition perturbation, conditioning mechanisms, and synthetic-pair construction of PairTrans. Appendix A reports dataset preparation and subset selection, generator training and sampling settings, baseline reproduction details, and the downstream training and evaluation protocol. It also provides random seeds, hyperparameters, the checkpoint-selection criterion, and stage-wise generator training costs. Appendix B presents additional qualitative examples and ablation results across all five downstream backbones.

REFERENCES

Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Comput. Vis. Image Underst.*, 187:102783, 2019. doi: 10.1016/j.cviu.2019.07.003. URL https://doi.org/10.1016/j.cviu.2019.07.003.

Hao Chen, Wenyuan Li, and Zhenwei Shi. Adversarial instance augmentation for building change detection in remote sensing images. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–16, 2022. doi: 10.1109/TGRS.2021.3066802. URL https://doi.org/10.1109/TGRS.2021.3066802.

Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. doi: 10.1109/TGRS.2022.3154390. URL https://doi.org/10.1109/TGRS.2022.3154390.

Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3362795. URL https://doi.org/10.1109/TGRS.2024.3362795.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5967–5976. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.632. URL https://doi.org/10.1109/CVPR.2017.632.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, 2024. doi: 10.1109/CVPR52733.2024.02282. URL https://doi.org/10.1109/CVPR52733.2024.02282.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 6007–6017. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00582. URL https://doi.org/10.1109/CVPR52729.2023.00582.

Zhenglai Li, Chang Tang, Xinwang Liu, Wei Zhang, Jie Dou, Lizhe Wang, and Albert Y Zomaya. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. doi: 10.1109/TGRS.2023.3241436. URL https://doi.org/10.1109/TGRS.2023.3241436.

Qiang Liu, Yang Kuang, Jun Yue, Pedram Ghamisi, Weiying Xie, and Leyuan Fang. Generating any changes in the noise domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3698–3713, 2026.

doi: 10.1109/TPAMI.2025.3643733. URL https://doi.org/10.1109/TPAMI.2025.3643733.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Dengsheng Lu, Paul Mausel, Eduardo Brondízio, and Emilio Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. doi: 10.1080/0143116031000139863. URL https://doi.org/10.1080/0143116031000139863.

Zhengyao Lv, Yuxiang Wei, Wangmeng Zuo, and Kwan-Yee K. Wong. PLACE: Adaptive layout-semantic fusion for semantic image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9264–9274, June 2024. doi: 10.1109/CVPR52733.2024.00885. URL https://doi.org/10.1109/CVPR52733.2024.00885.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: Contextually-aware high-resolution satellite image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pp. 460–470. IEEE, 2024. doi: 10.1109/CVPRW63382.2024.00051. URL https://doi.org/10.1109/CVPRW63382.2024.00051.

Hengtong Shen, Li Yan, Hong Xie, Yaxuan Wei, Xinhao Li, Wenfei Shen, Peixian Lv, and Fei Tan. Foundation model-driven semantic change detection in remote sensing imagery. *CoRR*, abs/2602.13780, 2026. URL https://arxiv.org/abs/2602.13780.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Jian Song, Hongruixuan Chen, and Naoto Yokoya. Syntheworld: A large-scale synthetic dataset for land cover mapping and building change detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 8272–8281. IEEE, 2024. doi: 10.1109/WACV57701.2024.00810. URL https://doi.org/10.1109/WACV57701.2024.00810.

Yingjie Tang, Shou Feng, Chunhui Zhao, Yongqi Chen, Zhiyong Lv, and Weiwei Sun. A semantic change detection network based on boundary detection and task interaction for high-resolution remote sensing images. *IEEE Trans. Neural Networks Learn. Syst.*, 36(9):17184–17198, 2025. doi: 10.1109/TNNLS.2025.3570425. URL https://doi.org/10.1109/TNNLS.2025.3570425.

Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *CoRR*, abs/2011.03247, 2020. URL https://arxiv.org/abs/2011.03247.

Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 27685–27695. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02615. URL https://doi.org/10.1109/CVPR52733.2024.02615.

Haoyu Wang, Lei Zhang, Wenrui Liu, Dengyang Jiang, Wei Wei, and Chen Ding. Jodiffusion: Jointly diffusing image with pixel-level annotations for semantic segmentation promotion. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 9775–9783. AAAI Press, 2026. doi: 10.1609/AAAI.V40I12.37941. URL https://doi.org/10.1609/aaai.v40i12.37941.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote. Sens.*, 60:1–18, 2022. doi: 10.1109/TGRS.2021.3113912. URL https://doi.org/10.1109/TGRS.2021.3113912.

Panli Yuan, Qingzhan Zhao, Xingbiao Zhao, Xuewen Wang, Xuefeng Long, and Yuchen Zheng. A transformer-based siamese network and an open optical dataset for semantic change detection of remote sensing images. *Int. J. Digit. Earth*, 15(1):1506–1525, 2022. doi: 10.1080/17538947.2022.2111470. URL https://doi.org/10.1080/17538947.2022.2111470.

Qi Zang, Jiayi Yang, Shuang Wang, Dong Zhao, Wenjun Yi, and Zhun Zhong. Changediff: A multi-temporal change detection data generator with flexible text prompts via diffusion model. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 9763–9771. AAAI Press, 2025. doi: 10.1609/aaai.v39i9.33058. URL https://doi.org/10.1609/aaai.v39i9.33058.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3813–3824. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00355. URL https://doi.org/10.1109/ICCV51070.2023.00355.

Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. Changemask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2021.10.015. URL https://doi.org/10.1016/j.isprsjprs.2021.10.015.

Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(2):725–741, 2025. doi: 10.1109/TPAMI.2024.3475824. URL https://doi.org/10.1109/TPAMI.2024.3475824.

APPENDIX OVERVIEW

Appendix A describes the datasets, PairTrans implementation, reproduced comparison methods, downstream training and evaluation protocol, computational costs, and hyperparameters. Appendix B presents additional qualitative examples, full ablation results on SECOND, and complete Landsat-SCD and HiUCD comparisons. Appendix C discusses generation characteristics, evaluation scope, efficiency and generalization, and broader impact.

A MORE IMPLEMENTATION DETAILS

A.1 DATASETS

We conduct experiments on three semantic change detection (SCD) benchmarks: SECOND (Yang et al., 2022), Landsat-SCD (Yuan et al., 2022), and HiUCD (Tian et al., 2020). These datasets cover different spatial resolutions, scene types, and change distributions, allowing us to evaluate PairTrans under diverse remote-sensing SCD settings. SECOND is used as the primary benchmark for the main comparisons and ablations, while Landsat-SCD and HiUCD are used to assess whether the augmentation gains remain consistent across datasets with different spatial resolutions and change distributions.

A.1.1 SECOND.

SECOND is a large-scale SCD benchmark containing 4,662 pairs of aerial images with dense semantic annotations for both temporal observations. Each image has a spatial size of 512×512, and the spatial resolution ranges from 0.5m to 3m. Following the commonly used split, SECOND contains 2,968 training pairs and 1,694 testing pairs. The semantic annotations include six land-cover categories, covering common urban and natural surface types such as non-vegetated ground surface, tree, low vegetation, water, building, and playground.

A.1.2 LANDSAT-SCD.

Landsat-SCD is a medium-resolution SCD benchmark constructed from Landsat satellite imagery. It contains 8,468 image pairs with a spatial size of 416×416 and a spatial resolution of 30m. The dataset is split into 6,053 training samples, 1,729 validation samples, and 686 testing samples. The 8,468-pair count refers to the released dataset including pre-generated augmentations. For our limited-supervision experiments, we remove these augmented samples and retain the 2,385 original pairs, which are re-split into 1,431/477/477 training/validation/testing pairs. Compared with SECOND, Landsat-SCD has lower spatial resolution and larger spatial coverage, making it useful for evaluating whether generative augmentation remains effective under different imaging resolutions and land-cover distributions.

A.1.3 HIUCD.

HiUCD is an ultra-high-resolution urban SCD benchmark built from aerial images of Tallinn, Estonia. It contains 40,800 image pairs with a spatial resolution of 0.1m. The dataset provides semantic annotations for nine land-cover categories and supports 48 semantic change categories. Compared with SECOND and Landsat-SCD, HiUCD has a more imbalanced change distribution: unchanged regions dominate a large portion of the images, and many changed categories occupy only a small fraction of pixels. Therefore, overall accuracy can be misleading on HiUCD, since a model may obtain high OA by predicting most pixels as unchanged.

A.2 IMPLEMENTATION DETAILS OF PAIRTRANS.

We instantiate the semantic–visual factorization in the main paper with two sequential latent-diffusion modules (Rombach et al., 2022). For a directed real pair, the target-state transition module takes the real source mask M_s and a perturbed transition specification $\tilde{\mathbf{T}}_{s\to t}$ and generates a target mask $\tilde{M}_t$. The target-image realization module then takes the real source image I_s and $\tilde{M}_t$ and generates the corresponding target image $\tilde{I}_t$. The generated mask is therefore an explicit interface

between semantic transition and visual realization. Each module is trained in two adaptation stages rather than executed as a 4-channel-to-8-channel cascade: the 4-channel stage is initialized from the COCO-pretrained weights released with PLACE (Lv et al., 2024) and adapted to the target SCD dataset, after which the adapted weights initialize the 8-channel source-conditioned model used for directional generation. The directional target-mask and target-image generators use DDIM sampling (Song et al., 2021); module-specific step counts and guidance scales are reported in Table 7.

A.2.1 TARGET-STATE TRANSITION MODEL.

Given a directed real mask pair (M_s, M_t), we first compute its row-normalized class-transition matrix. We perturb this matrix by selecting source-class rows and transferring transition mass between target-class entries, and then project the proposal under row-wise ℓ_1 and global change-ratio constraints. The resulting matrix remains anchored to the transition statistics of the real pair while specifying a modified target-class composition. Table 7 reports the shared perturbation parameters used for all three datasets.

The 4-channel LatentDiffusion model is initialized from the COCO-pretrained weights released with PLACE (Lv et al., 2024) and adapted to class-conditioned mask generation on each SCD dataset. Its denoiser operates on a noisy target-mask latent, with the target class composition supplied through CLIP-based cross-attention (Radford et al., 2021). The 8-channel RefMaskHybridLDM is initialized from the corresponding adapted 4-channel checkpoint. At each denoising step, the VAE-encoded real source mask is concatenated channel-wise with the noisy target-mask latent, yielding the 8-channel source-conditioned input. The perturbed source-to-target matrix is verbalized as a transition condition and simultaneously converted into a target-side category gate. During transition-conditioned denoising, the category gate is applied as a finite signed bias to category-token attention: allowed categories receive a positive bias and excluded categories receive a negative bias. We set the positive and negative attention-gate scales to 6.0 and 20.0, respectively. We treat the denoised output as the initial target mask and apply lightweight fragment removal, boundary cleanup, and simple smoothing to obtain the final synthetic mask. The resulting mask $\tilde{M}_t$ therefore remains conditioned on the source mask while providing a spatially coherent realization of the requested semantic transition.

A.2.2 SOURCE-CONDITIONED TARGET IMAGE REALIZATION.

We use PLACE (Lv et al., 2024) as the layout-semantic fusion backbone for the image realization module. This module follows the same two-stage adaptation scheme while assigning separate pathways to semantic layout and source appearance. The 4-channel PLACE LDM is initialized from the COCO-pretrained weights released with PLACE and adapted for mask-conditioned remote-sensing image generation. The target mask is provided as a spatial hint, and its class identities are injected through class-token PLACE attention. The 8-channel PLACE editing LDM is then initialized from the corresponding adapted 4-channel checkpoint. Its denoising input concatenates the noisy target-image latent with the VAE-encoded real source image, while the generated target mask remains the spatial condition throughout denoising. Thus, I_s makes source-scene geometry, appearance, and imaging context available as conditioning cues, whereas $\tilde{M}_t$ specifies the semantic layout to be realized on the generated side. The same source-conditioned renderer is applied in both temporal directions, and its two directional outputs are subsequently paired. The complete backbone, initialization, optimization, checkpoint, and sampling configurations are listed in Table 7.

A.2.3 TRAINING SUBSETS AND PSEUDO-PAIR CONSTRUCTION.

All four generator branches are adapted exclusively on the real subset used in the corresponding limited-supervision experiment. For SECOND, we use either 5% or 20% of the 2,968 training pairs according to the evaluated setting; both ratios follow the same seeded subset-construction, generator-adaptation, sampling, and downstream-training protocol, and differ only in the number of selected real pairs. The 5% setting contains 148 pairs. For Landsat-SCD, we remove the pre-generated augmented samples, retain 2,385 original pairs, split them into 1,431/477/477 training/validation/testing pairs, and select 71 training pairs. For HiUCD, we select 5% of the change-containing training-pair pool, yielding 87 pairs. Subset construction uses random seed 42. Both temporal observations con-

tribute to generator adaptation: the 4-channel models are adapted with samples from both sides, and the 8-channel models use directed training examples in both the $1 \to 2$ and $2 \to 1$ directions.

At generation time, PairTrans is applied in both temporal directions. We produce eight target-mask candidates for each directed transition proposal, apply lightweight fragment removal, boundary cleanup, and simple smoothing, and match candidates from the two directions with reference to the semantic change-area ratio of the corresponding real pair. The forward direction generates a new target side from the first observation, and the reverse direction generates the complementary side from the second observation. Each selected target image is realized from its processed target mask and the corresponding real source observation, after which the two generated sides are paired to form a bi-temporal training sample; neither real side is copied into the final pair. Both temporal directions use the same semantic and visual generators adapted on the selected real subset. We retain five synthetic pairs per selected real pair and mix them with the original labeled pairs for downstream SCD training, giving the real-to-synthetic ratio of $1 : 5$.

A.3 Implementation Details of Compared Methods

We reproduce the compared generative augmentation methods, including ChangeDiff (Zang et al., 2025) and Noise2Change (Liu et al., 2026), under the same limited-real-supervision protocol. For a controlled comparison, each method uses the same selected real subsets and retains exactly five pseudo pairs per real pair (a real-to-synthetic ratio of $1 : 5$). We then retrain the same downstream SCD models with matched optimization settings and evaluation protocols. All results are obtained from our reproductions under the shared downstream evaluation protocol.

A.3.1 ChangeDiff reproduction.

We use the official ChangeDiff implementation and retain its two-stage text-to-layout (T2L) and layout-to-image (L2I) pipeline. For each benchmark, the ChangeDiff training set is constructed exclusively from the real subset selected for our limited-supervision experiment. The two temporal image–mask sides are converted to the input format expected by ChangeDiff, and category names and pixel-area ratios are serialized into the original multi-class distribution-guided text prompts. Both stages operate at 512×512 resolution; Landsat-SCD samples are resized from 416×416 for generator training and sampling, and the generated images and masks are converted back to the native resolution when the downstream dataset is packaged. We retain exactly $5N$ generated pairs for a subset containing N real pairs.

We initialize both stages from Stable Diffusion v1.4 (Rombach et al., 2022), following the released configuration. T2L is optimized with AdamW (Loshchilov & Hutter, 2019) using a learning rate of 5×10^{-6}, a per-device batch size of 1, four-step gradient accumulation (effective batch size 4), a constant learning-rate schedule without warmup, and gradient checkpointing. The token- and pixel-level refinement losses are weighted by 10^{-3} and 5×10^{-5}, respectively. L2I uses a batch size of 1 and a base learning rate of 10^{-6}. We train both stages for 600 epochs on every dataset, and use the resulting generators to produce the ChangeDiff samples in the downstream comparison.

A.3.2 Noise2Change reproduction.

We reproduce Noise2Change using the authors' released code and its real-source-to-synthetic-target generation route. Images and semantic masks from both temporal observations of every selected real pair are treated as independent training samples and encoded at 512×512 resolution with the released preprocessing pipeline. We then train separate EDM2-XS (Karras et al., 2024) mask and image generators with 128 base channels and FP16 computation. Both generators use a learning rate of 10^{-3} and training seed 0. The mask generator uses an effective batch size of 256, a per-forward microbatch of 8, no dropout, and the EDM2 noise distribution $(P_{\mathrm{mean}}, P_{\mathrm{std}}) = (-0.4, 1.0)$. The image generator uses an effective batch size of 128, a microbatch of 16, dropout 0.1, and $(P_{\mathrm{mean}}, P_{\mathrm{std}}) = (-0.8, 1.6)$.

Noise2Change specifies training length by the number of processed images rather than by epochs. The released training schedule processes 50 Mi images for the mask generator and 200 Mi images for the image generator. We convert these official budgets into equivalent dataset passes using the size of the original training set, and then scale the number of processed images in proportion to the number

of single-temporal training samples obtained from both observations of each selected subset. Consequently, every reproduction follows a comparable number of equivalent dataset passes, while the absolute kimg values vary with dataset size. After rounding to valid batch and snapshot boundaries, the mask/image training budgets are approximately 5,222/20,890 kimg on SECOND, 2,454/9,974 kimg on Landsat-SCD, and 3,028/12,272 kimg on HiUCD; these correspond to 20,400/163,200, 9,586/77,920, and 11,828/95,876 optimizer updates, respectively. During generation, we apply the released noise-domain change simulation and label-refinement procedure and use the fixed seeds $\{42, 410, 1101, 1024, 3407\}$. We finally retain five pseudo pairs for each real pair, yielding the same $1:5$ augmentation ratio used by PairTrans and ChangeDiff.

A.4 Downstream SCD Training Protocol

Following ChangeDiff (Zang et al., 2025) and its released SCD implementation, we evaluate the generated supervision using five representative architectures: SSCDL (Ding et al., 2022), BiSRNet (Ding et al., 2022), TED (Ding et al., 2024), A2Net (Li et al., 2023), and SCanNet (Ding et al., 2024). For each benchmark and backbone, all compared methods use identical downstream configurations. Each model is optimized for 50 epochs using AdamW with an initial learning rate of 5×10^{-4}, a weight decay of 10^{-2}, and a polynomial learning-rate schedule with linear warmup over the first 10% of optimization. We use a batch size of 8, automatic mixed-precision training, and gradient clipping with a maximum norm of 0.1.

Each SCD model predicts semantic maps for both temporal observations together with a binary change map. Following the ChangeDiff formulation, the downstream objective combines the semantic classification losses for the two observations, the binary change-detection loss, and the temporal change-consistency loss:

$$\mathcal{L}_{\mathrm{SCD}} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{sem}}^{t_1} + \mathcal{L}_{\mathrm{sem}}^{t_2}\right) + \mathcal{L}_{\mathrm{bin}} + \mathcal{L}_{\mathrm{con}}. \tag{17}$$

We also follow the checkpoint-selection rule of the released ChangeDiff implementation. Let $\mathrm{Score}_e^{\mathrm{val}}$ denote the SCD Score on the validation split at epoch e. The checkpoint used for test-set metric reporting is selected exclusively by validation performance:

$$e^* = \arg\max_{1 \le e \le E} \mathrm{Score}_e^{\mathrm{val}}, \qquad \theta^* = \theta_{e^*}, \tag{18}$$

where $E = 50$ and $\mathrm{Score} = 0.3\,\mathrm{mIoU} + 0.7\,\mathrm{SeK}$. The held-out test split is evaluated only after checkpoint selection.

A.5 Evaluation Metrics

We follow the standard SCD evaluation protocol used in SECOND-style benchmarks (Yang et al., 2022). All metrics are computed on real test images; generated samples are used only as additional training data. Pixels marked as ignored by the dataset annotation, if any, are excluded before constructing the confusion matrix. All reported values in the tables are multiplied by 100.

Let $Q = \{q_{ij}\}_{i,j=0}^{C}$ denote the confusion matrix of the SCD prediction, where q_{ij} is the number of pixels predicted as class i and annotated as class j. Class 0 denotes unchanged pixels, and classes $1, \ldots, C$ denote semantic change types. Let $N = \sum_{i=0}^{C}\sum_{j=0}^{C} q_{ij}$ be the number of evaluated pixels.

A.5.1 Overall semantic-change metrics.

Overall accuracy (OA) measures the fraction of correctly classified pixels over all SCD classes:

$$\mathrm{OA} = \frac{\sum_{i=0}^{C} q_{ii}}{N}. \tag{19}$$

Kappa measures the agreement between prediction and annotation after accounting for chance agreement:

$$\kappa = \frac{\rho - \eta}{1 - \eta}, \quad \rho = \mathrm{OA}, \quad \eta = \frac{\sum_{j=0}^{C} q_{j+}q_{+j}}{N^2}, \tag{20}$$

where q_{j+} and q_{+j} are the row sum and column sum of Q, respectively. Following the SECOND evaluation protocol, we compute the binary non-change IoU and changed-region IoU from the SCD confusion matrix:

$$\mathrm{IoU}_0 = \frac{q_{00}}{\sum_{i=0}^{C} q_{i0} + \sum_{j=0}^{C} q_{0j} - q_{00}}, \tag{21}$$

$$\mathrm{IoU}_c = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij}}{N - q_{00}}. \tag{22}$$

The reported mIoU is the mean of the unchanged-region IoU and changed-region IoU:

$$\mathrm{mIoU} = \frac{1}{2}(\mathrm{IoU}_0 + \mathrm{IoU}_c). \tag{23}$$

Separated Kappa (SeK) reduces the dominance of unchanged pixels by removing the true unchanged count q_{00} from the agreement computation and emphasizing changed-region IoU. Let $\widehat{Q}$ be the matrix obtained by setting q_{00} to zero while keeping the other entries unchanged. We define

$$\widehat{\rho} = \frac{\sum_{i=1}^{C} q_{ii}}{N - q_{00}}, \quad \widehat{\eta} = \frac{\sum_{j=0}^{C} \widehat{q}_{j+}\widehat{q}_{+j}}{(N - q_{00})^2}. \tag{24}$$

Then SeK is computed as

$$\mathrm{SeK} = \exp(\mathrm{IoU}_c - 1) \cdot \frac{\widehat{\rho} - \widehat{\eta}}{1 - \widehat{\eta}}. \tag{25}$$

The overall Score follows the common SCD reporting protocol and combines binary region quality with separated semantic-change agreement:

$$\mathrm{Score} = 0.3\,\mathrm{mIoU} + 0.7\,\mathrm{SeK}. \tag{26}$$

A.5.2 Changed-category F-score.

F_{scd} evaluates whether changed pixels are assigned the correct semantic change type. It is the harmonic mean of SCD precision and SCD recall over changed categories:

$$P_{scd} = \frac{\sum_{i=1}^{C} q_{ii}}{\sum_{i=1}^{C}\sum_{j=0}^{C} q_{ij}}, \quad R_{scd} = \frac{\sum_{i=1}^{C} q_{ii}}{\sum_{i=0}^{C}\sum_{j=1}^{C} q_{ij}}, \tag{27}$$

$$F_{scd} = \frac{2P_{scd}R_{scd}}{P_{scd} + R_{scd}}. \tag{28}$$

Compared with binary change F1, F_{scd} is stricter because a pixel contributes to the numerator only when it is changed and its semantic change type is correctly predicted.

A.5.3 Binary change-region metrics.

For binary change-region evaluation, all semantic change types are merged into a single changed class. We compute

$$\mathrm{TP} = \sum_{i=1}^{C}\sum_{j=1}^{C} q_{ij}, \quad \mathrm{FP} = \sum_{i=1}^{C} q_{i0}, \quad \mathrm{FN} = \sum_{j=1}^{C} q_{0j}. \tag{29}$$

The changed-region IoU, Precision, Recall, and F1 are

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \quad \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{30}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{31}$$

These binary metrics evaluate change localization regardless of the specific semantic transition label, whereas SeK and F_{scd} additionally account for semantic change-type correctness.

Table 6: Approximate stage-wise upstream generator training time under the 5% real-data setting. All runs use one NVIDIA GeForce RTX 3090 GPU with 24 GB memory. Values are reported up to the checkpoint used by the downstream experiment in hours:minutes.

	PairTrans				ChangeDiff		Noise2Change	
Dataset	**Mask 4ch**	**Mask 8ch**	**Image 4ch**	**Image 8ch**	**T2L**	**L2I**	**Mask**	**Image**
SECOND	6:22	8:52	3:23	9:53	29:31	57:01	40:28	104:48
Landsat-SCD	2:30	4:15	1:42	2:16	15:00	34:57	11:41	48:01
HiUCD	3:02	5:13	7:52	3:02	13:13	19:55	14:41	58:08

A.6 Computational Cost

All upstream generators in the three 5% real-data settings are trained on a single NVIDIA GeForce RTX 3090 GPU with 24 GB memory. Table 6 reports their approximate stage-wise training time up to the checkpoint used by the corresponding downstream experiment. Because every run uses one GPU, the reported hours are numerically equivalent to GPU-hours. The accounting includes generator training only; pseudo-data generation and downstream SCD training are excluded.

For SECOND, the Mask 8ch entry corresponds to the 300-epoch run, and the Image 8ch entry is measured up to the 200-epoch, step-29600 checkpoint used to generate the reported synthetic set; its later continuation is excluded. For HiUCD Image 8ch, the table reports the time to the 100-epoch checkpoint used by the downstream experiment.

The total upstream cost is obtained by summing the constituent stages of each method: four mask/image stages for PairTrans, T2L and L2I for ChangeDiff, and the mask and image generators for Noise2Change. For PairTrans, ChangeDiff, and Noise2Change, respectively, the approximate costs are 28:30, 86:32, and 145:16 GPU-hours on SECOND; 10:43, 49:57, and 59:42 GPU-hours on Landsat-SCD; and 19:09, 33:08, and 72:49 GPU-hours on HiUCD.

A.7 Hyperparameter Summary

Tables 7 and 8 summarize the hyperparameters used in PairTrans and downstream SCD training. We group the hyperparameters according to the main stages of the pipeline. The tables report the three 5% settings; the SECOND 20% setting uses the same hyperparameters and differs only in the size of the selected real subset.

B More Results

B.1 More Qualitative Results

We provide additional qualitative examples to complement the main-paper visualization. Each example displays a complete pseudo pair $(\tilde{M}_1, \tilde{I}_1, \tilde{M}_2, \tilde{I}_2)$, and all four entries are generated outputs. The real annotated pair that conditions the two generation directions is not displayed. These examples visualize pair-level semantic changes and the correspondence between each generated mask and its realized image.

Figure 3 covers scenes with different land-cover compositions, object densities, and spatial structures. In the displayed examples, the two generated semantic states contain visible differences in class occupancy and region geometry. Their corresponding images place visual structures within the regions specified by each mask and broadly follow its major boundaries.

Figure 4 complements this comparison by exposing the roles of the two generation factors and the two semantic controls. Panel (a) should be read jointly across all four entries of a pseudo pair: semantic-only and visual-only generation each addresses one side of the supervision problem, whereas the full model jointly introduces a target-state difference and realizes that state as a spatially corresponding image. Panel (b) compares the transition-matrix perturbation and category-constrained attention mechanisms. Perturbation supplies alternative transition proposals, while con-

Table 7: Hyperparameter summary for Stage I–IV. Dataset-specific settings are listed by column; shared settings span all three dataset columns.

Hyperparameter	SECOND	Landsat-SCD	HiUCD
Stage I: Dataset Sampling and Pseudo-Data Construction			
Real-training-data ratio	5%	5%	5%
Selected real pairs	148	71	87
Generation direction	$1 \to 2$, $2 \to 1$	$1 \to 2$, $2 \to 1$	$1 \to 2$, $2 \to 1$
Real-to-synthetic ratio	$1:5$	$1:5$	$1:5$
Random seed for subset sampling	42	42	42
Synthetic-data generation seed	42	42	42
Generator training seed (Mask 4ch; Mask 8ch; Image 4ch; Image 8ch)	42; 42; 3407; 42	42; 42; 3407; 3407	42; 42; 42; 42
Stage II: Real-Transition Matrix Perturbation			
Perturbation strategy	class-swap		
Perturbed source rows	2		
Move proposal range	0.20–0.35		
Row-wise ℓ_1 limit	0.35		
Total change-ratio scale	0.70–1.30		
Total change-ratio range	0.02–0.45		
Minimum new-target rows	1		
Stage III: Target-State Transition Model			
4ch backbone	SD 1.5 LatentDiffusion		
8ch backbone	SD 1.5 RefMaskHybridLDM		
4ch initialization	PLACE COCO-pretrained checkpoint		
8ch initialization	corresponding 4ch checkpoint		
Input resolution	512×512		
4ch conditioning	class-composition CLIP cross-attention		
8ch conditioning	source-mask latent concatenation + category-constrained text attention		
Category-constrained attention scales (positive / negative)	6.0 / 20.0		
Mask constraint and post-processing	hard allowed-class projection + illegal-only repair; fragment removal + boundary smoothing		
Target-mask sampling (sampler / steps / η)	DDIM / 30 / 0		
Target-mask CFG scale	1.0		
Training epochs (4ch; 8ch)	200; 300		
Optimizer updates (4ch; 8ch)	29,600; 44,400	14,200; 21,300	17,400; 26,100
Batch size (4ch; 8ch)	2; 2		
Learning rate (4ch; 8ch)	10^{-6}; 10^{-6}	10^{-6}; 5×10^{-6}	10^{-6}; 5×10^{-6}
Optimizer / warm-up	AdamW / 1,000 steps		
Stage IV: Source-Conditioned Target Image Realization			
4ch backbone	PLACE LDM		
8ch backbone	PLACE editing LDM		
4ch initialization	PLACE COCO-pretrained checkpoint		
8ch initialization	corresponding 4ch checkpoint		
Input resolution	512×512		
4ch conditioning	target-mask hint and class-token PLACE attention		
8ch conditioning	opposite-time image latent concatenation + target-mask PLACE attention		
Target-image sampling (sampler / steps / η)	DDIM / 50 / 0		
Target-image CFG scale	5.0	3.0	2.0 / 5.0
Training epochs to used checkpoint (4ch; 8ch)	100; 200	100; 100	300; 100
Batch size (4ch; 8ch)	2; 2	2; 2	1; 1
Gradient accumulation (4ch; 8ch)	1; 1	1; 1	8; 8
Learning rate (4ch; 8ch)	10^{-6}; 5×10^{-6}	10^{-6}; 5×10^{-6}	10^{-6}; 5×10^{-7}
Optimizer / warm-up	AdamW / 10,000 steps	AdamW / 10,000 steps	AdamW / 500 steps

strained attention limits semantic denoising to the intended target categories. The figure illustrates this division of roles; their average downstream contribution is evaluated quantitatively below.

B.2 More Ablation Results

The main paper reports ablation studies on SECOND with 5% real training data using SCanNet. Tables 9–12 report the same ablations over all downstream SCD backbones. All comparisons use

Table 8: Hyperparameter summary for downstream SCD training, organized by dataset.

Hyperparameter	SECOND	Landsat-SCD	HiUCD
Downstream SCD Training			
Training epochs	50	50	50
Batch size	8	8	8
Crop size	fixed-size 512×512	fixed-size 416×416	fixed-size 512×512
Input resolution	512×512	416×416	512×512
Optimizer	AdamW	AdamW	AdamW
Learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}
Learning-rate schedule	Poly	Poly	Poly
Weight decay	10^{-2}	10^{-2}	10^{-2}
Loss function	$0.5\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{con}}$	$0.5\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{con}}$	$0.5\mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{con}}$
Data augmentation	geometric + photometric + temporal swap	geometric + photometric + temporal swap	geometric + photometric + temporal swap

Table 9: Full component-level ablation on SECOND with 5% real training data over all downstream SCD backbones. Img. denotes source-image-conditioned target realization; Perturb. denotes real-transition matrix perturbation; Attn. denotes category-constrained mask attention.

Backbone	Img.	Perturb.	Attn.	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	×	×	×	85.29	28.68	66.51	12.46	53.09	56.63	47.39	64.30	59.62	69.78
SSCDL	×	✓	✓	**86.21**	30.24	67.82	14.14	54.93	58.74	49.08	65.84	59.67	73.44
SSCDL	✓	×	×	85.94	29.12	66.89	12.94	53.19	57.17	47.44	64.35	57.12	**73.67**
SSCDL	✓	✓	×	83.60	28.45	66.20	12.27	52.02	56.50	48.26	65.10	**67.08**	63.23
SSCDL	✓	×	✓	84.43	28.96	66.71	12.78	53.23	57.18	48.56	65.37	65.12	65.63
SSCDL	✓	✓	✓	86.18	**30.78**	**68.43**	**14.65**	**55.46**	**59.77**	**50.25**	**66.89**	61.99	72.62
BiSRNet	×	×	×	84.97	28.71	66.53	12.50	53.45	56.75	47.79	64.67	61.82	67.79
BiSRNet	×	✓	✓	85.44	30.31	68.03	14.15	55.08	59.24	50.19	66.83	64.88	68.90
BiSRNet	✓	×	×	85.45	30.21	67.97	14.03	54.23	59.10	49.93	66.60	63.77	69.71
BiSRNet	✓	✓	×	85.65	29.86	67.38	13.78	54.45	58.08	48.77	65.56	61.27	70.50
BiSRNet	✓	×	✓	85.68	29.63	67.30	13.49	53.90	57.91	48.50	65.32	60.29	**71.27**
BiSRNet	✓	✓	✓	**85.80**	**31.28**	**68.67**	**15.26**	**56.32**	**60.27**	**51.16**	**67.69**	**65.87**	69.61
TED	×	×	×	85.20	28.72	66.56	12.51	53.30	56.74	47.59	64.49	60.44	69.12
TED	×	✓	✓	**86.47**	30.81	68.23	14.77	**55.79**	59.39	49.67	66.37	60.07	**74.14**
TED	✓	×	×	85.10	29.23	67.51	12.83	52.57	58.34	49.18	65.93	62.77	69.43
TED	✓	✓	×	85.77	30.21	67.83	14.09	54.71	58.81	49.46	66.19	62.01	70.97
TED	✓	×	✓	85.69	30.40	68.07	14.25	54.86	59.24	49.99	66.65	**63.30**	70.38
TED	✓	✓	✓	86.22	**30.93**	**68.53**	**14.81**	55.30	**59.92**	**50.35**	**66.97**	61.82	73.07
A2Net	×	×	×	84.68	30.09	67.85	13.91	54.40	59.11	50.47	67.08	**68.14**	66.06
A2Net	×	✓	✓	**86.47**	30.79	68.17	14.77	55.54	59.30	49.54	66.25	59.69	**74.43**
A2Net	✓	×	×	85.78	29.68	67.62	13.42	53.66	58.43	48.92	65.70	60.35	72.10
A2Net	✓	✓	×	84.93	30.11	67.71	14.00	54.39	58.80	50.00	66.66	66.27	67.07
A2Net	✓	×	✓	85.41	30.35	68.09	14.18	54.83	59.34	50.28	66.91	64.95	69.00
A2Net	✓	✓	✓	85.73	**31.32**	**68.70**	**15.30**	**55.86**	**60.32**	**51.19**	**67.72**	65.85	69.70
SCanNet	×	×	×	85.28	29.57	67.04	13.51	54.51	57.58	48.54	65.35	62.51	68.47
SCanNet	×	✓	✓	85.81	31.22	68.33	15.31	56.35	59.72	50.58	67.18	64.89	69.63
SCanNet	✓	×	×	86.09	31.10	68.18	15.21	56.20	59.38	50.02	66.68	62.66	71.26
SCanNet	✓	✓	×	85.79	31.43	68.46	15.56	56.60	59.95	50.87	67.44	**65.70**	69.27
SCanNet	✓	×	✓	85.65	30.99	68.04	15.10	56.27	59.27	50.23	66.87	65.01	68.84
SCanNet	✓	✓	✓	**86.45**	**32.02**	**68.82**	**16.25**	**57.38**	**60.43**	**50.99**	**67.54**	63.50	**72.13**

the same selected real subset, downstream training schedule, and evaluation script. The generated-data amount is also fixed within each ablation, except in Table 11, where it is the variable under study.

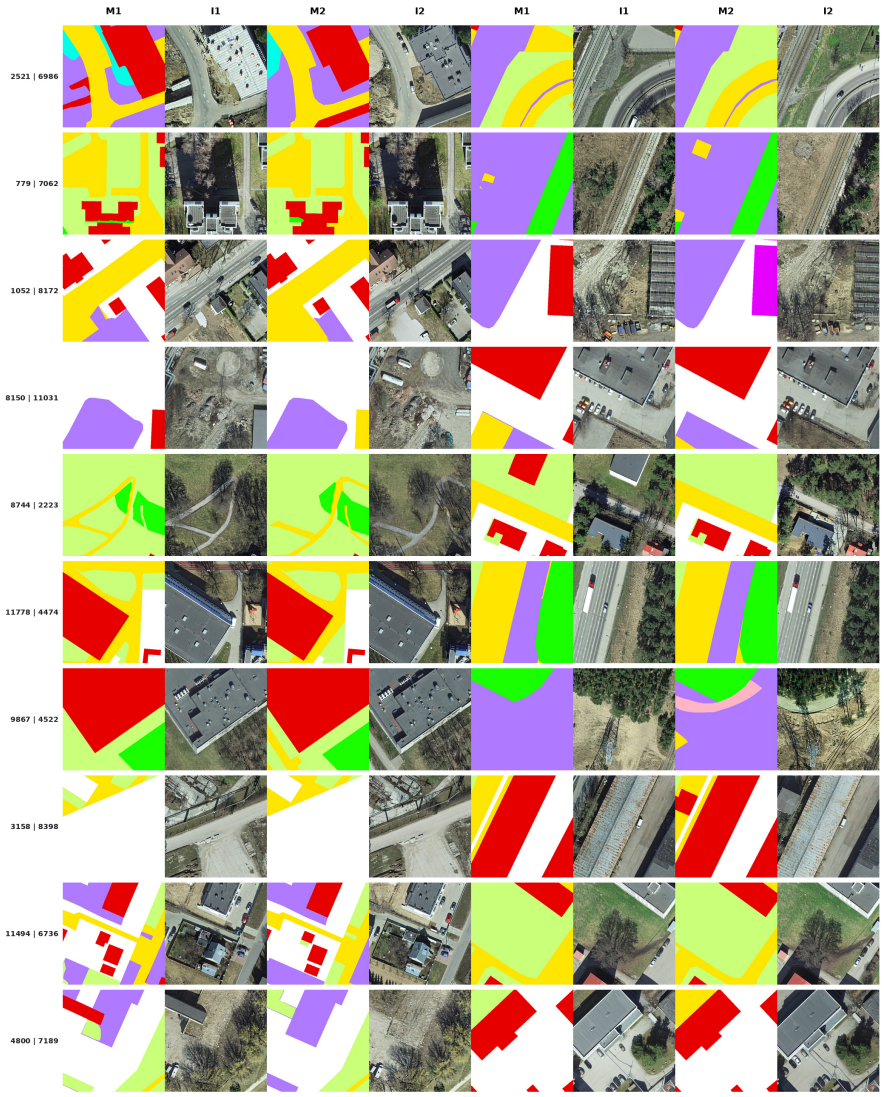

Figure 3: Additional complete pseudo pairs generated by PairTrans on HiUCD with 5% real training data. Each block shows $(\tilde{M}_1, \tilde{I}_1, \tilde{M}_2, \tilde{I}_2)$, all of which are generated.

Joint effect of the core components. Table 9 shows that the full configuration consistently leads Score, mIoU, and SeK, while intermediate configurations show that component benefits depend on the accompanying controls. For SSCDL, adding transition perturbation to image conditioning without category-constrained attention increases Recall while lowering Score. Enabling both semantic controls gives a stronger joint result. Source-image conditioning also adds value when the semantic controls are already enabled, while image conditioning alone cannot match the complete configuration. Thus, the controls affect the balance of downstream errors and their benefits depend on the combination used. Some partial variants achieve higher OA, Recall, or Precision.

Direction-symmetric generation. Table 10 shows that selecting a single generation direction does not recover the benefit of the complete construction. The reverse unidirectional variant has higher Score than the forward variant for every backbone, but both remain below real-only training. The complete construction exceeds real-only training in Score, mIoU, and SeK across architectures. With synthetic-data volume and downstream schedules matched, these results support the complete pair construction.

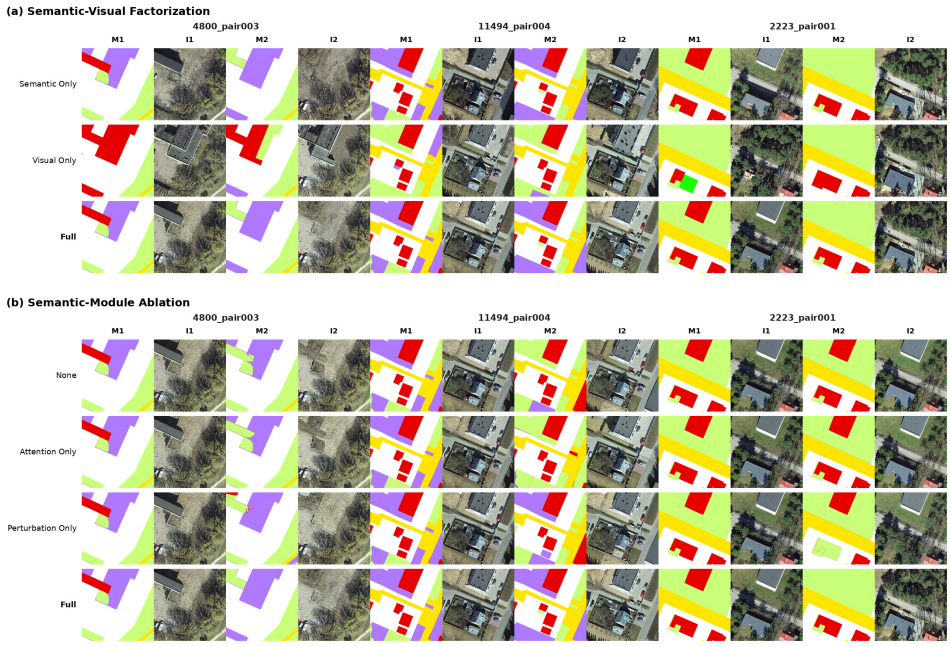

Figure 4: Qualitative ablations on HiUCD with 5% real training data. Panel (a) compares semantic-only generation, visual-only generation, and their full factorization. Panel (b) isolates transition-matrix perturbation and category-constrained attention in the semantic factor. Each group shows $(\tilde{M}_1, \tilde{I}_1, \tilde{M}_2, \tilde{I}_2)$.

Table 10: Full ablation on direction-symmetric generation on SECOND with 5% real training data over all downstream SCD backbones.

Backbone	Variant	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	Real only	85.29	28.68	66.51	12.46	53.09	56.63	47.39	64.30	59.62	69.78
SSCDL	Uni. $1 \to 2$	83.51	23.45	62.50	6.71	46.69	49.63	41.12	58.27	52.66	65.23
SSCDL	Uni. $2 \to 1$	84.58	24.62	62.67	8.31	48.74	49.80	40.72	57.87	49.76	69.14
SSCDL	Bidirectional	**86.18**	**30.78**	**68.43**	**14.65**	**55.46**	**59.77**	**50.25**	**66.89**	**61.99**	**72.62**
BiSRNet	Real only	84.97	28.71	66.53	12.50	53.45	56.75	47.79	64.67	61.82	67.79
BiSRNet	Uni. $1 \to 2$	83.42	23.91	62.32	7.44	46.77	49.32	40.90	58.05	52.67	64.67
BiSRNet	Uni. $2 \to 1$	83.87	24.36	62.91	7.84	48.00	50.34	41.68	58.83	53.02	66.07
BiSRNet	Bidirectional	**85.80**	**31.28**	**68.67**	**15.26**	**56.32**	**60.27**	**51.16**	**67.69**	**65.87**	**69.61**
TED	Real only	85.20	28.72	66.56	12.51	53.30	56.74	47.59	64.49	60.44	69.12
TED	Uni. $1 \to 2$	82.54	21.50	61.22	4.48	42.18	47.29	39.10	56.22	50.30	63.72
TED	Uni. $2 \to 1$	84.48	24.87	63.03	8.51	48.41	50.47	41.39	58.54	50.86	68.97
TED	Bidirectional	**86.22**	**30.93**	**68.53**	**14.81**	**55.30**	**59.92**	**50.35**	**66.97**	**61.82**	**73.07**
A2Net	Real only	84.68	30.09	67.85	13.91	54.40	59.11	50.47	67.08	**68.14**	66.06
A2Net	Uni. $1 \to 2$	85.17	28.06	66.41	11.63	51.27	56.40	47.02	63.96	58.28	70.87
A2Net	Uni. $2 \to 1$	**86.10**	28.91	66.55	12.78	52.77	56.56	46.63	63.61	55.05	**75.32**
A2Net	Bidirectional	85.73	**31.32**	**68.70**	**15.30**	**55.86**	**60.32**	**51.19**	**67.72**	65.85	69.70
SCanNet	Real only	85.28	29.57	67.04	13.51	54.51	57.58	48.54	65.35	62.51	68.47
SCanNet	Uni. $1 \to 2$	85.44	26.67	64.14	10.61	50.99	52.40	42.80	59.94	51.03	**72.63**
SCanNet	Uni. $2 \to 1$	84.91	27.87	65.49	11.74	52.30	54.90	45.85	62.88	58.27	68.27
SCanNet	Bidirectional	**86.45**	**32.02**	**68.82**	**16.25**	**57.38**	**60.43**	**50.99**	**67.54**	**63.50**	72.13

Sensitivity to the generated-data amount. Table 11 shows that even 1:1 improves Score over real-only training across backbones, while increasing generated data further does not produce a metric-independent ordering. The 1:5 setting favors Score for SSCDL, BiSRNet, and SCanNet; 1:10 favors TED and A2Net. A2Net illustrates an additional distinction: the larger ratio has higher Score,

Table 11: Full ablation on the real-to-synthetic ratio on SECOND with 5% real training data over all downstream SCD backbones.

Backbone	Ratio	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	Real only	85.29	28.68	66.51	12.46	53.09	56.63	47.39	64.30	59.62	69.78
SSCDL	1:1	85.20	29.41	67.36	13.15	53.82	58.11	49.04	65.81	63.11	68.75
SSCDL	1:5	**86.18**	**30.78**	**68.43**	14.65	55.46	**59.77**	**50.25**	**66.89**	61.99	**72.62**
SSCDL	1:10	85.86	30.73	68.18	**14.67**	**55.73**	59.42	50.16	66.81	**63.53**	70.45
BiSRNet	Real only	84.97	28.71	66.53	12.50	53.45	56.75	47.79	64.67	61.82	67.79
BiSRNet	1:1	85.54	29.53	67.37	13.31	54.15	58.06	48.79	65.59	61.53	70.22
BiSRNet	1:5	85.80	**31.28**	**68.67**	**15.26**	**56.32**	**60.27**	**51.16**	**67.69**	**65.87**	69.61
BiSRNet	1:10	**86.14**	31.25	68.65	15.22	55.82	60.14	50.71	67.30	63.18	**71.99**
TED	Real only	85.20	28.72	66.56	12.51	53.30	56.74	47.59	64.49	60.44	69.12
TED	1:1	85.45	29.44	67.24	13.24	53.77	57.86	48.60	65.41	61.33	70.08
TED	1:5	**86.22**	30.93	68.53	14.81	**55.30**	59.92	50.35	66.97	61.82	**73.07**
TED	1:10	85.85	**31.00**	**68.62**	**14.87**	54.77	**60.12**	**50.77**	**67.34**	**63.73**	71.39
A2Net	Real only	84.68	30.09	67.85	13.91	54.40	59.11	50.47	67.08	**68.14**	66.06
A2Net	1:1	85.13	30.68	68.43	14.50	54.50	59.95	51.05	67.60	67.19	68.00
A2Net	1:5	85.73	31.32	68.70	15.30	55.86	60.32	**51.19**	**67.72**	65.85	69.70
A2Net	1:10	**86.63**	**31.57**	**68.89**	**15.57**	**56.12**	**60.46**	50.71	67.30	61.19	**74.75**
SCanNet	Real only	85.28	29.57	67.04	13.51	54.51	57.58	48.54	65.35	62.51	68.47
SCanNet	1:1	85.31	30.48	67.57	14.58	55.54	58.53	49.61	66.32	**64.93**	67.77
SCanNet	1:5	86.45	**32.02**	**68.82**	**16.25**	**57.38**	**60.43**	**50.99**	**67.54**	63.50	72.13
SCanNet	1:10	**86.65**	31.66	68.46	15.88	56.87	59.77	50.08	66.73	60.74	**74.03**

Table 12: Full ablation on real-pair grounding on SECOND with 5% real training data over all downstream SCD backbones. Mismatch-ref breaks the correspondence between the source-side observation and the target-side transition structure by using a mismatched reference.

Backbone	Variant	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	Real only	85.29	28.68	66.51	12.46	53.09	56.63	47.39	64.30	59.62	69.78
SSCDL	Mismatch-ref	86.02	30.33	67.96	14.20	54.96	59.00	49.50	66.22	61.19	72.16
SSCDL	Ours	**86.18**	**30.78**	**68.43**	**14.65**	**55.46**	**59.77**	**50.25**	**66.89**	**61.99**	**72.62**
BiSRNet	Real only	84.97	28.71	66.53	12.50	53.45	56.75	47.79	64.67	61.82	67.79
BiSRNet	Mismatch-ref	**85.96**	30.22	67.84	14.09	54.75	58.80	49.31	66.05	60.96	**72.07**
BiSRNet	Ours	85.80	**31.28**	**68.67**	**15.26**	**56.32**	**60.27**	**51.16**	**67.69**	**65.87**	69.61
TED	Real only	85.20	28.72	66.56	12.51	53.30	56.74	47.59	64.49	60.44	69.12
TED	Mismatch-ref	85.90	30.62	68.25	14.49	54.87	59.50	50.10	66.75	**62.53**	71.59
TED	Ours	**86.22**	**30.93**	**68.53**	**14.81**	**55.30**	**59.92**	**50.35**	**66.97**	61.82	**73.07**
A2Net	Real only	84.68	30.09	67.85	13.91	54.40	59.11	50.47	67.08	**68.14**	66.06
A2Net	Mismatch-ref	**85.87**	31.09	68.53	15.05	55.64	60.00	50.74	67.32	64.31	**70.62**
A2Net	Ours	85.73	**31.32**	**68.70**	**15.30**	**55.86**	**60.32**	**51.19**	**67.72**	65.85	69.70
SCanNet	Real only	85.28	29.57	67.04	13.51	54.51	57.58	48.54	65.35	62.51	68.47
SCanNet	Mismatch-ref	86.13	31.26	68.32	15.38	56.22	59.62	50.23	66.87	62.80	71.51
SCanNet	Ours	**86.45**	**32.02**	**68.82**	**16.25**	**57.38**	**60.43**	**50.99**	**67.54**	**63.50**	**72.13**

but binary F1 peaks at 1:5. For SCanNet, the larger ratio increases Precision while reducing Recall and F1. The preferred generation budget therefore depends on both the backbone and evaluation metric. These are comparisons under a fixed-epoch protocol; sample volume and optimization steps vary together, limiting attribution to data composition alone.

Contribution of real-pair grounding. Table 12 shows that reference correspondence changes the utility of a fixed amount of synthetic data. Matched conditioning improves Score, mIoU, SeK, F_{scd}, and F1 for all five backbones, although Mismatch-ref already outperforms real-only training in Score. The matched-versus-mismatched comparison evaluates the contribution of source–transition correspondence at a fixed synthetic-data volume. Metric trade-offs remain: for BiSRNet and A2Net,

Table 13: Full augmentation results on Landsat with 5% real training data. All metrics are reported on real test images.

Backbone	Training data	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	Real only	86.32	33.15	69.80	17.45	62.01	62.35	53.69	69.86	75.86	64.75
SSCDL	+ ChangeDiff	87.34	33.71	70.37	18.00	63.15	63.00	53.85	70.01	71.70	68.39
SSCDL	+ Noise2Change	83.97	32.62	68.82	17.11	59.58	61.59	53.88	70.03	**86.20**	58.97
SSCDL	+ Ours	**88.96**	**37.35**	**72.91**	**22.11**	**66.94**	**66.77**	**57.31**	**72.86**	72.25	**73.49**
BiSRNet	Real only	86.48	33.01	69.66	17.30	62.18	62.06	53.31	69.54	74.45	65.24
BiSRNet	+ ChangeDiff	86.84	33.30	69.96	17.59	62.60	62.46	53.54	69.74	73.34	66.48
BiSRNet	+ Noise2Change	83.03	31.39	68.08	15.66	57.45	60.60	53.09	69.36	**86.92**	57.70
BiSRNet	+ Ours	**88.96**	**37.68**	**72.87**	**22.59**	**67.56**	**66.73**	**57.35**	**72.89**	73.03	**72.76**
TED	Real only	88.31	37.08	72.68	21.83	66.14	66.55	57.43	72.96	75.75	70.37
TED	+ ChangeDiff	88.31	37.39	73.02	22.12	66.32	67.08	58.05	73.46	77.07	70.16
TED	+ Noise2Change	85.63	35.19	71.16	19.78	62.38	64.90	56.78	72.44	**86.31**	62.41
TED	+ Ours	**89.95**	**40.68**	**75.01**	**25.97**	**69.96**	**69.87**	**60.53**	**75.41**	74.99	**75.83**
A2Net	Real only	87.95	35.74	71.53	20.40	65.55	64.80	55.69	71.54	74.08	69.16
A2Net	+ ChangeDiff	88.09	36.02	72.04	20.58	65.39	65.56	56.40	72.12	74.34	70.03
A2Net	+ Noise2Change	86.70	37.37	72.84	22.17	64.51	67.28	59.02	74.23	**87.11**	64.66
A2Net	+ Ours	**89.51**	**39.56**	**74.01**	**24.79**	**69.32**	**68.43**	**59.12**	**74.31**	74.77	**73.86**
SCanNet	Real only	88.72	39.24	73.64	24.50	68.40	68.04	59.12	74.31	79.06	70.09
SCanNet	+ ChangeDiff	89.12	39.09	73.70	24.25	68.75	68.04	58.88	74.12	76.32	72.05
SCanNet	+ Noise2Change	85.95	36.55	71.20	21.71	64.51	65.03	57.00	72.61	**87.69**	61.95
SCanNet	+ Ours	**90.40**	**42.80**	**76.02**	**28.57**	**72.16**	**71.35**	**62.25**	**76.73**	77.90	**75.60**

Table 14: Full augmentation results on HiUCD with 5% real training data. All metrics are reported on real test images.

Backbone	Training data	OA	Score	mIoU	SeK	F_{scd}	Kappa	IoU	F1	Recall	Precision
SSCDL	Real only	**98.44**	**14.77**	49.22	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00
SSCDL	+ ChangeDiff	89.31	14.49	49.49	-0.51	9.61	14.73	9.32	17.06	**67.43**	9.76
SSCDL	+ Noise2Change	92.36	13.42	51.36	-2.83	9.50	16.08	10.04	18.24	51.82	11.07
SSCDL	+ Ours	95.72	12.21	**54.55**	-5.94	**12.85**	**21.51**	**13.14**	**23.23**	38.89	**16.56**
BiSRNet	Real only	**98.44**	**14.76**	49.22	**-0.00**	0.00	-0.00	0.00	0.00	0.00	0.96
BiSRNet	+ ChangeDiff	91.16	14.09	50.46	-1.49	9.59	15.00	9.44	17.25	56.40	10.18
BiSRNet	+ Noise2Change	90.08	13.80	49.75	-1.61	8.12	14.26	9.03	16.57	**60.06**	9.61
BiSRNet	+ Ours	95.21	12.08	**53.46**	-5.66	**10.96**	**18.73**	**11.47**	**20.57**	37.50	**14.18**
TED	Real only	95.62	13.19	55.26	-4.84	14.46	23.83	14.63	25.53	44.82	17.85
TED	+ ChangeDiff	87.76	13.93	48.12	-0.72	7.04	12.51	8.07	14.93	**65.80**	8.42
TED	+ Noise2Change	**98.43**	**14.52**	49.29	**-0.38**	0.20	0.28	0.16	0.31	0.16	14.53
TED	+ Ours	96.00	13.01	**55.90**	-5.37	**14.69**	**25.24**	**15.51**	**26.85**	43.41	**19.44**
A2Net	Real only	93.08	15.37	53.56	-1.00	13.77	22.06	13.67	24.05	**65.75**	14.71
A2Net	+ ChangeDiff	93.40	**15.59**	53.77	**-0.78**	14.08	22.28	13.79	24.24	63.45	14.98
A2Net	+ Noise2Change	93.81	14.49	53.51	-2.24	13.34	20.91	12.91	22.86	55.36	14.41
A2Net	+ Ours	**96.11**	14.54	**57.19**	-3.73	**19.22**	**28.96**	**18.00**	**30.51**	50.64	**21.83**
SCanNet	Real only	95.08	13.01	54.91	-4.94	11.61	23.35	14.37	25.13	48.52	16.95
SCanNet	+ ChangeDiff	91.91	13.50	50.73	-2.46	7.85	14.62	9.19	16.83	**49.75**	10.13
SCanNet	+ Noise2Change	**98.44**	**14.77**	49.22	**-0.00**	0.00	0.00	0.00	0.00	0.00	**38.88**
SCanNet	+ Ours	97.06	11.57	**57.81**	-8.24	**13.65**	**29.59**	**18.26**	**30.88**	37.42	26.29

matched conditioning sacrifices some Precision while increasing Recall and F1. Correct pairing improves joint semantic and change-region prediction quality.

B.3 Full Main Results on Landsat and HiUCD

Tables 13 and 14 report all ten evaluation metrics for the Landsat and HiUCD comparisons summarized in Table 3. All values are measured on real test images with 5% real training data.

C Limitations and Broader Impact

PairTrans targets SCD augmentation from a limited set of annotated real pairs. Its reliance on real-pair anchors is therefore both a scope choice and a source of grounding: the anchors provide scene context and observed transition structure, while requiring an initial labeled set and dataset-specific generator adaptation. PairTrans is consequently most relevant to the limited-supervision regime studied in this work, where a small amount of reliable paired annotation is available but insufficient for training high-capacity SCD models on its own.

Generation Characteristics. The semantic and visual factors implement learned conditional generation rather than hard pixel-level constraints. This choice leaves room to construct meaningful variation around an anchor instead of simply copying an observed target. Exact target-class occupancy and unchanged-region preservation are therefore encouraged through conditioning and candidate selection rather than imposed as deterministic equalities. PairTrans supports these properties through feasible transition proposals, category-constrained attention, allowed-class repair, change-ratio filtering, and anchor-compatible candidate matching. The component and grounding ablations further show that these conditions improve the downstream value of the resulting supervision. Fine-grained correspondence nevertheless remains governed by the learned generator, and explicit local consistency objectives offer a useful direction for further refining the balance between diversity and correspondence.

Evaluation Scope. Our downstream comparisons use a fixed-epoch protocol. Equal epoch counts do not ensure equal optimization steps when training-set sizes differ, so comparisons with real-only training do not isolate augmentation effects from training-budget differences. The reported results are point estimates; variability across independently sampled low-data subsets and training seeds remains unquantified.

Efficiency and Generalization. The staged adaptation of the semantic and visual generators requires a one-time upstream cost during pseudo-data construction, after which the same generated supervision can be reused across different downstream SCD architectures. Our evaluation across three datasets with distinct spatial resolutions and change distributions, together with five downstream models, provides evidence across varied settings. Extending this evaluation to additional sensors, geographic regions, and land-cover taxonomies is a natural direction for future work. The architecture-dependent ratio results further indicate that the amount of pseudo supervision is not universally transferable and should be validated for a new deployment domain.

Broader Impact. PairTrans outputs are intended as additional training supervision, not as observational evidence of real-world land-cover change. As with other generative augmentation methods, generated samples may reflect biases in the annotated anchors or artifacts of the generative model. Applications involving operational monitoring or consequential decisions should therefore retain evaluation on real held-out data and appropriate domain-expert oversight. Used within this scope, PairTrans can reduce the effective annotation burden while preserving the role of real observations in model validation.