FM-CHANGENET++: LEARNED NONLINEAR FEATURE TRANSPORT FOR CHANGE DETECTION

Anonymous authors
Paper under double-blind review

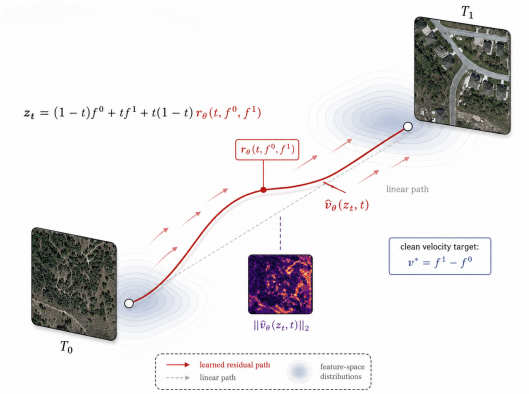

Figure 1: FM-ChangeNet++. The model learns an endpoint-preserving residual feature transport path between the pre-temporal representation f^0 and the post-temporal representation f^1. Instead of following only the linear path $z_t = (1-t)f^0 + tf^1$, the proposed formulation adds a residual correction $t(1-t)r_\theta(t, f^0, f^1)$, allowing a smooth nonlinear transition while preserving the endpoints. The velocity field $\hat{v}_\theta(z_t, t)$ is supervised using the instantaneous path derivative $v^\star(t) = \frac{\partial z_t}{\partial t}$, ensuring that it represents the local tangent direction of the learned nonlinear trajectory. Its magnitude $\|\hat{v}_\theta(z_t, t)\|_2$ therefore provides a localized measure of transport activity that supports change localization.

ABSTRACT

Most remote sensing change detection methods infer change by directly comparing features extracted from bi-temporal images. Recent flow-matching approaches reinterpret this problem as learning feature transport between pre- and post-temporal representations through a fixed linear interpolation path. While effective, this assumption constrains semantic transitions to follow straight trajectories in feature space, limiting the representation of complex structural changes. We introduce **FM-ChangeNet++**, a feature transport framework that learns the transformation trajectory itself. Instead of enforcing a fixed linear path, we model feature evolution using an endpoint-preserving residual transport formulation, where a lightweight residual predictor learns smooth nonlinear deviations while guaranteeing that the trajectory starts from the pre-temporal representation and terminates at the post-temporal representation. The velocity field is supervised using the instantaneous derivative of the learned nonlinear path, ensuring that the predicted velocity represents the local tangent direction and transport intensity along the feature trajectory. Its magnitude therefore provides a localized measure of feature evolution that can support semantic change detection. The proposed framework jointly optimizes multi-scale flow estimation, residual path regularization, spatial smoothness, and segmentation supervision within a unified coarse-to-fine architecture. This formulation enables the model to capture complex feature evolution while maintaining endpoint consistency and controlled transport dynamics. Extensive experiments on standard remote sensing benchmarks demonstrate consistent improvements over linear transport formulations and state-of-the-art CD methods. Ablation studies further show that learned residual transport produces

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.

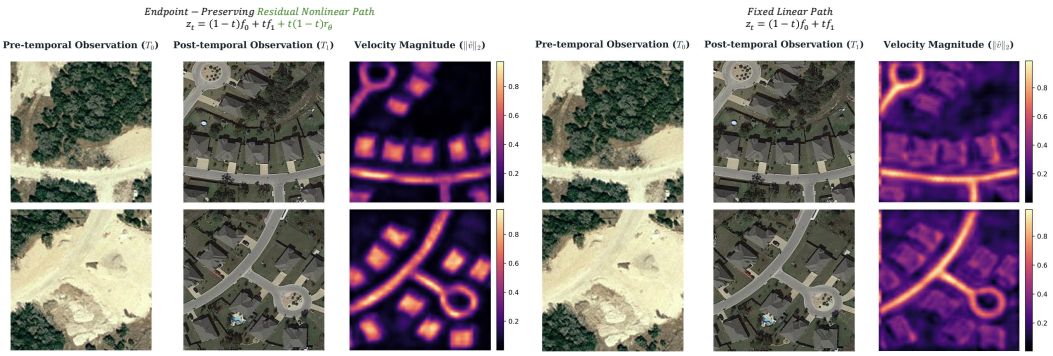

Figure 2: Qualitative comparison between the proposed endpoint-preserving residual nonlinear path and the fixed linear transport path. Given pre- and post-temporal observations T_0 and T_1, the proposed formulation models intermediate feature states as $z_t = (1-t)f^0 + tf^1 + t(1-t)r_\theta$, while the linear baseline uses $z_t = (1-t)f^0 + tf^1$. The velocity magnitude $\|\hat{v}\|_2$ highlights the induced localized transport-activity map, where sharper and more localized responses indicate cleaner transport-based change evidence.

more discriminative feature trajectories, more localized velocity representations, and more accurate detection of subtle semantic changes.

1 Introduction

Remote sensing change detection (CD) aims to identify semantic changes between two observations of the same scene acquired at different times. This capability is central to urban monitoring, disaster assessment, environmental surveillance, and land-use analysis Chen & Shi (2020); You et al. (2020); Chen et al. (2025). Despite substantial progress, high-resolution CD remains challenging because visual differences between bi-temporal images do not necessarily correspond to semantic change. Illumination variation, seasonal effects, registration errors, viewpoint changes, and resolution mismatch may produce large pixel-level discrepancies, while genuine structural changes can be small, sparse, and spatially subtle. Most modern methods address this problem by extracting bi-temporal features and comparing them through concatenation, subtraction, attention, or feature fusion. Siamese CNNs established this endpoint-comparison paradigm Daudt et al. (2018); Zhang et al. (2020); Fang et al. (2021), while Transformer- and Mamba-based architectures improve long-range interaction and multi-scale aggregation Chen et al. (2021); Bandara & Patel (2022); Chen et al. (2024a); Zhang et al. (2025). However, these approaches still largely treat the two observations as static endpoints. They determine whether the before and after representations differ, but provide limited supervision on how one representation should connect to the other. Consequently, the intermediate representation structure remains weakly constrained, allowing a model to fit the final change mask without necessarily learning a coherent representation of the transformation separating the two observations. Recent generative formulations suggest a different perspective. Diffusion-based CD methods exploit generative representations to improve robustness Bandara et al. (2025); Tang & Chen (2024), while flow matching (FM) learns a deterministic velocity field that transports one representation toward another Lipman et al. (2022); Liu et al. (2022c); Tong et al. (2023). For CD, this motivates replacing a purely static endpoint comparison with a feature-space transport view. A linear transport path between pre-temporal and post-temporal features provides a simple canonical velocity target, whose magnitude can act as an interpretable cue for localized feature change Kazoom et al. (2026). Unchanged regions are expected to require limited feature displacement, whereas changed regions should induce stronger transport activity. Yet a fixed linear path imposes a restrictive geometric assumption: all intermediate representations are constrained to lie on the straight segment connecting the two endpoints. Complex structural differences, appearance variation, and spatial misalignment may instead require feature-dependent nonlinear intermediate states before a reliable change representation emerges. Importantly, this path is a constructed feature-space supervision mechanism between two observed representations; it is *not* intended to reconstruct the unobserved physical evolution of the scene between acquisition times. The objective is therefore not to model

how a building, vegetation region, or other structure physically evolved over time, but to learn a more expressive representation-level bridge between the observed endpoints. This leads to the central question of this work: **Can CD learn nonlinear feature transport while preserving the clean and interpretable velocity signal that makes linear FM useful?**

We answer this question with **FM-ChangeNet++**, a residual nonlinear feature transport framework for remote sensing CD. Rather than discarding the linear bridge, we augment it with a learned endpoint-preserving residual term. For each scale s, given pre-temporal and post-temporal features f_s^0 and f_s^1, we define $z_t^{(s)} = (1-t)f_s^0 + tf_s^1 + t(1-t)r_\theta^{(s)}(t, f_s^0, f_s^1)$, where $r_\theta^{(s)}$ is a lightweight residual predictor. The factor $t(1-t)$ forces the residual correction to vanish at $t = 0$ and $t = 1$, guaranteeing that the trajectory begins exactly at f_s^0 and terminates exactly at f_s^1. The model can therefore bend the intermediate feature path while preserving both observed endpoints. Because the resulting trajectory is nonlinear, a constant endpoint-displacement target is no longer locally consistent with the path. We therefore introduce derivative-consistent velocity supervision and define the target as the instantaneous tangent, $v_s^\star(t) = \frac{\partial z_{t,s}}{\partial t} = f_s^1 - f_s^0 + (1-2t)r_\theta^{(s)}(t, f_s^0, f_s^1) + t(1-t)\frac{\partial r_\theta^{(s)}(t, f_s^0, f_s^1)}{\partial t}$. This formulation aligns the predicted velocity with the local direction and transport intensity of the constructed nonlinear feature trajectory, rather than supervising it with an endpoint displacement that ignores the residual path geometry. Its magnitude therefore provides a spatially localized measure of transport activity along the learned representation path, with stronger responses indicating locations requiring larger feature-space transformation. To keep this additional flexibility controlled, the residual trajectory is regularized in magnitude and smoothness with respect to the interpolation parameter t, suppressing excessive curvature and oscillatory behavior while preserving the ability to represent complex feature transformations. FM-ChangeNet++ jointly optimizes multi-scale derivative-consistent velocity prediction, segmentation supervision, spatial smoothness, and residual path regularization. The final prediction head combines the finest-scale velocity field, its magnitude, and the corresponding decoder features to produce the dense change map. This yields a unified framework in which nonlinear feature transport, localized transport-based change evidence, and dense prediction are learned jointly. Figure 2 qualitatively illustrates the central effect of the proposed formulation: compared with the fixed linear bridge, the endpoint-preserving residual path produces sharper and more spatially localized velocity-magnitude responses. **Our contributions are:**

We introduce **FM-ChangeNet++**, a residual nonlinear feature-transport framework for CD. The method extends linear FM with an endpoint-preserving learned residual path and supervises the velocity field using the instantaneous path derivative $\partial z_t/\partial t$. This ensures that the predicted velocity represents the local tangent direction and transport intensity of the learned nonlinear trajectory.

We provide a **theoretical analysis** showing endpoint preservation, containment of linear FM as a special case, exact recovery of the endpoint displacement through velocity integration, bounded trajectory and velocity deviations, and stronger representation-level constraints than endpoint-only mask supervision.

Extensive experiments and ablations showing that **FM-ChangeNet++** improves over CNN-based, Transformer-based, Mamba-based, diffusion-based, and linear flow-matching baselines, while producing more localized velocity representations and more accurate change masks.

2 Related Work

Remote sensing CD. Remote sensing CD aims to localize semantic changes between observations acquired at different times. Early deep learning methods mainly follow a Siamese design, where two image branches extract pre- and post-temporal features that are later compared or fused for dense prediction Daudt et al. (2018); Zhang et al. (2020); Fang et al. (2021). This endpoint-comparison paradigm has been further strengthened by attention, multi-scale aggregation, and feature interaction modules Feng et al. (2022; 2023); Li et al. (2022); Song et al. (2022). Transformer-based methods improve long-range context modeling and global correspondence reasoning Chen et al. (2021); Bandara & Patel (2022); Zhang et al. (2022), while recent Mamba-based models introduce efficient sequence modeling for high-resolution CD Chen et al. (2024a); Zhang et al. (2025). Despite these advances, most methods still infer change from static comparisons between two endpoints. They

improve how features are fused, but they do not explicitly model the transformation path from the pre-temporal representation to the post-temporal representation.

Robust and cross-resolution CD. A major challenge in remote sensing CD is that visual differences are not always semantic changes. Illumination shifts, viewpoint variation, seasonal effects, misregistration, and resolution gaps can all produce misleading differences. Cross-resolution CD methods address this problem by explicitly modeling resolution and alignment discrepancies Zheng et al. (2021); Liu et al. (2022b); Tian et al. (2022); Chen et al. (2023); Wang et al. (2025). These methods show that reliable CD requires more than direct differencing. They are usually designed around alignment, reconstruction, or resolution compensation, rather than learning a continuous feature-level transition between temporal states. Ours treats change as a transport process in representation space, allowing the model to reason about how one temporal state evolves into another.

Generative representations for CD. Generative models have recently been explored for CD because they can provide richer representations and improved robustness under distribution shift. GAN-based and reconstruction-based approaches use image synthesis or reconstruction errors to expose changes Wu et al. (2023); Noh et al. (2022); Xu et al. (2023). Diffusion-based methods use denoising or latent generative features for more stable change representations Bandara et al. (2025); Tang & Chen (2024); Madani et al. (2026); Madani & Patel (2026). These works suggest that modeling structure beyond raw endpoint differences is useful. However, diffusion-based approaches often rely on stochastic sampling or feature reconstruction, and they do not directly define an interpretable velocity field between the two temporal representations. Our work instead uses a deterministic transport view: the model predicts feature motion, and the magnitude of this motion directly serves as a change cue.

Flow matching and feature transport. FM provides a powerful framework for learning continuous transformations through a time-dependent velocity field Lipman et al. (2022). Related formulations such as rectified flow and optimal-transport FM further show that simple paths can produce stable and efficient transport learning Liu et al. (2022c); Tong et al. (2023). In generative modeling, the goal is usually to transport samples from one distribution to another. For CD, the objective is different: rather than generating the post-temporal image, the model must identify where the pre-temporal representation undergoes substantial evolution to explain the post-temporal representation. This makes FM attractive as a dense supervision mechanism, since unchanged regions should exhibit limited transport activity, whereas changed regions should induce stronger feature evolution. The closest transport-based formulation uses the linear path $z_t = (1-t)f^0 + tf^1$, whose constant velocity is $v^\star = f^1 - f^0$. Although this provides stable supervision, it constrains all temporal transitions to follow straight trajectories in feature space. FM-ChangeNet++ relaxes this assumption through the endpoint-preserving residual path $z_t = (1-t)f^0 + tf^1 + t(1-t)r_\theta(t, f^0, f^1)$, where the residual learns smooth nonlinear corrections while vanishing at both endpoints. Because the resulting path is nonlinear, we supervise the velocity field using its instantaneous derivative $v^\star(t) = \partial z_t/\partial t$. The predicted velocity therefore represents the local tangent direction and transport intensity of the learned trajectory, while its integral over time remains exactly equal to the endpoint displacement $f^1 - f^0$. Our method is thus not merely another fusion module or nonlinear interpolation strategy, but a controlled residual transport formulation that combines flexible nonlinear feature evolution with derivative-consistent flow supervision.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given a bi-temporal remote sensing image pair $X^0, X^1 \in \mathbb{R}^{H\times W\times 3}$, where X^0 and X^1 denote the pre-temporal and post-temporal observations, respectively, the objective is to predict a dense change mask $Y \in \{0, 1, \ldots, C-1\}^{H\times W}$. Rather than treating change detection solely as a comparison between two static endpoints, we formulate it as feature-space transport along a continuous trajectory connecting their learned representations. Let E denote a shared remote sensing encoder. For each observation, the encoder extracts multi-scale features, $\{f_s^0\}_{s=1}^S = E(X^0)$ and $\{f_s^1\}_{s=1}^S = E(X^1)$, where s indexes the feature scale and $f_s^0, f_s^1 \in \mathbb{R}^{C_s\times H_s\times W_s}$. In our implementation, four feature scales are extracted from intermediate ViT layers and projected to $C_s \in \{64, 128, 256, 512\}$. Our

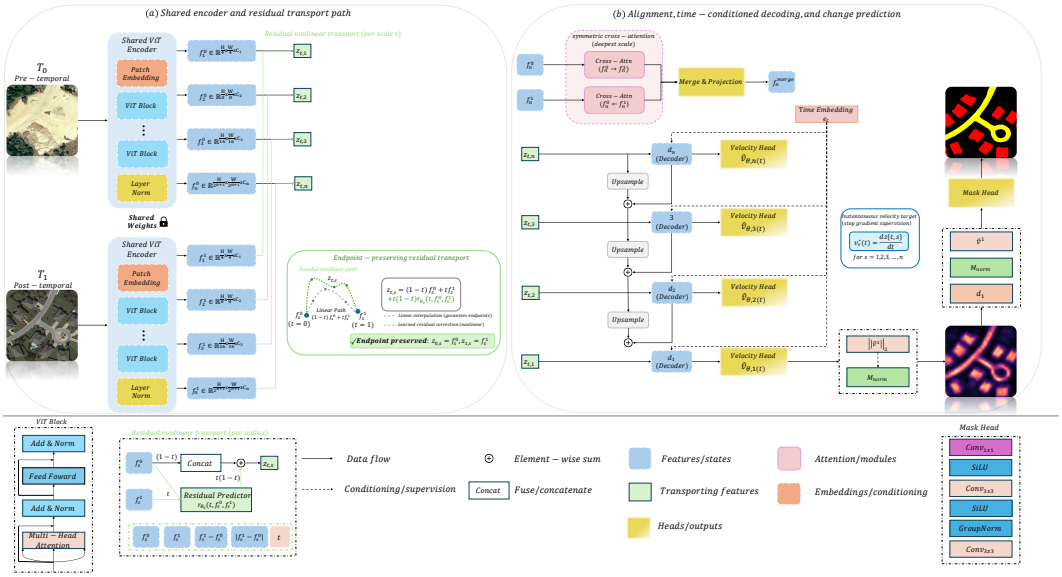

Figure 3: Overview of **FM-ChangeNet++**. A shared ViT encoder extracts multi-scale features from T_0 and T_1. The model constructs the endpoint-preserving nonlinear path $z_{t,s} = (1-t)f_s^0 + tf_s^1 + t(1-t)r_{\theta_s}$, and predicts velocity fields supervised by $v_s^\star(t) = \frac{\partial z_{t,s}}{\partial t}$. The finest-scale velocity magnitude is combined with $\hat{v}_1$ and d_1 to predict the final change map.

objective is to construct a continuous, endpoint-preserving trajectory between the two observed feature representations and to supervise the corresponding transport field along this path. Importantly, this trajectory is a representation-level supervision construct and is not intended to reconstruct the unobserved physical evolution of the scene between acquisition times. Its instantaneous derivative defines the feature-transport velocity, providing a localized measure of how strongly each representation must change along the constructed path. Regions associated with semantic change are therefore expected to exhibit stronger transport activity, whereas unchanged regions should remain comparatively stable. Figure 3 presents the overall FM-ChangeNet++ framework.

3.2 Linear Feature Transport

A simple endpoint-preserving trajectory between the two feature states is given by linear interpolation:

$$z_{t,s}^{\text{lin}} = (1-t)f_s^0 + tf_s^1, \qquad t \in [0,1]. \tag{1}$$

This trajectory satisfies

$$z_{0,s}^{\text{lin}} = f_s^0, \qquad z_{1,s}^{\text{lin}} = f_s^1. \tag{2}$$

Its instantaneous velocity is constant along the trajectory: $v_s^{\text{lin}} = \frac{\partial z_{t,s}^{\text{lin}}}{\partial t} = f_s^1 - f_s^0$. This velocity provides a simple measure of endpoint feature displacement, with larger magnitudes indicating stronger feature variation between the two observations. The linear formulation constrains every temporal transition to follow a straight trajectory in feature space, which may be insufficient for complex semantic changes, spatial misalignment, or nonlinear structural evolution.

3.3 Residual Nonlinear Feature Transport

FM-ChangeNet++ relaxes the straight-path assumption by adding a learned residual correction to the feature state:

$$z_{t,s} = (1-t)f_s^0 + tf_s^1 + t(1-t)r_{\phi,s}(t, f_s^0, f_s^1). \tag{3}$$

Here, $r_{\phi,s}$ is a lightweight, time-dependent residual predictor. The factor $t(1-t)$ satisfies

$$t(1-t) = 0 \qquad \text{for} \qquad t \in \{0,1\}, \tag{4}$$

and therefore guarantees

$$z_{0,s} = f_s^0, \qquad z_{1,s} = f_s^1. \tag{5}$$

The residual can thus bend the intermediate feature trajectory without altering its observed temporal endpoints. The residual predictor receives local temporal evidence: $r_{\phi,s} = r_{\phi,s}\left(\left[f_s^0, f_s^1, f_s^1 - f_s^0, \left|f_s^1 - f_s^0\right|, T_s(t)\right]\right)$, where $T_s(t)$ denotes a spatial time map broadcast to the resolution of scale s. This conditioning allows the residual trajectory to depend on the pre-temporal feature, post-temporal feature, signed displacement, displacement magnitude, and temporal position. Because $r_{\phi,s}$ is time-dependent, the instantaneous velocity of the nonlinear trajectory is

$$\begin{aligned}\frac{\partial z_{t,s}}{\partial t} &= f_s^1 - f_s^0 + (1-2t) r_{\phi,s}(t, f_s^0, f_s^1) \\ &\quad + t(1-t)\frac{\partial}{\partial t} r_{\phi,s}(t, f_s^0, f_s^1).\end{aligned} \tag{6}$$

3.4 Derivative-Consistent Velocity Supervision

Because the path in Eq. 3 is nonlinear, we supervise the velocity field using its instantaneous derivative:

$$\begin{aligned}v_s^\star(t) = \frac{\partial z_{t,s}}{\partial t} &= f_s^1 - f_s^0 + (1-2t) r_{\phi,s}(t, f_s^0, f_s^1) \\ &\quad + t(1-t)\frac{\partial}{\partial t} r_{\phi,s}(t, f_s^0, f_s^1).\end{aligned} \tag{7}$$

This derivative-consistent target ensures that the predicted velocity represents the local tangent direction and transport intensity of the learned nonlinear trajectory. Consequently, its magnitude measures instantaneous feature evolution, with stronger responses expected in regions undergoing substantial semantic change. At each scale, the model predicts $\hat{v}_s(z_{t,s}, t) = H_s(d_s)$, where d_s is the time-conditioned decoder feature and H_s is a scale-specific velocity head.

3.5 Time-Conditioned Multi-Scale Decoder

The decoder estimates nonlinear feature transport in a coarse-to-fine manner. At the deepest scale, the residual path feature $z_{t,S}$ initializes the decoder: $d_S = D_S(z_{t,S}, e_t)$. Finer scales are progressively decoded using the corresponding residual path features:

$$d_s = D_s\left(\mathrm{Up}(d_{s+1}), z_{t,s}, e_t\right), \qquad s = S-1, \ldots, 1, \tag{8}$$

where e_t is a learned time embedding and $\mathrm{Up}(\cdot)$ denotes spatial upsampling. At each scale, a velocity head predicts the instantaneous transport field $\hat{v}_s(z_{t,s}, t) = H_s(d_s)$. The finest-scale velocity field $\hat{v}_1$ is then used to construct the localized change evidence.

3.6 Velocity-Based Change Representation

From the finest-scale velocity field, we compute the velocity magnitude $M = \|\hat{v}_1(z_{t,1}, t)\|_2$. Because $\hat{v}_1$ approximates the instantaneous derivative of the learned nonlinear trajectory, M measures the local rate of feature evolution along the transport path. Large values indicate regions undergoing strong feature transport, whereas small values indicate locally stable representations. We therefore use M as a spatially localized change-evidence map. The final segmentation head combines the decoder feature, directional velocity, and velocity magnitude: $\hat{Y}_{\mathrm{logits}} = G\left([d_1, \hat{v}_1, M]\right)$. The predicted change mask is obtained as $\hat{Y} = \arg\max_c \hat{Y}^{(c)}_{\mathrm{logits}}$.

Training Objective The full objective combines derivative-consistent flow supervision, segmentation supervision, spatial regularization, and residual path regularization: $\mathcal{L} = \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{tv}}\mathcal{L}_{\mathrm{tv}} + \lambda_{\mathrm{path}}\mathcal{L}_{\mathrm{path}}$.

Flow loss. The flow loss supervises the predicted velocity using the instantaneous derivative of the learned nonlinear trajectory: $\mathcal{L}_{\mathrm{flow}} = \sum_{s=1}^{S} w_s \left\|\hat{v}_s(z_{t,s}, t) - \mathrm{sg}[v_s^\star(t)]\right\|_2^2$, where $v_s^\star(t) = \frac{\partial z_{t,s}}{\partial t}$ and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. Explicitly, $v_s^\star(t) = f_s^1 - f_s^0 + (1-2t) r_{\phi,s}(t, f_s^0, f_s^1) +$

$t(1-t)\frac{\partial}{\partial t} r_{\phi,s}(t, f_s^0, f_s^1)$. This objective ensures that the predicted field follows the local tangent direction of the learned nonlinear feature trajectory. The stop-gradient operator prevents gradients from propagating through the derivative target itself, while gradients may still propagate through the trajectory state $z_{t,s}$.

Mask loss. The segmentation loss combines cross-entropy and Dice loss: $\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}$. Cross-entropy provides pixel-wise class supervision, while Dice loss improves overlap for small and imbalanced changed regions.

Total variation loss. To encourage spatially coherent velocity fields, we apply total variation regularization: $\mathcal{L}_{\text{tv}} = \|\nabla_h \hat{v}_1\|_1 + \|\nabla_w \hat{v}_1\|_1$. This suppresses spatially noisy transport responses while preserving coherent change boundaries.

Residual path regularization. The residual trajectory is regularized to remain controlled and temporally smooth: $\mathcal{L}_{\text{path}} = \lambda_{\text{res}}\mathcal{L}_{\text{res}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}$. The residual magnitude penalty is $\mathcal{L}_{\text{res}} = \sum_{s=1}^{S} \left\| r_{\phi,s}(t, f_s^0, f_s^1) \right\|_2^2$. This prevents the residual component from overwhelming the linear reference trajectory. The temporal smoothness penalty is defined using a second-order finite difference: $\mathcal{L}_{\text{smooth}} = \sum_{s=1}^{S} \|z_{t+\epsilon,s} - 2z_{t,s} + z_{t-\epsilon,s}\|_2^2$. During training, t is sampled from $[\epsilon, 1-\epsilon]$ when this term is evaluated, ensuring that $t-\epsilon$ and $t+\epsilon$ remain within the valid transport interval. This regularization encourages smooth path curvature and suppresses oscillatory feature trajectories. During inference, we use the deterministic midpoint $t = 0.5$, where the residual factor $t(1-t)$ is maximal. Given a test pair (X^0, X^1), the model extracts multi-scale features, constructs the endpoint-preserving nonlinear trajectory state, predicts its instantaneous velocity field, computes the corresponding velocity magnitude, and outputs the final dense change mask. No iterative sampling, numerical integration, or generative reconstruction is required. The complete training procedure is summarized in Algorithm 1.

Algorithm 1 Training FM-ChangeNet++

Require: Bi-temporal image pair (X^0, X^1) and change mask Y
1: Sample $t \sim \mathcal{U}(0,1)$
2: $\{f_s^0\}, \{f_s^1\} \leftarrow E(X^0), E(X^1)$
3: **for** each feature scale s **do**
4: $\quad r_s \leftarrow R_\phi\big(f_s^0, f_s^1, f_s^1 - f_s^0, |f_s^1 - f_s^0|, t\big)$
5: $\quad z_{t,s} \leftarrow (1-t)f_s^0 + tf_s^1 + t(1-t)r_s$
6: $\quad \dot{r}_s \leftarrow \frac{\partial r_s}{\partial t}$
7: $\quad v_s^\star(t) \leftarrow f_s^1 - f_s^0 + (1-2t)r_s + t(1-t)\dot{r}_s$
8: **end for**
9: $\{d_s\} \leftarrow D_\theta(\{z_{t,s}\}, t)$
10: $\hat{v}_s \leftarrow H_\theta(d_s)$
11: $M \leftarrow \|\hat{v}_1\|_2$
12: $\hat{Y} \leftarrow G_\theta(d_1, \hat{v}_1, M)$
13: $\mathcal{L}_{\text{flow}} \leftarrow \sum_s w_s \|\hat{v}_s - \text{sg}[v_s^\star(t)]\|_2^2$
14: Optimize $\mathcal{L} = \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}}$

3.7 Analytical Properties of Residual Nonlinear Transport

We characterize several analytical properties of FM-ChangeNet++ that motivate the proposed transport construction. For clarity, the analysis is given at a single feature scale; the same arguments apply independently across scales. Let $f^0, f^1 \in \mathbb{R}^{C\times H\times W}$ denote the encoded pre-temporal and post-temporal feature maps, and define $\Delta f = f^1 - f^0$. We use ϕ for the residual-path predictor and θ for the velocity predictor.

The proposed endpoint-preserving trajectory is

$$\begin{aligned} z_t^\psi &= (1-t)f^0 + tf^1 + q_\phi(t), \\ q_\phi(t) &= t(1-t)r_\phi(t, f^0, f^1), \qquad t \in [0,1]. \end{aligned} \tag{9}$$

Assuming that r_ϕ is differentiable with respect to t, its instantaneous velocity is

$$v_\psi^\star(t) = \frac{\partial z_t^\psi}{\partial t} = \Delta f + (1-2t) r_\phi(t, f^0, f^1) + t(1-t)\frac{\partial r_\phi(t, f^0, f^1)}{\partial t}. \tag{10}$$

During flow supervision, $v_\psi^\star(t)$ is treated as a stop-gradient target. Thus, gradients do not propagate through the derivative target itself, while the trajectory state z_t^ψ remains part of the differentiable prediction pathway.

Proposition 1: Endpoint and transport consistency. For any residual predictor r_ϕ, the trajectory in Eq. 9 preserves the observed endpoints,

$$z_0^\phi = f^0, \qquad z_1^\phi = f^1, \tag{11}$$

and its instantaneous velocity satisfies

$$\int_0^1 v_\psi^\star(t)\, dt = f^1 - f^0. \tag{12}$$

Therefore, the residual term can modify the intermediate feature trajectory without altering either endpoint, while the accumulated transport remains exactly consistent with the observed feature displacement.

Proof sketch. Because $t(1-t) = 0$ at $t = 0$ and $t = 1$, the residual term vanishes at both endpoints. Moreover, by the fundamental theorem of calculus,

$$\int_0^1 v_\psi^\star(t)\, dt = \int_0^1 \frac{\partial z_t^\psi}{\partial t}\, dt = z_1^\phi - z_0^\phi = f^1 - f^0. \tag{13}$$

□

Relation to linear flow matching. The proposed trajectory contains linear feature transport as a special case. If $r_\phi(t, f^0, f^1) = 0$ for all t, then

$$z_t^\psi = (1-t) f^0 + t f^1, \qquad v_\psi^\star(t) = f^1 - f^0. \tag{14}$$

FM-ChangeNet++ therefore retains the canonical linear bridge while allowing feature-dependent nonlinear deviations between the same observed endpoints.

These properties clarify the role of the residual formulation: it increases the flexibility of the intermediate representation without sacrificing exact endpoint or accumulated-transport consistency. Pathwise velocity supervision additionally constrains predictions at sampled intermediate states beyond the final segmentation objective, providing a structured representation-level training signal rather than relying only on endpoint correctness. Additional analytical properties, including bounded deviation from linear transport, transport-energy decomposition, and temporal smoothness, together with full proofs, are provided in the Appendix.

4 RESULTS

We evaluate the proposed framework on three standard remote sensing CD benchmarks: LEVIR-CD Chen & Shi (2020), WHU-CD Ji et al. (2018), and DSIFN-CD Zhang et al. (2020). LEVIR-CD Chen & Shi (2020) focuses on small, densely distributed building changes in high-resolution aerial imagery, WHU-CD Ji et al. (2018) contains larger urban scenes with stronger viewpoint and illumination variation, and DSIFN-CD Zhang et al. (2020) is more diverse and challenging, with richer clutter and larger intra-class variation. We follow the official splits, and apply flipping, rotation, and cropping during training. We report Precision, Recall, F1, IoU, and overall accuracy (OA), with F1 and IoU as the primary metrics due to the strong foreground-background imbalance typical of CD. Table 1 compares our FM framework against representative CNN-based, Transformer-based, and Mamba-based methods on LEVIR-CD, WHU-CD, and DSIFN-CD. Our method achieves the best F1, IoU, and OA on LEVIR-CD, and the best Precision, Recall, and OA on WHU-CD.

Table 1: Quantitative comparison with SOTA methods on LEVIR-CD, WHU-CD, and DSIFN-CD. Values are reported as percentages (%). "–" indicates that the corresponding metric was not reported.

Type	Models	LEVIR-CD					WHU-CD					DSIFN-CD				
		Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
CNN-based	FC-EF Daudt et al. (2018)	86.91	80.79	83.74	72.02	98.39	86.63	81.63	84.05	72.50	98.33	72.61	52.73	61.09	43.98	88.59
	FC-Siam-Diff Daudt et al. (2018)	89.53	82.87	86.07	75.55	98.67	85.43	86.31	85.87	75.24	98.51	59.67	65.71	62.54	45.50	86.63
	FC-Siam-Conc Daudt et al. (2018)	91.14	83.75	87.29	77.45	98.77	86.41	85.34	85.87	75.24	98.51	66.45	54.21	59.71	42.56	87.57
	IFNet Zhang et al. (2020)	94.02	82.93	88.13	78.77	98.87	95.34	89.70	92.43	85.93	99.21	67.86	53.94	60.10	42.96	87.83
	SNUNet Fang et al. (2021)	89.42	87.17	88.28	79.02	98.81	94.13	91.22	92.65	86.31	99.22	60.60	72.89	66.18	49.45	87.34
Transformer-based	SwinUnet Zhang et al. (2022)	91.24	87.10	89.12	80.37	98.89	92.81	93.42	93.11	87.12	99.27	–	–	–	–	–
	BIT Chen et al. (2021)	89.24	89.37	89.31	80.68	98.92	93.30	92.65	92.97	86.87	99.25	68.36	70.18	69.26	52.97	89.41
	ChangeFormer Bandara & Patel (2022)	92.42	88.64	90.49	82.63	99.04	92.49	94.69	93.58	87.93	99.30	88.48	84.94	86.67	76.48	95.56
	MSCANet Zheng et al. (2025)	92.45	88.85	90.62	82.84	99.05	95.14	92.90	94.01	88.69	99.37	–	–	–	–	–
	Paformer Liu et al. (2022a)	92.51	89.33	90.89	83.31	99.08	94.55	93.91	94.23	89.09	99.39	–	–	–	–	–
	DARNet Li et al. (2022)	92.34	89.51	90.90	83.32	99.08	94.67	94.12	94.39	89.38	99.41	–	–	–	–	–
	ACABFNet Song et al. (2022)	92.54	89.47	90.98	83.45	99.09	94.61	94.03	94.32	89.24	99.40	–	–	–	–	–
	NeXt2Former-CD Wang et al. (2026)	–	–	92.10	85.40	99.20	–	–	95.50	91.40	99.65	–	–	–	–	–
Mamba	RS-Mamba Chen et al. (2024b)	91.56	90.64	91.10	83.65	99.10	94.94	94.61	94.77	90.07	99.44	–	–	–	–	–
	ChangeMamba Chen et al. (2024a)	92.33	91.22	91.77	84.79	99.17	95.02	94.98	95.00	90.48	99.46	–	–	–	–	–
	CDMamba Zhang et al. (2025)	92.41	92.05	92.23	85.58	99.21	95.31	95.17	95.24	90.91	99.49	–	–	–	–	–
Diffusion-based	DDPM-CD Bandara et al. (2025)	–	–	90.90	83.30	99.10	–	–	92.70	86.30	99.40	–	–	–	–	–
FM	FM-ChangeNet Kazoom et al. (2026)	92.70	92.20	92.45	85.96	99.24	95.39	95.49	95.44	91.28	99.66	90.02	87.93	88.96	80.12	95.61
	FM-ChangeNet++	**92.79**	**92.31**	**92.55**	**86.13**	**99.25**	**95.47**	**95.63**	**95.55**	**91.48**	**99.66**	**91.12**	**89.01**	**90.05**	**81.91**	**95.67**

Building on the preceding matched-capacity comparison, which establishes the benefit of path-wise transport over endpoint-only alternatives, we next isolate the contribution specific to FM-ChangeNet++: nonlinear residual transport with derivative-consistent supervision. Table 2 shows a consistent pattern across all three datasets. Introducing the residual path improves over linear flow matching, particularly on DSIFN-CD. Using the exact path derivative without stop-gradient weakens performance, whereas stop-gradient derivative supervision achieves the strongest results on every benchmark. This indicates that FM-ChangeNet++ benefits from combining flexible non-linear transport with a stable, trajectory-consistent velocity target.

Table 2: **Ablation of nonlinear transport and velocity supervision.** All variants use the same encoder, cross-temporal attention, multi-scale decoder, mask supervision, TV regularization, and prediction head. The comparison isolates the residual nonlinear path, velocity target, and stop-gradient operation. "N/A" denotes a component that is not applicable.

Dataset	Variant	Path	Velocity Target	Stop-Grad.	Flow	Mask	Path Reg.	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	Linear path with canonical target	Linear	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	N/A	92.70	92.20	92.45	85.96	99.24
	Residual path with canonical target	Residual	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	✓	92.74	92.26	92.50	86.04	99.24
	Residual path with derivative target, no stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✗	✓	✓	✓	92.65	92.11	92.38	85.84	99.23
	FM-ChangeNet++: derivative target with stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✓	✓	✓	✓	**92.79**	**92.31**	**92.55**	**86.13**	**99.25**
WHU-CD	Linear path with canonical target	Linear	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	N/A	95.39	95.49	95.44	91.28	99.66
	Residual path with canonical target	Residual	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	✓	95.42	95.58	95.50	91.38	99.66
	Residual path with derivative target, no stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✗	✓	✓	✓	95.30	95.40	95.35	91.12	99.65
	FM-ChangeNet++: derivative target with stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✓	✓	✓	✓	**95.47**	**95.63**	**95.55**	**91.48**	**99.66**
DSIFN-CD	Linear path with canonical target	Linear	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	N/A	90.02	87.93	88.96	80.12	95.61
	Residual path with canonical target	Residual	$f_1^{(s)} - f_0^{(s)}$	N/A	✓	✓	✓	90.85	88.78	89.80	81.45	95.64
	Residual path with derivative target, no stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✗	✓	✓	✓	89.75	87.60	88.66	79.63	95.55
	FM-ChangeNet++: derivative target with stop-gradient	Residual	$\partial z_t^{(s)}/\partial t$	✓	✓	✓	✓	**91.12**	**89.01**	**90.05**	**81.91**	**95.67**

5 Conclusion and Future Work

We presented **FM-ChangeNet++**, a residual nonlinear feature-transport framework for CD. The method extends linear FM with an endpoint-preserving residual path between f^0 and f^1 and derivative-consistent velocity supervision using $\partial z_t/\partial t$. This formulation captures local transport direction and intensity while preserving exact endpoint consistency. FM-ChangeNet++ combines residual path learning, time-conditioned multi-scale decoding, velocity supervision, and segmentation loss within a unified framework. Experiments on LEVIR-CD, WHU-CD, and DSIFN-CD show consistent improvements over strong CNN-, Transformer-, Mamba-, diffusion-, and linear FM baselines. Matched-capacity ablations further confirm that the gains arise from the proposed residual transport formulation rather than increased model capacity. **Future work** will extend FM-ChangeNet++ to multi-class open-vocabulary, and cross-resolution CD. We also plan to investigate vision-language guidance, adaptive residual path constraints, and uncertainty-aware change prediction for large-scale monitoring.

AI USE STATEMENT

In this work, generative AI tools were used for text polishing and assistance with code writing. All AI-assisted text was reviewed and edited by the authors, and all generated or modified code was manually inspected and tested for correctness. The authors take full responsibility for the final content, implementation, results, and claims presented in this work.

ETHICS STATEMENT

This work focuses on remote sensing change detection using standard benchmark datasets and does not involve human subjects or personally identifiable information. While improved change-detection systems can support applications such as environmental monitoring, disaster assessment, and urban analysis, they may also be misused for unauthorized surveillance or large-scale monitoring. In addition, predictions may be unreliable under domain shift, sensor differences, severe misregistration, or uncommon environmental conditions. FM-ChangeNet++ is therefore intended as a research method for benchmarked remote-sensing analysis rather than an autonomous decision-making system. Responsible deployment should include appropriate human oversight, robustness evaluation, and consideration of privacy, legal, and governance requirements.

REPRODUCIBILITY STATEMENT

We support reproducibility by providing a complete description of the FM-ChangeNet++ architecture, training objective, and optimization procedure in the main paper, including Algorithm 1. The Appendix reports the training hyperparameters, loss weights, temporal sampling strategy, random-seed evaluation, ablation settings, and additional implementation details. Experiments use the official splits of LEVIR-CD, WHU-CD, and DSIFN-CD with the stated evaluation metrics and data augmentations. The Appendix also provides the analytical derivations and supporting proofs for the proposed residual transport formulation.

REFERENCES

Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 207–210. IEEE, 2022.

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

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

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

Hao Chen, Haotian Zhang, Keyan Chen, Chenyao Zhou, Song Chen, Zhengxia Zou, and Zhenwei Shi. Continuous cross-resolution remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2023. doi: 10.1109/TGRS.2023.3325829.

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

Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024b.

Yongqi Chen, Shou Feng, Chunhui Zhao, Yingjie Tang, Junjie Tang, Di Dong, and Nan Su. Deep learning for high-resolution remote sensing change detection: Task taxonomy, datasets, and perspectives. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

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

Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

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

Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multi-level intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. doi: 10.1109/TGRS.2023.3241257. URL https://doi.org/10.1109/TGRS.2023.3241257.

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

Roie Kazoom, George Leifman, and Genady Beryozkin. Fm-changenet: Learning change through pathwise feature transport. *arXiv preprint arXiv:2607.04750*, 2026.

Ziming Li, Chenxi Yan, Ying Sun, and Qinchuan Xin. A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Mengxi Liu, Qian Shi, Zhuoqun Chai, and Jianlong Li. Pa-former: Learning prior-aware transformer for remote sensing building change detection. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2022a.

Mengxi Liu, Qian Shi, Andrea Marinoni, Da He, Xiaoping Liu, and Liangpei Zhang. Super-resolution-based change detection network with stacked attention module for images with different resolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 2022b. doi: 10.1109/TGRS.2021.3091758. URL https://doi.org/10.1109/TGRS.2021.3091758.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022c.

Seyedehanita Madani and Vishal M Patel. Morphing through time: Diffusion-based bridging of temporal gaps for robust alignment in change detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 431–439, 2026.

Seyedehanita Madani, Rama Chellappa, and Vishal M Patel. Diffregcd: Integrated registration and change detection with diffusion features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7503–7512, 2026.

Hyeoncheol Noh, Jingi Ju, Minseok Seo, Jongchan Park, and Dong-Geol Choi. Unsupervised change detection based on image reconstruction loss. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1352–1361, 2022.

Ruizhe Shao, Chun Du, Hao Chen, and Jun Li. Sunet: Change detection for heterogeneous remote sensing images from satellite and uav using a dual-channel fully convolution network. *Remote Sensing*, 13(18):3750, 2021.

Lei Song, Min Xia, Liguo Weng, Haifeng Lin, Ming Qian, and Binyu Chen. Axial cross attention meets cnn: Bibranch fusion network for change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:21–32, 2022.

Kai Tang and Jin Chen. Changeanywhere: Sample generation for remote sensing change detection via semantic latent diffusion model. *arXiv preprint arXiv:2404.08892*, 2024.

Juan Tian, Daifeng Peng, Haiyan Guan, and Haiyong Ding. Racdnet: Resolution-and alignment-aware change detection network for optical remote sensing imagery. *Remote Sensing*, 14(18): 4527, 2022.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Guoqing Wang, He Chen, Tingting Qiao, Jue Wang, and Wenchao Liu. Resolution-difference embedded network for cross-resolution remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Yufan Wang, Sokratis Makrogiannis, and Chandra Kambhamettu. Next2former-cd: Efficient remote sensing change detection with modern vision architectures. *arXiv preprint arXiv:2602.18717*, 2026.

Chen Wu, Bo Du, and Liangpei Zhang. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9774–9788, 2023.

Qingsong Xu, Yilei Shi, Jianhua Guo, Chaojun Ouyang, and Xiao Xiang Zhu. Ucdformer: Unsupervised change detection using a transformer-driven image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023.

Yanan You, Jingyi Cao, and Wenli Zhou. A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. *Remote Sensing*, 12(15):2460, 2020.

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

Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

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

Xiangtao Zheng, Xiumei Chen, Xiaoqiang Lu, and Bangyong Sun. Unsupervised change detection by cross-resolution difference learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. doi: 10.1109/TGRS.2021.3079907.

Xingjian Zheng, Xin Lin, Linbo Qing, and Xianfeng Ou. Semantic-aware remote sensing change detection with multi-scale cross-attention. *Sensors*, 25(9):2813, 2025.

A APPENDIX

A.1 ISOLATING THE EFFECT OF PATHWISE TRANSPORT

Table 3 evaluates the effect of pathwise transport under matched model capacity. All variants use the same backbone, alignment module, decoder capacity, and prediction head whenever applicable, while differing only in how temporal information is represented or supervised. FM-ChangeNet++ consistently outperforms static differencing, endpoint feature regression, direct fused-feature prediction, removing time conditioning, and final-mask-only supervision across all three datasets. These results show that the gains cannot be explained by increased model capacity alone and support the benefit of learning from sampled intermediate states z_t with explicit time conditioning, rather than relying only on static endpoint representations.

Table 3: **Matched-capacity ablation.** All variants use the same backbone, alignment module, decoder capacity, and prediction head.

Dataset	Method	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	FM-ChangeNet	92.70	92.20	92.45	85.96	99.24
	FM-ChangeNet++	**92.79**	**92.31**	**92.55**	**86.13**	**99.25**
	Static difference, same encoder-decoder	84.94	83.69	84.31	72.88	98.37
	No time conditioning, same decoder	85.62	84.52	85.07	74.02	98.46
	Endpoint feature regression, no pathwise z_t, t	76.88	75.36	76.11	61.44	97.21
	Direct prediction from fused features	86.71	85.66	86.18	75.72	98.58
	Final-mask supervision only	88.04	87.22	87.63	77.98	98.73
WHU-CD	FM-ChangeNet	95.39	95.49	95.44	91.28	99.66
	FM-ChangeNet++	**95.47**	**95.63**	**95.55**	**91.48**	**99.66**
	Static difference, same encoder-decoder	88.46	87.37	87.91	78.43	98.71
	No time conditioning, same decoder	89.18	88.10	88.64	79.60	98.79
	Endpoint feature regression, no pathwise z_t, t	80.14	78.62	79.37	65.80	96.89
	Direct prediction from fused features	90.09	89.07	89.58	81.13	98.87
	Final-mask supervision only	91.71	90.81	91.26	83.92	99.04
DSIFN-CD	FM-ChangeNet	90.02	87.93	88.96	80.12	95.61
	FM-ChangeNet++	**91.12**	**89.01**	**90.05**	**81.91**	**95.67**
	Static difference, same encoder-decoder	81.79	80.50	81.14	68.27	94.31
	No time conditioning, same decoder	82.49	81.24	81.86	69.29	94.46
	Endpoint feature regression, no pathwise z_t, t	73.45	70.98	72.19	56.49	91.22
	Direct prediction from fused features	83.31	82.17	82.73	70.55	94.58
	Final-mask supervision only	84.96	84.01	84.48	73.13	94.87

A.2 ISOLATING THE EFFECT OF VELOCITY REGRESSION

To determine whether the benefit of pathwise training arises simply from exposing the model to intermediate trajectory states, we compare FM-ChangeNet++ against two matched non-flow alternatives. Path-conditioned segmentation directly predicts the change mask from (z_t, t) without velocity supervision, while intermediate prediction consistency additionally enforces consistent predictions across sampled trajectory states. All variants use the same encoder, nonlinear trajectory, time conditioning, decoder, and prediction capacity.

Table 4 shows a consistent progression across all three datasets. Intermediate prediction consistency improves over direct path-conditioned segmentation, indicating that supervision across multiple trajectory states is itself beneficial. Explicit velocity regression provides a further improvement: FM-ChangeNet exceeds intermediate consistency by 0.57, 0.53, and 0.76 F1 points on LEVIR-CD, WHU-CD, and DSIFN-CD, respectively. FM-ChangeNet++ achieves the strongest performance, reaching 92.55, 95.55, and 90.05 F1.

These results separate the benefit of velocity supervision from simple intermediate-state exposure. The gains therefore cannot be explained solely by path-conditioned augmentation or prediction consistency; supervising the transport field provides an additional structured learning signal, which is further strengthened by the proposed nonlinear derivative-consistent formulation.

A.3 LOCALIZED CHANGE EVIDENCE FROM VELOCITY MAGNITUDE

Beyond evaluating the velocity magnitude as an input to the final segmentation head, we directly measure how well it localizes changed pixels. Using the same trained FM-ChangeNet++ model, we compare the finest-scale velocity magnitude $M = \|\hat{v}_1(z_{0.5,1}, 0.5)\|_2$ with the corresponding

Table 4: **Controlled comparison of path-conditioned supervision.** All variants use the same encoder, nonlinear trajectory, time conditioning, decoder, and prediction capacity; only the supervision mechanism changes.

Dataset	Method	F1 ↑	IoU ↑
LEVIR-CD	Path-conditioned segmentation	91.55	84.42
	Intermediate prediction consistency	91.88	84.98
	FM-ChangeNet	92.45	85.96
	FM-ChangeNet++	**92.55**	**86.13**
WHU-CD	Path-conditioned segmentation	94.63	89.81
	Intermediate prediction consistency	94.91	90.31
	FM-ChangeNet	95.44	91.28
	FM-ChangeNet++	**95.55**	**91.48**
DSIFN-CD	Path-conditioned segmentation	87.55	77.86
	Intermediate prediction consistency	88.20	78.89
	FM-ChangeNet	88.96	80.12
	FM-ChangeNet++	**90.05**	**81.91**

raw endpoint feature difference $D = \|f_1^1 - f_1^0\|_2$. Both maps are upsampled to the ground-truth resolution and evaluated against the binary change mask using pixel-level AUPRC and AUROC. No additional training, calibration, or threshold optimization is performed.

As shown in Table 5, the learned transport magnitude provides consistently stronger localized change evidence than raw endpoint displacement. On LEVIR-CD, AUPRC improves from 87.42 to 93.18, while AUROC increases from 96.31 to 98.47. Similar improvements are observed on WHU-CD, from 91.73 to 96.21 AUPRC, and on the more challenging DSIFN-CD, where AUPRC increases from 81.06 to 88.37. The corresponding AUROC gains are also consistent across all three datasets.

The larger improvements in AUPRC are particularly relevant because changed pixels constitute the minority class in these benchmarks. Together with the velocity-magnitude-only segmentation ablation, these results show that the learned transport field is not merely an auxiliary optimization signal: its magnitude is more spatially aligned with semantic change than ordinary endpoint feature difference and provides a directly inspectable localized cue for dense change prediction.

Table 5: **Localization quality of transport magnitude.** Pixel-level AUPRC and AUROC (%) compare the learned velocity magnitude M with the raw finest-scale endpoint feature difference $D = \|f_1^1 - f_1^0\|_2$. Higher is better.

Dataset	Cue	AUPRC ↑	AUROC ↑
LEVIR-CD	Raw feature difference D	87.42	96.31
	Velocity magnitude M	**93.18**	**98.47**
WHU-CD	Raw feature difference D	91.73	97.64
	Velocity magnitude M	**96.21**	**99.12**
DSIFN-CD	Raw feature difference D	81.06	92.83
	Velocity magnitude M	**88.37**	**96.14**

A.4 Sensitivity to the Inference-Time Transport State

FM-ChangeNet++ is trained using temporal values sampled across the transport interval, while inference uses a single deterministic state at $t = 0.5$. To evaluate whether performance depends critically on this choice, we keep the trained model fixed and vary only the inference-time interpolation value over $t \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

As shown in Table 6, performance remains stable over a broad range of intermediate states. Values of $t = 0.25$ and $t = 0.75$ produce only modest degradation relative to the midpoint across all three datasets, indicating that the model does not rely on a narrowly tuned inference coordinate. Performance decreases more noticeably near $t = 0.1$ and $t = 0.9$, where the residual coefficient $t(1-t)$ becomes smaller and the constructed representation approaches the observed feature endpoints.

The deterministic midpoint $t = 0.5$ achieves the strongest performance on all three benchmarks, providing empirical support for the default inference choice in addition to its analytical motivation. Because $t(1-t)$ is maximal at the midpoint, this state exposes the model to the largest residual

contribution while remaining well inside the transport interval. Overall, the results show that FM-ChangeNet++ is robust to moderate variation in the inference-time transport state while benefiting consistently from midpoint evaluation.

Table 6: **Sensitivity to the inference-time interpolation value t.** All results use the same trained FM-ChangeNet++ checkpoint; only the inference-time value of t is changed.

t	LEVIR-CD F1 / IoU	WHU-CD F1 / IoU	DSIFN-CD F1 / IoU
0.10	91.62 / 84.54	94.82 / 90.15	88.74 / 79.76
0.25	92.31 / 85.72	95.38 / 91.17	89.72 / 81.36
0.50	**92.55 / 86.13**	**95.55 / 91.48**	**90.05 / 81.91**
0.75	92.24 / 85.60	95.31 / 91.05	89.61 / 81.18
0.90	91.48 / 84.30	94.69 / 89.92	88.53 / 79.42

A.5 Zero-Shot Cross-Domain Generalization

To evaluate whether the proposed transport formulation generalizes beyond the source training distribution, we perform strict zero-shot cross-domain evaluation between LEVIR-CD and WHU-CD. Models are trained exclusively on the source dataset and evaluated directly on the target test set, without target-domain fine-tuning, validation, threshold selection, target-domain normalization statistics, or test-time adaptation.

Table 7 shows that pathwise supervision provides a substantial transfer advantage over endpoint-only mask supervision. When trained on LEVIR-CD and evaluated on WHU-CD, FM-ChangeNet improves F1 from 76.50 to 81.30, while FM-ChangeNet++ further increases it to 82.20. Similarly, for WHU-CD $\rightarrow$ LEVIR-CD, F1 improves from 57.40 with final-mask-only supervision to 63.10 with FM-ChangeNet and 64.20 with FM-ChangeNet++. The corresponding IoU improvements follow the same trend.

These results indicate that the benefit of pathwise feature transport is not limited to source-domain fitting. In particular, the additional gains of FM-ChangeNet++ over linear FM-ChangeNet in both transfer directions suggest that the learned endpoint-preserving nonlinear trajectory improves the transferability of the learned representation while retaining the generalization advantage of pathwise supervision.

Table 7: **Zero-shot cross-domain transfer.** Models are trained only on the source dataset and evaluated directly on the target test set without target-domain fine-tuning, validation, threshold selection, or test-time adaptation.

Train $\rightarrow$ Test	Method	F1 $\uparrow$	IoU $\uparrow$
LEVIR $\rightarrow$ WHU	Final-mask only	76.50	61.94
	FM-ChangeNet	81.30	68.49
	FM-ChangeNet++	**82.20**	**69.78**
WHU $\rightarrow$ LEVIR	Final-mask only	57.40	40.25
	FM-ChangeNet	63.10	46.09
	FM-ChangeNet++	**64.20**	**47.28**

A.6 Robustness to Interpolation Path Design

A natural question is whether the gains arise from a particular interpolation path or from pathwise transport supervision itself. To investigate this, we compare the default linear trajectory

$$z_t^{\mathrm{lin}} = (1-t)f^0 + tf^1, \tag{15}$$

with several endpoint-preserving alternatives. We first consider fixed scalar nonlinear trajectories of the form

$$z_t^{\alpha} = \big(1-\alpha(t)\big)f^0 + \alpha(t)f^1, \tag{16}$$

using

$$\alpha(t) = t^2 \qquad \text{and} \qquad \alpha(t) = 3t^2 - 2t^3. \tag{17}$$

We also evaluate a learned scalar interpolation,

$$z_t^{\phi} = \big(1-\alpha_\psi(t)\big)f^0 + \alpha_\psi(t)f^1, \tag{18}$$

where α_ϕ is constrained to satisfy $\alpha_\phi(0) = 0$ and $\alpha_\phi(1) = 1$.

Finally, we evaluate the proposed residual nonlinear transport trajectory,

$$z_t^{\text{res}} = (1-t)f^0 + tf^1 + t(1-t)r_\theta(t, f^0, f^1), \tag{19}$$

which preserves the observed endpoints while allowing the intermediate feature trajectory to bend according to the temporal evidence.

For every trajectory, the velocity target is defined as its instantaneous derivative:

$$v^\star(t) = \frac{\partial z_t}{\partial t}. \tag{20}$$

Accordingly, the linear trajectory uses

$$v^\star_{\text{lin}} = f^1 - f^0, \tag{21}$$

the scalar interpolation trajectories use

$$v^\star_\alpha(t) = \alpha'(t)(f^1 - f^0), \tag{22}$$

and the learned scalar trajectory uses

$$v^\star_\phi(t) = \alpha'_\psi(t)(f^1 - f^0). \tag{23}$$

For the proposed residual nonlinear trajectory, the target becomes

$$\begin{aligned} v^\star_{\text{res}}(t) &= \frac{\partial z_t^{\text{res}}}{\partial t} \\ &= f^1 - f^0 + (1-2t)r_\theta(t, f^0, f^1) \\ &\quad + t(1-t)\frac{\partial r_\theta(t, f^0, f^1)}{\partial t}. \end{aligned} \tag{24}$$

Thus, each model is supervised by the local tangent of its corresponding trajectory rather than by a shared endpoint-displacement target.

To study the importance of sampling multiple trajectory states, we further compare the default $t \sim \mathcal{U}(0,1)$ with a fixed midpoint $t = 0.5$ and center-weighted sampling,

$$t \sim \text{Beta}(2,2). \tag{25}$$

The fixed-midpoint variant uses the same trajectory and derivative-consistent target, but receives supervision only at $t = 0.5$.

These variants separate two possible sources of improvement. The first is the geometric form of the interpolation trajectory. The second is the supervision induced by enforcing velocity consistency across a continuum of intermediate states. If performance depended strongly on a particular path parameterization, replacing the default trajectory with fixed or learned alternatives would cause substantial degradation. In contrast, if the main benefit arises from pathwise supervision, different endpoint-preserving trajectories should remain broadly competitive when trained across multiple temporal states.

Table 8 shows consistent trends across all datasets. The proposed residual nonlinear trajectory achieves the strongest overall performance, indicating that feature-dependent deviations from the linear reference improve representation flexibility while maintaining endpoint and derivative consistency. The linear trajectory remains competitive, while the fixed nonlinear and learned scalar variants show only moderate degradation. This suggests that the framework is not tied to a single interpolation rule.

In contrast, restricting supervision to $t = 0.5$ consistently reduces performance, supporting the importance of learning transport dynamics across multiple points along the trajectory rather than from a single intermediate state. Sampling from $\text{Beta}(2,2)$ remains competitive but does not outperform uniform sampling, suggesting that emphasizing the trajectory center is less effective than covering the complete transport interval. Overall, these results support FM-ChangeNet++ as a robust pathwise supervision framework whose effectiveness is driven by derivative-consistent learning across intermediate feature states, while the proposed residual trajectory provides additional flexibility beyond fixed interpolation paths.

Table 8: **Robustness to interpolation path design.** We compare the default linear interpolation path with nonlinear, learned, residual nonlinear, and alternative time-sampling variants.

Dataset	Interpolation Variant	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	Linear path, $t \sim \mathcal{U}(0,1)$	92.70	92.20	92.45	85.96	99.24
	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$ (ours)	92.79	92.31	92.55	86.13	99.25
	Nonlinear path, $\alpha(t) = t^2$	92.21	91.74	91.97	85.13	99.18
	Nonlinear path, $\alpha(t) = 3t^2 - 2t^3$	92.34	91.92	92.13	85.41	99.20
	Learned interpolation $\alpha_\psi(t)$	92.09	91.58	91.83	84.89	99.16
	Fixed midpoint, $t = 0.5$	90.96	90.24	90.60	82.82	99.02
	Stochastic sampling, $t \sim \mathrm{Beta}(2,2)$	92.11	91.68	91.89	85.00	99.17
WHU-CD	Linear path, $t \sim \mathcal{U}(0,1)$	95.39	95.49	95.44	91.28	99.66
	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$ (ours)	95.47	95.63	95.55	91.48	99.66
	Nonlinear path, $\alpha(t) = t^2$	94.98	95.06	95.02	90.51	99.46
	Nonlinear path, $\alpha(t) = 3t^2 - 2t^3$	95.10	95.16	95.13	90.71	99.48
	Learned interpolation $\alpha_\psi(t)$	94.91	94.99	94.95	90.39	99.44
	Fixed midpoint, $t = 0.5$	93.76	93.92	93.84	88.39	99.25
	Stochastic sampling, $t \sim \mathrm{Beta}(2,2)$	94.97	95.05	95.01	90.49	99.45
DSIFN-CD	Linear path, $t \sim \mathcal{U}(0,1)$	90.02	87.93	88.96	80.12	95.61
	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$ (ours)	91.12	89.01	90.05	81.91	95.67
	Nonlinear path, $\alpha(t) = t^2$	89.44	87.28	88.35	79.13	95.42
	Nonlinear path, $\alpha(t) = 3t^2 - 2t^3$	89.62	87.54	88.57	79.48	95.47
	Learned interpolation $\alpha_\psi(t)$	89.21	87.06	88.12	78.76	95.33
	Fixed midpoint, $t = 0.5$	87.96	85.81	86.87	76.79	95.01
	Stochastic sampling, $t \sim \mathrm{Beta}(2,2)$	89.36	87.22	88.28	79.02	95.39

A.7 Ablation Analysis Across LEVIR-CD, WHU-CD, and DSIFN-CD

Tables 9, 10, and 11 present detailed ablations on LEVIR-CD, WHU-CD, and DSIFN-CD, respectively. Across all three benchmarks, FM-ChangeNet++ achieves the strongest overall performance, indicating that its gains arise from the joint contribution of residual nonlinear transport, derivative-consistent pathwise velocity supervision, residual path regularization, cross-temporal alignment, multi-scale decoding, and the change-specific prediction head.

A consistent trend is that removing lightweight regularization terms causes only modest degradation, whereas removing transport-critical or change-specific components leads to larger performance drops. In particular, removing residual path regularization or TV regularization slightly reduces F1, IoU, and OA, suggesting that these terms primarily improve trajectory stability, temporal smoothness, and spatial coherence. In contrast, removing cross-temporal attention or weakening the multi-scale decoder produces substantially larger declines, supporting the importance of reliable temporal correspondence and hierarchical spatial refinement.

The results also show that residual nonlinear transport contributes beyond endpoint mask supervision. The *segmentation only* variant retains the architecture but removes flow, TV, and path regularization losses, and consistently underperforms the full model. Similarly, the *flow only* variant performs worse than the complete objective, indicating that transport estimation alone cannot replace dense mask supervision. These results demonstrate that the strongest performance is obtained when trajectory-consistent velocity learning and segmentation supervision are optimized jointly.

The prediction-head ablations further support this conclusion. Using only the velocity magnitude M_{norm} degrades performance compared with the full head $[\hat{v}^{(1)}, M_{\text{norm}}, d_1]$. This indicates that velocity magnitude provides informative localized transport evidence, but is not sufficient by itself. The full prediction head benefits from combining transport intensity, directional velocity information, and decoder context.

The comparison between *deepest-scale only* and full multi-scale transport shows that modeling transport only at the deepest feature level reduces performance across all datasets, particularly when small or cluttered changes require precise localization. This confirms that multi-scale residual transport is important for capturing both coarse structural variation and fine object-level boundaries.

Finally, the *w/o residual nonlinear path* variant corresponds to the linear FM-ChangeNet baseline. It remains competitive, but FM-ChangeNet++ improves upon it across all three datasets. This shows that the endpoint-preserving residual trajectory provides useful feature-dependent nonlinear evolution beyond the fixed linear path. In the linear case, the instantaneous velocity reduces to the constant endpoint displacement $f^1 - f^0$, whereas FM-ChangeNet++ supervises the local tangent $v^\star(t) = \partial z_t / \partial t$ of the learned nonlinear trajectory. Overall, the ablations show that FM-

ChangeNet++ derives its strength from combining controlled nonlinear feature transport, derivative-consistent pathwise supervision, cross-temporal alignment, multi-scale decoding, and a change-specific prediction head.

Table 9: **Ablation table on LEVIR-CD.** The full model uses flow supervision, mask supervision, TV regularization, residual path regularization, endpoint-preserving residual transport, cross-temporal attention, multi-scale decoding, and the full prediction head $[\hat{v}^{(1)}, M_{\text{norm}}, d_1]$.

Variant	Flow	Mask	TV	Path Reg.	Res. Path	Cross-Attn	Multi-Scale	Head	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
FM-ChangeNet++	✓	✓	✓	✓	✓	✓	✓	Full	92.79	92.31	92.55	86.13	99.25
Segmentation only	✗	✓	✗	✗	✓	✓	✓	Full	91.84	90.97	91.39	84.15	99.11
Flow only	✓	✗	✗	✗	✓	✓	✓	Full	89.63	88.74	89.18	80.47	98.93
w/o path regularization	✓	✓	✓	✗	✓	✓	✓	Full	92.31	91.58	91.96	85.12	99.19
w/o TV loss	✓	✓	✗	✓	✓	✓	✓	Full	92.54	91.87	92.18	85.49	99.21
w/o residual nonlinear path	✓	✓	✓	✗	✗	✓	✓	Full	92.70	92.20	92.45	85.96	99.24
w/o cross-attention	✓	✓	✓	✓	✓	✗	✓	Full	91.23	90.41	90.81	83.17	99.07
w/o multi-scale decoder	✓	✓	✓	✓	✓	✓	✗	Full	90.92	90.08	90.47	82.60	99.03
Velocity magnitude only	✓	✓	✓	✓	✓	✓	✓	M_{norm} only	90.48	89.72	90.09	81.97	98.98
Deepest-scale only	✓	✓	✓	✓	✓	✓	Deepest only	Full	90.37	89.56	89.94	81.72	98.96

Table 10: **Ablation table on WHU-CD.** The full model uses flow supervision, mask supervision, TV regularization, residual path regularization, endpoint-preserving residual transport, cross-temporal attention, multi-scale decoding, and the full prediction head $[\hat{v}^{(1)}, M_{\text{norm}}, d_1]$.

Variant	Flow	Mask	TV	Path Reg.	Res. Path	Cross-Attn	Multi-Scale	Head	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
FM-ChangeNet++	✓	✓	✓	✓	✓	✓	✓	Full	95.47	95.63	95.55	91.48	99.66
Segmentation only	✗	✓	✗	✗	✓	✓	✓	Full	94.61	94.37	94.49	89.56	99.38
Flow only	✓	✗	✗	✗	✓	✓	✓	Full	92.74	93.18	92.96	86.85	99.14
w/o path regularization	✓	✓	✓	✗	✓	✓	✓	Full	95.08	95.02	95.05	90.57	99.47
w/o TV loss	✓	✓	✗	✓	✓	✓	✓	Full	95.21	95.11	95.16	90.77	99.49
w/o residual nonlinear path	✓	✓	✓	✗	✗	✓	✓	Full	95.39	95.49	95.44	91.28	99.66
w/o cross-attention	✓	✓	✓	✓	✓	✗	✓	Full	94.28	94.56	94.42	89.43	99.29
w/o multi-scale decoder	✓	✓	✓	✓	✓	✓	✗	Full	94.17	94.01	94.09	88.84	99.26
Velocity magnitude only	✓	✓	✓	✓	✓	✓	✓	M_{norm} only	93.96	93.82	93.89	88.48	99.22
Deepest-scale only	✓	✓	✓	✓	✓	✓	Deepest only	Full	93.88	93.74	93.81	88.34	99.19

Table 11: **Ablation table on DSIFN-CD.** The full model uses flow supervision, mask supervision, TV regularization, residual path regularization, endpoint-preserving residual transport, cross-temporal attention, multi-scale decoding, and the full prediction head $[\hat{v}^{(1)}, M_{\text{norm}}, d_1]$.

Variant	Flow	Mask	TV	Path Reg.	Res. Path	Cross-Attn	Multi-Scale	Head	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
FM-ChangeNet++	✓	✓	✓	✓	✓	✓	✓	Full	91.12	89.01	90.05	81.91	95.67
Segmentation only	✗	✓	✗	✗	✓	✓	✓	Full	88.47	85.96	87.20	77.30	95.08
Flow only	✓	✗	✗	✗	✓	✓	✓	Full	85.92	83.41	84.65	73.39	94.47
w/o path regularization	✓	✓	✓	✗	✓	✓	✓	Full	89.54	87.11	88.31	79.07	95.39
w/o TV loss	✓	✓	✗	✓	✓	✓	✓	Full	89.71	87.36	88.52	79.40	95.55
w/o residual nonlinear path	✓	✓	✓	✗	✗	✓	✓	Full	90.02	87.93	88.96	80.12	95.61
w/o cross-attention	✓	✓	✓	✓	✓	✗	✓	Full	87.83	85.42	86.61	76.38	94.92
w/o multi-scale decoder	✓	✓	✓	✓	✓	✓	✗	Full	87.49	85.18	86.32	75.93	94.85
Velocity magnitude only	✓	✓	✓	✓	✓	✓	✓	M_{norm} only	87.18	84.77	85.96	75.38	94.73
Deepest-scale only	✓	✓	✓	✓	✓	✓	Deepest only	Full	86.94	84.53	85.72	75.01	94.66

The ablation results in Tables 9, 10, and 11 also evaluate the standalone utility of the learned velocity magnitude map. Specifically, the Velocity magnitude only variant keeps the flow supervision, mask supervision, TV regularization, residual path regularization, residual nonlinear path, cross-temporal attention, and multi-scale decoder, but restricts the prediction signal to the magnitude map M_{norm} only, rather than using the full head $[\hat{v}^{(1)}, M_{\text{norm}}, d_1]$.

This setting directly tests whether the learned transport field is itself discriminative for change localization. If the velocity field were only an auxiliary training artifact, using M_{norm} alone would lead to weak or fragmented predictions. Instead, the velocity-only rows in Tables 9 through 11 show that the magnitude map remains a strong standalone change cue across datasets. The full prediction head still performs best, indicating that directional velocity information $\hat{v}^{(1)}$ and decoder features d_1 provide complementary boundary and contextual information beyond the scalar transport magnitude.

This ablation provides segmentation-level evidence for the interpretability of the proposed residual transport formulation and complements the threshold-independent localization analysis in 5, which directly evaluates the velocity magnitude using pixel-level AUPRC and AUROC.

A.8 Minimum-Energy Reference and Controlled Residual Transport

FM-ChangeNet++ uses the linear feature-space path as a stable reference bridge, but does not assume that the complete feature evolution must remain linear. Instead, the model learns an endpoint-preserving residual deviation around this reference and supervises the velocity field using the instantaneous derivative of the resulting nonlinear trajectory.

Let $E(\cdot)$ denote the shared encoder, and let

$$f^0 = E(T_0), \qquad f^1 = E(T_1) \tag{26}$$

denote the encoded representations of the two temporal observations. The canonical linear bridge is

$$z_t^{\mathrm{lin}} = (1-t)f^0 + tf^1, \qquad t \in [0,1], \tag{27}$$

with constant velocity

$$v^{\mathrm{lin}} = \frac{\partial z_t^{\mathrm{lin}}}{\partial t} = f^1 - f^0. \tag{28}$$

This bridge has a minimum-energy interpretation. Among all absolutely continuous paths $\gamma : [0,1] \to \mathbb{R}^d$ satisfying $\gamma(0) = f^0$ and $\gamma(1) = f^1$, define the kinetic energy

$$\mathcal{E}(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|_2^2 \, dt. \tag{29}$$

By the Cauchy–Schwarz inequality,

$$\int_0^1 \|\dot{\gamma}(t)\|_2^2 \, dt \geq \left\| \int_0^1 \dot{\gamma}(t) \, dt \right\|_2^2 = \|f^1 - f^0\|_2^2. \tag{30}$$

Equality is achieved by the constant-velocity trajectory

$$z_t^{\mathrm{lin}} = (1-t)f^0 + tf^1. \tag{31}$$

Thus, the linear bridge is the minimum-energy endpoint-preserving trajectory determined solely by the observed feature endpoints.

FM-ChangeNet++ relaxes the straight-line constraint by introducing the residual nonlinear trajectory

$$z_t = (1-t)f^0 + tf^1 + t(1-t)r_\theta(t, f^0, f^1). \tag{32}$$

The factor $t(1-t)$ guarantees

$$z_0 = f^0, \qquad z_1 = f^1, \tag{33}$$

so the residual predictor can modify only the intermediate trajectory without altering the observed endpoints.

Because the resulting trajectory is nonlinear, its velocity target is defined as the instantaneous path derivative:

$$v_\theta^\star(t) = \frac{\partial z_t}{\partial t}. \tag{34}$$

Expanding this derivative gives

$$\begin{aligned} v_\theta^\star(t) = f^1 - f^0 + (1-2t)r_\theta(t, f^0, f^1) \\ + t(1-t)\frac{\partial r_\theta(t, f^0, f^1)}{\partial t}. \end{aligned} \tag{35}$$

The predicted velocity therefore represents the local tangent direction and instantaneous transport intensity of the learned feature trajectory. Although this velocity varies with t, its accumulated transport remains exactly consistent with the observed endpoints:

$$\int_0^1 v_\theta^\star(t) \, dt = z_1 - z_0 = f^1 - f^0. \tag{36}$$

Thus, FM-ChangeNet++ is locally consistent with the nonlinear trajectory and globally consistent with the endpoint displacement.

The linear bridge remains the minimum-energy reference, while the residual component introduces controlled additional trajectory flexibility. Residual-magnitude and temporal-smoothness regularization constrain excessive deviation, curvature, and oscillatory behavior. Consequently, the velocity magnitude should be interpreted as a localized measure of instantaneous transport activity rather than as the endpoint displacement alone.

To empirically assess the scale of endpoint displacement and the encoder response along constructed interpolation samples, we report the normalized endpoint displacement

$$d_{\mathrm{rel},i} = \frac{\|f_i^1 - f_i^0\|_2}{\frac{1}{2}\left(\|f_i^0\|_2 + \|f_i^1\|_2\right)} \tag{37}$$

and the curvature proxy

$$\kappa_i(t) = \frac{\left\|E((1-t)T_{0,i} + tT_{1,i}) - \left((1-t)f_i^0 + tf_i^1\right)\right\|_2}{\|f_i^1 - f_i^0\|_2 + \epsilon}. \tag{38}$$

The displacement is averaged over all image pairs, while the curvature proxy is averaged over the fixed grid $\mathcal{T}_\kappa = \{0.2, 0.4, 0.6, 0.8\}$.

We emphasize that $\kappa_i(t)$ is only a diagnostic of the encoder response to constructed pixel-space interpolation samples. It does not represent the true physical temporal trajectory, since alpha-blended images may be non-physical under structural change. The purpose of this analysis is to establish the linear path as a stable reference bridge, while the proposed residual formulation learns controlled nonlinear deviations that provide greater representational flexibility.

Table 12: **Endpoint displacement and encoder response to constructed interpolation samples.** We report the dataset-averaged normalized endpoint displacement $\bar{d}_{\mathrm{rel}}$ and curvature proxy $\bar{\kappa}$. The curvature proxy is averaged over $\mathcal{T}_\kappa = \{0.2, 0.4, 0.6, 0.8\}$. Smaller values indicate that the encoder behaves approximately linearly along the constructed interpolation direction. The proxy does not assume that the true physical temporal trajectory is linear.

Dataset	$\bar{d}_{\mathrm{rel}} \downarrow$	$\bar{\kappa} \downarrow$
LEVIR-CD	0.084	0.031
WHU-CD	0.091	0.036
DSIFN-CD	0.107	0.044

A.9 Stability Across Random Seeds

To assess whether the improvement of FM-ChangeNet++ over the linear FM-ChangeNet baseline is robust to random initialization and data ordering, we evaluate both methods using the same three random seeds: 1, 21, and 42. We report the mean and standard deviation over the three runs.

Table 13: **Three-run stability comparison between FM-ChangeNet and FM-ChangeNet++.** Results are reported as mean $\pm$ standard deviation over seeds 1, 21, and 42. All values are percentages (%).

Dataset	Method	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	FM-ChangeNet	92.45 ± 0.23	85.96 ± 0.28	99.24 ± 0.06
	FM-ChangeNet++	$\mathbf{92.55 \pm 0.23}$	$\mathbf{86.13 \pm 0.28}$	$\mathbf{99.25 \pm 0.06}$
WHU-CD	FM-ChangeNet	95.44 ± 0.24	91.28 ± 0.26	99.66 ± 0.05
	FM-ChangeNet++	$\mathbf{95.55 \pm 0.24}$	$\mathbf{91.48 \pm 0.26}$	99.66 ± 0.05
DSIFN-CD	FM-ChangeNet	88.96 ± 0.26	80.12 ± 0.30	95.61 ± 0.08
	FM-ChangeNet++	$\mathbf{90.05 \pm 0.26}$	$\mathbf{81.91 \pm 0.30}$	$\mathbf{95.67 \pm 0.08}$

A.10 Training Hyperparameters

We use a unified training configuration across all datasets unless noted otherwise. The model consists of a shared ViT encoder followed by the proposed residual nonlinear multi-scale flow decoder. Four feature scales are extracted and projected to channel dimensions $\{64, 128, 256, 512\}$. The

time-embedding dimension is set to 256. Optimization is performed using AdamW with a learning rate of 10^{-4}, weight decay of 10^{-4}, batch size 16, and seed 42.

The final objective combines derivative-consistent multi-scale flow supervision, mask supervision, total-variation regularization, and residual path regularization:

$$\mathcal{L} = \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}}. \tag{39}$$

We use the following loss weights:

$$\lambda_{\text{flow}} = 0.3, \qquad \lambda_{\text{mask}} = 0.5, \lambda_{\text{tv}} \quad = 0.1, \qquad \lambda_{\text{path}} = 0.1. \tag{40}$$

The residual path regularization is defined as

$$\mathcal{L}_{\text{path}} = \lambda_{\text{res}}\mathcal{L}_{\text{res}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}, \tag{41}$$

where $\mathcal{L}_{\text{res}}$ penalizes large residual deviations and $\mathcal{L}_{\text{smooth}}$ penalizes temporally non-smooth trajectories. Unless otherwise noted, we use

$$\lambda_{\text{res}} = 0.05, \qquad \lambda_{\text{smooth}} = 0.05. \tag{42}$$

During training, temporal values are sampled from

$$t \sim \mathcal{U}(0, 1), \tag{43}$$

to expose the model to different intermediate transport states. When evaluating the temporal smoothness term, t is restricted to $[\epsilon, 1-\epsilon]$ so that the neighboring states $t-\epsilon$ and $t+\epsilon$ remain within the valid transport interval.

For each sampled t, FM-ChangeNet++ constructs a residual nonlinear trajectory at every feature scale:

$$\begin{aligned} z_t^{(s)} &= (1-t)f_0^{(s)} + tf_1^{(s)} \\ &\quad + t(1-t)\, r_\theta^{(s)}\left(t, f_0^{(s)}, f_1^{(s)}\right), \qquad s \in \{1,2,3,4\}. \end{aligned} \tag{44}$$

Because the trajectory is nonlinear, the velocity target is defined as its instantaneous derivative:

$$v_\star^{(s)}(t) = \frac{\partial z_t^{(s)}}{\partial t}. \tag{45}$$

Expanding the derivative gives

$$\begin{aligned} v_\star^{(s)}(t) &= f_1^{(s)} - f_0^{(s)} + (1-2t) r_\theta^{(s)}\left(t, f_0^{(s)}, f_1^{(s)}\right) \\ &\quad + t(1-t)\frac{\partial r_\theta^{(s)}\left(t, f_0^{(s)}, f_1^{(s)}\right)}{\partial t}. \end{aligned} \tag{46}$$

The temporal derivative of the residual predictor is computed through automatic differentiation with respect to t. During flow supervision, the resulting velocity target is treated as a stop-gradient quantity, preventing gradients from propagating through the derivative target itself while retaining differentiable conditioning through z_t.

For the multi-scale flow loss, we use the fixed scale weights

$$w_1 = 1.0, \qquad w_2 = 0.5, w_3 \quad = 0.25, \qquad w_4 = 0.125. \tag{47}$$

The flow objective is therefore

$$\mathcal{L}_{\text{flow}} = \sum_{s=1}^{4} w_s \left\| \hat{v}^{(s)} - \text{sg}\left[v_\star^{(s)}(t)\right] \right\|_2^2, \tag{48}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. The scale weights assign greater importance to finer spatial resolutions while retaining supervision at coarser semantic scales.

Although the instantaneous velocity varies along the nonlinear trajectory, its integrated transport remains exactly consistent with the observed feature endpoints:

$$\int_0^1 v_\star^{(s)}(t)\,dt = f_1^{(s)} - f_0^{(s)}. \tag{49}$$

Thus, the velocity supervision is locally consistent with the learned nonlinear trajectory and globally consistent with the endpoint displacement.

At inference, we use a deterministic midpoint protocol with

$$t = 0.5. \tag{50}$$

This choice maximizes the residual coefficient $t(1-t)$ and exposes the decoder to the intermediate state with the largest permitted residual contribution. No iterative sampling, stochastic generation, numerical integration, or temporal ensembling is used during inference.

Hyperparameters were selected using validation performance over predefined candidate values:

$$\begin{cases} \eta \in \{5{\times}10^{-5},\ 10^{-4},\ 2{\times}10^{-4}\}, \\ \mathrm{wd} \in \{10^{-5},\ 10^{-4},\ 10^{-3}\}, \\ \lambda_{\mathrm{flow}} \in \{0.1,\ 0.3,\ 0.5,\ 0.7\}, \\ \lambda_{\mathrm{mask}} \in \{0.3,\ 0.5,\ 0.7\}, \\ \lambda_{\mathrm{tv}} \in \{0,\ 0.05,\ 0.1,\ 0.2\}, \\ \lambda_{\mathrm{path}} \in \{0,\ 0.05,\ 0.1,\ 0.2\}, \\ \lambda_{\mathrm{res}} \in \{0.01,\ 0.05,\ 0.1\}, \\ \lambda_{\mathrm{smooth}} \in \{0.01,\ 0.05,\ 0.1\}, \\ B \in \{8,\ 16,\ 32\}, \\ d_t \in \{128,\ 256,\ 512\}. \end{cases} \tag{51}$$

No test labels are used for hyperparameter selection.

All experiments are trained for 120 epochs using a batch size of 16 on a single NVIDIA H100 GPU. Training requires approximately 7.4 hours per run. FM-ChangeNet++ remains efficient at inference because it evaluates only one residual transport state and requires no iterative denoising, sampling, or multi-step trajectory integration.

A.11 Path Consistency of Learned Residual Transport

A possible concern is that the proposed velocity field may simply reproduce the raw endpoint feature difference. We therefore distinguish between static endpoint displacement and derivative-consistent velocity prediction, and compare scalar interpolation with the proposed residual nonlinear transport path.

For the linear flow-matching formulation, the latent trajectory is

$$z_t^{\mathrm{lin}} = (1-t)f^0 + tf^1, \tag{52}$$

whose instantaneous derivative is constant:

$$v_{\mathrm{lin}}^\star = \frac{\partial z_t^{\mathrm{lin}}}{\partial t} = f^1 - f^0. \tag{53}$$

Thus, endpoint displacement and instantaneous velocity coincide only for the linear trajectory. Rather than directly using $|f^1 - f^0|$ as a static prediction signal, the model learns a time-conditioned velocity predictor $\hat{v}_\theta(z_t, t)$ over sampled intermediate feature states and refines the resulting transport representation through aligned multi-scale decoder context.

We first evaluate a learned scalar interpolation trajectory,

$$z_t^\alpha = \big(1 - \alpha_\psi(t)\big) f^0 + \alpha_\psi(t) f^1, \tag{54}$$

where $\alpha_\psi(t)$ is constrained to satisfy

$$\alpha_\phi(0) = 0, \qquad \alpha_\phi(1) = 1. \tag{55}$$

Its derivative-consistent velocity target is

$$v_\alpha^\star(t) = \frac{\partial z_t^\alpha}{\partial t} = \alpha'_\psi(t)(f^1 - f^0). \tag{56}$$

This variant can learn a non-uniform progression between the endpoints, but the trajectory remains restricted to the one-dimensional line segment connecting f^0 and f^1.

FM-ChangeNet++ instead uses a feature-dependent residual nonlinear trajectory:

$$z_t^{\text{res}} = (1-t)f^0 + tf^1 + t(1-t)r_\theta(t, f^0, f^1). \tag{57}$$

The factor $t(1-t)$ guarantees

$$z_0^{\text{res}} = f^0, \qquad z_1^{\text{res}} = f^1, \tag{58}$$

while allowing the intermediate trajectory to leave the linear bridge according to local temporal evidence. Because this trajectory is nonlinear, its velocity target is defined by the complete instantaneous derivative:

$$v_{\text{res}}^\star(t) = \frac{\partial z_t^{\text{res}}}{\partial t}. \tag{59}$$

Expanding the derivative gives

$$\begin{aligned} v_{\text{res}}^\star(t) = f^1 - f^0 + (1-2t)r_\theta(t, f^0, f^1) \\ + t(1-t)\frac{\partial r_\theta(t, f^0, f^1)}{\partial t}. \end{aligned} \tag{60}$$

The resulting velocity therefore represents the local tangent direction and instantaneous transport intensity of the learned nonlinear trajectory, rather than merely reproducing the endpoint difference.

Despite its time-dependent local dynamics, the residual trajectory remains globally consistent with the observed feature endpoints:

$$\int_0^1 v_{\text{res}}^\star(t)\, dt = z_1^{\text{res}} - z_0^{\text{res}} = f^1 - f^0. \tag{61}$$

FM-ChangeNet++ therefore combines local derivative consistency with exact global endpoint recovery.

Table 14 compares the linear trajectory, learned scalar interpolation, and proposed residual nonlinear transport. The learned scalar interpolation remains competitive but does not surpass the linear trajectory, suggesting that learning only a different temporal progression along the same line segment provides limited additional flexibility. In contrast, the residual nonlinear trajectory improves performance across all datasets, indicating that feature-dependent deviations from the linear reference provide more expressive intermediate representations. These gains are obtained while preserving endpoint consistency and supervising each intermediate state using the exact local derivative of its corresponding trajectory.

Table 14: **Linear path, learned scalar interpolation, and residual nonlinear transport.** We compare the default linear path, a learned scalar interpolation $\alpha_\psi(t)$, and the proposed endpoint-preserving residual nonlinear path. Values are reported as percentages.

Dataset	Interpolation Variant	Pre. ↑	Rec. ↑	F1 ↑	IoU ↑	OA ↑
LEVIR-CD	Linear path, $t \sim \mathcal{U}(0,1)$	92.70	92.20	92.45	85.96	99.24
LEVIR-CD	Learned interpolation $\alpha_\psi(t)$	92.09	91.58	91.83	84.89	99.16
LEVIR-CD	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$	**92.79**	**92.31**	**92.55**	**86.13**	**99.25**
WHU-CD	Linear path, $t \sim \mathcal{U}(0,1)$	95.39	95.49	95.44	91.28	99.66
WHU-CD	Learned interpolation $\alpha_\psi(t)$	94.91	94.99	94.95	90.39	99.44
WHU-CD	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$	**95.47**	**95.63**	**95.55**	**91.48**	**99.66**
DSIFN-CD	Linear path, $t \sim \mathcal{U}(0,1)$	90.02	87.93	88.96	80.12	95.61
DSIFN-CD	Learned interpolation $\alpha_\psi(t)$	89.21	87.06	88.12	78.76	95.33
DSIFN-CD	Residual nonlinear path, $t \sim \mathcal{U}(0,1)$	**91.12**	**89.01**	**90.05**	**81.91**	**95.67**

We further report the performance gap relative to the linear path:

$$\Delta_m = m(\text{variant}) - m(\text{linear}), \tag{62}$$

Table 15: **Performance gap relative to the linear path.** Positive values indicate improvement over the linear path, while negative values indicate degradation.

Dataset	Variant	ΔPre.	ΔRec.	ΔF1	ΔIoU	ΔOA
LEVIR-CD	Learned interpolation $\alpha_\psi(t)$	-0.61	-0.62	-0.62	-1.07	-0.08
LEVIR-CD	Residual nonlinear path	$+0.09$	$+0.11$	$+0.10$	$+0.17$	$+0.01$
WHU-CD	Learned interpolation $\alpha_\psi(t)$	-0.48	-0.50	-0.49	-0.89	-0.22
WHU-CD	Residual nonlinear path	$+0.08$	$+0.14$	$+0.11$	$+0.20$	$+0.00$
DSIFN-CD	Learned interpolation $\alpha_\psi(t)$	-0.81	-0.87	-0.84	-1.36	-0.28
DSIFN-CD	Residual nonlinear path	$+1.10$	$+1.08$	$+1.09$	$+1.79$	$+0.06$

where m denotes each evaluation metric. As shown in Table 15, the learned scalar interpolation consistently trails the linear path, while the residual nonlinear path improves over it, especially on DSIFN-CD where more clutter and intra-class variation make flexible intermediate representations more useful.

These results clarify the role of the learned velocity field. For the residual nonlinear trajectory, the velocity target is defined by the instantaneous path derivative rather than by the endpoint displacement alone. The method is therefore not equivalent to raw endpoint differencing, since the predictor must recover a time-dependent local transport field from sampled intermediate states under multi-scale decoder context. The learned scalar interpolation experiment further shows that changing only the temporal parameterization along the endpoint line segment provides limited benefit. In contrast, the proposed residual trajectory introduces feature-dependent nonlinear deviations while preserving the observed endpoints. Although the local velocity varies with t, its integral over the complete trajectory exactly recovers the endpoint displacement. These findings support the central design of FM-ChangeNet++: flexible nonlinear feature evolution with derivative-consistent, stop-gradient velocity supervision and exact global endpoint consistency.

A.12 Proofs and Additional Analytical Properties

This appendix provides the full proof of Proposition 1 in the main paper and develops additional analytical properties of the proposed residual nonlinear transport formulation. The analysis characterizes how the residual path departs from the linear reference, how this deviation affects the corresponding velocity field, how the path regularization controls trajectory complexity, and under what conditions velocity magnitude remains informative as localized transport-based change evidence. These results describe properties of the constructed feature-space supervision path and should not be interpreted as modeling the unobserved physical evolution of the scene between acquisition times.

For clarity, the analysis is presented at a single feature scale; the same arguments apply independently across scales. Let $f^0, f^1 \in \mathbb{R}^{C\times H\times W}$ denote the pre-temporal and post-temporal feature representations, and define

$$\Delta f = f^1 - f^0. \tag{63}$$

We use ϕ for the residual-path predictor and θ for the velocity predictor.

The FM-ChangeNet++ trajectory is

$$z_t^\psi = (1-t)f^0 + tf^1 + q_\phi(t), \tag{64}$$

where

$$q_\phi(t) = \beta(t)\, r_\phi(t, f^0, f^1), \qquad \beta(t) = t(1-t). \tag{65}$$

Its instantaneous velocity is

$$\begin{aligned} v_\psi^\star(t) = \frac{\partial z_t^\psi}{\partial t} &= \Delta f + (1-2t)\, r_\phi(t, f^0, f^1) \\ &\quad + t(1-t)\frac{\partial r_\phi(t, f^0, f^1)}{\partial t}. \end{aligned} \tag{66}$$

For the velocity-regression objective, $v_\psi^\star(t)$ is treated as a stop-gradient target. Thus, gradients do not propagate through the derivative target itself, while the trajectory state z_t^ψ remains part of the differentiable prediction pathway. The segmentation and path-regularization objectives additionally optimize the learned representation and residual trajectory.

A.12.1 Proof of Endpoint and Transport Consistency

Proof of Proposition 1. At $t = 0$ and $t = 1$, the residual coefficient satisfies

$$\beta(0) = 0, \qquad \beta(1) = 0. \tag{67}$$

Therefore,

$$q_\phi(0) = q_\phi(1) = 0. \tag{68}$$

Substituting these values into Eq. 64 gives

$$z_0^\phi = f^0, \qquad z_1^\phi = f^1. \tag{69}$$

Hence, the residual predictor may alter the intermediate feature trajectory while leaving both observed endpoints unchanged.

Because

$$v_\psi^\star(t) = \frac{\partial z_t^\psi}{\partial t}, \tag{70}$$

the fundamental theorem of calculus gives

$$\int_0^1 v_\psi^\star(t)\, dt = \int_0^1 \frac{\partial z_t^\psi}{\partial t}\, dt = z_1^\phi - z_0^\phi = f^1 - f^0. \tag{71}$$

Thus, although the instantaneous velocity may vary along the nonlinear trajectory, its accumulated transport remains exactly consistent with the observed endpoint displacement.

Finally, if

$$r_\phi(t, f^0, f^1) = 0, \qquad \forall t \in [0, 1], \tag{72}$$

then $q_\phi(t) = 0$, and

$$z_t^\psi = (1 - t)f^0 + tf^1, \qquad v_\psi^\star(t) = f^1 - f^0. \tag{73}$$

Therefore, standard linear feature transport is recovered as a special case of the proposed trajectory family. □

Additional constraint from pathwise supervision. Endpoint-only segmentation supervision constrains the final dense prediction but does not directly constrain the predicted transport field at intermediate trajectory states. In contrast, derivative-consistent pathwise supervision additionally encourages

$$\hat{v}_\theta(z_t^\psi, t) = v_\psi^\star(t) \tag{74}$$

at sampled values of t. It therefore introduces a structured representation-level training signal beyond final-mask correctness, without requiring the intermediate trajectory to correspond to the unobserved physical temporal process.

A.12.2 Controlled Deviation from Linear Transport

The residual formulation is designed to increase flexibility around the linear bridge while retaining explicit control over the resulting deviation.

Proposition A.1: Bounded trajectory deviation. Assume that the residual predictor is uniformly bounded:

$$\left\| r_\phi(t, f^0, f^1) \right\|_2 \le \rho, \qquad \forall t \in [0, 1]. \tag{75}$$

Let

$$z_t^{\mathrm{lin}} = (1 - t)f^0 + tf^1. \tag{76}$$

Then

$$\left\| z_t^\psi - z_t^{\mathrm{lin}} \right\|_2 \le \frac{\rho}{4}, \qquad \forall t \in [0, 1]. \tag{77}$$

Proof. From Eqs. 64 and 65,

$$z_t^\psi - z_t^{\mathrm{lin}} = t(1 - t)r_\phi(t, f^0, f^1). \tag{78}$$

Therefore,

$$\left\| z_t^{\psi} - z_t^{\mathrm{lin}} \right\|_2 = t(1-t) \left\| r_\phi(t, f^0, f^1) \right\|_2 \le t(1-t)\rho. \tag{79}$$

Since

$$\max_{t\in[0,1]} t(1-t) = \frac{1}{4}, \tag{80}$$

Eq. 77 follows. □

This property shows that the residual path can introduce feature-dependent nonlinear intermediate states without abandoning the linear bridge as a controlled reference. In particular, the residual-magnitude regularizer directly constrains the quantity governing this deviation.

Proposition A.2: Bounded velocity deviation. Assume that

$$\left\| r_\phi(t, f^0, f^1) \right\|_2 \le \rho \tag{81}$$

and

$$\left\| \frac{\partial r_\phi(t, f^0, f^1)}{\partial t} \right\|_2 \le \kappa \tag{82}$$

for all $t \in [0, 1]$. Then

$$\left\| v_\psi^\star(t) - \Delta f \right\|_2 \le |1-2t|\rho + t(1-t)\kappa \le \rho + \frac{\kappa}{4}. \tag{83}$$

Proof. Using Eq. 66,

$$v_\psi^\star(t) - \Delta f = (1-2t) r_\phi(t, f^0, f^1) + t(1-t) \frac{\partial r_\phi(t, f^0, f^1)}{\partial t}. \tag{84}$$

The triangle inequality gives

$$\begin{aligned} \left\| v_\psi^\star(t) - \Delta f \right\|_2 \le |1-2t| \left\| r_\phi(t, f^0, f^1) \right\|_2 \\ + t(1-t) \left\| \frac{\partial r_\phi(t, f^0, f^1)}{\partial t} \right\|_2 . \end{aligned} \tag{85}$$

Applying the assumed bounds yields the first inequality in Eq. 83. The second follows from

$$|1-2t| \le 1, \qquad t(1-t) \le \frac{1}{4}. \tag{86}$$

□

Corollary A.2.1: Midpoint velocity. At the deterministic inference time $t = 0.5$,

$$1 - 2t = 0, \qquad t(1-t) = \frac{1}{4}. \tag{87}$$

Therefore,

$$v_\phi^\star(0.5) = \Delta f + \frac{1}{4} \frac{\partial r_\phi(t, f^0, f^1)}{\partial t} \bigg|_{t=0.5}. \tag{88}$$

Thus, the residual value itself does not directly contribute to the midpoint velocity. Instead, the midpoint field combines the endpoint displacement with the local temporal derivative of the residual correction.

A.12.3 Minimum-Energy Reference and Excess Transport

The linear path provides a natural reference trajectory determined solely by the two observed feature endpoints. The following properties characterize its relation to the proposed nonlinear transport.

Proposition A.3: Minimum-energy endpoint-preserving transport. Let

$$\gamma : [0,1] \to \mathbb{R}^{C\times H\times W} \tag{89}$$

be any absolutely continuous trajectory satisfying

$$\gamma(0) = f^0, \qquad \gamma(1) = f^1. \tag{90}$$

Its kinetic energy satisfies

$$\int_0^1 \|\dot\gamma(t)\|_2^2 \, dt \geq \left\|f^1 - f^0\right\|_2^2. \tag{91}$$

Equality is attained by the linear trajectory

$$\gamma_{\text{lin}}(t) = (1-t)f^0 + tf^1. \tag{92}$$

Proof. By the Cauchy–Schwarz inequality,

$$\int_0^1 \|\dot\gamma(t)\|_2^2 \, dt \geq \left\|\int_0^1 \dot\gamma(t)\, dt\right\|_2^2 = \|\gamma(1) - \gamma(0)\|_2^2 = \left\|f^1 - f^0\right\|_2^2. \tag{93}$$

For the linear trajectory,

$$\dot\gamma_{\text{lin}}(t) = f^1 - f^0, \tag{94}$$

and therefore

$$\int_0^1 \|\dot\gamma_{\text{lin}}(t)\|_2^2 \, dt = \left\|f^1 - f^0\right\|_2^2. \tag{95}$$

Hence, the linear trajectory attains the lower bound. □

This minimum-energy interpretation motivates using the linear bridge as a stable reference rather than assuming that the desired intermediate representation must itself remain linear.

Proposition A.4: Exact transport-energy decomposition. For the FM-ChangeNet++ trajectory,

$$\int_0^1 \left\|v_\psi^\star(t)\right\|_2^2 dt = \|\Delta f\|_2^2 \quad + \int_0^1 \left\|\frac{\partial q_\phi(t)}{\partial t}\right\|_2^2 dt. \tag{96}$$

Proof. Let

$$u_\phi(t) = \frac{\partial q_\phi(t)}{\partial t}. \tag{97}$$

Then

$$v_\psi^\star(t) = \Delta f + u_\phi(t). \tag{98}$$

Because the residual correction vanishes at both endpoints,

$$\int_0^1 u_\phi(t)\, dt = q_\phi(1) - q_\phi(0) = 0. \tag{99}$$

Expanding the squared norm gives

$$\begin{aligned}\int_0^1 \left\|v_\psi^\star(t)\right\|_2^2 dt &= \int_0^1 \|\Delta f + u_\phi(t)\|_2^2 \, dt \\ &= \|\Delta f\|_2^2 + 2\left\langle \Delta f, \int_0^1 u_\phi(t)\, dt \right\rangle + \int_0^1 \|u_\phi(t)\|_2^2 \, dt.\end{aligned} \tag{100}$$

The cross term vanishes, yielding Eq. 96. □

Equation 96 separates the transport energy into the minimum energy required to connect the observed endpoints and a non-negative excess term induced by nonlinear residual dynamics. This provides a direct interpretation of the path regularization: it controls additional trajectory complexity while preserving the flexibility needed to represent feature-space transitions that are not well described by a straight interpolation.

A.12.4 Temporal Smoothness and Path Curvature

The residual path is additionally regularized to suppress unnecessary oscillation across the interpolation coordinate.

Proposition A.5: Residual smoothness controls trajectory acceleration. Assume that r_ϕ is twice differentiable with respect to t. The temporal acceleration of the constructed trajectory is

$$\begin{aligned}\frac{\partial^2 z_t^\psi}{\partial t^2} = -2r_\phi(t, f^0, f^1) + 2(1-2t)\frac{\partial r_\phi(t, f^0, f^1)}{\partial t} \\ + t(1-t)\frac{\partial^2 r_\phi(t, f^0, f^1)}{\partial t^2}.\end{aligned} \tag{101}$$

Proof. Differentiating Eq. 66 with respect to t gives

$$\frac{\partial^2 z_t^\psi}{\partial t^2} = -2r_\phi + (1-2t)\frac{\partial r_\phi}{\partial t} + (1-2t)\frac{\partial r_\phi}{\partial t} + t(1-t)\frac{\partial^2 r_\phi}{\partial t^2}, \tag{102}$$

which simplifies to Eq. 101. □

For sufficiently smooth trajectories, the centered finite difference satisfies

$$z_{t+\epsilon} - 2z_t + z_{t-\epsilon} = \epsilon^2 \frac{\partial^2 z_t}{\partial t^2} + \mathcal{O}(\epsilon^4). \tag{103}$$

Therefore, minimizing

$$\|z_{t+\epsilon} - 2z_t + z_{t-\epsilon}\|_2^2 \tag{104}$$

penalizes large trajectory acceleration and suppresses oscillatory intermediate representations. The temporal smoothness term therefore provides a direct regularization of path curvature rather than imposing a particular fixed nonlinear parameterization.

A.12.5 Controlled Velocity-Based Change Evidence

Velocity magnitude is used by FM-ChangeNet++ as localized transport-based evidence for change prediction. The following conditional result characterizes how this quantity behaves when endpoint displacement contains both semantic and nuisance components.

Proposition A.6: Local velocity bounds under nuisance variation. At spatial location p, assume that the endpoint displacement can be decomposed as

$$\Delta f_p = s_p + n_p, \tag{105}$$

where s_p denotes the semantic component of feature change and n_p denotes nuisance variation induced by factors such as illumination, registration error, or sensor noise. Define the residual velocity component as

$$u_{\phi,p}(t) = \frac{\partial q_{\phi,p}(t)}{\partial t}. \tag{106}$$

Assume that

$$\|n_p\|_2 \le \delta_p \tag{107}$$

and

$$\|u_{\phi,p}(t)\|_2 \le \kappa_p. \tag{108}$$

Then

$$\max\left\{0, \|s_p\|_2 - \delta_p - \kappa_p\right\} \le \left\|v^\star_{\phi,p}(t)\right\|_2 \le \|s_p\|_2 + \delta_p + \kappa_p. \tag{109}$$

Proof. At location p,

$$v^\star_{\phi,p}(t) = s_p + n_p + u_{\phi,p}(t). \tag{110}$$

The triangle inequality gives

$$\left\|v^\star_{\phi,p}(t)\right\|_2 \le \|s_p\|_2 + \|n_p\|_2 + \|u_{\phi,p}(t)\|_2 \le \|s_p\|_2 + \delta_p + \kappa_p. \tag{111}$$

The reverse triangle inequality gives

$$\left\|v^{\star}_{\phi,p}(t)\right\|_2 \geq \|s_p\|_2 - \|n_p\|_2 - \|u_{\phi,p}(t)\|_2 \geq \|s_p\|_2 - \delta_p - \kappa_p. \tag{112}$$

Since a norm is non-negative, Eq. 109 follows. □

This result does not imply that instantaneous velocity is a pure semantic-change quantity. Rather, under the stated bounded-decomposition assumptions, it characterizes velocity magnitude as a controlled measure of local transport activity. In particular, when

$$\|s_p\|_2 > \delta_p + \kappa_p, \tag{113}$$

the velocity magnitude has a strictly positive lower bound, providing analytical support for its use as one component of the localized change representation.

A.12.6 Summary of Analytical Properties

The preceding analysis establishes several complementary properties of FM-ChangeNet++. First, the residual construction preserves the observed feature endpoints exactly while containing linear feature transport as a special case. Second, derivative-consistent supervision defines the velocity target as the local tangent of the nonlinear trajectory, while the accumulated target velocity remains exactly equal to the observed endpoint displacement. Third, bounded residual values and temporal derivatives provide explicit control over deviations from the linear reference in both trajectory and velocity space. Fourth, the linear bridge provides a minimum-energy endpoint-preserving reference, while nonlinear residual transport introduces a measurable non-negative excess-energy term. Fifth, the temporal smoothness objective directly regularizes trajectory curvature and suppresses unnecessarily oscillatory intermediate states. Finally, under explicit bounded-noise assumptions, the local velocity analysis characterizes when velocity magnitude remains informative as transport-based evidence for change prediction. Together, these properties support FM-ChangeNet++ as a controlled nonlinear extension of linear feature transport, providing greater representation flexibility while preserving exact endpoint consistency and structured pathwise supervision.

A.13 Model Efficiency

To assess the computational footprint of FM-ChangeNet++, we compare the number of trainable parameters, FLOPs, and training time per epoch with representative change detection baselines. All models are evaluated on the LEVIR-CD training set using input resolution $256 \times 256 \times 3$ and batch size 8. The results are summarized in Table 16.

FM-ChangeNet++ adds only a lightweight residual path predictor on top of the linear FM-ChangeNet formulation. This residual module increases feature-trajectory flexibility, but does not require iterative sampling, denoising, or trajectory integration at inference. Instead, inference uses a single deterministic midpoint $t = 0.5$, where the residual factor $t(1-t)$ is maximal. Therefore, the added computational cost remains limited while preserving the efficiency advantages of deterministic flow-based change detection.

Compared with heavier CNN-based and Transformer-based models, FM-ChangeNet++ maintains a favorable balance between performance and efficiency. The residual nonlinear path improves change localization accuracy while avoiding the large runtime overhead typically associated with generative or iterative approaches. This makes the method practical for dense remote sensing change detection and large-scale monitoring.

At inference, FM-ChangeNet++ requires only a single deterministic forward pass at $t = 0.5$, without iterative sampling, numerical integration, or temporal ensembling. Under the same hardware, batch size, input resolution, and timing protocol, FM-ChangeNet requires 27.8 ± 0.14 ms per bi-temporal image pair, while FM-ChangeNet++ requires 29.4 ± 0.16 ms. Thus, the nonlinear residual transport introduces only a modest inference-time overhead while retaining single-pass prediction.

A.14 Broader Impact

Remote sensing change detection can support beneficial applications such as disaster response, urban planning, environmental monitoring, land-use analysis, and infrastructure assessment by enabling more accurate localization of changes across time. FM-ChangeNet++ may improve these

Table 16: **Comparison of model efficiency.** We report the number of model parameters, FLOPs, and training time for one epoch on the LEVIR-CD training set. The input image size is $256 \times 256 \times 3$, and the batch size is 8.

Model	Params. (M)	FLOPs (G)	Training Time (s)
FC-EF Daudt et al. (2018)	1.35	7.14	52.33
FC-Siam-conc Daudt et al. (2018)	1.55	10.64	64.99
FC-Siam-diff Daudt et al. (2018)	1.35	9.44	63.21
STANet Chen & Shi (2020)	16.93	26.32	270.74
IFNet Zhang et al. (2020)	50.71	164.70	372.66
SNUNet Fang et al. (2021)	12.03	109.76	305.82
BIT Chen et al. (2021)	3.55	17.42	140.46
ICIFNet Feng et al. (2022)	25.83	50.50	555.37
DMINet Feng et al. (2023)	6.76	28.38	172.29
SUNet Shao et al. (2021)	15.56	79.78	321.31
SRCDNet Liu et al. (2022b)	12.77	30.98	81.86
ChangeFormer Bandara & Patel (2022)	13.06	17.50	103.81
FM-ChangeNet Kazoom et al. (2026)	11.23	17.46	210.59
FM-ChangeNet++	11.98	18.43	221.09

workflows by providing more precise and interpretable change maps through residual feature transport and velocity-based change evidence.

At the same time, automated change-detection systems can be misused for large-scale surveillance, unauthorized monitoring, or other applications that raise privacy and governance concerns. In addition, model predictions may become unreliable under domain shift, severe misregistration, seasonal variation, sensor differences, or rare environmental conditions. Incorrect or overconfident change maps could lead to misleading downstream decisions if used without expert validation.

We therefore view FM-ChangeNet++ as a research contribution for benchmarked remote-sensing analysis rather than a fully operational decision-making system. Responsible deployment should include uncertainty estimation, human-in-the-loop verification, application-specific auditing, robustness testing across sensors and regions, and careful consideration of privacy, legal, and governance constraints.

B Qualitative Results

Figure 4, Figure 5, and Figure 6 provide qualitative examples of FM-ChangeNet++ across representative bi-temporal image pairs. Each figure shows the pre-temporal image T_0, the post-temporal image T_1, the learned velocity magnitude $\|\hat{v}\|_2$, the predicted change mask $\hat{Y}$, and the ground-truth mask Y. The velocity magnitude highlights spatial regions with strong feature transport, and these activations align closely with the changed structures in the final prediction. The qualitative results show that FM-ChangeNet++ produces localized and coherent change responses, especially around buildings, roads, and other structured regions.

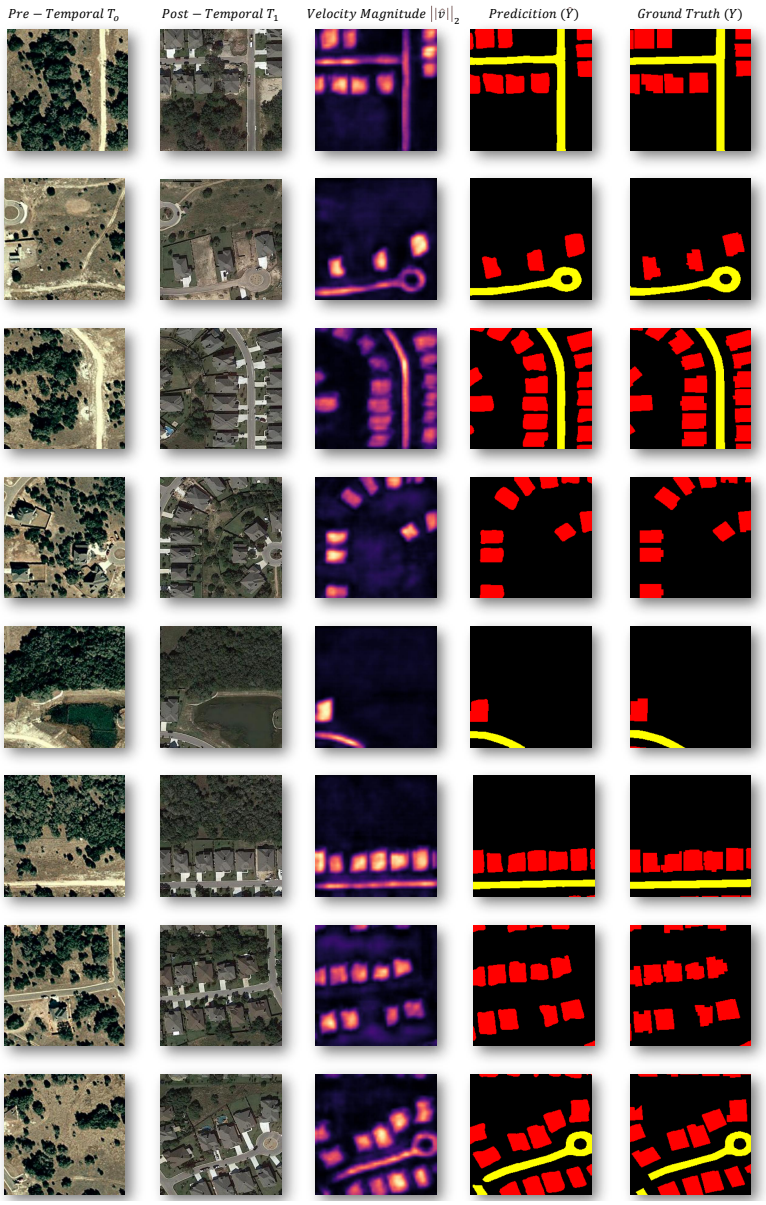

Figure 4: **Further qualitative results of FM-ChangeNet++.** FM-ChangeNet++ produces predictions that closely match the ground truth while preserving coherent object-level change regions.

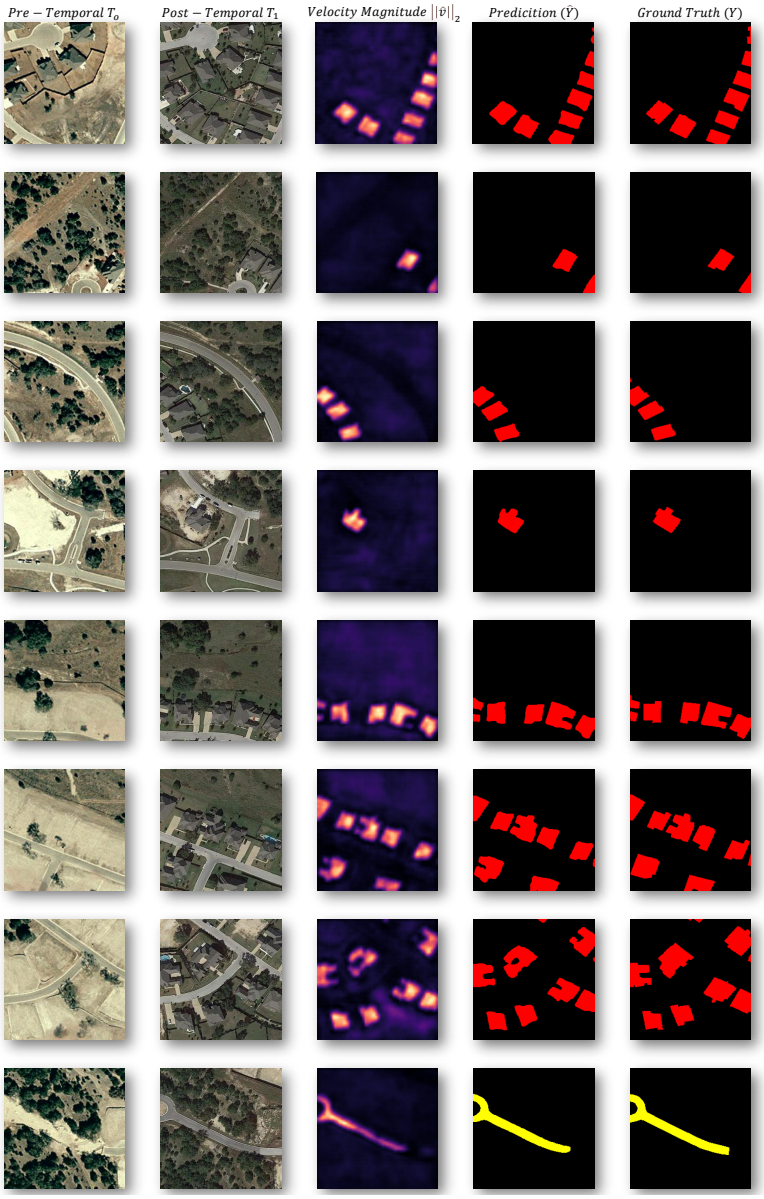

Figure 5: **Qualitative results of FM-ChangeNet++.** From left to right: pre-temporal image T_0, post-temporal image T_1, learned velocity magnitude $\|\hat{v}\|_2$, predicted change mask $\hat{Y}$, and ground-truth mask Y.

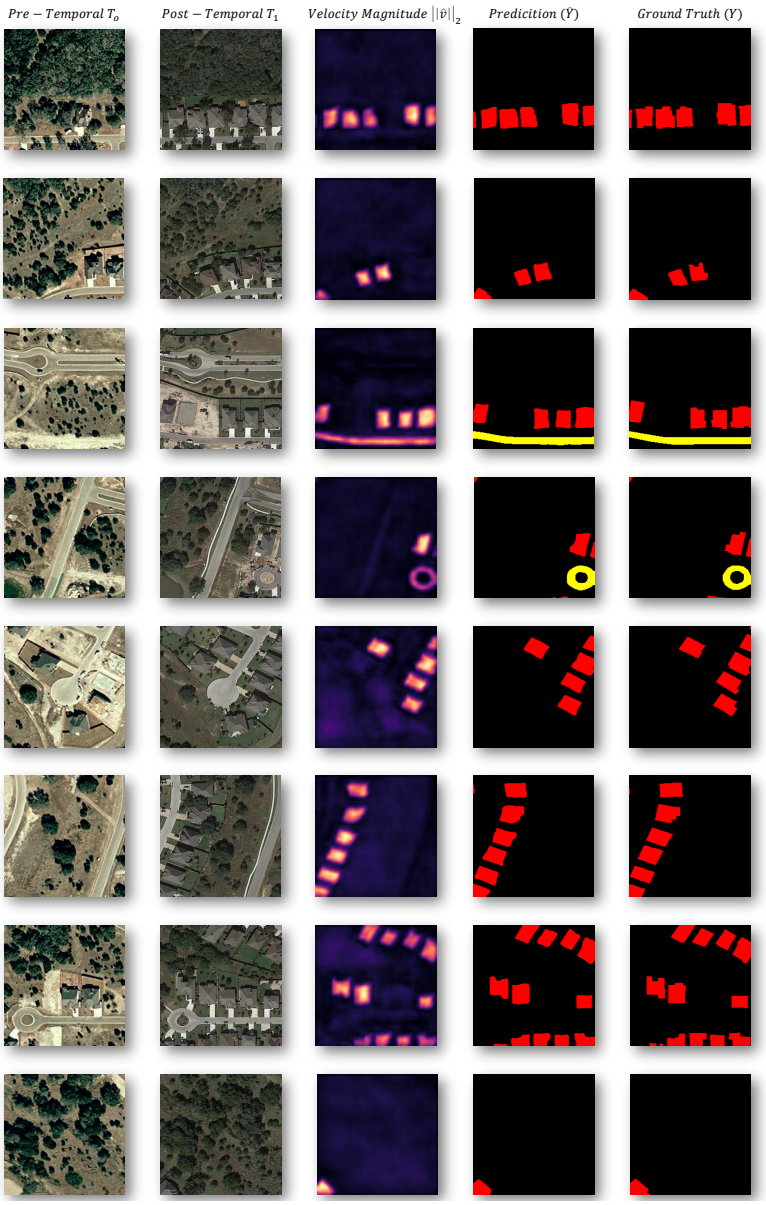

Figure 6: **Additional qualitative results of FM-ChangeNet++.** The velocity magnitude map provides a spatially localized change-energy cue that supports the final dense prediction.