# PTF-Net: Pseudo-Temporal Feature Fusion Network for Bi-Temporal Semantic Change Detection

**Xin Li**[1] , **Xin Dong**[1]
[1]China University of Petroleum (East China)
lix@upc.edu.cn, z24070037@s.upc.edu.cn

## Abstract

Bi-temporal Semantic Change Detection (SCD) is a fundamental analytical framework for capturing state transitions across various domains, with remote sensing being a key application area in this work. However, the prevailing paradigm of comparing two discrete temporal snapshots suffers from inherent limitations due to temporal discontinuity. These limitations lead to pronounced radiometric inconsistencies and pseudo-change noise in Bi-temporal SCD. To address these, we propose the Pseudo-Temporal Feature Fusion Network (PTF-Net), which reconceptualizes bi-temporal SCD as continuous pseudo-temporal evolution modeling. This novel paradigm facilitates the recovery of the missing temporal context between observation points. At its core lies an innovative Non-Linear Style Interpolation Mechanism, which projects discrete bi-temporal observations onto a smooth latent manifold to simulate plausible surface evolution dynamics. The mechanism synthesizes a sequence of semantically coherent intermediate representations, which inherently bridge the temporal gap and disentangles style variations from semantic changes. To fully exploit this synthesized sequence, we design a dedicated Temporal Dynamics Perception Branch. This branch employs efficient temporal interactions to robustly discriminate structural mutations from transient noise, effectively acting as a semantic filter. Extensive experiments on three public datasets reveal that PTF-Net consistently outperforms state-of-the-art methods, achieving comprehensive superiority across $mIoU$, $F_{scd}$, and $SeK$ metrics.

## 1 Introduction

As a fundamental task in remote sensing image analysis, Semantic Change Detection (SCD) is essential for monitoring land-cover evolution and understanding surface dynamics [Lu *et al.*, 2004; Hussain *et al.*, 2013; Xiang *et al.*, 2021]. Unlike traditional Binary Change Detection (BCD) [Lv *et al.*, 2018; Daudt *et al.*, 2018; Liu *et al.*, 2021; Chen *et al.*, 2021], SCD aims to simultaneously identify change regions and recognize the semantic categories of land-cover transitions between bi-temporal images [Daudt *et al.*, 2019; Ding *et al.*, 2022; Niu *et al.*, 2023]. It has proved to be highly effective in critical applications such as urban expansion monitoring [Doxani *et al.*, 2012], disaster assessment [Pitts and So, 2017; Li *et al.*, 2018], and ecosystem resource management [Rokni *et al.*, 2015] using multi-temporal remote sensing images.

The mainstream SCD paradigm is fundamentally built on bi-temporal image analysis [Jiang *et al.*, 2023; Wang *et al.*, 2025]. As illustrated in Fig. 1 (A), existing Siamese network-based architectures primarily focus on extracting high-level features from bi-temporal images independently and then comparing them through difference or concatenation operations [Daudt *et al.*, 2018; Daudt *et al.*, 2019]. However, this "compare-and-decide" paradigm inherently treats the temporal dimension as two discrete snapshots ($T_1$ and $T_2$), ignoring the continuous nature of surface evolution. In reality, land-cover changes are dynamic processes, not sudden jumps [Zhai *et al.*, 2021; Wang *et al.*, 2023]. The reliance on discrete bi-temporal data results in severely sparse temporal information, fundamentally limiting the model's ability to perceive the evolution of ground objects.

Although recent advances [Bandara and Patel, 2022; Liu *et al.*, 2024] have attempted to introduce attention mechanisms to model spatial dependencies, current SCD methods still face the following significant challenges caused by temporal discontinuity.

- **Temporal Information Scarcity.** The bi-temporal paradigm is limited to two discrete snapshots, providing insufficient context to model continuous dynamics or infer intermediate states.
- **Style-Structure Entanglement.** Radiometric style is inherently conflated with semantic content in feature space, and current methods lack explicit mechanisms for their disentanglement.
- **The Compounded Problem.** The temporal semantic void hinders dependency establishment. Compounding this, unresolved style entanglement allows irrelevant spectral shifts to corrupt invariant learning, directly resulting in false alarms(Fig. 1 A).

To address these challenges, we introduce the Pseudo-Temporal Feature Fusion Network (PTF-Net). Drawing

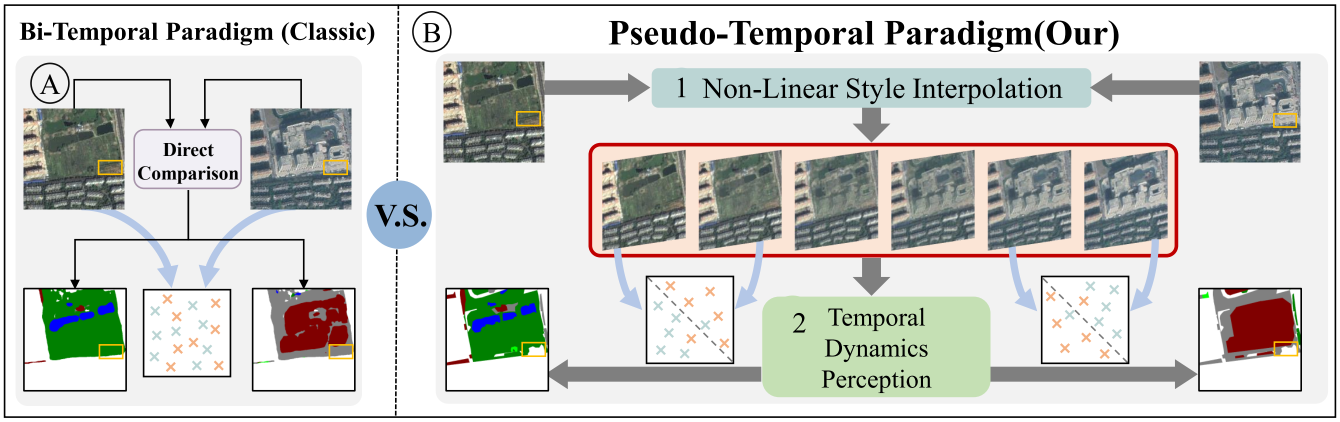


Figure 1: (A) Bi-Temporal Paradigm (Classic): Treats SCD as a static difference computing task between discrete snapshots. This often leads to false alarms (e.g., within the yellow box) caused by phenological pseudo-changes. (B) Pseudo-Temporal Paradigm: Reformulates SCD as continuous video evolution. It employs (1) Non-Linear Style Interpolation to synthesize intermediate states, and (2) Temporal Dynamics Perception to disentangle genuine semantic changes from stylistic noise (visualized as feature separation).

inspiration from efficient video understanding and spatio-temporal consistency modeling , our design is motivated by a key insight: SCD should model the continuous evolutionary process between observations rather than treating them as isolated snapshots. As depicted in Fig. 1 (B), this is realized through a novel Non-Linear Style Interpolation Mechanism. It projects the bi-temporal images onto a joint latent manifold to synthesize a smooth, pseudo-temporal sequence, which explicitly recovers the missing temporal context and provides a continuous bridge between the discrete observations.

Subsequently, to resolve style-structure entanglement, we devise a Temporal Dynamics Perception Branch that treats the generated sequence as a dynamic video stream. This branch synergizes the Temporal Shift Module (TSM) for efficient local interaction with a customized ConvGRU-based Temporal Gated Fusion (TGF) unit. Functioning as a semantic filter, the TGF utilizes reset gates to suppress transient style noise while employing update gates to accumulate genuine structural mutations, ensuring robust discrimination against significant spectral shifts.

In summary, our main contributions are as follows:

- We propose PTF-Net, which reconceptualizes the SCD problem from a discrete bi-temporal comparison task into a continuous pseudo-multi-temporal evolution modeling task. This framework fundamentally addresses the information gap inherent in discrete bi-temporal data by recovering the missing temporal context.
- We design a Non-Linear Style Interpolation Mechanism. Leveraging Non-Linear function mapping (e.g., Sigmoid) to simulate natural dynamics, it constructs smooth pseudo-temporal sequences. This mechanism recovers missing intermediate states, providing the necessary continuity context to bridge the temporal gap.
- We devise a Temporal Dynamics Perception Branch. By synergizing TSM with a customized TGF unit, this module treats the interpolated sequence as a dynamic video stream. Its selective gating mechanism acts as a semantic filter, suppressing transient style noise while accumulating genuine structural mutations.
- Experiments on three public datasets show that PTF-Net achieves superior performance and robustness against temporal discontinuity compared to advanced methods.

## 2 Related Work

### 2.1 Spatial Semantic Representation and Structural Prior Modeling

Spatial semantic representation and structural prior modeling are fundamental to SCD, establishing Multi-task Learning as the mainstream paradigm [Li *et al.*, 2024; Fang *et al.*, 2025; Li *et al.*, 2025].In terms of feature extraction and architecture, researchers strive to enhance interpretation capabilities for complex scenes. Liu et al. [2024] combined the local extraction strengths of CNNs with the long-range dependency modeling of Transformers, designing a hybrid architecture to capture global context while preserving texture details. To mitigate task interference in bi-temporal interpretation, Li et al. [2024] proposed a decoder-centric strategy that explicitly separates semantic and change streams. Similarly, Fang et al. [2024] developed BT-HRSCD with three independent decoding branches, significantly enhancing the parsing of fine-grained features in high-resolution scenarios.

Beyond architectural improvements, incorporating structural priors is crucial for addressing blurred boundaries and semantic confusion. Jiang et al. [2025] and Tang et al. [2024] focused on edge utilization, introducing boundary guidance mechanisms to inject edge priors into semantic branches, thereby refining the integrity of change regions. Furthermore, Jiang et al. [2024] adopted a topological perspective, utilizing hierarchical graph interaction networks to model semantic correlations between objects in complex distributions.

However, these methods typically treat the bi-temporal images $T_1$ and $T_2$ as isolated static moments. By ignoring the intermediate evolutionary process, they struggle to dis-

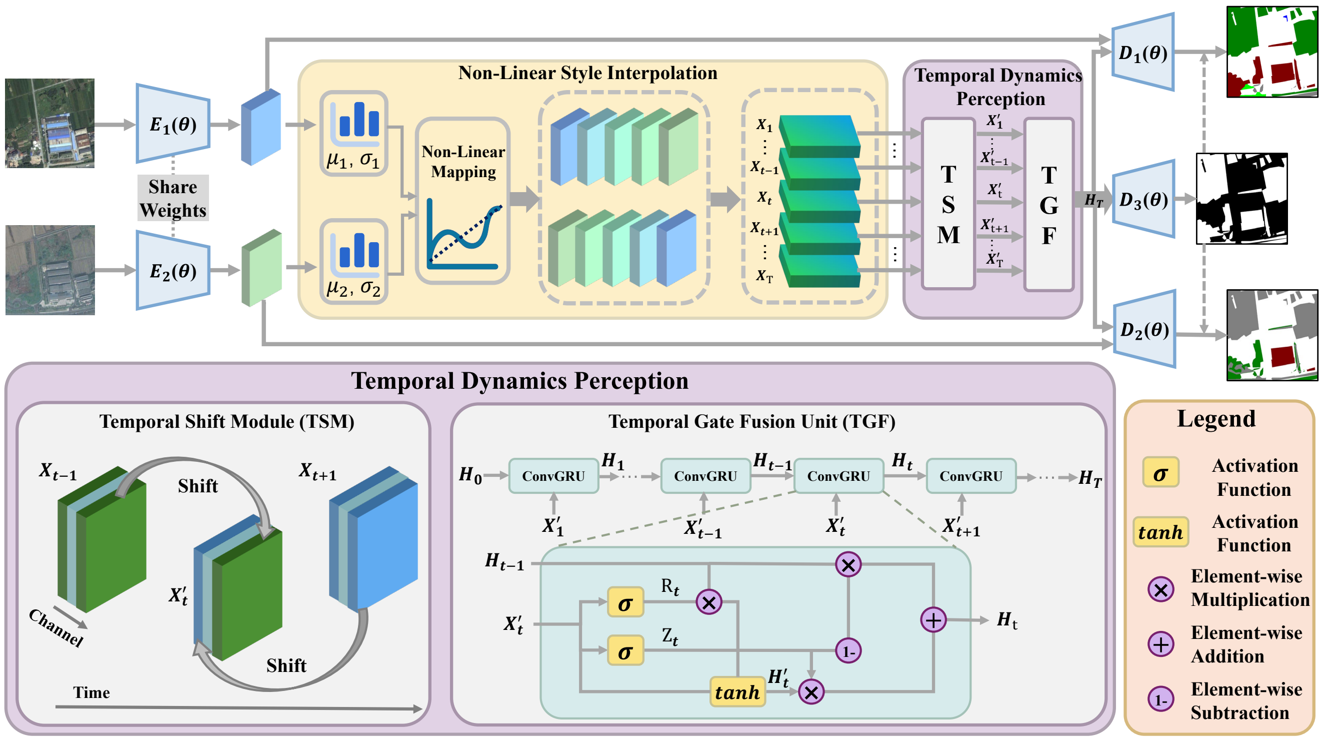


Figure 2: The overall framework of PTF-Net. A weight-sharing Siamese encoder extracts features, which are transformed into continuous pseudo-temporal sequences by the Non-Linear Style Interpolation module. Subsequently, the Temporal Dynamics Perception branch (integrating TSM and TGF) captures evolutionary dynamics, enabling parallel decoders to predict the final semantic and change maps.

tinguish genuine changes from background noise caused by non-structural phenological differences.

### 2.2 Spatio-temporal Feature Interaction and Consistency Modeling

Constructing robust spatio-temporal interaction mechanisms is essential to distinguish real semantic changes from “pseudo-changes” caused by lighting or seasonal fluctuations. Recently, research has shifted from simple feature subtraction to modeling “semantic consistency” between bi-temporal features [Yang *et al.*, 2024; Chen and Jiang, 2025; Zhou *et al.*, 2025]. Tang et al. [2024] proposed ClearSCD, which leverages the intrinsic logic between semantic categories and change regions to correct contradictory predictions via explicit relationship modeling. Wang et al. [2024] introduced a cross-difference semantic consistency network, designing specific loss functions to constrain semantic consistency in unchanged regions, effectively suppressing interference from imaging variations.

To further exploit the temporal dimension, Tian et al. [2025] introduced the concept of “temporal consistency”, aiming to automatically mine potential temporal dependencies in high-resolution imagery to optimize detection accuracy. These methods primarily rely on imposing logical constraints between $T_1$ and $T_2$ to filter pseudo-changes.

Although existing consistency modeling approaches suppress noise to some extent, they remain limited by static constraints on discrete states and fail to restore the dynamic essence of land cover evolution. Consequently, we shift the SCD paradigm from discrete comparison to continuous evolution perception, achieving effective decoupling of complex spatio-temporal dynamics.

## 3 Methodology

This section details the proposed Pseudo-Temporal Feature Fusion Network (PTF-Net). Section 3.1 introduces the overall framework and data flow. Section 3.2 elaborates on the Non-Linear Style Interpolation Mechanism for constructing high-fidelity pseudo-temporal sequences. Section 3.3 presents the Temporal Dynamics Perception Branch.

### 3.1 Overall Framework of PTF-Net

As illustrated in Fig. 2, PTF-Net utilizes a weight-shared Siamese dilated ResNet-34 [He *et al.*, 2016] encoder to extract features, which are then processed by the Non-Linear Style Interpolation Mechanism. This module synthesizes dense pseudo-temporal sequences on a continuous manifold, effectively bridging the temporal void to disentangle high-frequency style variations from structural changes [Dai *et al.*, 2024]. Subsequently, the Temporal Dynamics Perception Branch treats these sequences as dynamic video streams, synergizing the Temporal Shift Module (TSM) for local interaction and Temporal Gated Fusion (TGF) for global dependency modeling. This collaborative design filters transient environmental noise and extracts robust representations,

| Dataset | Size | Class | Total | Split (Tr / Val / Te) |
|---|---|---|---|---|
| Second | $512^2$ | 7 | 4,662 | 2,968 / 847 / 847 |
| Landsat | $416^2$ | 5 | 2,425 | 1,455 / 485 / 485 |
| JL1H | $512^2$ | 6 | 6,000 | 4,000 / 1,000 / 1,000 |

Table 1: Summary of the datasets.

which are finally projected into semantic and change maps via a multi-task decoder.

## 3.2 Non-Linear Style Interpolation Mechanism

The core challenge in processing discrete bi-temporal data lies in the temporal void, where intermediate evolutionary states are missing. Standard linear interpolation in the feature space often traverses low-probability regions of the data manifold, leading to unrealistic artifacts. To address this, we propose the Non-Linear Style Interpolation Mechanism to construct a smooth geodesic path between time points $t_1$ and $t_2$.

**Latent Style Statistics Modeling**

We adopt the Instance Normalization paradigm where the style of a remote sensing image is encoded in the global statistics of its feature channels. Given high-level feature maps $F$, we explicitly model their style representations by computing channel-wise statistics. Specifically, the mean $\mu_c(F)$ quantifies the central tendency of pixel values to represent global brightness or dominant tone, while the standard deviation $\sigma_c(F)$ measures the dispersion around the mean to capture contrast intensity. These statistics define a compact and low-dimensional embedding of the image style, effectively decoupling high-frequency environmental variations from low-frequency semantic content.

**Non-Linear Temporal Mapping Strategy**

Natural land surface transitions (e.g., crop growth curves, seasonal vegetation decay) inherently exhibit Non-Linear dynamics. A simple linear interpolation of style statistics assumes a constant rate of change, which fails to capture the acceleration and deceleration phases typical of natural processes. To rectify this, we employ a learnable Sigmoid-based warping function $\Phi(\cdot)$ to transform the linear time index $\tau \in [0, 1]$ into a Non-Linear evolutionary time $\tau'$:

$$\tau' = \Phi(\tau) = \frac{\text{Sigmoid}(k \cdot (2\tau - 1)) - \text{Sigmoid}(-k)}{\text{Sigmoid}(k) - \text{Sigmoid}(-k)} \quad (1)$$

Here, the learnable curvature parameter $k$ is the sole free parameter governing the evolution trajectory. A larger $k$ simulates an abrupt transition (e.g., rapid construction), while a smaller $k$ models a gradual evolution (e.g., slow vegetation growth). Using the warped time $\tau'$, we synthesize the intermediate style codes $\mu_t$ and $\sigma_t$ via weighted aggregation, ensuring the interpolated styles remain on a valid statistical manifold.

**Bi-directional Sequence Synthesis**

To fully recover the temporal context, we generate sequences in two complementary directions: Forward ($S^{fwd}$), evolving from $t_1$ to $t_2$, and Backward ($S^{bwd}$), evolving from $t_2$ to $t_1$. We utilize Adaptive Instance Normalization [Huang and Belongie, 2017] to inject the interpolated style statistics into the content features. For the forward sequence at time $t$, the feature $S_t^{fwd}$ is generated by:

$$S_t^{fwd} = \sigma_t(F_{t_1}, F_{t_2}) \cdot \frac{F_{t_1} - \mu_c(F_{t_1})}{\sigma_c(F_{t_1})} + \mu_t(F_{t_1}, F_{t_2}) \quad (2)$$

In the feature transformation, the subtraction $(F_{t_1} - \mu_c(F_{t_1}))$ is implemented via a broadcasting mechanism for element-wise normalization. This operation effectively "repaints" the normalized structural framework of $F_{t_1}$ with the evolving styles $\mu_t$ and $\sigma_t$. By bridging the stylistic gap across the pseudo-temporal sequence, the mechanism ensures the recovery of missing temporal context while strictly maintaining structural integrity. Finally, the forward and backward sequences are concatenated along the channel dimension, yielding a dense, information-rich pseudo-temporal tensor $\mathcal{X}_{raw} \in \mathbb{R}^{T \times 2C \times H \times W}$.

## 3.3 Temporal Dynamics Perception Branch

The generated tensor $\mathcal{X}_{raw}$ contains dense temporal information, but the forward and backward flows are initially isolated in the channel dimension. To extract robust change features, we design the Temporal Dynamics Perception Branch, which treats the sequence as a dynamic video stream. This branch employs a hierarchical "Local-to-Global" perception strategy, synergizing efficient local interaction with deep global fusion.

**Efficient Local Interaction via TSM**

Directly convolving over $\mathcal{X}_{raw}$ would fail to capture the interaction between the forward and backward evolution streams. To enable cross-stream information exchange, we first implement a Channel Shuffle mechanism [Zhang *et al.*, 2018]. By reshaping and transposing the channel dimension, we interleave the channel groups from $S^{fwd}$ and $S^{bwd}$. This ensures that complementary features from both evolutionary directions are adjacent in the channel space, facilitating the fusion of bi-directional contexts.

Following the alignment, we deploy the Temporal Shift Module (TSM) [Lin *et al.*, 2019], a parameter-free operator that shifts a fraction of channels along the temporal dimension, to model local temporal dependencies. Specifically, for a feature map $X_t$ at time $t$, we shift the $1/8$ channels to $t-1$ to access past information, and the next $1/8$ channels to $t+1$ to access future information. The operation can be formulated as:

$$X'_{t,c} = \begin{cases} X_{t-1,c} & \text{if } c \in \mathcal{C}_{bwd} \\ X_{t+1,c} & \text{if } c \in \mathcal{C}_{fwd} \\ X_{t,c} & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{C}_{bwd}$ and $\mathcal{C}_{fwd}$ denote the channel indices corresponding to the first and second $1/8$ splits, respectively.

By exchanging features between adjacent frames, TSM effectively expands the temporal receptive field of 2D convolutions without introducing any additional parameters or computational overhead (FLOPs). This results in a locally context-aware sequence $X' = \{X'_1, \ldots, X'_T\}$.

| Method | SECOND Dataset | | | | JL1H Dataset | | | | Landsat-SCD Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $OA$ | $mIoU$ | $Sek$ | $F_{scd}$ | $OA$ | $mIoU$ | $Sek$ | $F_{scd}$ | $OA$ | $mIoU$ | $Sek$ | $F_{scd}$ |
| HRSCD4 | 85.92 | 71.24 | 18.70 | 58.72 | 80.89 | 67.73 | 18.93 | 61.49 | 91.41 | 78.21 | 32.79 | 74.74 |
| SSCD-l | 87.41 | 72.65 | 22.36 | 62.72 | 83.08 | 69.27 | 22.64 | 65.57 | 94.86 | 85.63 | 50.22 | 84.57 |
| BiSRNet | 87.58 | 72.85 | 22.61 | 63.23 | 83.06 | 70.46 | 25.42 | 67.88 | 95.08 | 86.19 | 51.61 | 85.41 |
| TED | 87.43 | 72.76 | 22.58 | 62.67 | 82.02 | 68.93 | 20.63 | 63.46 | 95.92 | 88.18 | 57.34 | 87.69 |
| SCanNet | 87.29 | 72.82 | 23.33 | 63.54 | 82.97 | 69.25 | 23.56 | 66.39 | 96.29 | 89.11 | 60.14 | 88.82 |
| BT-HRSCD | 87.34 | 72.94 | 22.80 | 62.96 | 83.84 | 71.17 | 25.43 | 67.75 | 96.06 | 88.62 | 58.63 | 88.17 |
| DEFO-MLTSCD | 87.36 | <u>73.37</u> | 23.44 | 63.61 | 84.04 | 71.98 | 23.32 | 68.14 | 96.14 | 88.74 | 59.03 | 88.38 |
| GSTM-SCD | <u>87.96</u> | 73.23 | <u>23.93</u> | <u>64.14</u> | <u>87.93</u> | <u>75.64</u> | <u>36.27</u> | <u>76.33</u> | <u>96.69</u> | <u>90.35</u> | <u>63.98</u> | <u>90.12</u> |
| **PTF-Net(Ours)** | **88.31** | **73.51** | **24.21** | **64.45** | **92.75** | **84.58** | **54.35** | **86.22** | **97.08** | **91.44** | **67.45** | **91.37** |

Table 2: Quantitative comparison on three datasets. **Bold** indicates the best performance, and <u>underline</u> indicates the second-best.

| Method | Params (M) | FLOPs (G) | Time (ms) |
|---|---|---|---|
| HRSCD4 | **13.71** | **43.54** | **48.02** |
| SSCD-l | 23.31 | 189.52 | 51.27 |
| BiSRNet | 23.39 | 199.24 | 52.53 |
| TED | 24.19 | 206.98 | 50.75 |
| SCanNet | 27.90 | 264.97 | 61.87 |
| BT-HRSCD | 33.33 | 130.17 | 108.43 |
| DEFO-MLTSCD | 26.53 | 400.65 | 68.79 |
| GSTM-SCD | 31.55 | 73.35 | 101.37 |
| **PTF-Net(Ours)** | 25.92 | 208.04 | 56.17 |

Table 3: Comparison of model complexity and inference efficiency. Input size is set to $512 \times 512$. The inference time is averaged over 100 runs on a NVIDIA GeForce RTX 4090.

**Temporal Gated Fusion Unit**

While TSM captures short-term local motion, complex change events often involve long-range dependencies and transient noise. To aggregate the sequence into a unified representation, we propose a TGF unit based on ConvGRU [Mou and Zhu, 2018]. Unlike standard pooling or 3D convolutions, TGF utilizes a sophisticated gating mechanism to selectively accumulate structural information.

At each time step $t$, the TGF unit updates its hidden state $H_t$ based on the current input $X'_t$ and the previous memory $H_{t-1}$. The internal transition dynamics are governed by:

$$\begin{aligned} Z_t &= \sigma(W_z * [H_{t-1}, X'_t] + b_z) \\ R_t &= \sigma(W_r * [H_{t-1}, X'_t] + b_r) \\ \tilde{H}_t &= \tanh(W_h * [R_t \odot H_{t-1}, X'_t] + b_h) \\ H_t &= (1 - Z_t) \odot H_{t-1} + Z_t \odot \tilde{H}_t \end{aligned} \tag{4}$$

This mechanism plays a dual role in change perception: the Reset Gate ($R_t$) acts as a trainable filter to suppress transient noise by discarding irrelevant history, while the Update Gate ($Z_t$) governs the accumulation of structural mutations by selectively integrating genuine changes. Ultimately, the final hidden state $H_T$ yields the fused change feature $F_c$, encapsulating the robust spatio-temporal consensus of the entire change event.

# 4 Experiments

## 4.1 Experimental Setup

To assess the generalization capability of our model across varying resolutions and land-cover types, we utilize three representative public benchmark datasets:SECOND[Yang *et al.*, 2024], Landsat-SCD [Bandara and Patel, 2022], and JL1H[Tang *et al.*, 2025]. The detailed statistics of these datasets are summarized in Table 1.

**Implementation Details**

All experiments were implemented in PyTorch on a single NVIDIA GeForce RTX 4090 GPU, utilizing an ImageNet-pretrained ResNet-34 backbone. The model was trained using the AdamW optimizer (initial learning rate 0.01, weight decay 0.0001) and a Cosine Annealing scheduler (minimum $1 \times 10^{-6}$) with a batch size of 4. To ensure full convergence, the total epochs were set to 50 for the SECOND and JL1H datasets, and 100 for Landsat-SCD.

To ensure rigorous and fair comparisons, all baseline methods were re-evaluated under a unified experimental protocol identical to PTF-Net, encompassing consistent dataset splits, preprocessing pipelines, and hardware environments. This standardized benchmark eliminates potential biases introduced by varying experimental configurations, ensuring objective comparability across all models.

**Evaluation Metrics**

To quantitatively evaluate performance, we employ four standard metrics: Overall Accuracy ($OA$), $mIoU$, $F_{scd}$, and Separated Kappa coefficient ($SeK$). Specifically, $OA$ measures the global pixel-wise classification accuracy across all categories. The $mIoU$ provides a balanced assessment by averaging the Intersection over Union scores for both non-change and binary change detection, mitigating the dominance of background pixels. Beyond binary localization, $F_{scd}$ evaluates semantic consistency by calculating the harmonic mean of precision and recall specifically for semantic change categories. Finally, $SeK$ is utilized to alleviate class imbalance bias; it computes the Kappa coefficient on a reduced confusion matrix that excludes the non-change class, weighted by the binary change IoU to focus strictly on the accuracy of semantic transitions.

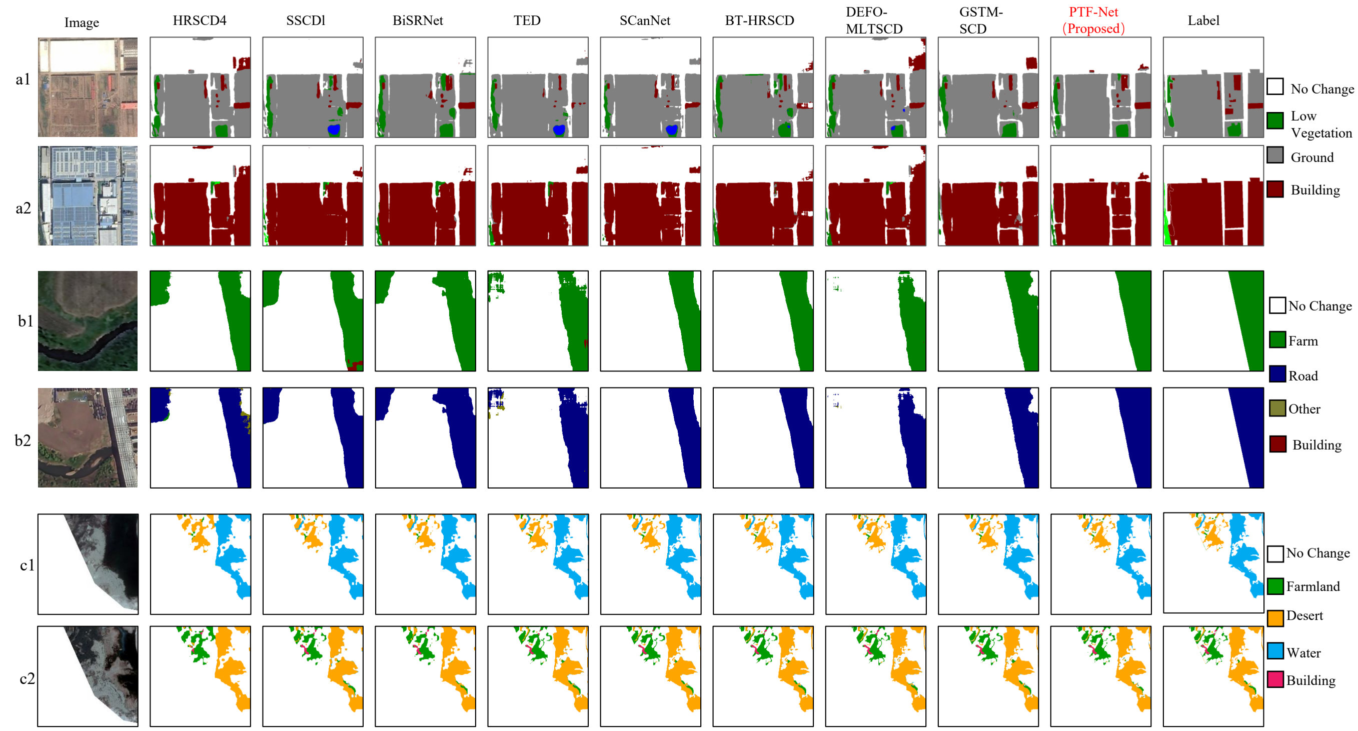


Figure 3: Qualitative visualization comparisons on three benchmark datasets. Rows (a1-a2), (b1-b2), and (c1-c2) correspond to samples from the SECOND, JL1H, and Landsat-SCD datasets, respectively. The proposed PTF-Net (red box) shows superior performance in noise suppression, detail preservation, and semantic consistency compared to other state-of-the-art methods.

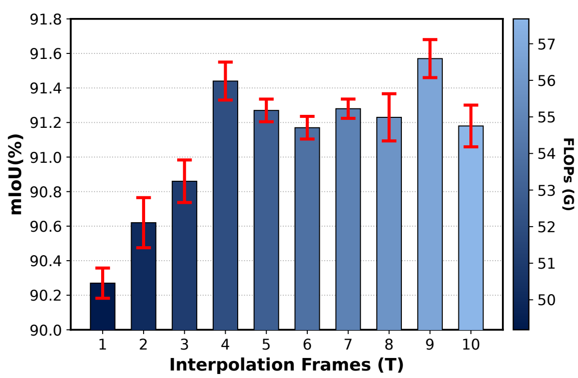


Figure 4: The impact of different pseudo-sequence lengths ($T$) on the Landsat-SCD dataset. The bar height represents $mIoU$, and the color intensity indicates FLOPs (darker is better/lower cost).The red error bars indicate the fluctuation range across multiple runs.

## 4.2 Comparison Experiments and Analysis

To verify the superiority of the proposed PTF-Net, eight representative methods are selected for comparison experiments, including HRSCD4 [Daudt *et al.*, 2019], BT-HRSCD [Fang *et al.*, 2024], DEFO-MLTSCD [Li *et al.*, 2024], SSCD-l [Ding *et al.*, 2022], BiSRNet [Ding *et al.*, 2022], TED [Ding *et al.*, 2024], SCanNet [Ding *et al.*, 2024], and GSTM-SCD [Liu *et al.*, 2025].

### Results of Quantitative Analysis

Regarding the heterogeneous SECOND dataset, Table 2 demonstrates that PTF-Net overcomes intrinsic style-structure entanglement to maintain steady improvements. By utilizing the Non-Linear Style Interpolation Mechanism to decouple stochastic style fluctuations from structural changes, our method exhibits superior robustness, highlighting its ability to distinguish semantic boundaries despite high intraclass variance and spectral diversity.

On the high-resolution JL1H dataset, PTF-Net achieves a decisive breakthrough compared to baseline methods, securing substantial improvements in the $SeK$ metric. This confirms the efficacy of the Temporal Dynamics Perception Branch in capturing minute structural mutations within complex urban scenes. By treating change as a continuous flow, our method preserves topological integrity more effectively than discrete bi-temporal comparisons.

Finally, Table 2 reveals that in the Landsat-SCD setting, characterized by drastic phenological shifts, PTF-Net significantly outperforms existing methods. Unlike baselines that often confuse seasonal changes with actual transitions, our method constructs a robust semantic filter via pseudo-temporal evolution. This allows the model to penetrate superficial seasonal interference to identify deep-seated land cover transitions, validating its effectiveness in handling complex temporal discontinuities.

| ID | Components | | | SECOND Dataset | | | JL1H Dataset | | | Landsat-SCD Dataset | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NL-SI | TSM | TGF | $mIoU$ | $Sek$ | $F_{scd}$ | $mIoU$ | $Sek$ | $F_{scd}$ | $mIoU$ | $Sek$ | $F_{scd}$ |
| M1 | | | | 72.39 | 21.30 | 61.55 | 69.24 | 26.16 | 68.32 | 86.60 | 53.54 | 86.41 |
| M2 | ✓ | | | 72.81 | 22.11 | 61.60 | 78.85 | 42.20 | 79.53 | 88.51 | 58.61 | 88.24 |
| M3 | ✓ | ✓ | | 73.29 | 22.69 | 62.47 | 81.21 | 47.26 | 82.72 | 90.24 | 63.67 | 90.03 |
| M4 | ✓ | | ✓ | 73.31 | 23.03 | 62.36 | 82.94 | 50.40 | 82.12 | 90.42 | 64.35 | 90.37 |
| M5 | ✓ | ✓ | ✓ | **73.51** | **24.21** | **64.45** | **84.58** | **54.35** | **86.22** | **91.44** | **67.45** | **91.37** |

Table 4: Ablation study of core components on three datasets. NL-SI: Non-Linear Style Interpolation; TSM: Temporal Shift Module; TGF: Temporal Gated Fusion. **Bold** indicates the best performance.

| Function | SECOND Dataset | | | JL1H Dataset | | | **Landsat-SCD Dataset** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $mIoU$ | $Sek$ | $F_{scd}$ | $mIoU$ | $Sek$ | $F_{scd}$ | $mIoU$ | $Sek$ | $F_{scd}$ |
| None | 72.39 | 21.30 | 61.55 | 69.24 | 26.16 | 68.32 | 86.60 | 53.54 | 86.41 |
| Linear | 72.75 | 22.08 | 61.61 | 78.52 | 41.50 | 79.83 | 89.58 | 61.81 | 89.48 |
| Cosine | 72.87 | 21.91 | 61.96 | 80.44 | 46.71 | 82.75 | 90.54 | 64.68 | 90.46 |
| Smoothstep | 73.23 | 22.81 | 62.06 | 82.85 | 50.61 | 84.61 | 91.16 | 66.64 | 91.15 |
| Exponential | 73.22 | 22.67 | 61.72 | 83.76 | 52.19 | 85.05 | 91.11 | 66.47 | 91.07 |
| **Sigmoid** | **73.51** | **24.21** | **64.45** | **84.58** | **54.35** | **86.22** | **91.44** | **67.45** | **91.37** |

Table 5: Comparison of different Non-Linear interpolation strategies on three datasets. **Bold** indicates the best performance.

| Shift Ratio | $mIoU$ (%) | $Sek$ (%) | $F_{scd}$ (%) |
|---|---|---|---|
| 1/2 | 82.97 | 50.48 | 84.44 |
| **1/4** | **84.58** | **54.35** | **86.22** |
| 1/8 | 83.56 | 51.62 | 84.86 |

Table 6: Ablation study on the channel shift ratio on the **JL1H** dataset. The default setting is highlighted in **bold**.

### Results of Qualitative Analysis

To intuitively evaluate our method, we visualize detection results in Fig. 3, empirically validating how PTF-Net resolves core SCD challenges. On the SECOND and JL1H datasets, baselines often succumb to high intraclass variance and inherent "information voids," leading to severe "salt-and-pepper" noise (Row a1) or fractured predictions for thin objects like roads (Row b2). In contrast, PTF-Net generates cohesive semantic regions and preserves topological integrity. This advantage stems from the synergy of our Non-Linear Style Interpolation, which recovers intermediate states to bridge temporal gaps, and the Temporal Dynamics Perception Branch, which treats image pairs as dynamic video streams to effectively decouple transient style fluctuations (e.g., shadows) from genuine structural signals.

Furthermore, on the Landsat-SCD dataset, competitors like GSTM-SCD struggle to distinguish seasonal spectral shifts from actual land-cover changes, resulting in jagged misclassifications (Row c2). Conversely, PTF-Net aligns accurately with the Ground Truth. By constructing pseudo-evolutionary sequences, our model successfully differentiates superficial phenological interference from deep-seated semantic transitions. This demonstrates that our Pseudo-Temporal Paradigm possesses superior robustness against pseudo-changes compared to discrete bi-temporal approaches, effectively mitigating the limitations of style-structure entanglement and temporal scarcity.

### Results of Computational Cost

We analyze efficiency using $512 \times 512$ inputs on an NVIDIA RTX 4090 (Table 3). Spatially, PTF-Net (25.92M) is significantly more compact than heavy-weight architectures like BT-HRSCD (33.33M) and GSTM-SCD (31.55M). Temporally, despite higher FLOPs (208.04G), it achieves a competitive 56.17 ms inference speed, nearly $1.8\times$ faster than GSTM-SCD. This efficiency arises from TSM's GPU-friendly zero-FLOP shifting, which avoids the irregular memory access inherent in GCNs, thus maximizing real-time capability without sacrificing modeling depth.

## 4.3 Ablation Experiments

### Effectiveness of Core Components

To validate the contribution of each component in PTF-Net, we conducted comprehensive ablation studies as summarized in Table 4. Integrating the Non-Linear Style Interpolation (NL-SI) module into the baseline yields a substantial performance leap across all metrics, particularly in high-resolution scenarios. This confirms that NL-SI effectively bridges the discrete temporal void by decoupling style from structure. Furthermore, the subsequent inclusion of the Temporal Shift Module (TSM) or Temporal Gated Fusion (TGF) brings continuous incremental improvements. This underscores a critical insight: perceiving the evolution is as vital as generating it. TSM enhances the model's sensitivity to local temporal dynamics, while TGF functions as a semantic filter to suppress transient noise, ensuring stability in long-span scenarios.

Ultimately, the complete architecture, which synergizes all components, achieves the optimal performance state. The decisive gains over the baseline validate our core hypothesis: reconceptualizing SCD from a discrete comparison into a continuous pseudo-multi-temporal evolution task provides a definitive solution for resolving complex spatio-temporal dynamics.

**Analysis of Interpolation Strategies**

To identify the optimal mapping for simulating land surface evolution, we benchmarked the Sigmoid strategy against linear and other Non-Linear functions. While linear interpolation offers improvements over the baseline, it falls short in capturing complex dynamics. In contrast, the Sigmoid strategy consistently secures the best performance across all settings. This superiority indicates that its S-shaped curve—modeling "gradual-rapid-gradual" phases—aligns more accurately with the physical laws governing vegetation growth and urban construction than other mathematical priors.

**Influence of Pseudo-sequence Length**

We investigate the impact of pseudo-sequence length $T$ on model performance, as illustrated in Fig. 4. Observations indicate that increasing the sequence length initially yields significant accuracy gains by enriching the temporal context. However, beyond a certain threshold, performance saturates into a plateau while computational overhead escalates substantially. This suggests that excessive sampling introduces information redundancy rather than meaningful features. Consequently, we identify $T = 4$ as the optimal trade-off, effectively balancing superior detection accuracy with computational efficiency to avoid diminishing returns.

**Influence of Shift Ratio**

We investigate the influence of TSM shift ratios to balance temporal modeling and spatial preservation. Results indicate that excessive shifting disrupts spatial feature integrity, whereas insufficient shifting constrains necessary temporal interaction, leading to sub-optimal outcomes. Consequently, the model achieves peak performance with a $1/4$ ratio. This configuration represents the most effective trade-off between robust temporal exchange and spatial detail retention, and is thus adopted as the default. This empirically confirms that controlled temporal mixing is essential for capturing evolutionary dynamics without overwhelming the original semantic representation.

## 5 Conclusion

In this paper, we propose PTF-Net. In contrast to the static bi-temporal paradigm that views images as isolated snapshots, our framework reconceptualizes Semantic Change Detection (SCD) as a continuous evolution modeling task. By synergizing Non-Linear style interpolation with temporal dynamics perception, we effectively resolve the intrinsic temporal void, reliably distinguishing genuine structural mutations from transient noise. Experiments validate PTF-Net's superior accuracy and efficiency, offering a transformative approach that reshapes the foundational paradigm of SCD and inspires future research in continuous evolution perception. Recognizing the training memory overhead and its sensitivity to extreme temporal intervals, future endeavors will focus on adaptive interpolation for long-term monitoring and heterogeneous multi-sensor fusion.

## References


[Bandara and Patel, 2022] Wele G. Chaminda Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 207–210, Kuala Lumpur, Malaysia, July 2022. IEEE.

[Chen and Jiang, 2025] Ming Chen and Wanshou Jiang. SCDVit: Semantic change detection based on Sam-Vit and semantic consistency. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:2269–2282, 2025.

[Chen *et al.*, 2021] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2021.

[Dai *et al.*, 2024] Anjin Dai, Jianyu Yang, Tingting Zhang, Bingbo Gao, Kaixuan Tang, and Xinyue Chen. DADR-HCD: A deep domain adaptation and disentangled representation network for unsupervised heterogeneous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Daudt *et al.*, 2018] Rodrigo C. Daudt, Bertrand L. Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *Proceedings of the 25th IEEE International Conference on Image Processing*, pages 4063–4067, Athens, Greece, October 2018. IEEE.

[Daudt *et al.*, 2019] Rodrigo C. Daudt, Bertrand L. Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019.

[Ding *et al.*, 2022] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[Ding *et al.*, 2024] Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Doxani *et al.*, 2012] Georgia Doxani, Konstantinos G. Karantzalos, and Maria Tsakiri-Strati. Monitoring urban changes based on scale-space filtering and object-oriented classification. *International Journal of Applied Earth Observation and Geoinformation*, 15:38–48, 2012.

[Fang *et al.*, 2024] Sheng Fang, Wen Li, Shuqi Yang, Zhe Li, Jianli Zhao, and Xiaoxin Wang. BT-HRSCD: High-resolution feature is what you need for a semantic change detection network with a triple-decoding branch. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Fang *et al.*, 2025] Sheng Fang, Wen Li, Yue Song, Zhe Li, and Jianli Zhao. Rethinking semantic change detection from a semantic alignment perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, Nevada, USA, June 2016. IEEE.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, Venice, Italy, October 2017. IEEE.

[Hussain *et al.*, 2013] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 80:91–106, 2013.

[Jiang *et al.*, 2023] Liangcun Jiang, Feng Li, Li Huang, Feifei Peng, and Lei Hu. TTNet: A temporal-transform network for semantic change detection based on bitemporal remote sensing images. *Remote Sensing*, 15(18):4555, 2023.

[Li *et al.*, 2018] Yu Li, Sandro Martinis, Simon Plank, and Ralf Ludwig. An automatic change detection approach for rapid flood mapping in Sentinel-1 SAR data. *International Journal of Applied Earth Observation and Geoinformation*, 73:123–135, 2018.

[Li *et al.*, 2024] Zhe Li, Xiaoxin Wang, Sheng Fang, Jianli Zhao, Shuqi Yang, and Wen Li. A decoder-focused multitask network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Li *et al.*, 2025] Zhihao Li, Yikun Liu, Minghao Liu, and Gongping Yang. Enhancing collaboration and mitigating conflict between subtasks for remote sensing semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, Seoul, Korea, October 2019. IEEE.

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

[Liu *et al.*, 2024] Wei Liu, Ziwen Kang, Jiawei Liu, Yiyuan Lin, Yongtao Yu, and Jonathan Li. A multitask CNN-transformer network for semantic change detection from bitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Liu *et al.*, 2025] Xuanguang Liu, Chenguang Dai, Lei Ding, Zhenchao Zhang, Yujie Li, Xibing Zuo, Mengmeng Li, Hanyun Wang, and Yuzhe Miao. GSTM-SCD: Graph-enhanced spatio-temporal state space model for semantic change detection in multi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:73–91, 2025.

[Long *et al.*, 2024] Jiang Long, Mengmeng Li, Xiaoqin Wang, and Alfred Stein. Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:318–335, 2024.

[Long *et al.*, 2025] Jiang Long, Sicong Liu, Mengmeng Li, Hang Zhao, and Yanmin Jin. BGSNet: A boundary-guided Siamese multitask network for semantic change detection from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225:221–237, 2025.

[Lu *et al.*, 2004] Dengsheng Lu, Paul Mausel, Eduardo S. Brondízio, and Emilio Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, January 2004.

[Lv *et al.*, 2018] Zhi Yong Lv, Wenzhong Shi, Xiaokang Zhang, and Jón Atli Benediktsson. Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5):1520–1532, 2018.

[Mou and Zhu, 2018] Lichao Mou and Xiao X. Zhu. A recurrent convolutional neural network for land cover change detection in multispectral images. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4363–4366, Valencia, Spain, July 2018. IEEE.

[Niu *et al.*, 2023] Yiting Niu, Haitao Guo, Jun Lu, Lei Ding, and Donghang Yu. SMNet: Symmetric multi-task network for semantic change detection in remote sensing images based on CNN and Transformer. *Remote Sensing*, 15(4):949, 2023.

[Pitts and So, 2017] Dilkushi A. De Alwis Pitts and Emily So. Enhanced change detection index for disaster response, recovery assessment and monitoring of accessibility and open spaces (camp sites). *International Journal of Applied Earth Observation and Geoinformation*, 57:49–60, 2017.

[Rokni *et al.*, 2015] Komeil Rokni, Anuar Ahmad, Karim Solaimani, and Sharifeh Hazini. A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques. *International Journal of Applied Earth Observation and Geoinformation*, 34:226–234, 2015.

[Tang *et al.*, 2024] Kai Tang, Fei Xu, Xuehong Chen, Qi Dong, Yuheng Yuan, and Jin Chen. The ClearSCD model: Comprehensively leveraging semantics and change relationships for semantic change detection in high spatial resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211:299–317, 2024.

[Tang *et al.*, 2025] Yingjie Tang, Shou Feng, Chunhui Zhao, Yongqi Chen, Zhiyong Lv, and Weiwei Sun. A semantic change detection network based on boundary detection and task interaction for high-resolution remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):17184–17198, 2025.

[Tian *et al.*, 2025] Shiqi Tian, Ailong Ma, Zhuo Zheng, Xicheng Tan, and Yanfei Zhong. Learning temporal consistency for high spatial resolution remote sensing imagery semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–25, 2025.

[Wang *et al.*, 2023] Chengyuan Wang, Shenghui Liu, Shuo Z. Zhou, Jian Zhou, Shichao Jiang, Yongkang Zhang, Tiantian Feng, Hanliang Zhang, Yuhang Zhao, Zhaoquan Lai, Shuai Cui, and Xuegang Mao. Spatial-temporal patterns of urban expansion by land use/ land cover transfer in China. *Ecological Indicators*, 155:111009, 2023.

[Wang *et al.*, 2024] Qi Wang, Wei Jing, Kaichen Chi, and Yuan Yuan. Cross-difference semantic consistency network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

[Wang *et al.*, 2025] Yinqing Wang, Liangjun Zhao, Yueming Hu, Hui Dai, and Yuanyang Zhang. Multitask semantic change detection guided by spatiotemporal semantic interaction. *Scientific Reports*, 15:16003, May 2025.

[Xiang *et al.*, 2021] Shao Xiang, Mi Wang, Xiaofan Jiang, Guangqi Xie, Zhiqi Zhang, and Peng Tang. Dual-task semantic change detection for remote sensing images using the generative change field module. *Remote Sensing*, 13(16):3336, 2021.

[Yang *et al.*, 2024] Wanying Yang, Haozhe Zhou, Yu Li, and Wenbo Xu. Semi-supervised semantic change detection based on weak-to-strong consistency. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 10241–10244, Athens, Greece, July 2024. IEEE.

[Zhai *et al.*, 2021] Han Zhai, Chaoquun Lv, Wanzeng Liu, Chao Yang, Dasheng Fan, Zikun Wang, and Qingfeng Guan. Understanding spatio-temporal patterns of land use/land cover change under urbanization in Wuhan, China, 2000–2019. *Remote Sensing*, 13(16):3331, 2021.

[Zhang *et al.*, 2018] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, Salt Lake City, Utah, USA, June 2018. IEEE.

[Zhou *et al.*, 2025] Ziyu Zhou, Keyan Hu, Yutian Fang, and Xiaoping Rui. SChanger: Change detection from a semantic change and spatial consistency perspective. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:10186–10203, 2025.