Bridging Semantic Filtering and Structural Alignment for Remote Sensing Change Detection

Ziming Song, Xueqin Liu, Jingpeng Yu, and Guocheng Yan*
Fuyang Normal University
China
{szm0371, lxq0312, 2023114966}@stu.fynu.edu.cn, yanguocheng54@fynu.edu.cn

***Abstract*—Remote sensing change detection aims to identify semantic changes by analyzing bi-temporal images. However, precision is constrained by semantic interference from seasonal or lighting variations and pseudo-changes caused by structural misalignments like viewpoint shifts. Existing methods often struggle to balance these issues, either employing expensive attention mechanisms to suppress noise or overlooking boundary false positives resulting from geometric deviations. In this paper, we propose a novel framework bridging semantic filtering and structural alignment. First, to overcome semantic interference, we introduce the Mono-temporal Semantic Purification (MSP) method. It synergizes a Spectral-Aware Spatial Filtering strategy to explicitly suppress background redundancy with a Semantic Filtering-Aware Enhancement (SFAE) mechanism. Unlike standard attention, SFAE leverages a filtering-aware prior to guide efficient global context modeling, enhancing semantic consistency while preserving high-frequency details. Second, to rectify geometric deviations, we propose the Bi-temporal Graph Alignment (BTGA) module. By constructing bidirectional KNN graphs in non-Euclidean space, this module dynamically registers bi-temporal features. Furthermore, a secondary purification strategy is applied to the fused difference representations, effectively eliminating boundary artifacts to ensure precise change delineation. Extensive experiments on the LEV, Google, and BCDD datasets demonstrate that our method outperforms state-of-the-art approaches, achieving superior robustness and accuracy with IoU scores of 79.50%, 77.12%, and 81.84%, respectively.**

***Index Terms*—Remote Sensing Change Detection, Semantic Filtering, Structural Alignment.**

I. Introduction

Remote Sensing Change Detection (RSCD) is designed to discern semantic changes in surface features by analyzing bi-temporal images acquired from the same geographical region [1]. As a pivotal task in Earth observation, RSCD is instrumental in various critical applications, including urban expansion monitoring, disaster assessment, resource management, and ecological preservation. The rapid advancement of high-resolution satellite technology has yielded a vast amount of high-quality Earth observation data, facilitating more fine-grained monitoring of surface dynamics. Concurrently, this abundance of data necessitates more stringent standards regarding the precision and robustness of change detection algorithms [1]–[4].

In recent decades, various change detection methods have been introduced. Traditional approaches, relying on algebraic operations or hand-crafted features, offered simplicity but lacked robustness against noise in complex environments.

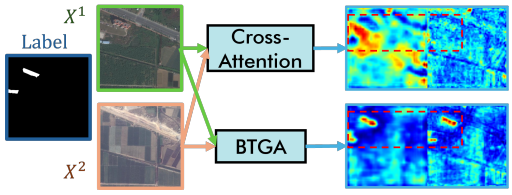


Fig. 1. Visual comparison with two level feature embedding of change-aware heatmaps in complex scenes using the proposed BTGA and cross-attention mechanisms.

Conversely, the recent emergence of deep learning has positioned CNN- and Transformer-based methods as the mainstream solution. Thanks to their superior feature extraction and context modeling capabilities, these models can derive deep semantic representations, resulting in substantially higher detection accuracy than their traditional counterparts [1], [2].

While deep learning has revolutionized change detection, its performance in complex high-resolution scenarios remains constrained by two critical issues. First, semantic interference leads to distributional shifts. Seasonal and lighting variations inject noise, resulting in divergent latent distributions across bi-temporal data [1], [2], [5], [6]. While current methods employ global modeling to mitigate this, they predominantly utilize heavy architectures with dense point-to-point attention mechanisms. This design not only imposes a heavy computational burden but also struggles to achieve efficient global interaction while filtering noise. Second, structural misalignment induces boundary false positives. Current approaches typically assume perfect registration [3]–[5], [7]. Yet, viewpoint shifts and topographic variations frequently cause geometric disparities. This misalignment can deceive RSCD models into flagging displaced but unchanged objects (such as vehicles) as true changes. Crucially, conventional methods operating in Euclidean space rely heavily on rigid pixel-wise alignment or fixed receptive-field patch matching, and thus fail to explicitly account for geometric distortions [3]. The intuitive visual comparison of the heatmaps generated by the proposed BTGA and cross-attention mechanisms is shown in Fig. 1.

In this paper, we propose a novel framework comprising Mono-temporal Semantic Purification (MSP) and Bi-temporal Graph Alignment (BTGA) to address semantic interference and structural misalignment in remote sensing change detection. First, to overcome semantic interference in complex environments, we design the Mono-temporal Semantic Purification (MSP) method, which adopts a "clean-then-enhance" paradigm. Specifically, this module incorporates a

Spatial Filtering Strategy that utilizes Spectral-Aware Group Normalization to explicitly decouple features into informative and redundant branches, effectively filtering background redundancy and seasonal noise. To further concentrate the denoising process at the semantic level, we introduce the Semantic Filtering-Aware Enhancement (SFAE) mechanism. Unlike simple sparse feedback, SFAE initializes semantic agents guided by a lightweight change prior map to perform efficient global context modeling. This mechanism captures data-driven global semantic distributions with linear computational complexity, effectively avoiding pseudo-changes induced by dense global interactions while preserving high-frequency structural details via edge-aware refinement. Second, to mitigate pseudo-changes stemming from structural misalignment, we develop the Bi-temporal Graph Alignment (BTGA) method. Transcending the constraints of Euclidean space, this module operates at the deepest feature level to construct bidirectional KNN graphs, which dynamically register bi-temporal features based on semantic similarity. This approach effectively corrects geometric deviations caused by viewpoint shifts, enabling the accurate discrimination of true changes from misaligned artifacts. Furthermore, we extend the explicit spatial filtering strategy by re-applying the Spatial Reconstruction Strategy (SRS) to the fused difference features. By performing this "secondary purification," we effectively eliminate residual boundary pseudo-changes caused by registration errors, ensuring the precise delineation of change regions.

II. Method

Fig. 2 illustrates the overall pipeline. For feature extraction and change generation (Section II-A and Section II-D), we explicitly denote the feature units with their respective level and time step subscripts. Conversely, for mono-temporal semantic enhancement (Section II-B) and bi-temporal interaction (Section II-C), we omit level subscripts as these operations are universally applied across all feature levels.

A. Siamese Mono-temporal Feature Encoder

To extract mono-temporal features using a Siamese architecture, let the input bi-temporal image pair be denoted as $X^1, X^2 \in \mathbb{R}^{H\times W\times 3}$. We employ a backbone network as the feature extraction function Φ_θ. By adopting a weight-sharing strategy, both branches share identical parameters θ. Consequently, the feature extraction process can be formulated as:

$$\mathcal{F}^t = \Phi_\theta(X^t) = \{F_l^t \mid l = 1, 2, 3, 4\}, \quad t \in \{1, 2\} \tag{1}$$

where, t denotes the time step or branch index. $\mathcal{F}^t$ is the multi-scale feature set corresponding to the t-th input image. F_l^t represents the feature map at the l-th level of the t-th branch. Specifically, the dimensions of these features are given by $F_l^t \in \mathbb{R}^{C_l\times \frac{H}{2^{l+1}}\times \frac{W}{2^{l+1}}}$, where the channel dimensions corresponding to levels $l = 1, 2, 3, 4$ are $C_l \in \{32, 64, 128, 256\}$, respectively.

B. Mono-temporal Semantic Purification

To mitigate semantic interference caused by seasonal or lighting variations, we propose the Mono-temporal Semantic Purification (MSP) method. This method incorporates an explicit spatial filtering strategy to suppress background redundancy and seasonal noise [5], [6], [8], while employing a computation-efficient global interaction mechanism to apply this purification process to high-level abstract semantics. As illustrated in Fig. 2, MSP adopts a "clean-then-enhance" paradigm, sequentially applying spatial filtering for explicit noise removal and an efficient attention-based mechanism for global context modeling.

1) Spatial Filtering Strategy: The Spatial Filtering Strategy suppresses spatial redundancy—such as background noise, seasonal variations, and sensor artifacts—while preserving semantically informative content. Let $F_{in} \in \mathbb{R}^{C\times H\times W}$ denote an encoder feature from the mono-temporal hierarchy $\mathcal{F}^t$ (Eq. 1). For brevity, layer and temporal indices are omitted, as the design is universally applicable.

Standard Group Normalization (GN) groups channels arbitrarily, often mixing spectrally distinct bands (e.g., RGB and NIR) and compromising statistical consistency. To preserve spectral integrity, we propose *Spectral-Aware Group Normalization*, which partitions channels into predefined spectral groups $\mathcal{G} = \{g_1, \ldots, g_k\}$ and normalizes each group independently. A sigmoid then converts the output into a spatial importance mask:

$$W_{gate} = \sigma(\mathrm{GN}_{\mathcal{G}}(F_{in})) \tag{2}$$

where $\mathrm{GN}_{\mathcal{G}}$ denotes the group normalization applied adaptively to the pre-defined spectral partitions. $W_{gate} \in \mathbb{R}^{C\times H\times W}$ serves as the importance mask.

Based on the spatial attention map W_{gate} , we apply a threshold-based decoupling strategy to split features into an informative branch and a residual background branch. Unlike standard soft attention, this approach explicitly retains high-confidence regions in one branch while suppressing them in the other. Specifically, we define a binary mask $M \in \{0, 1\}^{C\times H\times W}$ using a threshold $\tau = 0.5 : M = \mathbb{I}(W_{gate} > \tau)$, where $\mathbb{I}(\cdot)$ is the indicator function. The input features are thus decoupled into an informative component F_{info} and a redundant component F_{noise}, formulated as:

$$F_{info} = M \odot F_{in}, \quad F_{noise} = (1 - M) \odot F_{in} \tag{3}$$

where $\odot$ denotes element-wise multiplication.

Since W_{gate} encodes spatial importance via channel-wise normalization, we use it to modulate the two decoupled representations—enhancing salient semantics in F_{info} and suppressing background noise and stylistic variations in F_{noise}. The process is formulated as:

$$F_{info} = F_{info} \odot W_{gate}, \quad F_{noise} = F_{noise} \odot W_{gate} \tag{4}$$

Finally, to facilitate information interaction between the decoupled streams and prevent the gradient vanishing problem caused by hard suppression, we design a Cross-Channel Reconstruction strategy. We split both the modulated informative

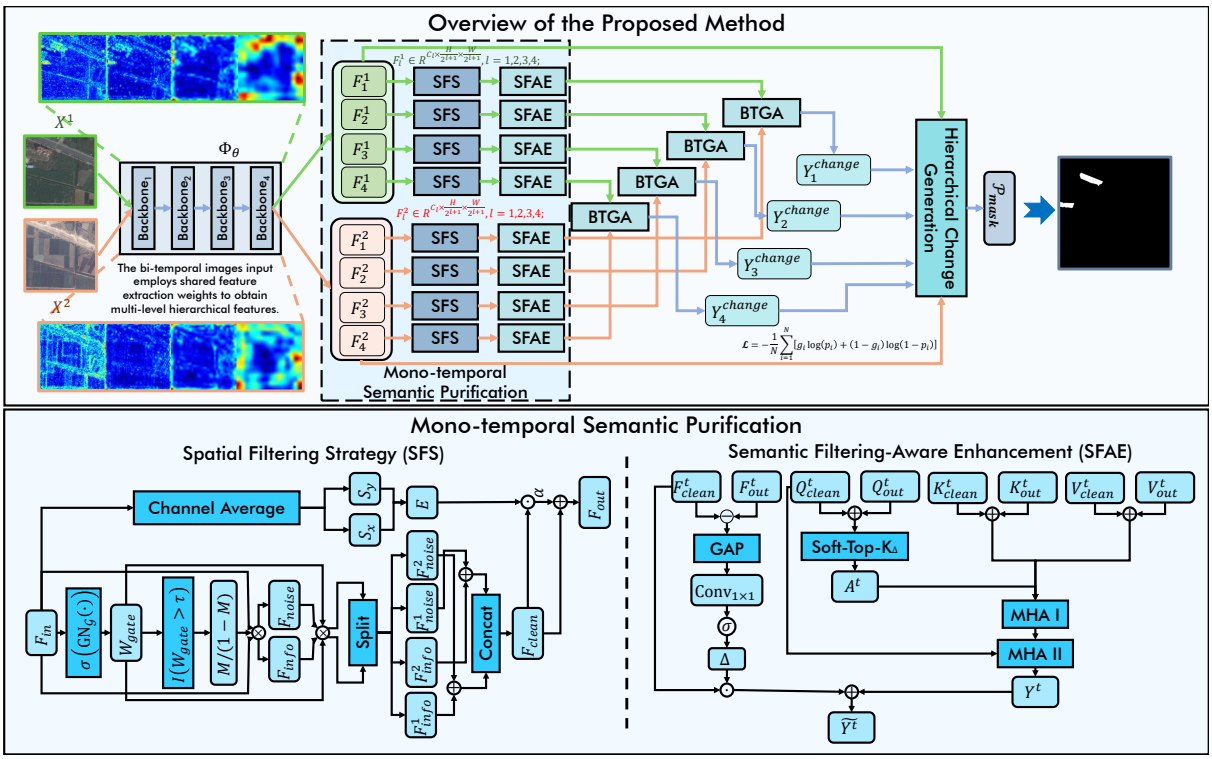


Fig. 2. Overview of the proposed framework. The upper part illustrates the overall pipeline, which comprises a Siamese encoder, the Mono-temporal Semantic Purification (MSP) module, the Bi-temporal Graph Alignment (BTGA) module, and a hierarchical decoder. The lower part details the two key components of MSP: the Spatial Filtering Strategy (SFS) for noise suppression and the Semantic Filtering-Aware Enhancement (SFAE) for global context modeling.

feature F_{info} and the suppressed noise feature F_{noise} into two subsets along the channel dimension C_l (Eq. 1), where l denotes the index of the current feature level. These subsets are denoted as $\{F^1_{info}, F^2_{info}\}$ and $\{F^1_{noise}, F^2_{noise}\}$, respectively. Then, a cross-fusion operation is performed to recombine these subsets, ensuring that the contextual information from the background branch aids the semantic representation, while the strong semantic cues help refine the background details. The final spatially purified feature F_{clean} is generated via concatenation:

$$F_{clean} = \text{Concat}(F^1_{info} + F^2_{noise}, ; F^2_{info} + F^1_{noise}) \tag{5}$$

where $\text{Concat}(\cdot)$ represents the concatenation operation along the channel dimension. Through this split-and-add mechanism, the module effectively suppresses spatial redundancy while maintaining the integrity of the feature space.

Despite the effective suppression of noise, the spatial filtering process may inadvertently weaken high-frequency structural details (e.g., boundaries of small targets). To address this, we introduce an edge-aware refinement mechanism as the final stage. First, we compute a global edge map $E \in \mathbb{R}^{1\times H\times W}$ by applying fixed Sobel operators (S_x, S_y) to the input feature:

$$E = \sqrt{(S_x * \bar{F}_{in})^2 + (S_y * \bar{F}_{in})^2} \tag{6}$$

where $\bar{F}_{in}$ denotes the channel-averaged input feature. To restore structural integrity, this edge map is injected back into the purified features via a learnable residual connection.

The definitions of S_x and S_y are as follows:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{7}$$

The final output F_{out} is formulated as:

$$F_{out} = F_{clean} + \alpha \cdot E \odot F_{clean} \tag{8}$$

where $\odot$ denotes element-wise multiplication, and α is a learnable parameter (initialized to 0.2) that adaptively controls the intensity of edge enhancement. This ensures that the network focuses on semantic purification while retaining critical geometric cues.

2) Semantic Filtering-Aware Enhancement: To capture subtle, localized changes while leveraging the benefits of the preceding purification stage, we propose a Semantic Filtering-Aware Enhancement (SFAE) mechanism. Unlike generic global attention, SFAE explicitly utilizes the spatially purified features to construct a filtering-aware prior. This prior guides semantic agents to focus on regions with high semantic discrepancy, thereby performing hierarchical interactions that preserve spatial fidelity.

Let F^t_{clean} and F^t_{out} denote the spatially purified features from two phases (output from the MSP module), both of shape $\mathbb{R}^{C\times H\times W}$. We first compute a lightweight *filtering-aware prior map* $\Delta \in \mathbb{R}^{H\times W}$. This map is derived from the channel-wise absolute difference of the purified features, processed by a 1×1 convolution and Sigmoid activation:

$$\Delta = \sigma\left(\text{Conv}_{1\times1}\left(\text{GAP}(|F^t_{clean} - F^t_{out}|)\right)\right) \tag{9}$$

where GAP denotes Global Average Pooling over channels, and σ is the Sigmoid function. This map effectively highlights regions with high inter-temporal discrepancy after semantic filtering.

The flattened features $X^t_{clean}, X^t_{out} \in \mathbb{R}^{N\times C}$ $(N = HW)$ are linearly projected into queries Q, keys K, and values V as in standard attention. Crucially, agent tokens are no longer derived from average pooling alone. Instead, we initialize N_A agents by sampling spatial locations weighted by the filtering-aware prior Δ, implemented via differentiable top-k or soft selection:

$$A = \text{Soft-Top-K}_{\Delta}(Q^t_{clean} \oplus Q^t_{out},\, k = N_A) \in \mathbb{R}^{N_A\times C} \quad (10)$$

where $\oplus$ denotes addition, and Soft-Top-K$_\Delta$ selects high-discrepancy regions in a differentiable manner based on the filtering outcomes.

Global context modeling proceeds in two stages. First, agents aggregate information from both temporal keys:

$$V_{agent} = \text{Softmax}\left(\frac{A(K^t_{clean} \oplus K^t_{out})^T}{\sqrt{d_k}}\right)(V^t_{clean} \oplus V^t_{out}) \quad (11)$$

Second, the refined agent representation broadcasts back to each time phase separately, producing enhanced features Y^t:

$$Y^t = \text{Softmax}\left(\frac{Q^t_{clean}A^T}{\sqrt{d_k}}\right)V_{agent} \quad (12)$$

Finally, to prevent over-smoothing of change boundaries, we fuse the enhanced features with the original purified inputs via a residual connection modulated by the filtering-aware prior Δ:

$$\tilde{Y}^t = Y^t + \Delta \odot F^t_{clean} \quad (13)$$

This ensures that high-change regions retain local details while benefiting from global semantics context.

C. Bi-temporal Graph Alignment for Semantic Interaction

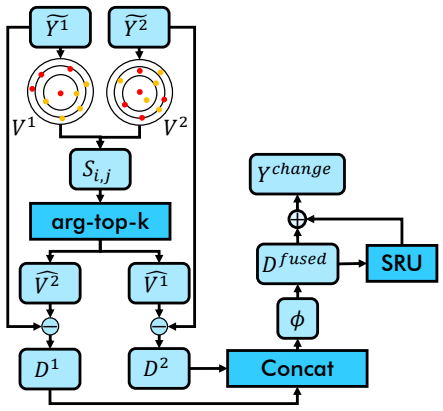


Fig. 3. Illustration of the Bi-temporal Graph Alignment (BTGA).

Traditional Euclidean subtraction struggles with structural misalignment caused by viewing angles in high-resolution imagery. To address this, we propose the Bi-temporal Graph Alignment (BTGA) module. Operating in a non-Euclidean domain, BTGA utilizes a bidirectional graph to explicitly align features via semantic similarity and refines differences through 'secondary purification (Fig. 3).

We flatten the purified features $\tilde{F}^1, \tilde{F}^2$ into sequences $V^1, V^2 \in \mathbb{R}^{C\times N}$ and apply alignment exclusively at the deepest layer ($l = 4$) for efficiency. By constructing bidirectional KNN graphs, we link semantically similar pixels across time to correct geometric misalignments. This high-level structural guidance implicitly refines the entire feature hierarchy during training.

Taking the $V^2 \to V^1$ alignment as an example, for each node $v^1_i \in V^1$, we retrieve its k-nearest neighbors in V^2 based on cosine similarity [9]. The semantic similarity $S_{i,j}$ is defined as:

$$S_{i,j} = \frac{(v^1_i)^\top v^2_j}{|v^1_i||v^2_j|} \quad (14)$$

The index set of the k-nearest neighbors for v^1_i, denoted as $\mathcal{N}_k(v^1_i)$, is then formally defined as:

$$\mathcal{N}_k(v^1_i) = \underset{j\in 1,\dots,N}{\arg\,\text{top-k}}(S_{i,j}) \quad (15)$$

Based on this neighborhood, the aligned feature $\hat{v}^2_i$ is synthesized via an aggregation function:

$$\hat{v}^2_i = \frac{1}{k}\sum_{j\in\mathcal{N}_k(v^1_i)} v^2_j \quad (16)$$

where v^2_j represents the feature representation of the corresponding neighbor in V^2. Effectively, this aggregation acts as a semantic transformation that warps the T_2 features in the latent domain to spatially register with T_1. By symmetry, in Direction 2, we synthesize $\hat{V}^1$ by aligning V^1 according to the topology of V^2.

Once the features are aligned in the latent space, we compute the absolute difference maps in both directions to capture the initial evidence of change:

$$D^1 = |V^1 - \hat{V}^2|, \quad D^2 = |V^2 - \hat{V}^1| \quad (17)$$

where $|\cdot|$ denotes the element-wise absolute difference. Here, D^1 represents the changes detected from the perspective of T_1, while D^2 represents those from T_2.

To synthesize these complementary cues, we concatenate them along the channel dimension and employ a fusion convolutional layer ϕ_{fuse} (comprising a 1×1 convolution, BN, and ReLU) to reduce the dimension back to C:

$$D^{fused} = \phi_{fuse}(\text{Concat}(D^1, D^2)) \quad (18)$$

While graph alignment corrects major geometric shifts, subtle boundary artifacts may remain. To address this, we re-apply the Spatial Reconstruction Strategy (SRS) for a "secondary purification" on the difference features. A residual connection is then added to preserve the original change magnitude:

$$Y^{change} = \text{SRU}(D^{fused}) + D^{fused} \quad (19)$$

This output, Y^{change}, serves as a robust difference representation, effectively suppressing pseudo-changes caused by registration errors while highlighting genuine semantic variations.

D. Hierarchical Change Generation

We employ a multi-level upsampling decoder that progressively aggregates multi-scale features. Crucially, the change representation Y_l^{change} acts as a structural guide, modulating the bi-temporal features F_l^1 and F_l^2 at each hierarchy level.

Specifically, the interaction between the change signals and mono-temporal representations is facilitated by a feature enhancement module $\text{MLP}^l(\cdot)$:

$$\begin{aligned} \hat{X}_l^1 &= \text{MLP}^l([X_l^1 \parallel Y_l^{change}]) \\ \hat{X}_l^2 &= \text{MLP}^l([X_l^2 \parallel Y_l^{change}]) \end{aligned} \tag{20}$$

where $[\cdot \parallel \cdot]$ denotes the concatenation operation along the channel dimension.

Subsequently, a fused difference representation $\mathcal{H}_l$ is synthesized through a residual summation to maintain semantic integrity:

$$\mathcal{H}_l = \hat{X}_l^1 + \hat{X}_l^2 + Y_l^{change} \tag{21}$$

To integrate these multi-level cues, the decoder adopts a top-down cascaded approach. Let M^l represent the decoded feature at level l. For the deepest level ($l = 4$), the feature is directly processed, while for subsequent levels ($l \in \{3, 2, 1\}$), the features from the higher level are upsampled and fused with the current level:

$$M^l = \begin{cases} \mathcal{C}^l(\mathcal{H}^l), & l = 4 \\ \mathcal{C}^l(\mathcal{H}^l + \text{Up}(M^{l+1})), & l \in \{3, 2, 1\} \end{cases} \tag{22}$$

where $\mathcal{C}^l$ denotes a convolutional block and $\text{Up}(\cdot)$ indicates bilinear interpolation.

Finally, the most refined feature map M^1 is mapped to the output space to generate the final change prediction $\mathcal{P}_{mask}$:

$$\mathcal{P}_{mask} = \text{MLP}^{final}(M^1) \tag{23}$$

To supervise the training process, we employ the Binary Cross-Entropy (BCE) loss to measure the discrepancy between the prediction $\mathcal{P}_{mask}$ and ground truth $G \in \{0, 1\}^{H \times W}$:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \tag{24}$$

where $N = H \times W$ is the total number of pixels, and g_i, p_i denote the ground truth and predicted probability for pixel i, respectively. Minimizing $\mathcal{L}$ enables the network to suppress pseudo-changes while accentuating genuine semantic variations.

III. Experiment

A. Experiment Setting

Datasets: Our experiments are performed on three public datasets: GoogleBuilding [10], LEVCD [11], and BCDD [12], where all images are cropped to 512×512 pixels for training and testing. *Implementation Details:* Our framework is implemented in PyTorch and trained on a single NVIDIA RTX 5090 GPU. We utilize the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 8, training for a total of 100 epochs. During training, standard data augmentation techniques, including random cropping, are employed to enhance robustness. *Baselines:* To validate the effectiveness of our approach, we benchmark it against a comprehensive set of state-of-the-art methods, including FCEF [1], FCSiam [1], DGMAANet [3], UNLLNet [13], BIFA [6], ELGCNet [14], RHighNet [2], STRobustNet [4], DAMFANet [3], ECTI [15] and STADECDNet [5]. *Evaluation Metrics:* Performance is assessed using four standard metrics: F1-score (%), Intersection over Union (IoU, %), Precision (%), and Recall (%). These metrics are selected for their robustness in handling class imbalance and their ability to simultaneously evaluate classification accuracy and geometric integrity. For the detailed mathematical definitions of these metrics, we follow the protocols described in [2].

B. Result Performance Quantitative Comparison

To comprehensively evaluate the effectiveness of the proposed method, we conduct quantitative comparisons against current state-of-the-art approaches on three widely used remote sensing change detection benchmarks: LEV, Google, and BCDD. The results reported in Table I summarize the performance of all competing models in terms of four standard metrics: Intersection over Union (IoU), F1-score, Precision (Pre), and Recall (Rec).

On the LEV dataset, our method achieves an IoU of 79.50% and an F1-score of 88.58%, significantly outperforming existing methods. Although ECTI demonstrates competitive results with an IoU of 79.31% and an F1-score of 88.46%, it exhibits a notable imbalance between precision 86.44% and recall 90.57%. In contrast, our approach maintains a high recall 88.55% while achieving even higher precision 88.61%, indicating a superior trade-off between detection completeness and localization accuracy. While DAMFANet and MSCANet also deliver strong overall performance, they fall slightly short in capturing fine-grained changes.

On the more challenging Google dataset, our method again attains the best results with an IoU of 77.12% and an F1-score of 87.08%. Notably, STADECDNet achieves the highest recall 93.08% but suffers from low precision 72.95%, resulting in a substantially lower F1-score 81.80%—a clear indication of excessive false positives. Our method, by comparison, achieves a high recall of 90.02% while simultaneously improving precision to 84.33%, effectively mitigating the noise typically associated with aggressive recall-oriented strategies. Moreover, although BIFA and MSCANet exhibit reasonable precision, their limited recall 74.46% and 78.58%, respectively, compromises their ability to ensure comprehensive change coverage, which is critical for practical applications.

On the BCDD dataset, our method further demonstrates exceptional generalization capability, setting a new state-of-the-art with an IoU of 81.84% and an F1-score of 90.01%. BIFA achieves the second-highest precision 91.33% but at the cost of low recall 84.09%, suggesting a conservative prediction behavior that tends to miss subtle changes. STADECDNet, while achieving strong recall 89.52%, is constrained by moderate precision 84.80%, yielding an F1-score of only 87.09%. In

TABLE I

QUANTITATIVE COMPARISON RESULTS ON THREE DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE. FOR THE SAKE OF BREVITY, PRECISION (%) AND RECALL (%) ARE DENOTED AS PRE AND REC.

Model	LEV				Google				BCDD				Efficiency
	IoU(%)	F1(%)	Pre(%)	Rec(%)	IoU(%)	F1(%)	Pre(%)	Rec(%)	IoU(%)	F1(%)	Pre(%)	Rec(%)	Flops(G)
FCEF [1]	26.63	42.05	26.84	**97.04**	52.09	68.50	76.44	62.06	69.39	81.93	78.33	85.88	12.44
FCSiam [1]	36.23	53.19	37.32	92.51	51.03	67.58	77.09	60.16	67.24	80.41	74.85	86.86	12.44
DAMFANet [3]	76.96	86.98	86.94	87.01	72.97	84.37	85.90	82.89	77.32	87.21	87.78	86.64	45.47
RHighNet [2]	72.90	84.33	82.10	86.67	64.23	78.22	83.26	73.75	74.22	85.20	84.09	86.35	64.89
STRobustNet [4]	73.75	84.89	83.15	86.70	65.35	79.05	81.03	77.16	72.59	84.12	85.82	82.48	13.21
UNLLNet [13]	65.33	79.03	74.46	84.20	72.62	84.14	82.52	85.83	81.02	89.51	**93.19**	86.12	49.24
ECTI [15]	79.31	88.46	86.44	90.57	71.95	83.68	84.17	83.20	79.48	88.56	88.65	88.48	43.52
BIFA [6]	76.87	86.92	86.25	87.60	65.71	79.31	84.83	74.46	77.87	87.56	91.33	84.09	42.31
STADECDNet [5]	76.42	86.63	80.20	94.19	69.20	81.80	72.95	**93.08**	77.14	87.09	84.80	89.52	103.83
Ours	**79.50**	**88.58**	**88.61**	88.55	**77.12**	**87.08**	84.33	90.02	**81.84**	**90.01**	90.10	**89.93**	34.00

TABLE II

ABLATION STUDY RESULTS ON LEV, GOOGLE, AND BCDD DATASETS.

Setting			LEV		Google		BCDD	
SFS	SFAE	BTGA	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)
✓	×	×	76.24	86.52	75.13	85.80	78.68	88.07
×	✓	×	76.33	86.58	74.17	85.17	79.50	88.58
×	×	✓	77.75	87.48	75.60	86.10	79.70	88.70
✓	✓	×	78.66	88.06	76.86	86.92	80.24	89.04
✓	×	✓	78.18	87.75	75.72	86.18	81.57	89.85
×	✓	✓	77.85	87.55	76.04	86.39	80.13	88.97
✓	✓	✓	79.50	88.58	77.12	87.08	81.84	90.01

stark contrast, our method achieves near-perfect balance, with both precision 90.10% and recall 89.93% approaching 90%, thereby delivering highly reliable and consistent detection performance under complex land-cover variations.

C. Ablation Experiment and Sensitivity Analysis

1) Ablation Experiment: We abbreviate the Spatial Filtering Strategy, Semantic Filtering-Aware Enhancement, and Bi-temporal Graph Alignment as SFS, SFAE, and BTGA, respectively, to facilitate the ablation analysis. Specifically, to conduct a rigorous comparative study rather than merely removing these modules, we substitute SFS with a Fourier filtering strategy, SFAE with CBAM enhancement, and BTGA with a cross-attention mechanism.

As illustrated in Table II, removing any individual module results in a marked performance degradation compared to the full model integrated with all components (Row 7). In single-module configurations (Rows 1–3), the model's feature extraction capability is notably constrained. For instance, utilizing SFS or SFAE alone yields IoU drops of 3.26% and 3.17% on the LEV dataset, respectively, relative to the full model (79.50%). Similarly, on the Google dataset, the decreases reach 1.99% and 2.95%. Even the BTGA module, which performs relatively better in isolation (Row 3), falls short by 2.14% in IoU on the BCDD dataset without the aid of SFS and SFAE. This strongly suggests that a single component is insufficient to address the diverse challenges inherent in complex scenes.

When deploying dual-module combinations (Rows 4–6), the performance gap narrows but remains suboptimal. Specifically, the configuration without BTGA (Row 4), despite decent performance on LEV, still lags behind the full model by 1.60% IoU on BCDD. Meanwhile, the absence of SFAE (Row 5) incurs a 1.40% IoU loss on the Google dataset. Ultimately, the full model (Row 7), leveraging the synergistic effects of SFS, SFAE, and BTGA, achieves the highest metrics across all datasets (LEV: 79.50%, Google: 77.12%, BCDD: 81.84%). This conclusively demonstrates that the complementarity of spatial filtering, semantic enhancement, and graph alignment is indispensable for achieving optimal segmentation accuracy.

2) Sensitivity Analysis: We conducted sensitivity analyses on three critical parameters: the spatial filtering threshold τ in Eq. 3, the sampling ratio in Eq. 10, and the neighbor count k in Eq. 15. As shown in Fig. 5 (a), as τ varies from 0.1 to 0.9, the mean IoU exhibits a fluctuating trend, reaching a peak of 79.49% at $\tau = 0.5$. This result suggests that $\tau = 0.5$ serves as the optimal equilibrium point for distinguishing between "high-confidence semantic information" and "background redundancy". Specifically, a lower threshold (e.g., 0.3) tends to introduce excessive background noise, thereby degrading detection accuracy, whereas a higher threshold (e.g., 0.7) risks inadvertently suppressing valid semantic features.

As shown in Fig. 5 (b), we further investigated the impact of the Soft-top-k sampling ratio within the SFAE module on model performance. The results demonstrate that as the sampling ratio increases from 10% to 30%, the mean IoU improves significantly, peaking at 79.49%. This indicates that an appropriate number of semantic agents can effectively capture critical change cues. However, a marked decline in performance is observed as the ratio further extends to 50% or 90%. This implies that an excessive sampling ratio introduces substantial background noise, which interferes with global context modeling guided by the difference prior.

As shown in Fig. 5 (c), we also investigated the impact of the nearest neighbor count k within the Bi-temporal Graph Alignment (BTGA) module on model performance. As illustrated in the figure, with a small k (e.g., $k = 2$), the IoU is limited to 77.12% due to insufficient semantic context for assisting feature alignment. Performance improves significantly as k increases, reaching a peak of 79.49% at $k = 5$. This validates that moderate neighborhood aggregation effectively rectifies geometric deviations. However, a performance regression is observed when k is further increased to 7 or 9. This indicates that incorporating excessive neighbors introduces semantically irrelevant noise pixels into the aggregation process, thereby undermining the accuracy of feature alignment. Consequently, $k = 5$ is identified as the optimal choice for balancing context richness with semantic consistency.

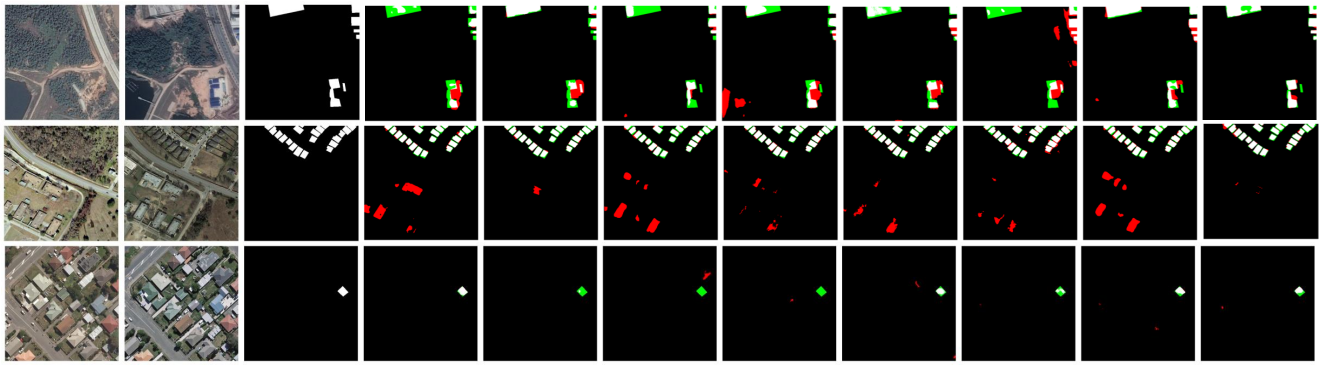
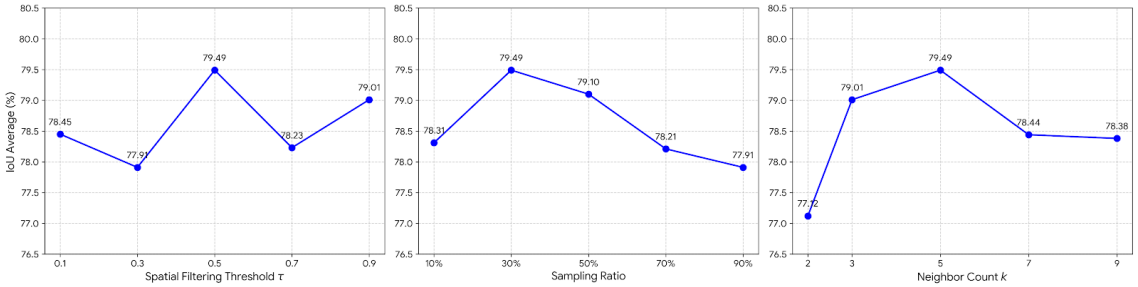


Fig. 4. The visual representation of scenarios prone to errors on three datasets. The rows from top to bottom correspond to the LEV, Google, and BCDD datasets, respectively. The columns from left to right display: Image T_1, Image T_2, Ground Truth (GT), and the change maps generated by DAMFANet, RHighNet, STRobustNet, UNLLNet, ECTI, BIFA, STADECDNet, and Ours.



(a) Impact of τ (b) Impact of Sampling Ratio (c) Impact of k

Fig. 5. (a) Impact of the spatial filtering threshold τ in the MSP. (b) Impact of the Soft-top-k sampling ratio in the SFAE. (b) Impact of the nearest neighbor count k in the BTGA.

IV. Conclusion

This paper proposes a unified framework bridging semantic filtering and structural alignment to address semantic interference and geometric misalignments in RSCD. We introduce the Mono-temporal Semantic Purification (MSP) module, which employs a "clean-then-enhance" paradigm to suppress background noise, and the Bi-temporal Graph Alignment (BTGA) module, which aligns features in non-Euclidean space to correct viewpoint distortions. Additionally, a secondary purification strategy effectively removes boundary artifacts. Extensive experiments demonstrate that our method achieves state-of-the-art accuracy and robustness on three benchmark datasets

References

[1] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 2018, pp. 4063–4067.

[2] H. Dong, X. Du, Z. Li, Z. Ma, Y. Wang, H. Zhu, and W. Ma, "Relation-aware high-order interaction network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[3] Z. Ying, Z. Tan, Y. Zhai, X. Jia, W. Li, J. Zeng, A. Genovese, V. Piuri, and F. Scotti, "Dgma 2-net: a difference-guided multiscale aggregation attention network for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[4] H. Zhang, Y. Teng, H. Li, and Z. Wang, "Strobustnet: Efficient change detection via spatial-temporal robust representations in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[5] Z. Li, S. Cao, J. Deng, F. Wu, R. Wang, J. Luo, and Z. Peng, "Stade-cdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[6] H. Zhang, H. Chen, C. Zhou, K. Chen, C. Liu, Z. Zou, and Z. Shi, "Bifa: Remote sensing image change detection with bitemporal feature alignment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[7] M. Liu, Z. Chai, H. Deng, and R. Liu, "A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 4297–4306, 2022.

[8] J. Li, Y. Wen, and L. He, "Scconv: Spatial and channel reconstruction convolution for feature redundancy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6153–6162.

[9] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision gnn: An image is worth graph of nodes," *Advances in neural information processing systems*, vol. 35, pp. 8291–8303, 2022.

[10] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5891–5906, 2020.

[11] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sensing*, vol. 12, no. 10, p. 1662, 2020.

[12] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on geoscience and remote sensing*, vol. 57, no. 1, pp. 574–586, 2018.

[13] B. Yang, J. Li, S. Yu, L. Liu, and X. Liu, "Uncertainty-aware noisy label learning for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[14] M. Noman, M. Fiaz, H. Cholakkal, S. Khan, and F. S. Khan, "Elgc-net: Efficient local–global context aggregation for remote sensing change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[15] Y. Liu, K. Wang, M. Li, Y. Huang, and G. Yang, "Exploring the cross-temporal interaction: Feature exchange and enhancement for remote sensing change detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 11 761–11 776, 2024.