Cascaded Spectral-Spatial Enhancement and Consistency Guided Interaction 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 (RSCD) aims to identify semantic changes from bi-temporal imagery but is often constrained by seasonal variations and structural misalignments. Existing methods struggle to balance noise suppression with geometric integrity. Their reliance on "blind pixel-level matching" conflates semantic invariance with observational reliability, leading to boundary false positives. To address these challenges, this paper proposes a unified framework based on Cascaded Spectral-Spatial Enhancement (CSSE) and Spectral Consistency-Guided Interaction (SCGI). First, we introduce the CSSE module. By integrating dynamic spectral gating with a local-global dual attention mechanism, CSSE explicitly models what land covers are and how they are spatially organized, effectively filtering non-semantic artifacts while maintaining structural coherence. Second, to handle misalignments, we propose the SCGI strategy. This strategy decouples global spectral consistency priors from local observation confidence, enabling robust change identification even under viewpoint shifts or partial occlusion. Finally, a Hierarchical Decoder (HD) is employed to progressively aggregate multi-scale cues for boundary refinement. 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 79.65%, 73.40%, and 81.02%, respectively.**

***Index Terms*—Remote Sensing Change Detection, Pseudo-change, Structural Alignment.**

I. Introduction

Remote Sensing Change Detection (RSCD) identifies semantic surface changes from bi-temporal imagery. It plays a critical role in Earth surface monitoring, serving vital applications such as urban planning, disaster assessment, and ecological monitoring [1].

Prior to the proliferation of deep learning, traditional change detection methods primarily relied on handcrafted features and elementary mathematical operations. Representative approaches include algebraic methods, such as Change Vector Analysis (CVA) and Principal Component Analysis (PCA), as well as post-classification comparison strategies based on traditional machine learning algorithms like Support Vector Machines (SVM) and Random Forests. However, in the face of high-resolution and complex remote sensing scenarios, these methods have exposed severe limitations, including extreme sensitivity to non-semantic noise, heavy reliance on strict pixel-level alignment, and a deficiency in contextual understanding [1]. With the rapid advancement of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), RSCD has witnessed breakthrough progress, establishing itself as the mainstream paradigm. For instance, cross-attention-based methods, leveraging their capability for bi-temporal interaction modeling, facilitate semantic-level learning across time steps, thereby effectively mapping learned representations to change identification. Similarly, fusion network-based approaches incorporate diverse architectural designs to synergistically address both spatial feature semantics and temporal change perception [2]–[4].

Nevertheless, two persistent challenges continue to impede the achievement of accurate and robust RSCD. First, *Semantic Interference and Pseudo-changes*: Non-semantic variations induced by seasonal shifts, illumination disparities, cloud occlusion, or shadows are frequently misclassified as genuine changes. For instance, cloud cover can render spectrally similar regions indistinguishable, while varying illumination may cause unchanged areas to appear significantly different [3]–[6]. Second, *Boundary Errors caused by Structural Misalignments*: Geometric discrepancies arising from viewpoint shifts or registration errors often result in imperfect object alignment. Existing methods relying on "blind pixel-level matching" are particularly prone to generating False Positives in these boundary regions [1], [2], [7]. To mitigate environmental noise (e.g., from seasons, lighting, or clouds), many existing methods employ complex attention mechanisms to enhance feature representation, attempting to filter out irrelevant interference. However, these approaches often struggle to strike a balance: they either over-suppress noise, leading to missed detections, or introduce excessive noise in an effort to preserve details. While some methods (e.g., BIFA [6], STADECDNet [7]) attempt to explicitly align bi-temporal features to correct geometric deviations, they typically operate under the assumption of perfect pixel correspondence, thereby conflating "semantic invariance" with "observational reliability".

In this paper, to resolve the persistent dilemma between suppressing environmental noise and preserving geometric integrity, we propose a unified framework based on Cascaded Spectral-Spatial Enhancement and Spectral Consistency-Guided Interaction. First, to tackle semantic interference (e.g., seasonal shifts and cloud occlusion), we introduce the Cascaded Spectral-Spatial Enhancement (CSSE)

module. By integrating dynamic spectral gating with a local-global dual attention mechanism, CSSE explicitly models what land covers are and how they are spatially organized, thereby filtering non-semantic artifacts while maintaining structural coherence. Second, to address structural misalignments and the limitations of "blind pixel-level matching", we propose the Spectral Consistency-Guided Interaction (SCGI) strategy. Unlike existing methods that conflate semantic invariance with observational reliability, SCGI decouples global spectral consistency priors from local observation confidence. This allows the model to robustly identify genuine changes even under viewpoint shifts or partial occlusion. Finally, a Hierarchical Decoder (HD) aggregates these multi-scale cues to progressively refine object boundaries.

II. Method

A. Overview

We first provide a brief overview of the overall pipeline. As detailed in Section II-B in Fig. 1, a backbone network is employed to extract multi-level feature sets, denoted as $\{X_l^t \mid l = 1,2,3,4\}$ for each time step $t \in \{1,2\}$. Subsequently, Section II-C performs mono-temporal feature enhancement on each individual feature map. The enhanced features from corresponding levels across different time steps (i.e., different t but identical l) are then fed into the Spectral Consistency-Guided Interaction (SCGI) in Section II-D for bi-temporal change perception, yielding the refined difference maps $D_l^{refined}$ where $l \in \{1,2,3,4\}$. Finally, the Hierarchical Decoder (HD) described in Section II-E integrates the enhanced mono-temporal features with $D_l^{refined}$ to generate the final change detection result.

B. Siamese Mono-temporal Feature Encoder

To extract mono-temporal features using a Siamese architecture, let the input bi-temporal image pair be denoted as $I^1, I^2 \in \mathbb{R}^{H\times W\times 3}$. 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:

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

where, t denotes the time step or branch index. $\mathcal{X}^t$ is the multi-scale feature set corresponding to the t-th input image. X_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 $X_1^t \in \mathbb{R}^{32\times \frac{H}{2}\times \frac{W}{2}}$, $X_2^t \in \mathbb{R}^{64\times \frac{H}{4}\times \frac{W}{4}}$, $X_3^t \in \mathbb{R}^{128\times \frac{H}{8}\times \frac{W}{8}}$, $X_4^t \in \mathbb{R}^{256\times \frac{H}{16}\times \frac{W}{16}}$, where feature maps from different time steps at the same level share identical spatial dimensions.

C. Cascaded Spectral-Spatial Enhancer

In remote sensing change detection, the core challenge of single-temporal semantic enhancement lies in simultaneously parsing both *what* the land cover is—such as buildings, vehicles, or vegetation—and *how* it is spatially organized, including regular grids, linear roads, and periodic farmlands. To this end, the proposed Cascaded Spectral-Spatial Enhancer, referred to as CSSE, consists of three tightly coupled components: a local-global dual attention mechanism, a dynamic spectral gating module, and a spatial-spectral interaction modulator.

First, to overcome the limited receptive field of CNNs and the geometric discontinuity induced by fixed patch partitioning in Vision Transformers, CSSE introduces a local-global dual attention mechanism, comprising two carefully designed operators. (i) the window attention operator reshapes the input feature X_l^t into $X_{l,\text{perm}}^t$ and partitions it into non-overlapping windows $\Omega_{i,j} \in \mathbb{R}^{w\times w\times C}$, with a default window size of $w = 7$. Within each local window, queries, keys, and values are computed via linear projections. For notational simplicity in the following formulation, we use X to denote the generic input feature. This operation is applied across all hierarchy levels $l \in \{1,2,3,4\}$ and time steps $t \in \{1,2\}$, as:

$$[Q,K,V] = \text{Linear}(\text{Flatten}(\Omega_{i,j})) \in \mathbb{R}^{w^2\times 3C}. \tag{2}$$

where $[Q,K,V]$ are split into h heads, yielding $Q_h, K_h, V_h \in \mathbb{R}^{w^2\times d_h}$ where $d_h = C/h$.

The intra-window self-attention is computed as [8]:

$$\text{Attn}_\Omega(Q_h,K_h,V_h) = \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}}\right)V_h \tag{3}$$

Outputs are merged and reshaped into $X_{perm} = \text{WindowAttn}(X)$, preserving fine details (e.g., roof textures) through isolated window processing.

(ii) Axial Attention models long-range dependencies along the two primary spatial axes. For the height axis, treating X_{perm} as sequences $\{\mathbf{x}_{i,:,:}\}_{i=1}^H$, the attention weight at position (i,j) along height is:

$$\alpha_{i,i'}^{(j)} = \text{Softmax}\left(\frac{q_{i,j}k_{i',j}^\top}{\sqrt{d_h}}\right),\ \text{Out}_H(i,j) = \sum_{i'=1}^{H}\alpha_{i,i'}^{(j)}v_{i',j} \tag{4}$$

where q,k,v are generated by independent linear layers. A similar procedure is applied along the width axis. The two outputs are summed and projected to yield the global contextual feature $\text{AxialAttn}(\mathbf{X})$, which effectively captures large-scale patterns such as urban road orientations and agricultural strip layouts.

The outputs of both operators are concatenated along the channel dimension and fused via a 1×1 convolution $\text{Conv}_{1\times 1}(\cdot)$ into a unified spatial representation:

$$F_s = \text{Conv}_{1\times 1}([\text{WindowAttn}(\mathbf{X}),\ \text{AxialAttn}(\mathbf{X})]) \tag{5}$$

Second, to explicitly leverage inherent spectral structural priors—such as high-frequency directional energy in buildings versus isotropic responses in forests—CSSE incorporates a dynamic spectral gating mechanism. Specifically, given F_s, we compute the 2D real-valued FFT $\hat{F}_s = \mathcal{F}(F_s)$ and utilize its magnitude $|\hat{F}_s|$ as a semantic-aware spectral fingerprint. A channel-wise gate is then generated via Global Average Pooling (GAP$(\cdot)$) followed by a lightweight MLP.

$$\mathcal{G}_f = \sigma\Big(\text{MLP}\big(\text{GAP}(|\hat{F}_s|)\big)\Big) \tag{6}$$

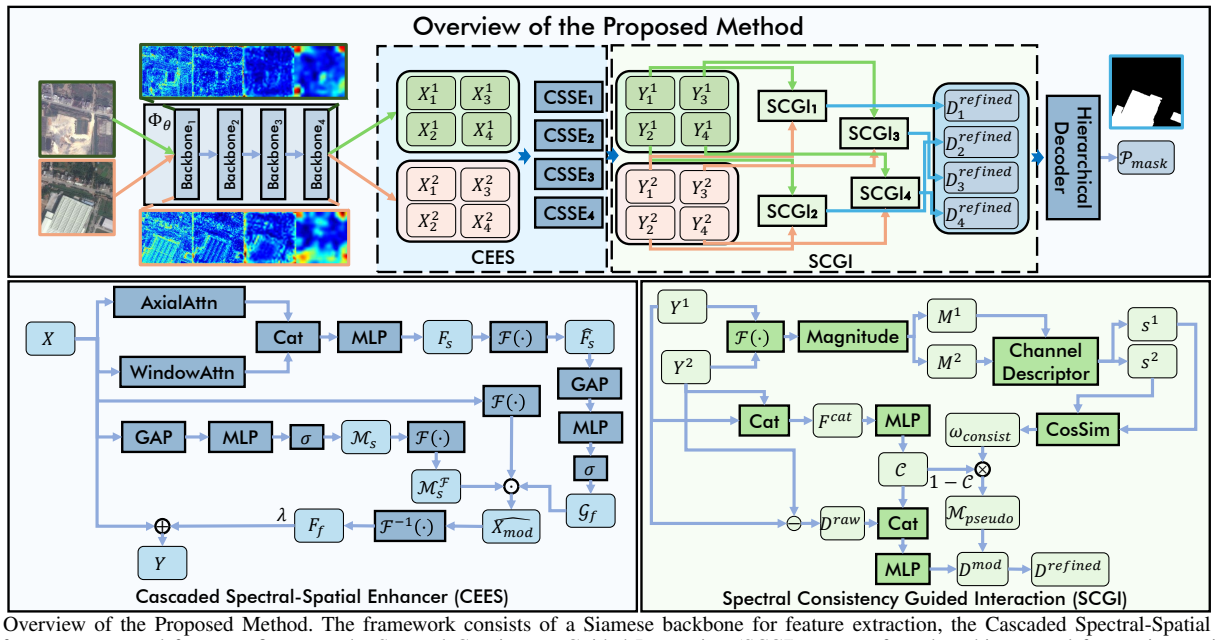

Fig. 1. Overview of the Proposed Method. The framework consists of a Siamese backbone for feature extraction, the Cascaded Spectral-Spatial Enhancer (CSSE) for mono-temporal feature refinement, the Spectral Consistency-Guided Interaction (SCGI) strategy for robust bi-temporal feature interaction, and a Hierarchical Decoder (HD) for generating the final change map.

where σ denotes the sigmoid function. This gate adaptively reweights the input spectrum $\mathcal{F}(X)$, amplifying frequency bands consistent with the current semantics and suppressing irrelevant noise.

Third, to mitigate contamination from non-land-cover artifacts such as clouds and shadows, CSSE introduces spatial-spectral interaction modulation. Specifically, a spatial attention mask is generated from the original input:

$$\mathcal{M}_s = \sigma\big(\mathrm{MLP}\big(\mathrm{GAP}(X)\big)\big) \tag{7}$$

which highlights reliable regions (e.g., high-albedo buildings, regular roads) and suppresses low-confidence areas like cloud shadows. This mask is then transformed into the spectral domain as $\mathcal{M}_s^{\mathcal{F}} = \mathcal{F}(\mathcal{M}_s)$ and used to modulate the gated spectrum.

Building upon these components, CSSE integrates them into a unified cascaded enhancement pipeline. The input spectrum undergoes dual modulation:

$$\hat{X}_{mod} = \big(\mathcal{G}_f \odot \mathcal{F}(X)\big) \odot \mathcal{M}_s^{\mathcal{F}} \tag{8}$$

followed by inverse FFT to obtain the enhanced feature $F_f = \mathcal{F}^{-1}(\hat{X}_{mod})$. The final output adopts a residual formulation:

$$Y = X + \lambda \cdot F_f \tag{9}$$

where λ denotes the weighting coefficient.

This bidirectional synergy aligns spatial semantic cues (*what*) with spectral physical priors (*how*) to mitigate seasonal shifts. The resulting features are thus semantically discriminative and geometrically robust, forming a solid basis for bi-temporal reasoning.

D. Spectral Consistency Guided Interaction

Recent approaches transcend simplistic comparisons to enhance bi-temporal feature interaction and identify genuine semantic changes. Conventional methods often fail due to: (i) *pseudo-changes* from non-semantic variations (e.g., seasons, shadows); and (ii) false positives induced by spatial misalignments or high-frequency noise.

Existing blind pixel-matching methods erroneously conflate semantic invariance with observational reliability. We address this via Spectral Consistency Guided Interaction (SCGI), which synergizes spectral structural consistency (a global semantic prior) and local observation confidence (a spatial reliability cue). Specifically, we extract global spectral fingerprints from Y_l^t via 2D real-valued FFT magnitude $\mathcal{F}(\cdot)$:

$$\hat{F}^t = \mathcal{F}(Y^t), \;\; M^t = |\hat{F}^t|. \tag{10}$$

where M^t denotes the magnitude of $\hat{F}^t$.

Considering different land covers exhibit characteristic spectral structures—for instance, buildings produce directional high-frequency patterns, while forests yield isotropic low-to-mid frequency responses. To capture this globally, we compress each magnitude spectrum into a spatially invariant channel descriptor via global average pooling:

$$s^t = \frac{1}{HW'} \sum_{h,w} M^t(h,w) \tag{11}$$

where HW' denotes the spatial resolution of M^t.

The cosine similarity $\mathrm{CosSim}(\cdot)$ between these two temporal descriptors,

$$\omega_{consist} = \mathrm{CosSim}(s^1, s^2) \tag{12}$$

where $\omega_{consist}$ serves as a robust measure of semantic consistency. Unlike pixel-wise metrics (e.g., L2 distance), this global fingerprint is insensitive to small misalignments and better reflects true invariance. Critically, high similarity ($\omega_{consist} \approx 1$) often signals pseudo-invariance—for example, when both

images are obscured by clouds—making it a valuable prior for pseudo-change suppression.

However, global spectral consistency alone is insufficient: consistently occluded regions (e.g., thick clouds in both images) may yield high $\omega_{consist}$ yet offer no reliable evidence of invariance, while clear regions with subtle spectral shifts can indicate true changes despite moderate consistency. To resolve this, we introduce a spatially varying confidence map $\mathcal{C}$ that quantifies local observability.

Specifically, we concatenate the bi-temporal features $F^{cat} = [F^1; F^2]$ and pass them through a convolutional head:

$$\mathcal{C} = \text{MLP}(F^{cat}) \tag{13}$$

where high values indicate reliable observations (e.g., cloud-free, well-aligned pixels), and low values flag uncertain regions (e.g., cloud-covered or noisy areas).

We then fuse these two cues into a pseudo-change suppression mask:

$$\mathcal{M}_{pseudo} = \omega_{consist} \cdot (1 - \mathcal{C}) \tag{14}$$

This formulation enforces a dual-condition criterion: only regions that are *both* spectrally consistent (*global prior*) *and* unobservable (*low local confidence*) are flagged as high-risk pseudo-changes. This prevents misinterpreting unreliable consistency as genuine invariance—a common failure mode in conventional methods.

Finally, we generate an adaptive change difference map in two stages. First, we compute the raw absolute difference $D^{raw} = |F^1 - F^2|$ and modulate it with the upsampled confidence map $\mathcal{C} \uparrow$ through a lightweight MLP, enabling the model to amplify changes in reliable regions while suppressing noise in uncertain ones, as:

$$D^{mod} = \text{MLP}([D^{raw}; \mathcal{C} \uparrow]). \tag{15}$$

Then, we suppress high-risk pseudo-changes using the learned mask:

$$D^{refined} = D^{mod} \odot \mathcal{M}_{pseudo}. \tag{16}$$

where $\odot$ denotes the Hadamard product.

E. Hierarchical Decoder

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

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

$$\begin{aligned} \hat{X}_l^1 &= \text{MLP}^l([Y_l^1 \parallel D_l^{refined}]) \\ \hat{X}_l^2 &= \text{MLP}^l([Y_l^2 \parallel D_l^{refined}]) \end{aligned} \tag{17}$$

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

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

$$\mathcal{H}^l = \hat{X}_l^1 + \hat{X}_l^2 + Y_l^{change} \tag{18}$$

We then adopt a top-down cascaded decoder to integrate multi-level cues. Letting M^l denote the feature at level l, the decoding process is recursively defined as:

$$M^l = \begin{cases} \mathcal{C}^l(\mathcal{H}^l), & l = 4 \\ \mathcal{C}^l(\mathcal{H}^l + \text{Up}(M^{l+1})), & l \in 3, 2, 1 \end{cases} \tag{19}$$

where $\mathcal{C}^l$ is a convolutional block and $\text{Up}(\cdot)$ applies bilinear interpolation. The highest-resolution feature M^l is then passed through an MLP to yield the final change mask $\mathcal{P}_{mask}$:

$$\mathcal{P}_{mask} = \text{MLP}^{final}(M^l) \tag{20}$$

The network is optimized using the Binary Cross-Entropy (BCE) loss, which measures the discrepancy between $\mathcal{P}_{mask}$ and the ground truth $G \in \{0, 1\}^{H \times W}$:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \tag{21}$$

Here, N is the total pixel count, and g_i, p_i denote the target and predicted values for the i-th pixel. This objective function drives the model to distinguish true semantic changes from background noise.

III. Experiment

A. Experiment Setting

Datasets: Our experiments are performed on three public datasets: GoogleBuilding [9], LEVCD [10], and BCDD [11], 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 [2], UNLLNet [5], BIFA [6], ELGCNet [12], RHighNet [3], and STRobustNet [4]. *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 [3].

B. Result Performance Quantitative Comparison

Overall, our method achieves the most balanced and consistently top-tier performance across all evaluation metrics and datasets, demonstrating superior robustness in diverse change detection scenarios, see Table I.

On the LEV dataset, DAMFANet and RHighNet deliver competitive F1-scores (85.98% and 85.32%, respectively) but suffer from lower precision due to false alarms in shadowed regions. STRobustNet and FCEP show improved precision yet

TABLE I

QUANTITATIVE COMPARISON RESULTS ON THREE DATASETS. **BOLD** INDICATES THE BEST PERFORMANCE, AND UNDERLINED INDICATES THE SECOND BEST. 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 [2]	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 [3]	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 [5]	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 [13]	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 [7]	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	**79.65**	**88.67**	**87.63**	89.74	**73.40**	**84.66**	74.92	**97.31**	**81.02**	**89.51**	**93.19**	86.12	25.48

TABLE II

ABLATION STUDY OF THE PROPOSED COMPONENTS ON LEV, GOOGLE, AND BCDD DATASETS. CSSE: CASCADED SPECTRAL-SPATIAL ENHANCER; SCGI: SPECTRAL CONSISTENCY GUIDED INTERACTION; HD: HIERARCHICAL DECODER.

Setting			LEV		Google		BCDD	
CSSE	SCGI	HD	IoU(%)	F1(%)	IoU(%)	F1(%)	IoU(%)	F1(%)
✓	×	×	70.43	82.65	66.76	80.07	74.17	85.17
×	✓	×	71.01	83.05	71.34	83.27	73.82	84.94
×	×	✓	73.44	84.69	68.01	80.96	75.13	85.80
✓	✓	×	74.52	85.40	72.62	84.14	80.61	89.26
✓	×	✓	74.07	85.10	73.34	84.62	79.97	88.87
×	✓	✓	75.67	86.15	70.04	82.38	76.86	86.92
✓	✓	✓	79.65	88.67	73.40	84.66	81.02	89.51

lag in recall, indicating overly conservative change prediction. In contrast, our approach simultaneously attains the highest IoU (79.65%), F1 (88.67%), and precision (87.63%), reflecting a better trade-off between sensitivity and reliability.

On Google, methods leveraging explicit alignment (e.g., BIFA, STADECDNet) achieve high recall (>92%) but at the cost of precision (<85%), resulting in suboptimal F1 scores. Meanwhile, Siamese-based models like FCSiam and MSCANet exhibit moderate performance across all metrics, lacking adaptability to local noise. Our method outperforms all competitors in both IoU and F1, confirming its effectiveness in handling registration errors without over-predicting changes.

On BCDD, ECTI and STADECDNet, despite strong performance on other datasets, fall below 80% F1, as their attention mechanisms mistakenly treat cloud-induced pseudo-changes as real variations. DAMFANet and BIFA perform relatively well but still trail behind in IoU. Our model, by explicitly modeling spectral consistency and local observability, achieves the best results (81.02% IoU, 89.51% F1), with notably higher precision and recall balance than all baselines. The visual representation can be seen in Fig. 2.

In summary, while individual baselines may excel in specific aspects (e.g., recall on Google or precision on LEV), none match our method's consistent dominance across metrics and datasets—validating its comprehensive design for real-world remote sensing change detection.

C. Ablation Experiment and Sensitivity Analysis

1) Ablation Experiment: To evaluate the effectiveness of each proposed component in our framework, we conduct a comprehensive ablation study on three benchmark datasets: LEV, Google, and BCDD (see Table II). The three key modules are: Cascaded Spectral-Spatial Enhancer (CSSE): enhances feature representation by fusing spectral and spatial information through cascaded cross-modal interactions. Spectral Consistency Guided Interaction (SCGI): enforces consistency across spectral channels during feature fusion to preserve fine-grained spectral semantics. Hierarchical Decoder (HD): refines segmentation maps progressively using multi-scale decoding strategies to improve boundary accuracy. The results are summarized in Table II. Each row corresponds to a different configuration of these components, allowing us to assess their individual and combined contributions.

The ablation experiments demonstrate the progressive benefits of each module. In single-module configurations, while CSSE and SCGI provide moderate baselines (e.g., 70.43% and 71.01% IoU on LEV, respectively), the HD module independently yields the most significant gains, reaching an IoU of 75.13% on BCDD. This indicates that hierarchical decoding is crucial for recovering spatial details. Moving to pairwise combinations, the integration of SCGI and HD proves particularly effective, achieving 75.67% IoU on LEV and surpassing both the CSSE+SCGI (74.52%) and CSSE+HD (74.07%) configurations. This suggests that enforcing spectral consistency alongside multi-scale refinement offers a distinct advantage in complex scenes.

Ultimately, the full model integrating all three components (CSSE + SCGI + HD) achieves the optimal performance, exhibiting a substantial leap over any partial configuration. The complete architecture secures the highest metrics across all datasets, peaking at 79.65% IoU on LEV and 81.02% IoU on BCDD. These results confirm that the synergistic combination of spectral-spatial enhancement, consistency enforcement, and hierarchical decoding is essential for maximizing segmentation accuracy and boundary precision.

D. Sensitivity Analysis

To investigate the impact of the residual weighting factor λ within the Cascaded Spectral-Spatial Enhancer (CSSE), as defined in Eq. 9, we conducted a sensitivity analysis across the three datasets. The parameter λ governs the intensity with which the spectrally and spatially modulated features F_f are injected into the original feature representation X. As illustrated in Fig. 3, we evaluated the average IoU performance in three datasets varying λ from 0.1 to 0.9. The

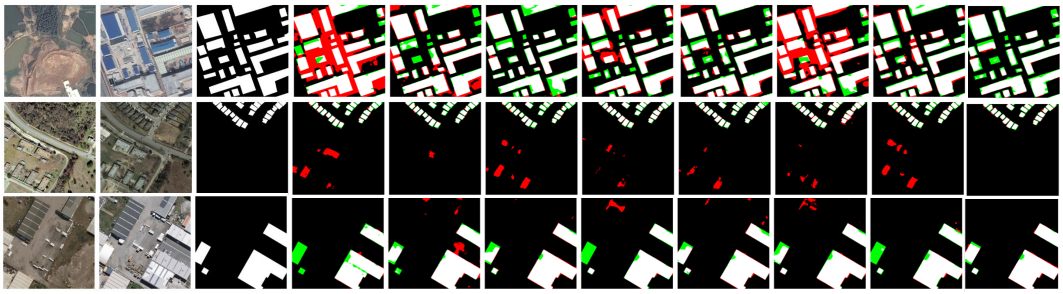

Fig. 2. Visual comparison results on three datasets. The rows from top to bottom correspond to the LEV, Google, and BCDD datasets, respectively. The columns from left to right display: Image T_1, Image T_2, Ground Truth (GT), and the change maps generated by DAMFANet, RHighNet, STRobustNet, UNLLNet, ECTI, BIFA, STADECDNet, and Ours.

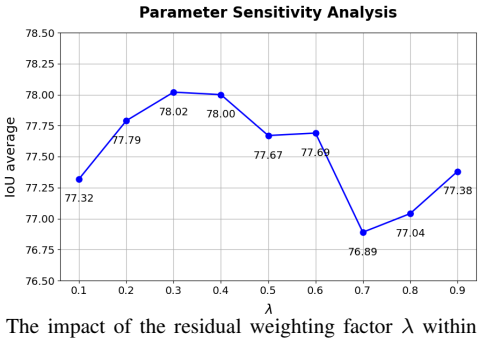

Fig. 3. The impact of the residual weighting factor λ within the Cascaded Spectral-Spatial Enhancer (CSSE).

results demonstrate that performance initially improves and then degrades as λ increases. With a small λ (e.g., 0.1), the injection of spectral-spatial priors is insufficient, limiting the model's ability to fully benefit from the semantic enhancement provided by CSSE. The performance peaks at $\lambda = 0.3$, achieving an optimal average IoU of 78.02%. This indicates that 0.3 represents the ideal trade-off, effectively incorporating "what" and "how" semantic cues without overwhelming the intrinsic feature representation. Conversely, performance drops significantly as λ exceeds 0.5, reaching a nadir of 76.89% at $\lambda = 0.7$. This decline suggests that an excessive weight may introduce high-frequency noise or distort the original feature distribution, thereby adhering to semantic interference rather than suppressing it. Consequently, we set $\lambda = 0.3$ for all subsequent experiments to ensure robust detection accuracy.

IV. Conclusion

In this work, we address two critical challenges in remote sensing change detection—semantic interference from non-land-cover variations and false positives caused by structural misalignments. We propose a unified framework that integrates Cascaded Spectral-Spatial Enhancement (CSSE) for joint semantic-spatial modeling and Spectral Consistency-Guided Interaction (SCGI) to decouple global semantic consistency from local reliability. Coupled with a hierarchical decoder for boundary refinement. Extensive experiments demonstrate that our method achieving superior robustness and accuracy with IoU of 79.65%, 73.40%, and 81.02% in the LEV, Google, and BCDD datasets, respectively.

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] 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.

[3] 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.

[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] 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.

[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] 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.

[8] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.