MISA-net: multi-scale interaction and supervised attention network for remote-sensing image change detection

Yin, H., Wang, J., Liu, S., Wang, Y., Liu, Y. ORCID: https://orcid.org/0000-0003-3056-7713, Guo, T. and Xia, M. ORCID: https://orcid.org/0000-0003-4681-9129 (2026) MISA-net: multi-scale interaction and supervised attention network for remote-sensing image change detection. Remote Sensing, 18 (2). 376. ISSN 2072-4292 doi: 10.3390/rs18020376 Available at https://centaur.reading.ac.uk/128310/

remote sensing

MDPI

Article

MISA-Net: Multi-Scale Interaction and Supervised Attention Network for Remote-Sensing Image Change Detection

Haoyu Yin [1], Junzhe Wang [1], Shengyan Liu [1], Yuqi Wang [2], Yi Liu [1], Tengyue Guo [1,2] and Min Xia [1,*]

[1] Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, No. 219, Ningliu Road, Nanjing 210044, China; 202383250028@nuist.edu.cn (H.Y.); 202412490580@nuist.edu.cn (J.W.); 202412492002@nuist.edu.cn (S.L.); 003766@nuist.edu.cn (Y.L.); wj835167@student.reading.ac.uk (T.G.)

[2] Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK; 202412491650@nuist.edu.cn

* Correspondence: xiamin@nuist.edu.cn

Highlights

What are the main findings?

- A novel bi-temporal interactive architecture for remote sensing change detection;
- The network effectively enhances multi-scale feature representation and bi-temporal interaction to improve change discrimination.

What are the implications of the main findings?

- Significantly improves boundary accuracy and robustness against pseudo-changes in complex remote sensing scenes;
- Achieves state-of-the-art performance on LEVIR-CD, SYSU-CD, and GZ-CD datasets, demonstrating strong robustness and generalization capability.

Abstract

Change detection in remote sensing imagery plays a vital role in land use analysis, disaster assessment, and ecological monitoring. However, existing remote sensing change detection methods often lack a structured and tightly coupled interaction paradigm to jointly reconcile multi-scale representation, bi-temporal discrimination, and fine-grained boundary modeling under practical computational constraints. To address this fundamental challenge, we propose a Multi-scale Interaction and Supervised Attention Network (MISANet). To improve the model's ability to perceive changes at multiple scales, we design a Progressive Multi-Scale Feature Fusion Module (PMFFM), which employs a progressive fusion strategy to effectively integrate multi-granular cross-scale features. To enhance the interaction between bi-temporal features, we introduce a Difference-guided Gated Attention Interaction (DGAI) module. This component leverages difference information between the two time phases and employs a gating mechanism to retain fine-grained details, thereby improving semantic consistency. Furthermore, to guide the model's focus on change regions, we design a Supervised Attention Decoder Module (SADM). This module utilizes a channel–spatial joint attention mechanism to reweight the feature maps. In addition, a deep supervision strategy is incorporated to direct the model's attention toward both fine-grained texture differences and high-level semantic changes during training. Experiments conducted on the LEVIR-CD, SYSU-CD, and GZ-CD datasets demonstrate the effectiveness of our method, achieving F1-scores of 91.19%, 82.25%, and 88.35%, respectively. Compared with the state-of-the-art BASNet model, MISANet achieves performance gains of 0.50% F1 and 0.85% IoU on LEVIR-CD, 2.13% F1 and 3.02% IoU on SYSU-CD, and 1.28% F1 and 2.03% IoU on GZ-CD. The proposed method demonstrates strong generalization capabilities and is applicable to various complex change detection scenarios.

Academic Editor: Zhenwei Shi

Received: 4 December 2025
Revised: 18 January 2026
Accepted: 19 January 2026
Published: 22 January 2026

Keywords: remote sensing image; change detection; deep learning; multi-scale feature fusion; attention mechanism; deep supervision

1. Introduction

Remote sensing change detection (RSCD) refers to the process of analyzing and identifying changes in the same geographical area from remotely sensed images acquired at different times [1]. It has been widely applied in various domains, including land-use monitoring [2], urban expansion [3], disaster assessment [4,5], and ecological and environmental surveillance [6]. Through these applications, RSCD provides crucial insights into natural and anthropogenic dynamics and supports decision-making in disaster prevention and management. However, due to variations in season, time, solar elevation angle, and atmospheric conditions, identical land-cover objects may exhibit significant spectral differences across acquisitions. Moreover, objects with similar textures can easily be confused, making the accurate identification of real changes a challenging task.

Over the past few decades, numerous studies have been conducted to achieve more accurate detection of land-cover changes across different periods. Generally, RSCD methods can be categorized into two groups: traditional methods and deep-learning-based methods. Traditional methods are typically divided into three types: algebra-based, transformation-based, and classification-based approaches. Algebra-based methods [7,8] compute image differences or ratios between multi-temporal images to highlight changed regions. These methods are computationally simple and interpretable but are sensitive to noise and inadequate for complex scenarios. Transformation-based methods employ principal component analysis [9], wavelet transform [10], or multi-scale geometric analysis [11] to map images into new feature spaces, enhancing change information while suppressing redundant background content. Classification-based methods rely on traditional machine-learning algorithms such as support vector machines [12] and k-nearest neighbors [13], performing supervised or unsupervised classification on individual temporal images and comparing results to produce change maps. Although effective in certain contexts, these methods depend heavily on handcrafted features and exhibit limited generalization. Furthermore, conventional classifiers struggle to capture the nonlinear and high-dimensional characteristics of complex land-cover transitions, which restricts detection accuracy.

With the emergence of deep learning, RSCD has undergone a paradigm shift [14,15]. Deep neural networks, particularly convolutional neural networks (CNNs), demonstrate strong capability in modeling nonlinear mappings and automatically extracting multi-scale representations [16]. By leveraging local receptive fields and parameter sharing, CNNs can effectively capture both low-level texture details and high-level semantic cues, substantially improving feature representation for RSCD [17–25]. Architectures such as ResNet alleviate gradient-vanishing issues through residual connections, enabling deeper models to learn more robust change features. DenseNet [26] promotes feature reuse, facilitating the detection of subtle and weak changes. HRNet [27,28] maintains a high-resolution branch throughout the network and continuously exchanges information with lower-resolution branches, proving advantageous for spatial detail preservation. Nevertheless, CNN-based methods remain constrained by their limited receptive fields, which hinder their ability to capture long-range contextual dependencies [29]. This limitation reduces their effectiveness in large-scale or structurally complex RSCD scenarios—for example, detecting extensive urban sprawl or fine-grained modifications such as rooftop additions or river course adjustments.

To overcome these shortcomings, Transformer-based architectures employing self-attention mechanisms have recently been introduced into RSCD [30]. The self-attention mechanism enables global dependency modeling across all pixels, thereby improving semantic understanding of large-scale spatial structures [31–34]. For instance, Yan et al. [35] proposed the Fully Transformer Network (FTN), which replaces CNN backbones with a pyramid-structured Transformer and incorporates a boundary-aware loss to enhance boundary integrity and accuracy. Similarly, Feng et al. [36] developed SMBCNet, reformulating change detection as a semantic segmentation problem by combining a hierarchical Transformer encoder with multi-scale feature fusion and aggregation modules to effectively model global bi-temporal dependencies. These studies highlight the potential of self-attention mechanisms for RSCD. However, the superior global modeling capability of Transformers often comes with substantial computational cost. The quadratic complexity of self-attention in both spatial and temporal dimensions leads to high memory consumption and limited scalability for large-scale or high-resolution imagery. Moreover, excessive reliance on global attention may weaken the model's ability to capture local boundary details, increasing false or missed detections in small-scale change regions.

Despite the remarkable progress achieved by CNN-based and Transformer-based remote sensing change detection methods, a fundamental challenge remains insufficiently addressed: how to achieve effective cross-scale semantic interaction and robust bi-temporal discrimination under limited computational budgets. In recent years, several studies have investigated the trade-off between detection accuracy and computational efficiency in RSCD. Bifa [37] adopts bitemporal feature alignment to suppress pseudo-changes with moderate computational cost; however, its alignment is performed at a single semantic level, limiting cross-scale information interaction. TaCo [38] enhances spatio-temporal semantic consistency between bi-temporal features, but relies on relatively complex interaction mechanisms, leading to increased architectural complexity and computational overhead.

This trade-off arises from the absence of a tightly coupled interaction paradigm among multi-scale features, bi-temporal discrimination, and attention modeling. Many RSCD methods adopt loosely connected designs, such as naive multi-scale fusion [39,40] and simple bi-temporal feature concatenation or subtraction [39,41], which overlook semantic discrepancies and fail to embed differential cues into the interaction process, making them vulnerable to noise and pseudo-changes [42]. Additionally, noise-sensitive attention mechanisms without adaptive gating [43,44] and the lack of intermediate supervision during decoding [45,46] often result in blurred boundaries and incomplete localization.

To address these challenges in a unified and efficiency-aware manner, we propose the Multi-Scale Interaction and Supervised Attention Network (MISANet), which constructs a tightly coupled yet lightweight interaction framework for remote sensing change detection. MISANet jointly models multi-scale fusion, bi-temporal interaction, and boundary refinement under a shared design philosophy, emphasizing effective information exchange with low computational cost. Specifically, a pretrained MobileNetV2 [47] is adopted as the encoder to extract hierarchical bi-temporal features with high efficiency. On this basis, the Progressive Multi-Scale Feature Fusion Module (PMFFM) progressively aligns and integrates cross-scale features to enable semantic interaction without heavy fusion overhead. The Difference-Guided Gated Attention Interaction (DGAI) module embeds differential cues directly into the attention mechanism, selectively enhancing change-relevant information while suppressing pseudo-changes. Finally, the Supervised Attention Decoder Module (SADM) combines channel–spatial attention with deep supervision to refine features during up-sampling and preserve fine-grained change boundaries.

The main contributions of this work are summarized as follows:

1. **A unified and efficiency-aware interaction framework for RSCD:** We formulate remote sensing change detection from the perspective of a tightly coupled interaction paradigm and propose MISANet to achieve the balance between cross-scale semantic interaction, robust bi-temporal discrimination, and computational efficiency.
2. **Progressive Multi-Scale Feature Fusion Module (PMFFM):** This module progressively aligns and fuses multi-level features, effectively bridging the semantic gap between shallow details and deep semantics.
3. **Difference-Guided Gated Attention Interaction (DGAI):** This module integrates differential features into the attention mechanism, improving the network's ability to distinguish true changes from pseudo-changes.
4. **Supervised Attention Decoder Module (SADM):** This module combines deep supervision with channel–spatial attention to enhance boundary perception and small-object detection accuracy.

2. Materials and Methods

2.1. Proposed Approach

2.1.1. Network Structure

In this study, we propose an end-to-end Siamese encoder–decoder network with shared weights for remote sensing change detection, as illustrated in Figure 1. A lightweight MobileNetV2 backbone is adopted to extract hierarchical bi-temporal features with high computational efficiency. Specifically, each branch produces five feature maps with spatial resolutions of 128 × 128, 64 × 64, 32 × 32, 16 × 16, and 8 × 8, respectively, where spatial downsampling is achieved through stride-2 depthwise separable convolution layers, following the standard MobileNetV2 architecture and no upsampling operations are involved in the backbone network. These multi-level features are progressively aligned and fused by the Progressive Multi-Scale Feature Fusion Module (PMFFM) to construct semantically consistent multi-scale representations.

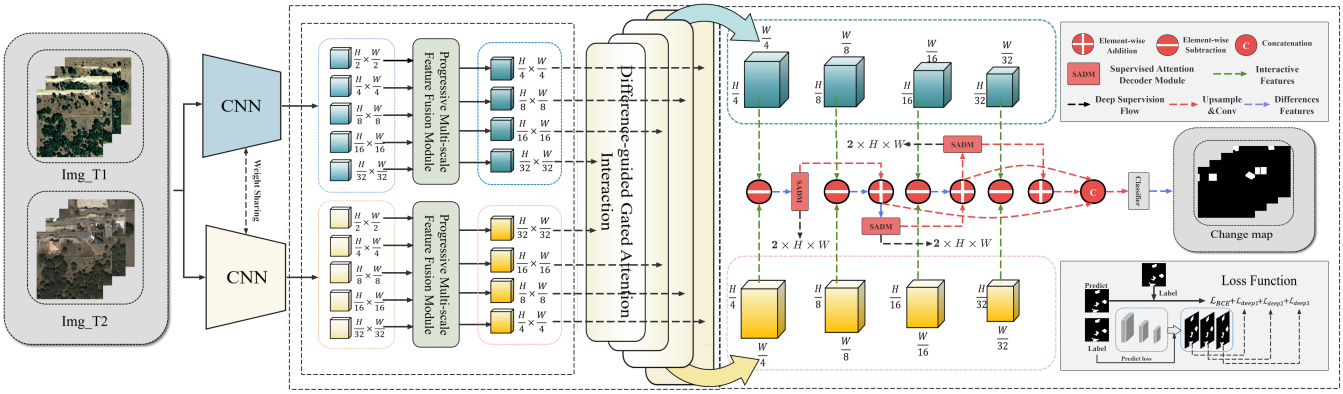

Figure 1. Proposed MISANet framework. MISANet adopts a Siamese encoder–decoder architecture with a shared MobileNetV2 backbone for bi-temporal feature extraction. Multi-level features are progressively fused by PMFFM to form semantically consistent representations, followed by DGAI, which integrates difference-guided attention with adaptive gating to enhance true changes and suppress pseudo-changes. In the decoder, SADM refines features via channel–spatial attention under deep supervision, and multi-scale outputs are fused to produce the final change map.

To enhance cross-temporal interaction, a Difference-Guided Attention Interaction (DGAI) module is introduced to explicitly incorporate bi-temporal difference cues and highlight change-relevant features while suppressing pseudo-changes. In the decoding stage, a Supervised Attention Decoder Module (SADM) performs channel–spatial attention

refinement under deep supervision. Finally, multi-scale decoder outputs are concatenated, transformed, and fed into a classifier to generate the final change map.

Although the individual modules of MISANet are based on established techniques, its core novelty lies in a tightly coupled interaction paradigm that integrates progressive multi-scale alignment, difference-guided attention, and supervision-aware decoding under efficiency constraints, enabling robust and computationally efficient change detection.

2.1.2. Progressive Multi-Scale Feature Fusion Module (PMFFM)

To fully integrate feature representations from different scales of the backbone network, we design a Progressive Multi-Scale Feature Fusion Module (PMFFM). This module receives five feature maps (denoted as X1, X2, X3, X4, X5) of different spatial resolutions generated by MobileNetV2 and progressively fuses them through four hierarchical fusion stages. In the first stage, the five feature maps are resampled to the same spatial resolution through upsampling and downsampling operations. These rescaled feature maps are then concatenated along the channel dimension and passed through a convolutional layer to compress the channel number, resulting in the first-level multi-scale fused feature representation.

$$\begin{aligned} X_1 &= \mathrm{Down}(X1) + \mathrm{Conv}_{1\times1}(X2) + \mathrm{Up}(X3) + \mathrm{Up}(X4) + \mathrm{Up}(X5) \\ X_1' &= \mathrm{Conv}_{1\times1}(X_1) \end{aligned} \tag{1}$$

In the second stage, the first-level fused feature is treated as the new input and fused again with the remaining original features at different scales. Through similar rescaling, concatenation, and channel compression operations, we obtain the second-level fused feature map:

$$\begin{aligned} X_2 &= \mathrm{Down}(X1) + \mathrm{Down}(X_1) + \mathrm{Conv}_{1\times1}(X3) + \mathrm{Up}(X4) + \mathrm{Up}(X5) \\ X_2' &= \mathrm{Conv}_{1\times1}(X_2) \end{aligned} \tag{2}$$

In the third stage, we further integrate more contextual and semantic information on top of the previous fused features. Following the same progressive strategy, the third-level fusion captures higher-level semantic features while maintaining shallow texture details:

$$\begin{aligned} X_3 &= \mathrm{Down}(X1) + \mathrm{Down}(X_1) + \mathrm{Down}(X_2) + \mathrm{Conv}_{1\times1}(X4) + \mathrm{Up}(X_5) \\ X_3' &= \mathrm{Conv}_{1\times1}(X_3) \end{aligned} \tag{3}$$

Finally, in the fourth stage, the outputs from the first three fusion stages are combined with the deepest feature map X5 to generate the final fused representation:

$$\begin{aligned} X_4 &= \mathrm{Down}(X1) + \mathrm{Down}(X_1) + \mathrm{Down}(X_2) + \mathrm{Down}(X_3) + \mathrm{Conv}_{1\times1}(X_5) \\ X_4' &= \mathrm{Conv}_{1\times1}(X_4) \end{aligned} \tag{4}$$

Through this progressive fusion mechanism, each stage effectively incorporates semantic information from multiple scales, gradually refining the representation from low-level texture to high-level semantics. As a result, the fourth-level fused feature map contains both fine-grained spatial details and abstract semantic cues, providing a robust representation for subsequent change detection tasks. The structure of the PMFFM is illustrated in Figure 2. Here, Down() and Up() denote downsampling and upsampling operations, respectively, while Conv() represents a convolutional layer equipped with batch normalization and ReLU activation.

Within the Progressive Multi-Scale Feature Fusion Module (PMFFM), all of the downsampling and upsampling is implemented using bilinear interpolation-based rescaling.

Downsampling is realized by bilinear interpolation with a scale factor of 0.5, while upsampling is performed via corresponding scale factors to match the target resolution. After rescaling, a 1 × 1 convolution is applied to each feature map to perform channel alignment and compression before fusion.

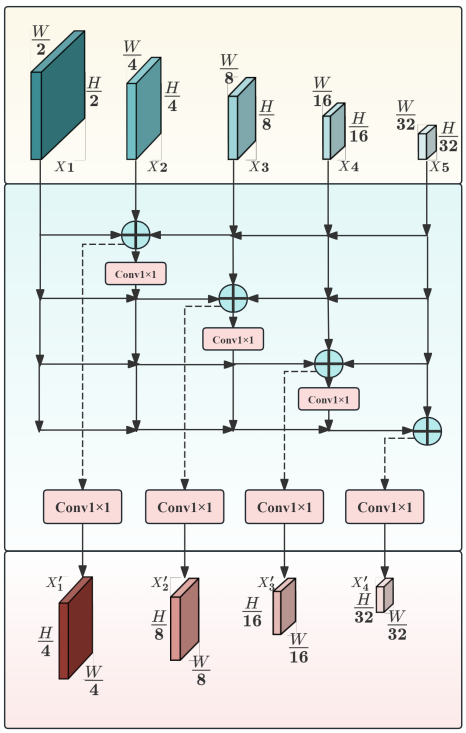

Figure 2. Progressive Multi-Scale Feature Fusion Module.

2.1.3. DGAI: Difference-Guided Gated Attention Interaction

In remote sensing change detection, images captured at different times often exhibit significant variations. These differences are not only caused by structural changes in the scene but also by non-structural factors such as illumination, seasonal variation, and atmospheric conditions. Directly comparing bi-temporal images may therefore introduce pseudo-changes and background noise, making it difficult to distinguish true changes. To suppress pseudo-changes and highlight genuine change regions, we propose a Difference-Guided Attention Interaction (DGAI) module. The structure of DGAI is shown in Figure 3.

Specifically, this module first computes the absolute difference between the bi-temporal feature pairs and concatenates it with the original input features. The concatenated feature is then processed through a global attention structure, which consists of a two-dimensional adaptive average pooling layer followed by a bottleneck convolution. This structure produces a global semantic guidance vector G that captures image-level contextual variation.

$$G = Conv_{1\times1}(Avgpool(Concat(X_1, X_2, |X_1 - X_2|))) \quad (5)$$

Formally, $Concat(\cdot)$ denotes channel-wise concatenation, $AvgPool(\cdot)$ represents two-dimensional adaptive average pooling, and $Conv(\cdot)$ indicates a one-dimensional convolution operation.

The generated guidance vector encodes holistic semantic dependencies across the entire spatial domain and, through the bottleneck convolution, forms a compact weight tensor. This tensor subsequently participates in the query–key interaction during the attention computation, providing a global semantic prior that enables the attention mechanism

to model cross-temporal global trends while maintaining structural awareness within each temporal feature.

To further enhance feature interaction, we design a pair of self-attention-based Transformer blocks for the bi-temporal features. Taking one input feature map as an example, it is passed through three successive one-dimensional convolution layers, projecting the feature into the query (Q), key (K), and value (V) spaces, respectively. The channel-transformed features can be expressed as $Q, K, V \in \mathbb{R}^{C \times H \times W}$. Each of these three feature tensors is then reshaped into a sequence form as follows:

$$Q' \in \mathbb{R}^{N \times \frac{C}{4}}, K' \in \mathbb{R}^{N \times \frac{C}{4}}, V' \in \mathbb{R}^{N \times C}, N = H \times W \tag{6}$$

where H and W denote the spatial dimensions, and C denotes the number of channels. Reshape(·) indicates the transformation from a tensor representation to a sequence of tokens. The standard self-attention mechanism is then formulated as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where Softmax(·) represents the normalization function, T denotes the transpose operation, and d_k is the key dimension.

In our design, the global semantic guidance vector G is transposed and multiplied with Q, similar to K, to obtain a global similarity representation. The local similarity QK^T and global similarity QG^T are then element-wise added and activated by a Softmax function to obtain the shared attention matrix. The final shared attention response can thus be formulated as

$$AttShare(Q, K, V, G) = Softmax(QK^T + QG^T)V \tag{8}$$

This shared attention mechanism adaptively reweights the features, amplifying the responses of true change regions while suppressing irrelevant background areas. However, during this process, some pseudo-changes or noise may also be inadvertently enhanced. To maintain a balance between the original and the attention-enhanced features [48], we introduce a gated fusion mechanism, which adaptively regulates the contribution of both components. The gating mechanism is defined as

$$\begin{aligned} F_{out} &= A \cdot F_{att} + (1 - A) \cdot F_{ori} \\ A &= \sigma(Conv_{1\times1}(Concat(F_{att}, F_{ori}))) \\ F_{att} &= AttShare(Q, K, V, G) \end{aligned} \tag{9}$$

where A represents the gating tensor and $\sigma(\cdot)$ is a nonlinear activation function that maps feature values to the range [0, 1]. F_{att} represents the input feature map. This design ensures a dynamic equilibrium between attention enhancement and original feature preservation. Through this mechanism, the network can effectively suppress pseudo-changes caused by illumination, seasonal, or shadow variations while emphasizing genuine structural changes, thereby providing more robust feature representations for subsequent detection processes.

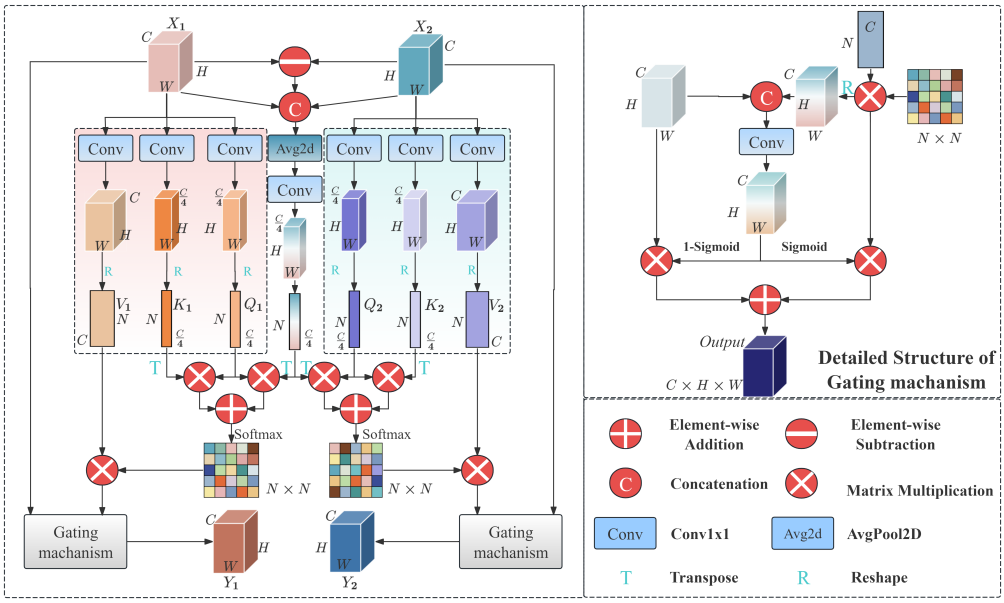

Figure 3. Difference-Guided Gated Attention Interaction.

2.1.4. SADM: Supervised Attention Decoder Module

In remote sensing change detection tasks, factors such as solar illumination variations [49], sensor noise [50], and seasonal changes may cause significant spectral or textural differences even in areas without real land-cover changes. These pseudo-changes reduce the accuracy of change detection. Moreover, objects in remote sensing imagery often exhibit large variations in spatial scale, while the fixed receptive field of conventional convolutions [51] limits the network's ability to simultaneously capture both large-scale and small-scale changes. This may lead to missed detections of small objects and blurred boundaries. To address these issues, we design a Supervised Attention Decoder Module (SADM) that enables the network to maintain multi-scale perception and suppress pseudo-changes during the upsampling process. The structure of the SADM is illustrated in Figure 4.

The input feature map $f_{in} \in \mathbb{R}^{C\times H\times W}$ is first refined using the Channel–Spatial Attention Mechanism (CSAM), which redistributes feature weights along both the channel and spatial dimensions to emphasize the true change regions. The CSAM consists of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM).

In the CAM, global average pooling and global max pooling are applied along the spatial dimension to aggregate channel-wise information, generating two 1D vectors. These vectors are then passed through a Multi-Layer Perceptron (MLP), followed by element-wise addition and a Sigmoid activation to produce the channel attention map. In the SAM, the input feature tensor is first processed by average pooling and max pooling along the channel dimension. The resulting two feature maps are concatenated and passed through a 7×7 convolutional block, followed by a Sigmoid activation, to generate the spatial attention map. The CAM and SAM operations can be formulated as

$$\begin{aligned} W_r &= \sigma((MLP(AvgPool(f_{in})) + MLP(MaxPool(f_{in}))) \\ f_r &= W_r \cdot f_{in} \\ f_{r'} &= f_r + Conv_{3\times3}(f_r) \\ W_s &= \sigma(Conv_{7\times7}(Concat(MaxPool(f_{r'}), AvgPool(f_{r'})))) \\ f_s &= W_s \cdot f_{r'} \end{aligned} \tag{10}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $MLP(\cdot)$ denotes the multilayer perceptron, and $f^{7\times7}$ represents a set of 7 × 7 convolutions, batch normalization, and ReLU activation functions. $AvgPool(\cdot)$ represents the average-pooling and $MaxPool(\cdot)$ represents the max-pooling. f_r represents the feature map refined by the CAM, while f_s represents the feature map refined by the SAM.

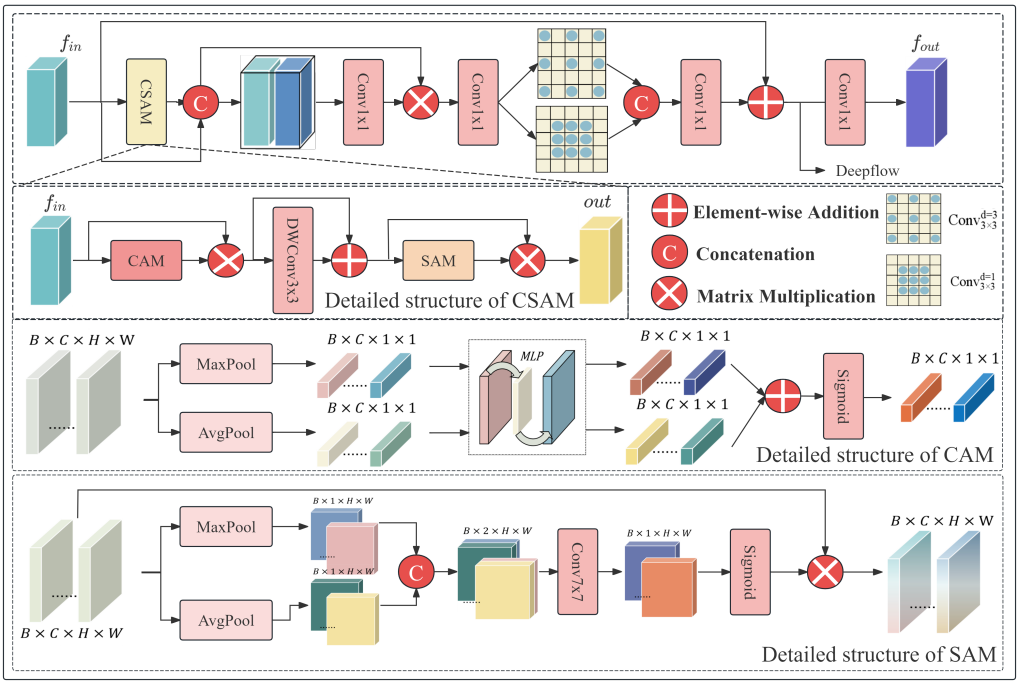

Figure 4. Supervised Attention Decoder Module.

To balance enhanced semantic information with the original feature representation, the output f_s and f_{in} are concatenated along the channel dimension. The concatenated feature is then passed through a 1 × 1 convolution followed by batch normalization, which performs pixel-wise weighting to suppress local responses caused by pseudo-changes while preserving stable responses to true changes. The process can be expressed as

$$\begin{aligned} f_t' &= Concat(f_{in}, f_s) \\ f_t &= Conv_{1\times1}(Conv_{1\times1}(f_t') \cdot f_t') \end{aligned} \tag{11}$$

where f_t represents the balanced feature map.

Next, we employ a dual-branch 3 × 3 convolutional structure with different dilation rates to simultaneously capture local and global contextual information. One branch uses a dilation rate of 1 to preserve fine-grained details, while the other uses a dilation rate of 3 to capture larger-scale contextual dependencies. The resulting feature maps are concatenated and compressed through a 1 × 1 convolution, after which a residual connection is introduced to integrate the original feature representation and maintain the consistency of change information. Formally, the process is defined as

$$\begin{aligned} f_u &= Conv_{1\times1}(Concat(Conv_{3\times3}^{d=1}(f_t), Conv_{3\times3}^{d=3}(f_t))) \\ f_{out} &= Conv_{1\times1}(f_u + f_{in}) \end{aligned} \tag{12}$$

where $Conv_{3\times3}^{d=1}(\cdot)$ and $Conv_{3\times3}^{d=3}(\cdot)$ denote 3 × 3 convolutions with dilation rates of 1 and 3, respectively, f_u and f_{out} represent the dual-branch fused feature map and the final output of the SADM, respectively.

In the decoder and deep supervision branches, upsampling operations are consistently implemented using bilinear interpolation to progressively restore spatial resolution, followed by convolutional refinement. This unified interpolation-based rescaling strategy ensures stable feature alignment across different scales, avoids checkerboard artifacts introduced by transposed convolutions, and improves reproducibility while maintaining computational efficiency.

Through the combination of supervised attention, multi-scale dilated convolutions, and residual enhancement, the proposed SADM effectively suppresses pseudo-changes induced by illumination and seasonal variations while preserving precise boundary details and small-scale variations. This design ensures that the decoder maintains strong multi-scale awareness and produces stable, high-fidelity change detection maps.

2.2. Datasets

To comprehensively evaluate the effectiveness of the proposed MISANet model, three benchmark datasets for remote sensing change detection were employed: LEVIR-CD [52], SYSU-CD [53], and GZ-CD [54].

2.2.1. LEVIR-CD

LEVIR-CD is a large-scale change detection dataset comprising 637 pairs of high-resolution remote sensing images. Each image pair has a spatial resolution of 0.5 m/pixel and a size of 1024 × 1024 pixels, collected from ultra-high-resolution Google Earth imagery. The dataset spans 5–14 years, covering 20 different regions across various cities in Texas, including residential areas, skyscrapers, large warehouses, and garages. In practical applications, each image is cropped into 256 × 256 patches. Following a 7:1:2 split ratio, the dataset is divided into 7120 training pairs, 1024 validation pairs, and 2048 testing pairs. Representative samples are illustrated in Figure 5.

Figure 5. Representative samples of the LEVIR-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.2.2. SYSU-CD

The SYSU-CD dataset was created and released by Sun Yat-sen University in 2022. It consists of 20,000 pairs of bi-temporal remote sensing image patches, each of size 256 × 256 with a spatial resolution of 0.5 m/pixel. This dataset contains diverse and complex change scenarios designed for training, validation, and testing. Specifically, it is divided into

12,000 pairs for training, 4000 pairs for validation, and 4000 pairs for testing, following a 6:2:2 ratio. Several representative examples are shown in Figure 6.

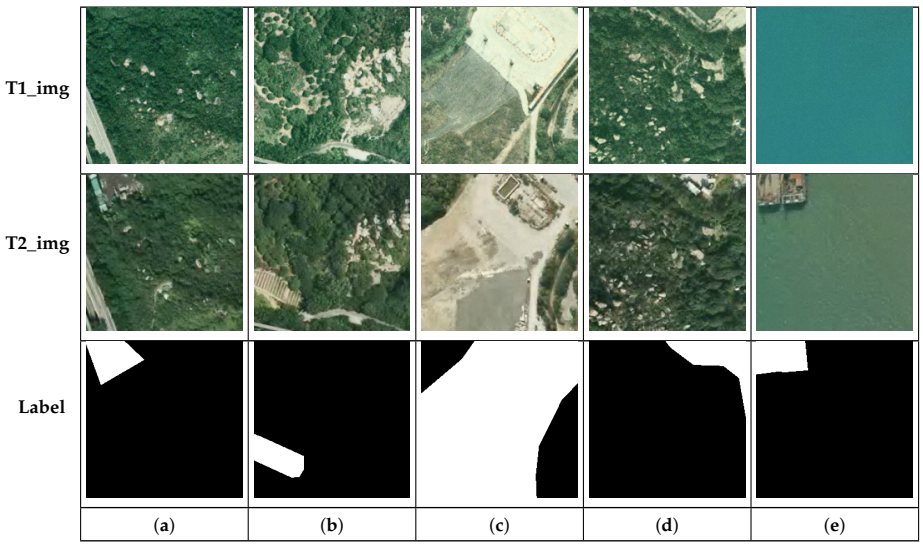

Figure 6. Representative samples of the SYSU-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.2.3. GZ-CD

The GZ-CD dataset is constructed for high-resolution remote sensing change detection. Its original images are obtained from Google Earth, covering the suburban areas of Guangzhou from 2006 to 2019. It consists of 19 pairs of bi-temporal images, with image sizes ranging from 1006 × 1168 to 4936 × 5224 pixels and a uniform spatial resolution of 0.55 m/pixel. Because the bi-temporal images maintain consistent resolutions, this dataset is particularly suitable for evaluating the resolution-invariant performance of change detection methods, assessing both model adaptability and generalization capability. Representative examples are displayed in Figure 7.

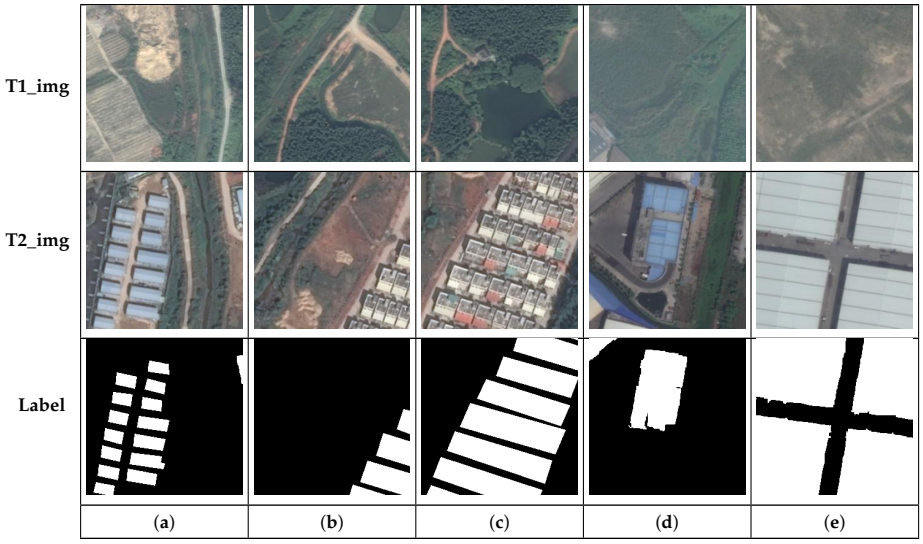

Figure 7. Representative samples of the GZ-CD dataset. (**a**–**e**) correspond to example images in the dataset.

2.3. Implementation Details

2.3.1. Experimental Environment

All experiments in this study were conducted on a workstation equipped with an NVIDIA GeForce RTX 4070 Super GPU(NVIDIA Corporation, Santa Clara, CA, USA), using the PyTorch 2.4.1 deep learning framework and Python 3.8 as the programming environment. Before training, input images were augmented through random scaling and cropping to improve the model's generalization ability. We adopted BCEWithLogitsLoss as the loss function and used the Adam optimizer for network optimization. To dynamically adjust the learning rate, the Poly learning rate policy was employed, where the learning rate at each epoch is updated as follows:

$$lr = lr_0 \times (1 - \frac{epoch}{num\_epoch})^p \tag{13}$$

The initial learning rate lr_0 was set to 1×10^{-5}, and the power was set to 0.9. The batch size was fixed at 8, and the network was trained for 500 epochs in total.

2.3.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed MISANet in remote sensing change detection tasks, we employed a set of widely used evaluation metrics, including Precision (PR), Recall (RC), Overall Accuracy (OA), Kappa Coefficient (KAPPA), Intersection over Union (IoU), and F1-score (F1). These metrics jointly provide a detailed assessment of both accuracy and robustness. Among them, F1-score serves as the primary indicator for assessing change-class prediction accuracy, while IoU evaluates the spatial overlap between predicted and ground-truth change regions. Additionally, PR, RC, OA, and KAPPA act as complementary metrics, providing a more comprehensive understanding of the model's reliability and precision. The calculation formulas for each metric are as follows:

$$PR = \frac{TP}{TP + FP} \tag{14}$$

$$RC = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2 \times PR \times RC}{PR + RC} \tag{16}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$E = \frac{(TP + FP) \times (TP + FN) + (FN + TN) \times (FP + TN)}{(TP + TN + FP + FN)^2} \tag{18}$$

$$KAPPA = \frac{OA - E}{1 - E} \tag{19}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{20}$$

2.3.3. Deep Supervision Training

In remote sensing change detection (RSCD) tasks, the model must simultaneously capture fine-grained texture variations and high-level semantic differences to accurately identify change regions. However, as network depth increases, the supervision signal from the final output tends to diminish during backpropagation, making it difficult for intermediate layers to receive sufficient gradient information. To address this problem, we introduce a deep supervision mechanism during decoding, enabling intermediate layers to receive direct supervision from the ground truth. We use BCEWithLogitsLoss as the loss function and assign weighted coefficients to emphasize learning in change regions.

Specifically, the total training loss consists of one main loss and three auxiliary losses, formulated as

$$L_{BCE}(c, g) = g \cdot \log c + (1 - g) \log(1 - c) \tag{21}$$

where c and g represent the predicted change map and corresponding ground truth, respectively. The overall training objective is expressed as

$$L_{total} = \lambda_1 L_{BCE}(c_1, g) + \lambda_2 L_{BCE}(c_2, g) + \lambda_3 L_{BCE}(c_3, g) + \lambda_4 L_{BCE}(c_4, g) \tag{22}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ denote the weights for the main and auxiliary branches, respectively. c_1 represents the prediction from the main decoder branch, while c_2, c_3, c_4 correspond to the outputs of the three auxiliary branches. This multi-level supervision effectively alleviates the gradient vanishing problem and guides the network to learn more discriminative representations for both semantic- and boundary-level change regions.

In Equation (22), the primary loss supervises the final prediction at full resolution and serves as the main optimization objective. The auxiliary losses are applied to intermediate decoder outputs to facilitate stable training by constraining intermediate representations and improving gradient propagation, rather than competing with the final objective. Based on the sensitivity analysis, the final loss weights are set to $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in all experiments. Accordingly, although the primary and auxiliary loss coefficients are numerically identical in the optimal setting, the primary loss plays the central optimization role, as it directly guides the final output, while the auxiliary losses act as regularization terms to support effective gradient propagation during decoding. All supervision terms adopt Binary Cross-Entropy (BCE) loss due to its stable gradients and consistent optimization behavior across decoding stages. Although class imbalance is common in RSCD, MISANet primarily mitigates this issue through architectural design, including difference-guided attention and supervised decoding. In this setting, BCE provides a reliable optimization objective without introducing potential instability associated with region-based or dynamically reweighted losses when applied to intermediate predictions.

3. Experiments and Results

3.1. Ablation Study

3.1.1. Ablation Study on the LEVIR-CD, SYSU-CD and GZ-CD Datasets

To comprehensively evaluate the effectiveness of the proposed MISANet, a series of ablation experiments were conducted on the LEVIR-CD, SYSU-CD, and GZ-CD datasets. Throughout all experiments, the training strategies and hyperparameters were kept consistent to ensure scientific fairness and reliability. We progressively integrated each proposed module into the backbone network in different combinations, and the results are summarized in Table 1. We primarily focus on the IoU and F1-score metrics to validate the individual contributions and overall effectiveness of each module.

Ablation of PMFFM: The proposed Progressive Multi-Scale Feature Fusion Module (PMFFM) effectively integrates semantic representations across multiple scales in a progressive manner, thereby enhancing the expressive capability of bi-temporal features. As shown in Table 1, incorporating the PMFFM into the backbone network leads to performance improvements, with the F1-score and IoU increasing by 1.04% and 1.70% on LEVIR-CD, 1.23% and 1.7% on SYSU-CD, 1.37% and 2.08% on GZ-CD, respectively. These results clearly demonstrate the effectiveness of PMFFM in improving multi-scale feature fusion and strengthening the representation of temporal variations.

Ablation of DGAI: The proposed Difference-Guided Attention Interaction (DGAI) module introduces difference features into the attention interaction process between bi-temporal features, aiming to enhance the model's ability to distinguish true changes from

pseudo-changes. As presented in Table 1, integrating DGAI into the backbone results in F1 and IoU gains of 1.20% and 1.97% on LEVIR-CD, 1.37% and 1.89% on SYSU-CD, 1.51% and 2.30% on GZ-CD, respectively. When both PMFFM and DGAI are incorporated, the model achieves 90.57% F1 and 82.77% IoU on LEVIR-CD, 81.05% F1 and 68.14% IoU on SYSU-CD, and 87.55% F1 and 77.86% IoU on GZ-CD, representing additional improvements of 0.53% F1 and 0.89% IoU on LEVIR-CD, 0.87% F1 and 1.22% IoU on SYSU-CD, and 1.56% F1 and 2.44% IoU on GZ-CD over the PMFFM-only configurations. This indicates that DGAI significantly enhances cross-temporal feature interaction and semantic discrimination capabilities within the network.

Ablation of SADM: In the decoding stage, we design a Supervised Attention Decoder Module (SADM) that effectively combines deep supervision with channel–spatial attention to enhance the network's sensitivity to object boundaries and fine-grained change regions. According to Table 1, adding SADM to the backbone improves the F1-score and IoU by 1.32% and 2.17% on LEVIR-CD, 1.48% and 2.05% on SYSU-CD, and 1.65% and 2.52% on GZ-CD, respectively. When SADM is introduced on top of the backbone equipped with PMFFM and DGAI, the F1-score and IoU further increase by 0.62% and 1.04% on LEVIR-CD, 1.20% and 1.71% on SYSU-CD, and 0.80% and 1.27% on GZ-CD, respectively. These results demonstrate that SADM substantially improves the accuracy, boundary precision, and overall robustness of the proposed MISANet model.

Table 1. Ablation results of the proposed modules on the LEVIR-CD, SYSU-CD, and GZ-CD. All experiments are conducted under identical training settings, and each experiment is repeated three times with different random seeds. The reported results are presented as mean ± standard deviation. The best performance is highlighted in bold.

Dataset	Methods	PMFFM	DGAI	SADM	IoU (%)	F1 (%)
LEVIR-CD	Baseline+				80.18 ± 0.12	89.00 ± 0.09
	Baseline+	✓			81.88 ± 0.11	90.04 ± 0.08
	Baseline+		✓		82.15 ± 0.14	90.20 ± 0.07
	Baseline+			✓	82.35 ± 0.13	90.32 ± 0.06
	Baseline+	✓	✓		82.77 ± 0.12	90.57 ± 0.09
	Baseline+	✓		✓	82.80 ± 0.13	90.59 ± 0.07
	Baseline+		✓	✓	83.02 ± 0.15	90.72 ± 0.08
	Baseline+	✓	✓	✓	**83.81 ± 0.14**	**91.19 ± 0.07**
SYSU-CD	Baseline+				65.22 ± 0.15	78.95 ± 0.10
	Baseline+	✓			66.92 ± 0.17	80.18 ± 0.11
	Baseline+		✓		67.11 ± 0.14	80.32 ± 0.09
	Baseline+			✓	67.27 ± 0.13	80.43 ± 0.00
	Baseline+	✓	✓		68.14 ± 0.17	81.05 ± 0.10
	Baseline+	✓		✓	68.72 ± 0.11	81.46 ± 0.06
	Baseline+		✓	✓	69.02 ± 0.10	81.67 ± 0.09
	Baseline+	✓	✓	✓	**69.85 ± 0.11**	**82.25 ± 0.07**
GZ-CD	Baseline+				73.34 ± 0.21	84.62 ± 0.15
	Baseline+	✓			75.42 ± 0.17	85.99 ± 0.13
	Baseline+		✓		75.64 ± 0.15	86.13 ± 0.12
	Baseline+			✓	75.86 ± 0.14	86.27 ± 0.12
	Baseline+	✓	✓		77.86 ± 0.16	87.55 ± 0.14
	Baseline+	✓		✓	77.59 ± 0.18	87.38 ± 0.15
	Baseline+		✓	✓	77.76 ± 0.15	87.49 ± 0.13
	Baseline+	✓	✓	✓	**79.13 ± 0.17**	**88.35 ± 0.14**

3.1.2. Ablation Study on DGAI

To analyze the individual and combined effects of the Difference-Guided Gated Attention Interaction (DGAI) module, an ablation study is conducted to disentangle the effects of the difference-guidance term and the gating mechanism in DGAI. By comparing four configurations—DGAI without G and Gating, G only, Gating only, and full DGAI—under identical experimental settings on three datasets, we assess whether the observed improvements arise from individual components or their synergistic interaction.

The quantitative results of the ablation study are summarized in Table 2. As shown in the table, the complete DGAI consistently achieves the best performance across all three datasets, demonstrating the effectiveness of jointly employing difference guidance and gating. On LEVIR-CD, enabling only the difference-guided tensor or the gating mechanism improves the F1-score to 90.62% and 90.73%, respectively, compared with the baseline, while their combined use further increases the F1-score to 91.19%. Similar trends are observed on SYSU-CD, where the full DGAI achieves the highest F1-score of 82.25%, as well as on GZ-CD, where the combined configuration improves the F1-score to 88.35%. These consistent improvements indicate that the two components contribute in a complementary manner rather than as independent enhancements. To further analyze their spatial effects, qualitative visualizations of gating responses under different ablation settings are provided in Figure 8.

Table 2. Ablation study of the two key components in the proposed DGAI module across three datasets: LEVIR-CD, SYSU-CD, and GZ-CD. G denotes the difference-guided tensor and Gating denotes the gating mechanism in DGAI. ✓ and ✗ indicate that the corresponding component is enabled or removed, respectively. All experiments are conducted under identical training settings. All reported values are F1-scores (%). The best performance is highlighted in bold.

G	Gating	LEVIR-CD	SYSU-CD	GZ-CD
✗	✗	90.54	80.47	86.98
✗	✓	90.62	81.23	87.52
✓	✗	90.73	81.06	87.64
✓	✓	**91.19**	**82.25**	**88.35**

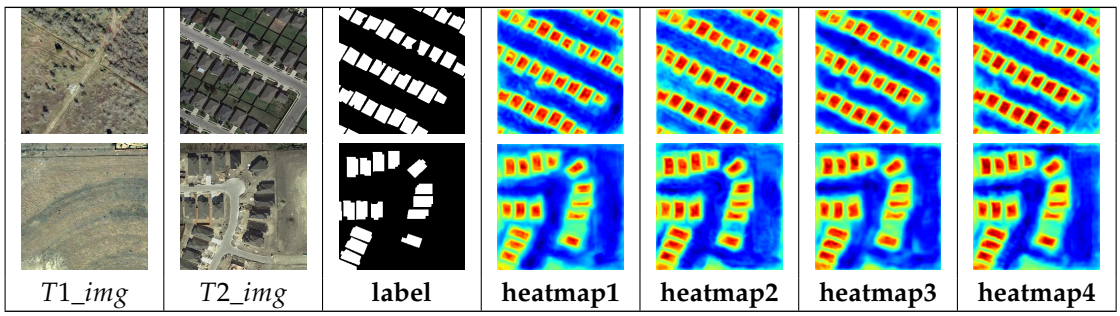

Figure 8. Heatmaps depicting the ablation of DGAI. Specifically, heatmap1 corresponds to the model with both the difference-guided tensor G and the gating mechanism removed, heatmap2 shows the response with the difference-guided tensor G removed but the gating mechanism retained, heatmap3 represents the configuration with the gating mechanism removed while keeping G, and heatmap4 depicts the complete DGAI model.

Figure 8 visualizes the attention heatmaps under different DGAI ablation settings, providing intuitive support for the quantitative results in Table 2. When both the difference-guided tensor and the gating mechanism are removed, the model exhibits diffuse and noisy activations over background regions, indicating poor discrimination between true changes and irrelevant variations. When only one component is enabled, complementary limitations

emerge. Using only the difference-guided tensor yields more focused responses around potential change regions but it still suffers from over-activation caused by background clutter or illumination variations. In contrast, using only the gating mechanism effectively suppresses background noise but also weakens responses to subtle or fine-grained changes due to the lack of explicit semantic guidance. By integrating both components, the complete DGAI module produces well-balanced activations that are consistently aligned with true change regions while suppressing irrelevant variations. This demonstrates that the difference-guided tensor and gating mechanism play complementary roles—change localization and adaptive noise suppression, respectively—and that the performance gains arise from their coordinated interaction rather than any single isolated component.

3.1.3. Ablation Study on Loss Weight Settings

To assess the sensitivity of MISANet to loss weight settings, we conduct an ablation study by varying the auxiliary supervision coefficients while fixing the primary loss weight. The three auxiliary coefficients are set equally and increased from 0 to 1 to ensure supportive rather than dominant supervision. As shown in Table 2, the results on three benchmark datasets demonstrate a consistent performance improvement with stronger auxiliary supervision.

The ablation results in Table 3 demonstrate that the performance of MISANet is not overly sensitive to the specific choice of auxiliary loss weights. As the auxiliary coefficients increase from 0 to 1 while keeping the primary loss weight fixed, the F1-score exhibits a stable and monotonic improvement across all three datasets, indicating that auxiliary supervision consistently facilitates optimization without interfering with the primary objective. This behavior confirms that the adopted weighting strategy effectively improves gradient propagation in intermediate layers while preserving stable convergence.

Table 3. Ablation study of loss function coefficients. All reported values are F1-scores (%). The best performance is highlighted in bold.

λ_1	$\lambda_2 = \lambda_3 = \lambda_4$	LEVIR-CD	SYSU-CD	GZ-CD
1	0	90.27	81.46	87.32
1	0.25	90.42	81.61	87.54
1	0.5	90.65	81.77	87.81
1	0.75	90.88	81.93	87.99
1	1	**91.19**	**82.25**	**88.35**

3.2. Comparative Experiments

To evaluate the effectiveness and superiority of the proposed MISANet, we conducted comparative experiments against several state-of-the-art (SOTA) change detection (CD) methods, including FC-EF [39], FC-Siam-Diff [39], FC-Siam-Conc [39], ChangeNet [55], Deeply Supervised Image Fusion Network (DSIFN) [56], Bitemporal Image Transformer (BIT) [43], Siamese Nested U-Net (SNUNet) [57], Intra-scale Cross Interaction and Inter-scale Feature Fusion Network (ICIFNet) [58], Dual-branch Multi-level Intertemporal Network (DMINet)[59], Siamese Attention-Guided Network (SAGNet) [60], Bitemporal Attention Sharing Network (BASNet) [61], Attention-based Multi-branch Fusion Network (ABMFNet) [62], AANet [63], and LCCDMamba [64]. All models were evaluated on three benchmark datasets: LEVIR-CD, SYSU-CD, and GZ-CD.

To ensure fairness and reliability, all comparative methods were trained under identical experimental settings, including the same optimizer, learning rate policy, and number of epochs. This guarantees that the observed performance differences arise solely from

the network architectures rather than training discrepancies. We further report several quantitative indicators for model comparison:

Params (M)—The total number of trainable parameters, measured in millions, representing the model's complexity and capacity.

FLOPs (G)—The number of floating-point operations required for a single forward pass, measured in billions, indicating the computational cost.

Time (ms)—The inference time for a single forward propagation, measured in milliseconds, providing a direct measure of model efficiency and speed. These metrics jointly provide a comprehensive assessment of accuracy, efficiency, and computational complexity, enabling an objective evaluation of the performance advantages of MISANet compared with existing methods.

FPS (Frames Per Second)—The inference throughput of the model, calculated based on the per-image inference time, reflecting the number of images processed per second under batch size set to one. This metric complements the reported inference time by providing a more intuitive assessment of real-time performance, which is particularly relevant for large-scale remote sensing change detection tasks and efficiency-sensitive deployment scenarios.

3.2.1. Comparative Experiments on LEVIR-CD

The quantitative results of various methods on the LEVIR-CD dataset are summarized in Table 4. As shown, the proposed MISANet demonstrates superior performance across multiple evaluation metrics. Specifically, our method achieves a Precision of 92.35%, an F1-score of 91.19%, and an IoU of 83.81%. Compared with the best-performing competing approach, MISANet achieves additional improvements of 0.30% in F1-score and 0.85% in IoU. When compared with several representative methods, our model consistently attains higher F1-scores while maintaining a more lightweight architecture in terms of parameter count. These results clearly indicate that MISANet outperforms existing methods in both detection accuracy and computational efficiency.

Table 4. Comparative experiments on the LEVIR-CD dataset (best results are highlighted in bold type).

Methods	PR (%)	RC (%)	OA (%)	KAPPA (%)	IoU(%)	F1 (%)	Params (M)	FLOPs (G)	Time (ms)	FPS
SNUNet	91.51	88.51	99.00	89.49	81.79	89.98	12.03	54.82	9.66	103.52
SAGNet	91.79	88.76	99.02	89.58	81.98	90.10	32.23	12.25	25.32	39.49
ICIFNet	91.31	87.23	98.56	89.16	81.24	89.18	23.84	24.51	49.53	20.18
FC_EF	85.58	80.89	98.33	82.30	71.19	83.17	1.35	3.57	7.59	131.75
FC_Diff	89.49	80.67	98.53	84.08	73.69	84.85	1.35	4.72	5.13	194.93
FC_CONC	86.76	85.83	98.61	85.56	75.89	86.29	1.55	5.32	5.22	191.57
DSIFN	91.53	85.70	98.87	87.75	79.12	88.34	35.73	82.26	12.13	82.44
DMINet	92.02	87.77	98.99	89.31	81.56	89.85	6.24	14.55	12.87	77.69
ChangeNet	91.63	86.88	98.93	88.63	80.49	89.19	47.20	10.91	17.01	58.79
BIT	91.26	88.51	98.98	89.33	81.59	89.86	3.49	10.63	16.12	62.03
BASNet	92.66	88.81	99.07	90.21	82.96	90.69	4.58	4.70	9.7	103.09
ABMFNet	83.02	82.06	96.53	80.68	68.88	81.57	29.56	66.17	22.35	44.75
AANet	91.85	88.93	99.03	89.30	82.42	90.36	15.82	24.21	14.57	68.63
LCCDMamba	91.87	89.03	97.85	88.79	82.53	90.43	93.90	38.20	5.70	175.44
Ours	**92.35**	**90.06**	**99.29**	**90.82**	**83.81**	**91.19**	8.53	3.49	19.66	50.87

The qualitative change-detection results on the LEVIR-CD dataset are illustrated in Figure 9. As shown in Figure 9I, methods such as FC-EF, FC-Diff, and FC-Conc tend to produce false detections in densely distributed small-scale buildings, whereas our approach identifies these fine-grained changes more precisely, effectively reducing both false-positive and missed-detection rates. In Figure 9II, for the vacant area in the lower-left region of the scene, illumination variations cause most comparison methods to miss newly constructed buildings to varying degrees. In contrast, MISANet achieves markedly higher detection

accuracy, significantly lowering omission and false-alarm rates. These findings provide further evidence of the robustness and effectiveness of the proposed network in handling complex urban scenes.

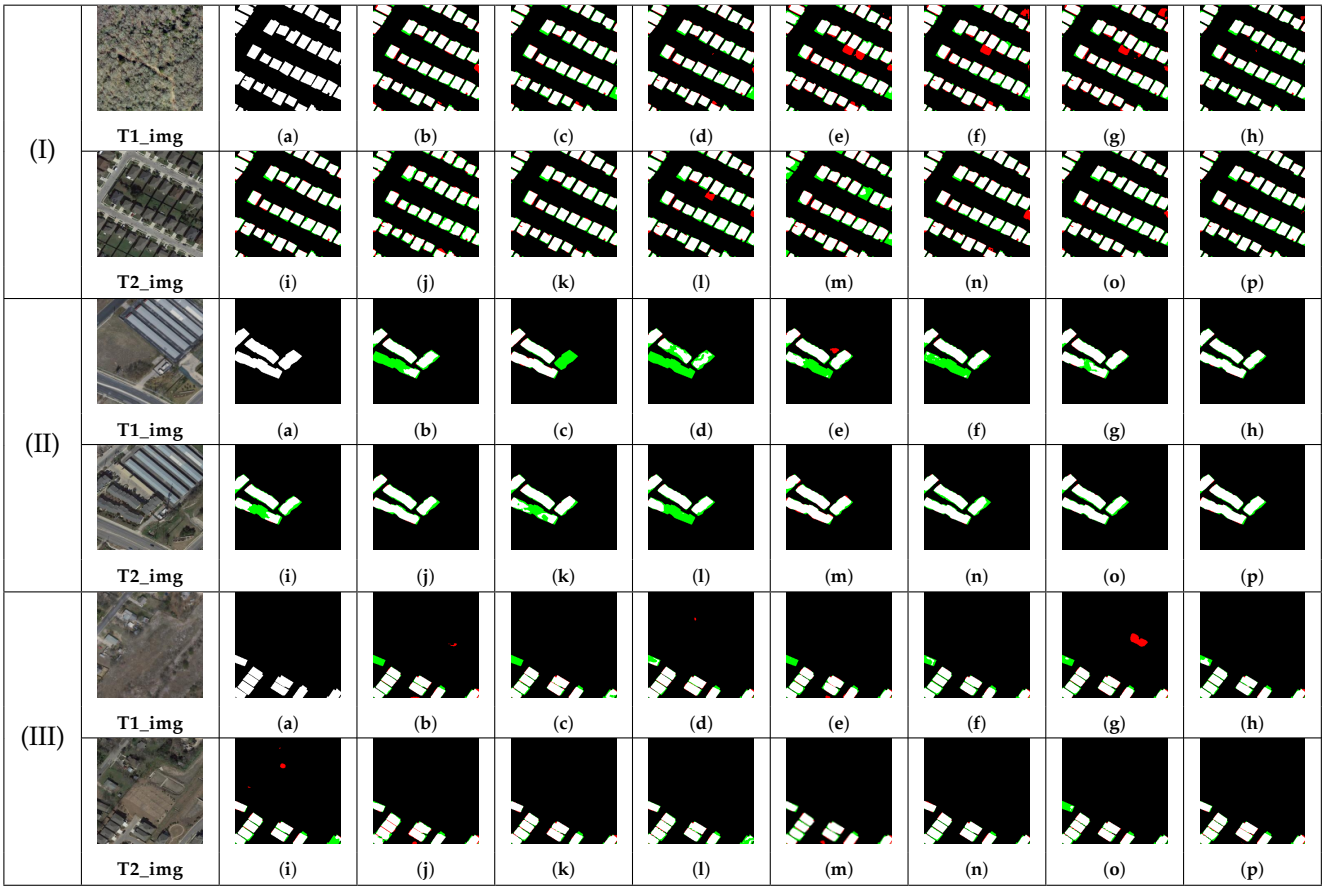

Figure 9. Three groups of comparative diagrams illustrating the performance of different algorithms on the LEVIR-CD dataset. (**I**) Densely distributed residential building areas. (**II**) Industrial areas with newly constructed large-scale buildings. (**III**) Complex scenes with sparse buildings and background interference. (**a**–**p**) correspond to the Ground-truth label SNUNet, SAGNet, ICIFNet, FC_EF, FC_Diff, FC_CONC, DSIFN, DMINet, ChangeNet, BIT, BASNet, ABMFNet, AANet, LCCDMamba, and our MISANet.

3.2.2. Comparative Experiments on SYSU-CD

To further validate the effectiveness of MISANet, we conducted comparative experiments on the SYSU-CD dataset. The quantitative results are presented in Table 5. As shown in the table, our method achieves the best overall performance among all compared approaches, with Precision, F1-score, and IoU reaching 88.23%, 82.25%, and 69.85%, respectively. Among the competing methods, SAGNet yields the second-best results; compared with it, MISANet improves the F1-score and IoU by 0.38% and 0.54%, respectively. These findings demonstrate that MISANet consistently attains higher F1-scores than other advanced methods, clearly highlighting its superiority and robustness on the SYSU-CD dataset. The qualitative comparison results are illustrated in Figure 10. In Figure 10I, for large vegetation-covered areas affected by illumination and shadow variations, most competing models (Figure 10I(b–e,g–i,m)) exhibit various degrees of false detection, whereas our method achieves higher accuracy and better visual consistency. Furthermore, our model produces smoother and more precise boundary predictions, maintaining robustness even in regions where targets and backgrounds share high visual similarity. As shown in Figure 10III , due to illumination changes and seasonal variations in

vegetation, several methods generate extensive false-positive and missed-detection regions. In contrast, MISANet, benefiting from its bi-temporal feature-interaction and multi-scale semantic-fusion modules, delivers significantly higher detection accuracy and reliability. These results collectively confirm the effectiveness and superiority of the proposed approach in handling complex environmental variations.

Table 5. Comparative experiments on the SYSU-CD dataset (best results are highlighted in bold type).

Methods	PR (%)	RC (%)	OA (%)	KAPPA (%)	IoU (%)	F_1 (%)
SNUNet	79.37	78.39	90.10	72.42	21.61	78.88
SAGNet	81.25	81.76	91.72	76.57	69.31	81.87
ICIFNet	78.23	74.38	89.08	69.17	61.62	76.25
FC_EF	78.78	76.69	89.63	70.97	63.56	77.72
FC_Diff	80.35	74.26	88.71	64.42	55.11	71.06
FC_CONC	81.51	75.11	90.11	71.80	64.17	78.18
DSIFN	78.82	81.30	90.44	73.76	66.72	80.04
DMINet	81.54	79.44	91.15	74.59	67.06	80.28
ChangeNet	79.91	71.11	88.97	68.19	60.33	75.25
BIT	81.22	73.87	89.81	70.81	63.09	77.37
BASNet	82.13	80.08	91.26	75.41	66.83	80.12
ABMFNet	75.68	74.41	88.42	67.50	57.87	73.31
AANet	82.48	79.73	86.50	70.50	68.18	81.08
LCCDMamba	86.03	77.20	89.51	74.81	68.60	81.37
Ours	**88.23**	**82.51**	**92.30**	**77.38**	**69.85**	**82.25**

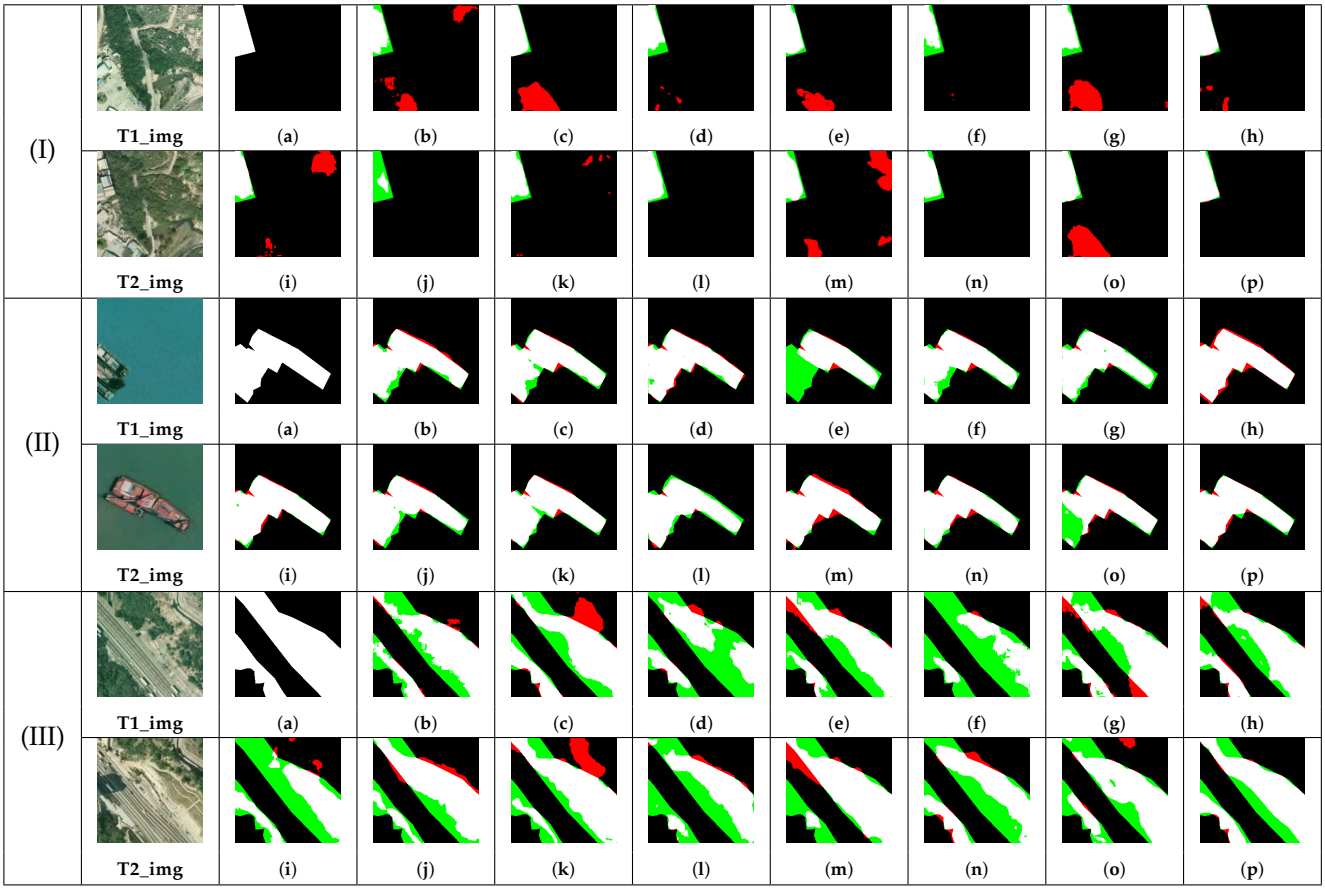

Figure 10. Three groups of comparative diagrams illustrating the performance of different algorithms on the SYSU-CD dataset. (**I**) Vegetation-covered areas affected by seasonal and illumination variations. (**II**) Waterbody scenes with ship-related changes. (**III**) Transportation infrastructure changes, such as roads or railways. (**a**–**p**) correspond to the Ground-truth label, SNUNet, SAGNet, ICIFNet, FC_EF, FC_Diff, FC_CONC, DSIFN, DMINet, ChangeNet, BIT, BASNet, ABMFNet, AANet, LCCDMamba, and our MISANet.

3.2.3. Comparative Experiments on GZ-CD

Finally, to further validate the effectiveness of the proposed MISANet, we conducted comparative experiments on the GZ-CD dataset. The quantitative results are summarized in Table 6. As observed, our method achieves the best overall performance among all compared approaches, with Precision, F1-score, and IoU reaching 90.21%, 88.35%, and 79.13%, respectively. Compared with the best-performing competing model, MISANet achieves improvements of 0.93% and 1.39% in F1-score and IoU, respectively. These results indicate that MISANet consistently delivers higher F1-scores than other advanced methods, highlighting its superiority and strong generalization capability on the GZ-CD dataset. The qualitative visual comparisons are presented in Figure 11. In Figure 11I , for the change regions on the left side of the scene, most models exhibit missed detections, while our method significantly reduces omission errors. In contrast, in the upper-right non-change area, several competing methods produce large-scale false alarms, whereas our model effectively suppresses them, yielding clean and precise results. In Figure 11II , for the vacant area adjacent to the factory, other methods show varying degrees of false detection, while MISANet accurately captures the true changes in that region. Moreover, our network produces smoother and more continuous boundaries around changed objects, indicating improved precision in capturing fine-grained edge variations. In Figure 11III , under complex land–water transformation scenarios, many models (Figure 11III(e–g)) fail to detect newly constructed buildings and even generate large false-alarm regions around them (Figure 11III(b–d,i–k)). In contrast, MISANet exhibits superior accuracy and robustness, precisely identifying real changes even in challenging environments with complex backgrounds. These findings further demonstrate the reliability and robustness of the proposed model in handling diverse and intricate real-world change detection scenarios.

In short, to comprehensively assess the trade-off between model complexity, computational cost, and detection accuracy, we compare several representative change detection methods in terms of their number of parameters (Params), floating-point operations (FLOPs), and F1-scores on three widely used benchmark datasets: LEVIR-CD, GZ-CD, and SYSU-CD. As shown in Figure 12, Figure 12a and Figure 12d correspond to the LEVIR-CD dataset, Figure 12b and Figure 12e to the SYSU-CD dataset, and Figure 12c and Figure 12f to the GZ-CD dataset, respectively.

Table 6. Comparative experiments on the GZ-CD dataset (best results are highlighted in bold type).

Methods	PR (%)	RC (%)	OA (%)	KAPPA (%)	IoU (%)	F_1 (%)
SNUNet	89.00	84.80	97.62	85.54	76.75	86.85
SAGNet	89.56	84.05	97.58	84.98	75.91	86.30
ICIFNet	88.09	81.31	97.25	83.05	73.25	84.56
FC_EF	79.86	65.53	95.28	69.44	56.24	71.99
FC_Diff	82.70	57.99	94.99	65.55	51.72	68.18
FC_CONC	82.16	62.80	95.29	68.67	55.26	71.19
DSIFN	89.35	75.46	96.91	79.83	68.76	81.49
DMINet	86.62	82.85	97.23	83.17	73.45	84.70
ChangeNet	88.63	82.99	97.44	84.32	75.01	85.72
BIT	86.80	82.04	97.18	82.80	72.94	84.35
BASNet	88.41	85.76	98.70	86.39	77.10	87.07
ABMFNet	85.43	77.77	96.71	79.62	68.66	81.42
AANet	89.00	85.70	95.20	84.50	77.40	87.28
LCCDMamba	89.53	85.46	93.26	82.83	77.74	87.42
Ours	**90.21**	**86.58**	**98.84**	**87.74**	**79.13**	**88.35**

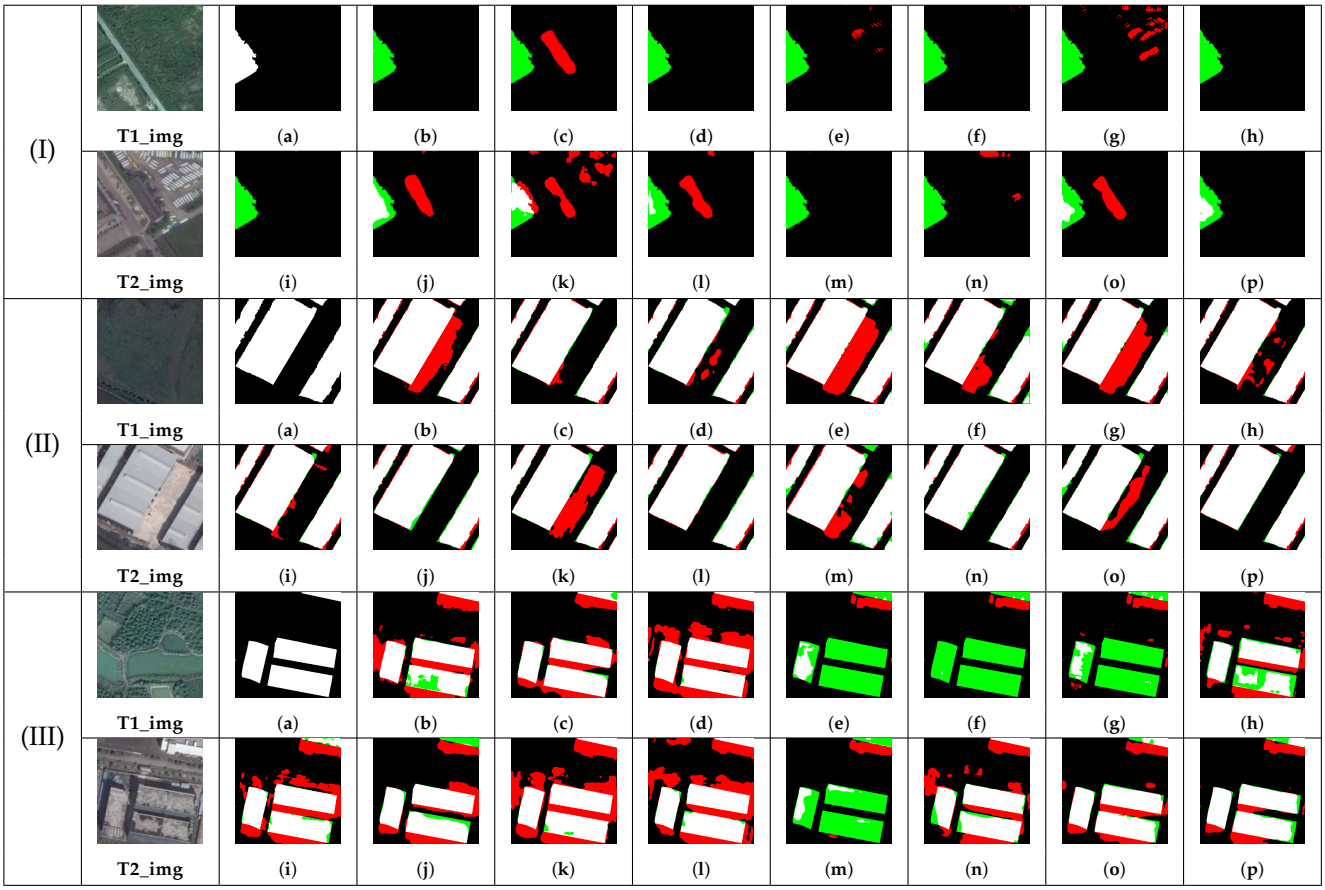

Figure 11. Three groups of comparative diagrams illustrating the performance of different algorithms on the GZ-CD dataset. (**I**) Road-adjacent construction and infrastructure changes. (**II**) Newly constructed industrial buildings. (**III**) Complex suburban scenes with scattered construction changes. (**a**–**p**) correspond to the Ground-truth label, SNUNet, SAGNet, ICIFNet, FC_EF, FC_Diff, FC_CONC, DSIFN, DMINet, ChangeNet, BIT, BASNet, ABMFNet, AANet, LCCDMamba and our MISANet.

The proposed MISANet consistently achieves a more favorable balance between accuracy and efficiency across all three datasets. Specifically, MISANet attains higher F1-scores while maintaining fewer parameters and lower FLOPs compared with other state-of-the-art methods. This demonstrates that MISANet effectively improves change detection performance without increasing computational burden. Its superior results can be attributed to the proposed multi-scale interaction and spatial attention mechanisms, which enable the network to capture fine-grained contextual cues and inter-scale dependencies more effectively, thereby enhancing discriminative feature representation with lightweight computation.

Overall, these experimental results confirm that MISANet achieves superior accuracy with compact model complexity and reduced computational cost, highlighting its potential for application in real-world remote sensing change detection tasks where both efficiency and accuracy are critical—such as onboard processing, edge deployment, and large-scale urban monitoring.

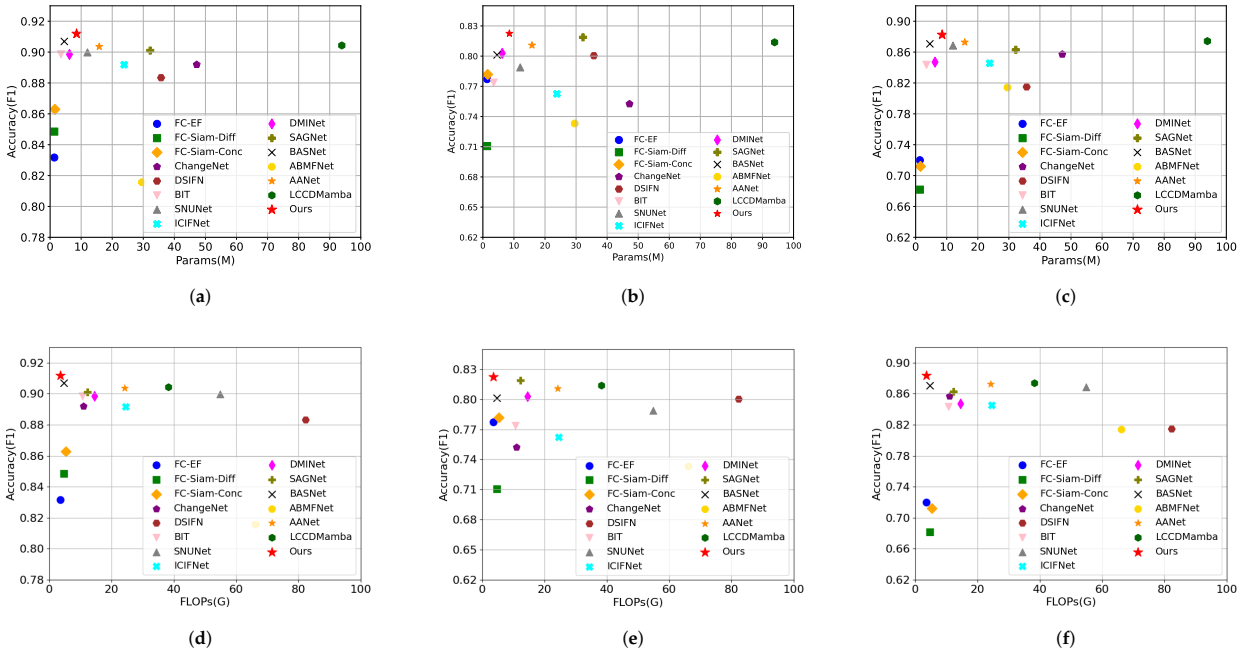

Figure 12. Model complexity comparison of different methods in terms of Params (memory cost), FLOPs (computational cost), and F1-score on the LEVIR-CD, GZ-CD, and SYSU-CD datasets, respectively. (**a**,**d**) LEVIR-CD dataset. (**b**,**e**) GZ-CD dataset. (**c**,**f**) SYSU-CD dataset. Our proposed MISANet method shows superior accuracy with fewer parameters and lower computation costs.

4. Discussion

4.1. Robustness of the DGAI Gating Mechanism Under Challenging Conditions

Although DGAI uses a gating mechanism to suppress pseudo-changes, gating weights derived from bi-temporal differences may be sensitive to noise or suppress real changes under weak-change or illumination-variant conditions. As shown in Figure 8, using gating alone reduces background activations but tends to over-suppress weak changes, whereas using only difference-aware guidance leads to over-activated background responses under low-contrast or illumination-induced variations, indicating that neither component alone is sufficient. By integrating difference-aware guidance with adaptive gating, DGAI mitigates these limitations; the difference-guided tensor provides explicit change-localization cues, while the gating mechanism adaptively suppresses noise and pseudo-changes. This complementary interaction yields balanced representations and explains the consistent performance gains in practical remote sensing scenarios.

4.2. Accuracy–Efficiency Trade-Off and Practical Applicability

In practical remote sensing change detection, accuracy must be balanced with computational efficiency, especially for large-scale and resource-constrained deployments. Although MISANet introduces dedicated interaction modules, the reported inference time and FPS show that its computational overhead remains moderate. The performance gains are therefore not achieved by increasing model size or complexity, but by more effective feature representation and interaction. This efficiency-aware design enables MISANet to achieve strong detection accuracy while maintaining near real-time inference, making it suitable for large-scale processing and efficiency-sensitive deployment scenarios.

4.3. Limitations and Potential Extensions of MISANet

Although the proposed MISANet demonstrates excellent performance in change detection tasks, several limitations should be acknowledged. Firstly, the current model

is specifically designed for bi-temporal remote sensing imagery. However, in real-world applications such as ecological environment monitoring and disaster assessment, it is often necessary to process multi-temporal data. A straightforward extension of the proposed bi-temporal interaction strategy to multi-temporal scenarios would inevitably lead to a significant increase in the number of parameters, thereby weakening the lightweight advantage of our network. Moreover, the proposed MISANet is primarily developed for optical imagery and does not incorporate multi-source data. In practical scenarios, optical images are easily affected by cloud cover, shadows, and seasonal variations, which may cause false spectral or semantic changes. Integrating multi-source data—such as SAR and hyperspectral imagery—could provide complementary information, effectively mitigating the influence of illumination and atmospheric variations. Such an extension would further enhance the robustness and accuracy of the model in detecting real land-cover changes under complex environmental conditions.

5. Conclusions

In this paper, we proposed a multi-scale interactive and deeply supervised network (MISANet) for remote sensing image change detection. The proposed framework integrates three key components—Progressive Multi-Scale Feature Fusion Module (PMFFM), Difference-Guided Attention Interaction Module (DGAI), and Supervised Attention Decoding Module (SADM)—to significantly enhance detection accuracy. Comprehensive ablation studies conducted on the LEVIR-CD dataset validate the effectiveness of each individual module. Furthermore, extensive comparative experiments on the LEVIR-CD, SYSU-CD, and GZ-CD datasets demonstrate the superior performance of our model, achieving F1-scores of 91.19%, 82.25%, and 88.35% and IoU values of 83.81%, 69.85%, and 79.13%, respectively. Compared with state-of-the-art (SOTA) methods, MISANet achieves higher accuracy and stronger robustness, confirming its effectiveness and generalization capability for real-world change detection tasks.

Author Contributions: Conceptualization, H.Y., J.W., S.L. and Y.W.; methodology, H.Y., J.W., Y.L. and T.G.; software, H.Y. and J.W.; validation, S.L., Y.W. and Y.L.; formal analysis, Y.L. and T.G.; investigation, S.L. and Y.W.; resources, M.X.; data curation, S.L.; writing—original draft preparation, H.Y.; writing—review and editing, Y.L.; visualization, H.Y. and T.G.; supervision, M.X.; project administration, M.X.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 42075130.

Data Availability Statement: All code, models, and data are publicly available at https://github.com/yhy-nj/RSCD-MISANet.git (accessed on 21 December 2025) for reproducibility and community-driven innovation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, M.; Jiao, L.; Liu, F.; Hou, B.; Yang, S.; Jian, M. DPFL-Nets: Deep pyramid feature learning networks for multiscale change detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 6402–6416. [CrossRef]
2. Chughtai, A.H.; Abbasi, H.; Karas, I.R. A review on change detection method and accuracy assessment for land use/land cover. *Remote Sens. Appl. Soc. Environ.* **2021**, *22*, 100482. [CrossRef]
3. Chen, J.; Xia, M.; Wang, D.; Lin, H. Double branch parallel network for segmentation of buildings and waters in remote sensing images. *Remote Sens.* **2023**, *15*, 1536. [CrossRef]
4. Qiao, H.; Wan, X.; Wan, Y.; Li, S.; Zhang, W. A novel change detection method for natural disaster detection and segmentation from video sequence. *Sensors* **2020**, *20*, 5076. [CrossRef] [PubMed]

5. Liu, Y.; Wang, J.; Song, Y.; Liang, S.; Xia, M.; Zhang, Q. Lightning Nowcasting Based on High-density Area and Extrapolation Utilizing Long-range Lightning Location Data. *Atmos. Res.* **2025**, *321*, 108070. [CrossRef]
6. Marin, C.; Bovolo, F.; Bruzzone, L. Building change detection in multitemporal very high resolution SAR images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2664–2682. [CrossRef]
7. Zhu, G.; Xu, X.; Ma, Z.; Xu, L. From land to sea: Land change analysis in the coastal zone of Bohai Bay based on RS and GIS. In Proceedings of the IEEE MTT-S International Microwave Symposium Digest, Baltimore, ML, USA, 5–10 June 2011; pp. 6243–6247. [CrossRef]
8. Singh, B.A. Change detection in the tropical forest environment of northeastern India using Landsat. *Remote Sens. Trop.* **1986**, *44*, 254–273.
9. Máckiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [CrossRef]
10. Bouhlel, N.; Rousseau, D. Multi-temporal SAR change detection using wavelet transforms. In Proceedings of the European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 28 August–2 September 2022; pp. 538–542. [CrossRef]
11. Schmitt, A.; Wessel, B.; Roth, A. Curvelet-based change detection on SAR images for natural disaster mapping. *Photogramm. Fernerkund. Geoinf.* **2010**, *2010*, 463–474. [CrossRef]
12. Habib, T.; Inglada, J.; Mercier, G.; Chanussot, J. Support vector reduction in SVM algorithm for abrupt change detection in remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 606–610. [CrossRef]
13. Sun, Y.; Lei, L.; Guan, D.; Kuang, G. Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images. *IEEE Trans. Image Process.* **2021**, *30*, 6277–6291. [CrossRef] [PubMed]
14. Dou, P.; Huang, C.; Han, W.; Hou, J.; Zhang, Y.; Gu, J. Remote sensing image classification using an ensemble framework without multiple classifiers. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 190–209. [CrossRef]
15. Jiang, S.; Dong, R.; Wang, J.; Xia, M. Credit card fraud detection based on unsupervised attentional anomaly detection network. *Systems* **2023**, *11*, 305. [CrossRef]
16. Yuan, S.; Mao, Y.; Tian, C.; Yu, F.; Guo, T.; Xia, M. GSTAformer: Graph-Guided Spatio-Temporal Autoformer for Mid-Term Wind Power Forecasting. *Energies* **2026**, *19*, 254. [CrossRef]
17. Chen, K.; Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H. MSFANet: Multi-scale strip feature attention network for cloud and cloud shadow segmentation. *Remote Sens.* **2023**, *15*, 4853. [CrossRef]
18. Wang, Z.; Xia, M.; Weng, L.; Hu, K.; Lin, H. Dual encoder–decoder network for land cover segmentation of remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2372–2385. [CrossRef]
19. Ding, L.; Xia, M.; Lin, H.; Hu, K. Multi-level attention interactive network for cloud and snow detection segmentation. *Remote Sens.* **2023**, *16*, 112. [CrossRef]
20. Dong, Y.; Liu, Q.; Du, B.; Zhang, L. Weighted feature fusion of CNN and graph attention network for hyperspectral image classification. *IEEE Trans. Image Process.* **2022**, *31*, 1559–1572. [CrossRef]
21. Feng, Y.; Zheng, J.; Qin, M.; Bai, C.; Zhang, J. 3D octave and 2D vanilla mixed CNN for hyperspectral image classification with limited samples. *Remote Sens.* **2021**, *13*, 4407. [CrossRef]
22. Chen, B.; Xia, M.; Qian, M.; Huang, J. MANet: A multi-level aggregation network for semantic segmentation of high-resolution remote sensing images. *Int. J. Remote Sens.* **2022**, *43*, 5874–5894. [CrossRef]
23. Cui, Y.; Yan, L.; Cao, Z.; Liu, D. TF-Blender: Temporal feature blender for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 8138–8147. [CrossRef]
24. Yang, K.; Xia, G.-S.; Liu, Z.; Du, B.; Yang, W.; Pelillo, M.; Zhang, L. Asymmetric Siamese networks for semantic change detection in aerial images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609818. [CrossRef]
25. Song, J.; Chen, H.; Yokoya, N. SyntheWorld: A large-scale synthetic dataset for land cover mapping and building change detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2024; pp. 8287–8296. [CrossRef]
26. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
27. Pan, F.; Wu, Z.; Jia, X.; Liu, Q.; Xu, Y.; Wei, Z. A temporal-reliable method for change detection in high-resolution bi-temporal remote sensing images. *Remote Sens.* **2022**, *14*, 3100. [CrossRef]
28. Peng, T.; Hu, L.; Huang, J.; Liu, J.; Zhu, P.; Hu, X.; He, R. A HRNet–Transformer network combining recurrent tokens for remote sensing image change detection. In *Advances in Computer Graphics—CGI 2023*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2024; Volume 14497, pp. 15–26. [CrossRef]
29. Ren, Z.; Weng, L.; Xia, M.; Lin, H. MCINet: Multi-attentive cross-level interaction network for cloud and snow segmentation. *J. Appl. Remote Sens.* **2026**, *20*, 021404. [CrossRef]
30. Dosovitskiy, A. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. [CrossRef]

31. Chen, H.; Nemni, E.; Vallecorsa, S.; Li, X.; Wu, C.; Bromley, L. Dual-task Siamese transformer framework for building damage assessment. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 1600–1603. [CrossRef]
32. Bandara, W.G.C.; Patel, V.M. A transformer-based Siamese network for change detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 207–210. [CrossRef]
33. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
34. Zhang, C.; Weng, L.; Ding, L.; Xia, M.; Lin, H. CRSNet: Cloud and cloud shadow refinement segmentation networks for remote sensing imagery. *Remote Sens.* **2023**, *15*, 1664. [CrossRef]
35. Yan, T.; Wan, Z.; Zhang, P. Fully Transformer Network for change detection of remote sensing images. In Proceedings of the Asian Conference on Computer Vision (ACCV), Macao, China, 4–8 December 2022. [CrossRef]
36. Feng, J.; Yang, X.; Gu, Z.; Zeng, M.; Zheng, W. SMBCNet: A transformer-based approach for change detection in remote sensing images through semantic segmentation. *Remote Sens.* **2023**, *15*, 3566. [CrossRef]
37. Zhang, H.; Chen, H.; Zhou, C.; Chen, K.; Liu, C.; Zou, Z.; Shi, Z. BIFA: Remote Sensing Image Change Detection with Bitemporal Feature Alignment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5614317. [CrossRef]
38. Guo, H.; Liu, C.; Zhang, H.; Chen, B.; Zou, Z.; Shi, Z. TaCo: Capturing Spatio-Temporal Semantic Consistency in Remote Sensing Change Detection. *arXiv* **2025**, arXiv:2511.20306. [CrossRef]
39. Daudt, R.C.; Saux, B.L.; Boulch, A. Fully convolutional Siamese networks for change detection. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067. [CrossRef]
40. Hou, H.; Wang, Y.; Qin, Q.; Tan, Y.; Liu, T. Multi-scale feature fusion based on difference enhancement for remote sensing image change detection. *Symmetry* **2025**, *17*, 590. [CrossRef]
41. Peng, X.; Zhong, R.; Li, Z; Li, Q. Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. *IEEE Trans. Geosci. Remote Sens.***2021**, *59*, 7296–7307. [CrossRef]
42. Gong, M.; Zhou, Z.; Ma, J. Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 375–386. [CrossRef]
43. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
44. Zhou, Y.; Wang, F.; Zhao, J.; Yao, R.; Chen, S.; Ma, H. Spatial–temporal-based multihead self-attention for remote sensing image change detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6615–6626. [CrossRef]
45. Wu, C.; Zhang, L.; Zhang, L. A scene change detection framework for multi-temporal very high-resolution remote sensing images. *Signal Process.* **2016**, *124*, 184–197. [CrossRef]
46. Wu, C.; Zhang, L.; Du, B. Kernel slow feature analysis for scene change detection. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2367–2384. [CrossRef]
47. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [CrossRef]
48. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection of high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**,*14*, 1194–1206. [CrossRef]
49. Ma, X.; Zhang, Y.; Wulamu, A.; Zhu, X. Sun-Angle Effects on Remote-Sensing Phenology Observed and Modelled Using Himawari-8. *Remote Sens.* **2020**, *12*, 1339. [CrossRef]
50. Shibata, T. Digital correction of solar illumination and viewing angle effects. In *Proceedings of the LARS Symposium*; Purdue Research Foundation: West Lafayette, IN, USA, 1981.
51. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
52. Chen, H.; Shi, Z. A spatial–temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
53. Shi, Q.; Liu, M.; Li, S.; Liu, X.; Wang, F.; Zhang, L. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5604816. [CrossRef]
54. Peng, D.; Bruzzone, L.; Zhang, Y.; Guan, H.; Ding, H.; Huang, X. SemiCDNet: A semisupervised convolutional neural network for change detection in high-resolution remote-sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5891–5906. [CrossRef]
55. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A deep learning architecture for visual change detection. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 1–16. [CrossRef]

56. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high-resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
57. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
58. Feng, Y.; Xu, H.; Jiang, J.; Liu, H.; Zheng, J. ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bi-temporal remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4410213. [CrossRef]
59. Feng, Y.; Jiang, J.; Xu, H.; Zheng, J. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4401015. [CrossRef]
60. Yin, H.; Weng, L.; Li, Y.; Xia, M.; Hu, K.; Lin, H.; Qian, M. Attention-guided Siamese networks for change detection in high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103206. [CrossRef]
61. Wang, Z.; Gu, G.; Xia, M.; Weng, L.; Hu, K. Bitemporal Attention Sharing Network for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 10368–10379. [CrossRef]
62. Li, Y.; Weng, L.; Xia, M.; Hu, K.; Lin, H. Multi-scale fusion Siamese network based on three-branch attention mechanism for high-resolution remote sensing image change detection. *Remote Sens.* **2024**, *16*, 1665. [CrossRef]
63. Hang, R.; Xu, S.; Yuan, P.; Liu, Q. AANet: An Ambiguity-Aware Network for Remote-Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5612911. [CrossRef]
64. Huang, J.; Yuan, X.; Lam, C.T.; Wang, Y.; Xia, M. LCCDMamba: Visual State Space Model for Land Cover Change Detection of VHR Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 5765–5779. [CrossRef]