remote sensing

MDPI

Article

MSDT-Net: A Multi-Scale Smoothing Attention and Differential Transformer Encoding Network for Building Change Detection in Coastal Areas

Weitong Ma [2], Lebao Yang [2], Yuxun Chen [2], Yangyu Zhao [2], Zheng Wei [1], Xue Ji [1,2,3,*] and Chengyao Zhang [2]

[1] Nansha Islands Coral Reef Ecosystem National Observation Research Station, Guangzhou 510300, China; weizheng@scs.mnr.gov.cn
[2] College of Geoexploration Science and Technology, Jilin University, Changchun 130026, China; mawt23@mails.jlu.edu.cn (W.M.); yanglb2322@mails.jllu.edu.cn (L.Y.); chenyx2322@mails.jlu.edu.cn (Y.C.); zhaoyy2322@mails.jllu.edu.cn (Y.Z.); chengyao23@mails.jlu.edu.cn (C.Z.)
[3] Key Laboratory of Ocean Geomatics, Ministry of Natural Resources, Qingdao 266590, China
* Correspondence: jixuesdqd@jlu.edu.cn

Highlights

What are the main findings?

- Introduced the Multi-Scale Smoothing Attention (MSA) and Differential Transformer Encoding Module (DTEM), which enhance the detection of boundary ambiguities and small-scale changes.
- Developed the Focal–Dice–IoU Unified Boundary Loss (FDUB-Loss) that effectively mitigates class imbalance and improves model performance.

What are the implications of the main findings?

- The experimental results indicate that integrating multi-scale attention with differential encoding significantly improves change detection accuracy in high-resolution remote sensing imagery. The proposed approach not only outperforms mainstream methods on the self-constructed island building dataset but also demonstrates practical value on other building datasets. This suggests that the proposed framework can serve as a new paradigm for high-precision change detection and provide technical support for intelligent remote sensing interpretation.
- In multiple island scenarios with highly imbalanced changed pixels the model still maintains high recall and precision, highlighting its robustness against challenges such as class imbalance and boundary ambiguity. This implies that the method has broad applicability for sparse target detection in island and coastal environments.

Abstract

Island building change detection is a critical technology for environmental monitoring, disaster early warning, and urban planning, playing a key role in dynamic resource management and sustainable development of islands. However, the imbalanced distribution of class pixels (changed vs. unchanged) undermines the detection capability of existing methods and severe boundary misdetection. To address issue, we propose the MSDT-Net model, which makes breakthroughs in architecture, modules, and loss functions; a dual-branch twin ConvNeXt architecture is adopted as the feature extraction backbone, and the designed Edge-Aware Smoothing Module (MSA) effectively enhances the continuity of the change region boundaries through a multi-scale feature fusion mechanism. The proposed Difference Feature Enhancement Module (DTEM) enables deep interaction and

check for updates

Academic Editors: Massimiliano Pepe and Wen Liu

Received: 19 September 2025
Revised: 28 October 2025
Accepted: 31 October 2025
Published: 5 November 2025

Citation: Ma, W.; Yang, L.; Chen, Y.; Zhao, Y.; Wei, Z.; Ji, X.; Zhang, C. MSDT-Net: A Multi-Scale Smoothing Attention and Differential Transformer Encoding Network for Building Change Detection in Coastal Areas. *Remote Sens.* **2025**, *17*, 3645. https://doi.org/10.3390/rs17213645

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

fusion between original semantic and change features, significantly improving the discriminative power of the features. Additionally, a Focal–Dice–IoU Boundary Joint Loss Function (FDUB-Loss) is constructed to suppress massive background interference using Focal Loss, enhance pixel-level segmentation accuracy with Dice Loss, and optimize object localization with IoU Loss. Experiments show that on a self-constructed island dataset, the model achieves an F1-score of 0.9248 and an IoU value of 0.8614. Compared to mainstream methods, MSDT-Net demonstrates significant improvements in key metrics across various aspects. Especially in scenarios with few changed pixels, the recall rate is 0.9178 and the precision is 0.9328, showing excellent detection performance and boundary integrity. The introduction of MSDT-Net provides a highly reliable technical pathway for island development monitoring.

Keywords: island building change detection; edge feature enhancement; class imbalance; multi-loss function optimization

1. Introduction

Island building change detection is a crucial technology for environmental monitoring [1], disaster response [2], and urban planning [3]. It plays a vital role in the scientific management and sustainable development of island resources [4]. With the advancement of remote sensing capabilities, high-resolution remote sensing images have become increasingly available, providing a solid data foundation for refined surface monitoring [5]. However, despite the strong potential of remote sensing images in detecting and characterizing buildings over large areas, accurately extracting and analyzing building change information from dual-temporal remote sensing images remains a highly challenging task [6].

Traditional change detection methods primarily rely on handcrafted feature engineering, where low-level visual features from bi-temporal remote sensing images are compared to identify changes. For instance, He et al. [7] proposed an improved superpixel clustering approach that integrates regional consistency and boundary information to extract large-scale building changes. While effective in low-density urban areas, this method often suffers from over-segmentation and error propagation in dense or high-rise regions due to the structural complexity and texture heterogeneity within superpixels, thereby reducing detection accuracy. Feature point-based matching strategies have also been widely applied in multi-temporal image registration and change extraction [8]. For example, Wang et al. [9] developed a matching algorithm constrained by neighborhood topology and affine geometry, which demonstrates strong performance in multi-source image registration but is prone to mismatches in repetitive or weak-texture regions, undermining the reliability of subsequent change analysis. To improve detection stability, some researchers have incorporated multi-temporal constraints. Zhang et al. [10] introduced a three-temporal logical constraint change vector analysis method, which enhances detection consistency and performs well in significant building demolition and construction scenarios. Nevertheless, this method remains insufficiently responsive to subtle or low-contrast changes and shows limited robustness against noise and pseudo-changes. Overall, while these traditional methods have expanded the technical pathways of change detection, their reliance on handcrafted features leads to weak generalization ability, complex parameter tuning, and low automation [11], making them inadequate for accurately extracting subtle building changes in high-resolution island imagery.

In recent years, deep learning has emerged as a dominant paradigm in change detection due to its powerful capabilities in image modeling, feature representation, and

pattern recognition [12,13]. Fully convolutional networks (FCNs) pioneered the end-to-end pipeline from raw images to change maps [14]. Subsequently, UNet introduced skip connections to achieve multi-scale feature fusion, demonstrating excellent performance in semantic segmentation and being widely adapted to building change detection (BCD) tasks [15]. However, due to the limited receptive fields and isotropic nature of conventional convolution kernels [16,17], FCN-based models still struggle with capturing complex building contours, precise boundary localization, and scale adaptability. To address these issues, various architectural improvements have been proposed. Zhao et al. [18] developed HRSCD-Net, which integrates a high-resolution semantic preservation module and a multi-scale context aggregation mechanism, significantly improving the detection of small buildings and achieving competitive results on datasets such as HRSCD and LEVIR-CD. Nonetheless, its convolution-dominated design remains less effective in representing irregular or complex building boundaries, particularly under challenging island conditions.

To further enhance feature discrimination, attention mechanisms have been incorporated into change detection networks [19]. For example, SNUNet integrates both channel and spatial attention modules into the UNet architecture [20], thereby improving the sensitivity and robustness of the model to change regions, and surpassing conventional methods across multiple benchmark datasets. More recently, Transformer architectures have attracted growing attention owing to their global modeling capability [21]. BIT-CD employs a dual-temporal Transformer framework to model temporal dependencies, achieving notable improvements in small-object change detection and boundary preservation, and is regarded as one of the current state-of-the-art models [22]. However, the quadratic growth of computational complexity with input resolution imposes substantial demands on computational resources and data scale, which significantly constrains the applicability of Transformer-based approaches in large-scale, high-resolution scenarios.

To address the widespread class imbalance and boundary blurring issues in island building change detection, this study proposes the MSDT-Net model based on multi-scale deep feature fusion. The model employs a twin network architecture for feature extraction, replacing traditional complex structures with ConvNeXt-Tiny with shared weights. Specifically, the model introduces the Transformer-based Difference Feature Enhancement Module (DTEM), which explicitly models the temporal and spatial difference features between dual-temporal images, effectively enhancing the small-target change detection capability and boundary localization accuracy. Based on this, the innovatively designed Edge-Aware Smoothing Module (MSA) integrates multi-scale convolutions and channel attention mechanisms to effectively enhance the representation of boundary features in change regions, overcoming the limitations of traditional methods in contour recognition.

The two modules work in synergy to address two key challenges in island scenes—namely, the difficulty of detecting small objects and the ambiguity of building boundaries—from the feature representation perspective. This collaborative design enables structured modeling and differential enhancement of complex island environments. Their joint integration not only enhances the model's sensitivity to subtle local variations but also strengthens the stability of cross-temporal feature alignment, thereby providing solid semantic support for subsequent change region segmentation.

In response to the class imbalance characteristic in island scenarios, the proposed FDUB-Loss joint loss function uses a dynamic weight adjustment mechanism to significantly improve the model's sensitivity to sparse change targets. Experimental results demonstrate that in island scenarios with a very low proportion of changed pixels, the model effectively addresses the issues of missed detections and false positives in traditional methods, significantly improving detection accuracy and boundary integrity.

The contributions of our work can be summarized as follows:

(1) A novel remote sensing change detection framework, termed MSDT-Net, is proposed, which creatively integrates the advantages of Siamese neural networks and the ConvNeXt architecture to achieve a balanced representation of local texture details and global semantic consistency. By effectively exploiting the spatiotemporal correlations of multi-temporal imagery and multi-scale contextual information, the proposed method significantly improves the continuity of change contours and successfully mitigates the problem of class imbalance.

(2) Two core functional modules are designed: the Multi-scale Smoothing Attention (MSA) module and the Differential Temporal Encoding Module (DTEM). These modules form a complementary structural relationship—MSA focuses on boundary smoothing and multi-scale perception, while DTEM explicitly models spatiotemporal discrepancies. Their cooperation enables highly sensitive detection of small-scale change boundaries in complex island environments.

(3) Furthermore, a Focal–Dice–IoU Boundary Unified Loss (FDUB-Loss) is introduced. This hybrid loss function adaptively optimizes the model with respect to both imbalanced sample distributions and boundary ambiguity, maintaining high accuracy and robustness even in low-change scenarios. It effectively alleviates the class imbalance issue and substantially enhances the model's sensitivity to subtle variations. The proposed loss function can be seamlessly integrated into existing deep learning models for change detection, markedly improving their performance in island building change detection tasks.

Overall, the innovations of this study lie not only in the novel architectural design but also in the systematic optimization strategies tailored for the specific challenge of island building change detection, achieving breakthroughs in both methodological and application aspects.

2. Related Work

Building change detection refers to the process of analyzing remote sensing imagery data from different temporal instances of the same geographic region to identify and extract building-related change information [23]. Traditional change detection methods often struggle with interference factors such as diverse building forms, varying lighting conditions, and seasonal vegetation differences. Recently, the rapid development of deep learning technologies has brought new breakthroughs to this field, particularly with the introduction of Siamese neural network architectures, which have significantly improved change detection performance. Yun et al. [24] were among the first to apply Siamese networks to building change detection in coastal environments. By using a shared-weight mechanism, they effectively alleviated the feature confusion between marine backgrounds and building targets. However, this method still exhibits notable scale sensitivity in detecting small buildings. To improve classification accuracy, Pan et al. [25] proposed a simplified object network that innovatively combines object-oriented segmentation with deep learning. By adopting a multi-level feature fusion strategy, they significantly improved classification accuracy for high-resolution images. However, the real-time performance of their change detection still requires improvement. Zheng et al. [26] developed HFA-Net, which introduced a high-frequency attention module to enhance the expression of structural edge features. It achieved significant results in building change detection on ultra-high-resolution imagery. However, misclassification still occurs in the presence of strong shadow interference. Notably, the application of Transformer architectures, such as the Siamese Transformer proposed by Rao et al. [27], achieved superior global semantic modeling through a hierarchical attention mechanism. Yet, its complex network structure and large parameter size limit its practical application.

Although these methods have made remarkable progress in multi-scale perception [28], boundary delineation [29], and context modeling [30], they still face two key technical bottlenecks: First, building edge features are easily ignored or blurred during detection, leading to insufficient boundary localization accuracy. This issue is especially prominent in detecting small buildings on islands. Second, the class imbalance problem, caused by the extremely low proportion of changed areas in the imagery, causes models to overly favor the "no change" class, severely restricting the detection performance of small change targets. To address the core challenge of class imbalance, recent research has proposed several innovative solutions. Alcantarilla et al. [31] introduced multi-scale feature fusion and dynamic weight adjustment mechanisms to significantly improve detection accuracy. However, the method still has limitations in adapting to heterogeneous data fusion. Wang et al. [32] designed a foreground-prior sampling strategy in their HANet to enhance the modeling of change areas, but this method is prone to sample bias in complex backgrounds. Mou et al. [33] developed a cascading attention network that strengthens long-term temporal dependency modeling, though its network complexity reduces computational efficiency. Kemker et al. [34] developed a progressive change detection framework that effectively reduces false alarm rates but exhibits noticeable latency in handling sudden changes. MASNet [35] employed a bidirectional attention mechanism to suppress false change interference, but it still faces limitations in boundary preservation in low-texture areas.

In addition, substantial progress has been made in recent years in attention-based feature fusion strategies. The SegFormer model [36] achieves efficient inter-layer feature aggregation through a lightweight multi-scale attention architecture, balancing global modeling capability and computational efficiency while maintaining strong semantic consistency. The FCIHMRT (Feature Cross-Layer Interaction Hybrid Method) [37] further introduces a cross-layer interaction mechanism that enhances semantic correlations among hierarchical features through a hybrid attention strategy, thereby improving feature fusion in complex scenes. These studies provide valuable insights for the design of the proposed Multi-scale Smoothing Attention (MSA) module, enabling it to better integrate multi-scale differential features while preserving boundary smoothness.

A comprehensive analysis of the existing studies reveals that current methods still have significant shortcomings in global and local feature collaborative modeling, distinguishing between true changes and interfering factors, and representing internal structures of change targets. In particular, in complex scenarios characterized by extreme sample imbalance, boundary blurring, and sparse changes, a unified robust detection framework has yet to be established.

3. Materials and Methods

To address the challenges of class imbalance and boundary blurring in island building change detection, this paper proposes a novel Edge-Aware Siamese Masking Network (MSDT-Net). Compared to traditional change detection networks, this model innovatively integrates two core modules: the Edge-Aware Smoothing Module (MSA), which enhances the feature representation of island building boundaries through a collaborative design of multi-scale convolutions and channel-space attention mechanisms, and the Difference Modeling Module (DTEM), which explicitly models the spatiotemporal difference features of dual-temporal imagery based on a Transformer architecture, significantly improving the model's ability to detect subtle changes in island buildings. Furthermore, the FDUB-Loss loss function, designed in this study, optimizes the model's sensitivity to sparse change regions in island scenes by integrating Focal, Dice, and IoU losses. The following sections will first provide an overview of the MSDT-Net architecture in Section 3.1, followed by detailed explanations of the MSA and DTEM modules in Sections 3.2 and 3.3, respectively.

3.1. MSDT-Net

The proposed MSDT-Net is a change detection model based on Siamese ConvNeXt, designed to model feature differences, enhance boundary awareness in complex backgrounds using the Edge-Aware Smoothing Module, and employ a progressive upsampling decoder to recover high-resolution masks for accurate pixel-level change detection.

The MSDT-Net model processes dual-temporal remote sensing images I_1 and I_2 of size $H_0 \times W_0$ as input, aiming to generate a binary change detection map $M \in \{0,1\}^{(H_0 \times W_0)}$ of the same size, where 1 represents the changed area (e.g., newly constructed/demolished buildings) and 0 represents the unchanged area. As shown in Figure 1, the model architecture consists of four key components: feature extraction network, difference modeling module, multi-scale attention module, and lightweight decoder. Based on the Siamese neural network architecture, the ConvNeXt-Tiny with shared weights is used as the backbone network to extract deep features from the dual-temporal remote sensing images, effectively representing the spatial semantic information of island buildings through a 768-dimensional feature map. To address the challenge of difference modeling in change detection, a Transformer-based difference encoding module is designed. The dual-temporal features and their absolute differences are concatenated and linearly projected into a 512-dimensional space, and a two-layer Transformer encoder is employed to capture long-range spatial dependencies, generating a difference feature map enriched with contextual information. To enhance boundary recognition, a Multi-Scale Attention Module (MSA) is introduced. This module extracts multi-scale local features through parallel 3×3, 5×5, and 7×3 convolutional kernels, and dynamically integrates information from different receptive fields via a channel attention mechanism, significantly improving the feature response strength at the boundaries. Finally, a lightweight transposed convolution decoder progressively upsamples to a resolution that is one-fourth of the original, producing the binary change detection map.

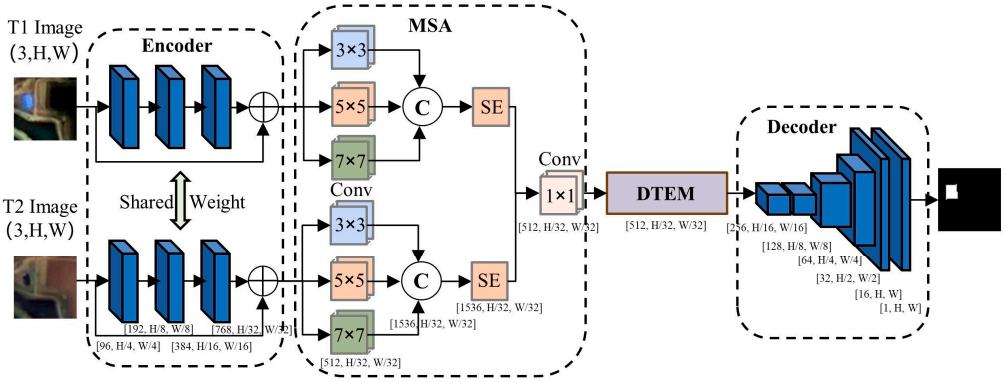


Figure 1. MSDT-Net Architecture Diagram.

3.2. Multi-Scale Smoothing Attention (MSA)

Boundary feature representation plays a key role in remote sensing image analysis, such as building extraction [38], land cover classification [39], and change detection [40]. Multi-scale boundary feature fusion has been proven to significantly improve boundary localization accuracy in previous studies [12,21]. Inspired by the multi-scale residual structure of MSRN [41], the proposed Edge-Aware Smoothing Module (MSA) integrates multi-scale convolutions and channel-space attention mechanisms to adaptively enhance the representation of building boundary features. Specifically, the parallel convolution kernels of 3×3, 5×5, and 7×7 cooperate to extract local detail features, while the attention mechanism dynamically weights important feature channels, jointly enhancing the model's ability to capture the complex boundaries of island buildings.

In addition, to further enhance the channel-wise attention modeling capability, the Squeeze-and-Excitation (SE) mechanism [42] is incorporated into the MSA module. The SE block adaptively learns the importance weights of different channels through the "Squeeze" and "Excitation" operations, thereby strengthening the response to critical boundary features and suppressing redundant information. This mechanism effectively improves the model's sensitivity to boundary regions and its feature representation capacity, providing more accurate boundary feature support for subsequent change detection.

The architecture of the MSA module, illustrated in Figure 2, systematically enhances feature representation by integrating a multi-scale convolution and attention fusion mechanism. Parallel convolutional kernels of sizes 3×3, 5×5, and 7×7 operate under different receptive fields to jointly capture both local details and global contextual information, thereby alleviating the limitation of single-scale convolutions in boundary characterization. Meanwhile, a channel attention module based on the SE structure adaptively reweights semantic channels, enabling the model to emphasize boundary-related features. In addition, spatial attention guides the network to focus on regions with significant boundary variations, effectively mitigating boundary blurring caused by background interference. This design jointly optimizes feature representation from the perspectives of scale, channel, and space, thereby addressing the insufficient extraction of building boundaries while improving the recognition accuracy of minority-class change regions.

$$X_k = Conv_{k \times k}(X) \tag{1}$$

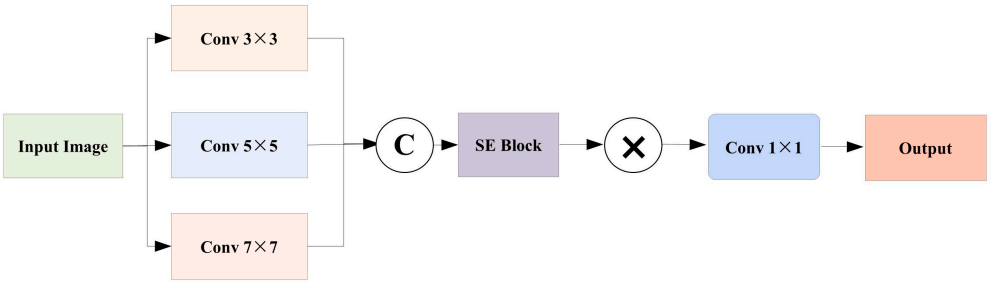


Figure 2. MSA Module Architecture Diagram.

Let X denote the feature map and k represent the size of the two-dimensional convolution kernel. X_k denotes the feature map extracted using a convolution kernel of size $k \times k$. Extracting multi-scale features is achieved by employing convolutional kernels of different sizes, where larger kernels capture broad contextual information suitable for large-scale buildings, while smaller kernels preserve fine-grained details, particularly beneficial for delineating the boundaries of small-scale buildings.

$$s = \sigma(W_2 \cdot ReLU(W_1 \cdot z + b_1) + b_2) \tag{2}$$

Let z denote the global descriptor vector and s represent the channel attention weight vector. The dimensionality of channels is first reduced through a fully connected layer W_1, which significantly decreases the number of model parameters while maintaining computational efficiency. Subsequently, a *ReLU* activation function is applied to introduce nonlinear transformation capability, thereby enhancing the representational power of the model. Finally, a fully connected layer W_2 is employed to restore the original channel dimensionality, and a Sigmoid activation function is used to generate normalized attention weights ranging from 0 to 1.

$$X_{att} = X_{cat} \odot s \tag{3}$$

The channel attention-weighted output feature map, denoted as $X_{(catt)}$, is obtained by applying the channel attention weight vector s to the input feature map $X_{(cat)}$. By performing element-wise multiplication between the channel attention weights and the multi-scale features, each feature map is scaled by its corresponding attention weight s, which strengthens the responses of important feature channels while suppressing those of less informative ones.

$$Y = Conv1 \times 1(Xatt) \tag{4}$$

A 1×1 convolution is applied to project the fused multi-scale features into the target output dimension, which not only accomplishes feature integration but also effectively controls the number of parameters and computational complexity. Finally, the output feature map Y is obtained.

3.3. Differential Transformer Encoding Module (DTEM)

In order to capture fine-grained change features between pre- and post-temporal remote sensing images more effectively, this paper designs a Transformer-based Difference Modeling Module. This module explicitly constructs the difference representation between images on the basis of the Siamese feature extraction structure and introduces a global modeling mechanism to enhance the model's ability to discriminate small-scale targets and boundary changes in complex scenes.

The core design of the DTEM lies in the integration of explicit feature differencing and global modeling to achieve precise characterization of bi-temporal image discrepancies. This design effectively highlights change regions while suppressing noise responses in non-change areas.

Specifically, as shown in Figure 3, the module first computes the pixel-wise absolute difference between the two input feature maps to explicitly characterize the change regions across the temporal pair. During the differential modeling stage, a pixel-wise absolute difference operation is employed, defined as $D = | F_1 - F_2 |$, to explicitly describe the spatial and semantic feature differences between two temporal images. Compared with simple concatenation or weighted summation, this operation provides a more intuitive representation of change regions and effectively reduces interference from unchanged areas. The original bi-temporal features and the difference map are then unfolded into sequences and concatenated into a unified representation. This fusion strategy enhances the model's sensitivity and discriminative ability toward change regions while preserving the original semantic representations. This representation not only preserves the original semantic information but also strengthens sensitivity to change regions.

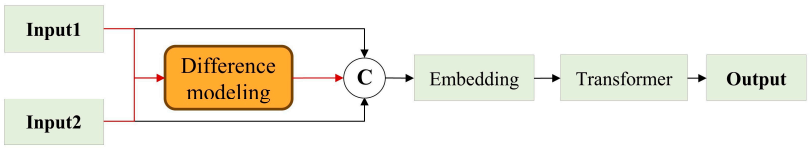


Figure 3. DTEM Architecture Diagram.

During multi-scale feature fusion, simple channel-wise concatenation may overlook semantic discrepancies across different scales, potentially introducing conflicts or redundancy. To address this issue, the DTEM leverages stacked Transformer encoders with self-attention mechanisms to perform global modeling on the fused features. Through global attention modeling implemented by a multi-layer Transformer encoder, DTEM can adaptively select complementary information during the feature fusion process and effectively suppress semantic conflicts and feature redundancy across scales. As a result, the model maintains strong discriminative capability for small objects and sparse change regions. This enables

adaptive selection of complementary information and suppression of conflicting representations, thereby ensuring that multi-scale features retain strong discriminability for small objects and sparse change regions.

To adapt the fused features to the Transformer input structure, they are projected into fixed-dimensional embeddings before being processed by stacked Transformer encoders. Through the self-attention mechanism, DTEM captures long-range spatial dependencies and cross-temporal global change relations, significantly enhancing the separability of small-scale objects and sparse changes. This dual modeling strategy, based on differential and global representations, enables DTEM to more effectively distinguish building changes from environmental noise in complex island scenes, thereby achieving more robust change detection performance. Finally, the encoded features are reconstructed into spatial structures to support subsequent change region segmentation tasks.

$$D = |F_1 - F_2| \tag{5}$$

The absolute difference between the bi-temporal feature maps is computed, where F_1 and F_2 are extracted from different time phases. Their difference D directly reflects pixel-level change information and serves as the fundamental input for subsequent change detection. The differential operation defined in this equation represents the key step of the entire module, explicitly mapping the changes between bi-temporal features into a difference tensor that provides essential input for subsequent global modeling via the Transformer.

$$F_1^{seq} = Reshape(F_1) \in R^{B \times L \times C} \tag{6}$$

The spatial feature maps are flattened into a sequence of length L, transforming the two-dimensional spatial structure into a one-dimensional sequence, which enables the Transformer to process image data while keeping the batch dimension B and channel dimension C unchanged. This serialization process ensures the spatial continuity of positional information within the Transformer input, facilitating the capture of long-range dependencies.

$$X = [F_1^{seq} \oplus F_2^{seq} \oplus D^{seq}] \tag{7}$$

Three serialized features are concatenated along the channel dimension, where the concatenation operation preserves both the original feature information (${F_1}^{seq}$, ${F_2}^{seq}$) and the explicit difference information D^{seq}, thereby providing the Transformer with enriched contextual representations. The tri-feature concatenation strategy enables the model to simultaneously incorporate three types of information—original semantics, temporal differences, and contextual awareness—thus enhancing the completeness and robustness of differential modeling.

3.4. Focal–Dice–IoU Boundary Joint Loss Function

To improve the model's detection ability for small-scale changes in island buildings, this paper designs the Focal–Dice–IoU boundary joint loss function, which effectively mitigates the class imbalance problem and improves the precision of detecting subtle changes. The function is formed by weighted fusion of Focal Loss (L_{focal}), Dice Loss (L_{dice}), and IoU Loss (L_{iou}):

$$L = \lambda_1 L_{focal} + \lambda_2 L_{\text{dice}} + \lambda_3 L_{\text{iou}} \tag{8}$$

$$L_{focal} = -\sum_{n=1}^{N} \left(\alpha q_n (1 - p_n)^{\gamma} log(p_n) + (1 - \alpha)(1 - q_n) p_n^{\gamma} log(1 - p_n)\right) \tag{9}$$

$$L_{dice} = 1 - \frac{2\sum_{n=1}^{N} p_n q_n + \in}{\sum_{n=1}^{N} p_n + \sum_{n=1}^{N} q_n + \in} \tag{10}$$

$$L_{iou} = 1 - \frac{\sum_{n=1}^{N} p_n q_n + \in}{\sum_{n=1}^{N} p_n + \sum_{n=1}^{N} q_n - \sum_{n=1}^{N} p_n q_n + \in} \tag{11}$$

where the weights are set as $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$. Lfocal reduces the impact of easily classified samples using a balancing factor $\alpha = 0.25$ and a modulation factor $\gamma = 2$, focusing on optimizing the recognition of hard-to-classify samples. Ldice enhances the model's sensitivity to small targets and boundary features based on the overlap metric, which measures regional overlap. Liou, through the intersection-over-union (IoU) constraint, enhances the completeness of the predicted region. In the specific implementation, let N be the total number of pixels in the image, $q_n \in \{0,1\}$ represent the true label, $p_n \in [0,1]$ be the predicted probability, and introduce a smoothing term of $\epsilon = 1 \times 10^{-6}$ to ensure the stability of numerical calculation. This multi-objective optimization strategy improves the model's performance in detecting fine-grained changes in island buildings while maintaining computational robustness. It also shows significant advantages in addressing key issues like class imbalance and boundary blurring.

4. Experiments and Analysis

4.1. Island Change Detection Dataset

Figure 4 presents the overall framework and spatial distribution of our dataset. As illustrated, this study employs high-resolution multispectral imagery acquired from the GaoFen-1 (GF-1) satellite, which provides a spatial resolution of 2 m. Dual-temporal data collected in 2020 and 2024 were utilized, covering representative island and coastal regions including Chongming Island, Longxue Island, Lushun (Liaoning Province), and Liaocheng (Shandong Province). The original dataset consisted of 1420 image patches with a spatial size of 256 × 256 pixels. Following rigorous quality control and preprocessing—comprising valid-area cropping, cloud removal, and the exclusion of irrelevant regions—a total of 2728 high-quality change detection samples of 64 × 64 pixels were generated. These samples encompass diverse change types such as building additions, demolitions, and other structural modifications. To ensure experimental reliability and data representativeness, a stratified sampling strategy was applied to partition the dataset into training, testing, and validation subsets at a ratio of 7:2:1, resulting in 1909, 546, and 273 samples, respectively. The proportion of change types was maintained consistently across subsets, ensuring balanced sample distribution and a robust foundation for model training and evaluation. The data preprocessing pipeline was carefully designed to guarantee both representativeness and class balance, thereby meeting the quality requirements for deep learning-based change detection tasks.

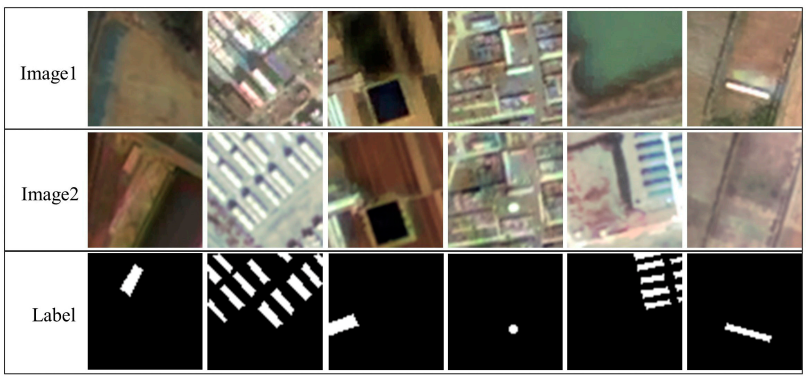


Figure 4. Dataset Overview Diagram.

4.2. Implementation Details and Evaluation Metrics

The proposed model was implemented in the PyTorch3.7 framework and trained on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. A batch size of 8 was adopted to optimize computational efficiency during the comparative experiments. Model optimization was conducted using the Adam optimizer with an initial learning rate of 0.001, a setting empirically verified to balance convergence speed and training stability. Input samples were uniformly resized to 64 × 64 pixels, and tensor normalization was applied to improve numerical stability. The training process was executed for 400 epochs, and the model's probability outputs were binarized using a fixed threshold of 0.5 to obtain the final change maps. To enhance data diversity and generalization capability, several data augmentation strategies were employed during training, including random horizontal and vertical flips, ±45° rotations, and brightness/contrast adjustments. All experiments were performed under identical hardware and software configurations to ensure result comparability. Model performance was quantitatively evaluated using Precision, Recall, F1-score, and Intersection over Union (IoU), which are standard evaluation metrics in remote sensing change detection. The metrics are defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

TP (True Positive) represents the number of pixels correctly predicted as change; *FP* (False Positive) denotes pixels incorrectly predicted as change; and *FN* (False Negative) indicates actual change pixels missed by the model. Precision measures the exclusion of irrelevant or unchanged regions from predictions, Recall reflects the detection rate of true changes, F1-score provides an overall harmonic assessment of Precision and Recall, and IoU quantifies the overlap between the predicted and reference change regions.

4.3. Ablation Experiments

To validate the effectiveness and collaborative benefits of the proposed FDUB-Loss and the newly integrated MSA module, two groups of ablation experiments were conducted. These experiments quantitatively assess their respective and joint contributions to mitigating class imbalance and improving detection accuracy in building change detection.

4.3.1. Determination of Loss Function Weights

To systematically investigate the influence of different weight configurations in the composite loss function, a series of ablation studies were performed on the weights of Focal Loss (λ_1), Dice Loss (λ_2), and IoU Loss (λ_3). Six representative configurations were examined: λ_1:λ_2:λ_3 = 0.5:1.0:1.0, 1.0:1.0:1.0, 1.0:0.5:1.0, 0.5:0.5:1.0, 0.25:1.0:1.0, and 0.5:1.0:0.5. The model performance was comprehensively evaluated using four metrics—F1 Score, IoU, Recall, and Precision—to jointly characterize the accuracy, completeness, and boundary consistency of the change detection results.

As summarized in Table 1, the model achieves the best overall performance when the loss weights are set to λ_1:λ_2:λ_3 = 0.5:1.0:1.0, with an F1 score of 0.9248, IoU of 0.8614, Recall of 0.9178, and Precision of 0.9319. Compared with other configurations, moderately reducing the weight of Focal Loss (λ_1 = 0.5) leads to a substantial improvement in overall

performance, while further increasing or decreasing λ_1 degrades the results. This indicates that this weight ratio achieves an effective balance between learning from hard samples and optimizing the global objective.

Table 1. Ablation Study on Loss Weight Configurations for Island Building Change Detection.

λ_1: Focal, λ_2: Dice, λ_3: IoU	Recall	F1	IoU	Precision
0.5:1.0:1.0	**0.9250**	**0.9273**	**0.8702**	**0.9296**
1.0:1.0:1.0	0.7826	0.7899	0.6598	0.7974
1.0:0.5:1.0	0.8504	0.8620	0.7621	0.8739
0.5:0.5:1.0	0.8877	0.8973	0.8168	0.9071
0.25:1.0:1.0	0.7388	0.7455	0.6039	0.7523
0.5:1.0:0.5	0.4027	0.2454	0.1467	0.1737

The analysis further reveals that Dice Loss and IoU Loss contribute comparably to the model's performance. Reducing λ_2 to 0.5 results in a modest decrease of approximately 1.5% in both F1 and IoU, and similar performance trends are observed when λ_2 varies while λ_3 remains constant. In contrast, halving λ_3 yields a significant drop across all metrics, with IoU decreasing by more than 4%, underscoring the essential role of IoU Loss in refining boundary localization and ensuring model stability. Overall, the coordinated configuration of the three loss components is critical, where the IoU Loss weight is particularly influential, while a moderate reduction in Focal Loss weight enhances model generalization and prevents overfitting.

Based on these findings, the configuration λ_1:λ_2:λ_3 = 0.5:1.0:1.0 is identified as the optimal setting for the composite loss function. This configuration consistently delivers superior performance across multiple quantitative metrics, effectively balancing accuracy, completeness, and boundary precision. The weight assignment also reflects the relative importance of each component: Dice and IoU Losses, both weighted equally, contribute to stable optimization and improved spatial alignment, while a reduced Focal Loss weight (0.5) alleviates the overemphasis on hard samples and improves overall generalization. This setting thus provides a well-balanced and robust loss design for building change detection tasks.

4.3.2. Ablation Experiment with Different Loss Function Combinations

To further assess the contribution of individual and composite loss components, we evaluated several loss function combinations within the unified MSDT-Net framework for the building change detection task. Specifically, we investigated three configurations: Cross-Entropy Loss (*CE*) combined with Dice Loss (*DL*), Boundary-Aware Loss (*BAL*), and their full combination (*CE* + *DL* + *BAL*). All experiments were conducted under identical training strategies and hyperparameter settings to ensure fairness and comparability. Evaluation was based on key indicators such as Precision, Recall, and F1 Score, providing empirical evidence for the optimal selection of loss components that improve the accuracy and boundary delineation of detected changes.

The experimental results demonstrate that combining multiple loss terms yields significant synergistic effects, as shown in the experiments presented in Table 2. The combination of Focal Loss and IoU Loss achieves a preliminary balance between Precision and Recall, resulting in an F1 score of 87.63% and an IoU of 78.22%. The Dice + IoU configuration exhibits stronger stability and consistency, reaching an F1 score of 89.95%. Notably, integrating all three losses yields a progressive improvement across all metrics. The Focal + Dice combination excels in Recall (88.12%, F1 = 88.89%), whereas Focal + IoU attains superior Precision (88.29%). Ultimately, the proposed FDUB-Loss—which fuses Focal, Dice, and IoU losses—achieves optimal results across all key metrics: Recall of 91.78%, F1 Score

of 92.48%, IoU of 86.14%, and Precision of 93.28%. These results confirm the complementary advantages of the three components: Focal Loss effectively mitigates class imbalance, Dice Loss enhances region-level consistency, and IoU Loss refines boundary localization. The collaborative optimization mechanism of FDUB-Loss enables robust detection of multi-scale changes and complex boundary structures, while suppressing false changes. By integrating sensitivity to sparse change regions (via Focal Loss), structural awareness of entire change areas (via Dice Loss), and boundary precision (via IoU Loss), FDUB-Loss provides a generalizable and effective loss paradigm for high-precision building change detection in remote sensing applications.

Table 2. Performance Comparison of Different Loss Function Combinations.

Loss Function Combination	Recall	F1	IoU	Precision
Focal + Dice +IoU	**0.9250**	**0.9273**	**0.8702**	**0.9296**
Dice + IoU	0.8926	0.8995	0.8194	0.9068
Focal + Dice	0.8812	0.8889	0.8019	0.8956
Focal + IoU	0.8686	0.8763	0.7822	0.8829
Dice	0.8468	0.8521	0.7455	0.8582
IoU	0.8395	0.8624	0.7608	0.8806
Focal	0.8171	0.8400	0.7272	0.8677

4.3.3. Joint Ablation Experiment of Modules and Loss Functions

To further verify the synergistic optimization effect between the MSA module and the FDUB-Loss, a second group of ablation experiments was designed. This experiment quantitatively assessed the complementary performance improvement achieved by their individual and joint contributions, evaluating the results after sequentially integrating the MSA module into the optimal loss configuration. The specific results are shown in Table 3.

Table 3. Ablation Experiment Results of Module and Loss Function Combinations.

Model Architecture	Recall	F1	IoU	Precision
Base	0.8802	0.8928	0.8106	0.8680
Base + MSA	0.9051	0.9121	0.8402	0.8982
Base + DTEM	0.9155	0.922	0.8567	0.9091
Ours	**0.9250**	**0.9273**	**0.8702**	**0.9296**

A comprehensive series of ablation experiments was conducted to rigorously examine the effectiveness of each module within the proposed architecture. The quantitative results, illustrated in Figure 5, demonstrate that the complete MSDT-Net consistently outperforms all comparative variants across four key performance indicators. Specifically, the full model achieves an F1 score of 0.9318 and an IoU of 0.8734, representing gains of 4.9 and 6.28 percentage points, respectively, relative to the baseline model—highlighting its substantial performance improvement.

Further analysis reveals distinct contributions from individual modules. The Edge-Aware Module notably enhances Recall, improving it by 2.49 percentage points, while the Difference Modeling Module contributes a 4.11 percentage point increase in Precision. These complementary effects enhance the model's overall capability to capture genuine change regions while effectively suppressing false positives, thereby improving detection accuracy.

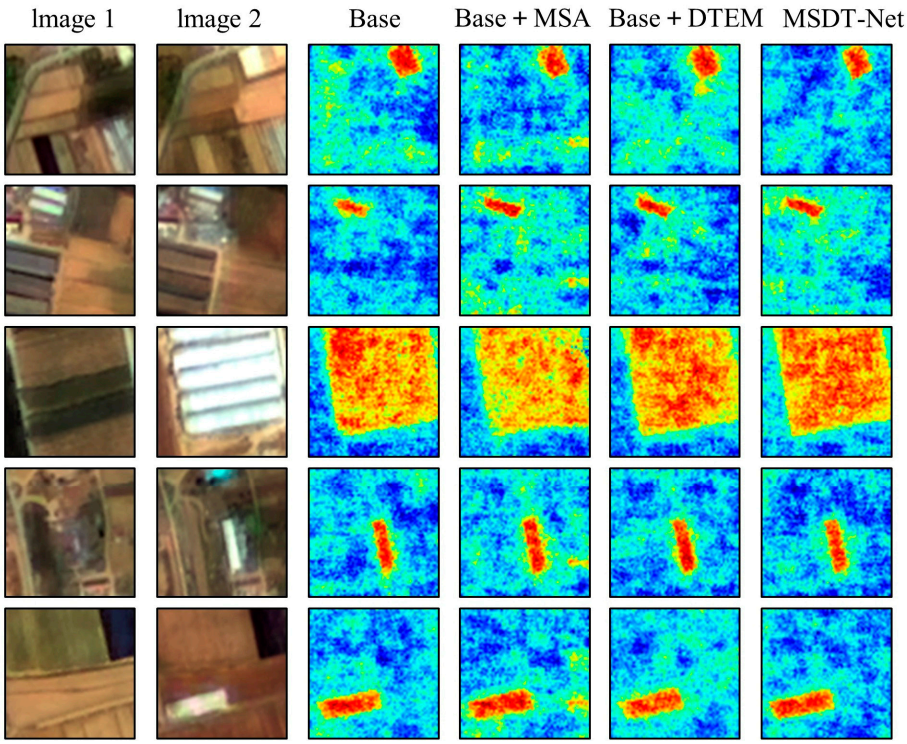


Figure 5. Island Building Change Detection Ablation Experiment Heatmap.

Heatmap visualizations further elucidate the functional advantages of each component in feature enhancement and spatial localization. The baseline model (Base) exhibits limited responsiveness to change regions, with dispersed activation areas, blurred boundaries, and pronounced background noise, resulting in frequent false and missed detections. Incorporating the Edge-Aware Module (Base + MSA) significantly enhances responses to true change areas, producing more concentrated heatmaps, although some residual false activations persist. With the subsequent integration of the Difference Modeling Module (Base + DTEM), the model demonstrates improved discrimination capability, yielding sharper object boundaries and stronger suppression of non-change regions. The final complete model, MSDT-Net, delivers the most precise and consistent localization, with response regions closely aligned with actual changes. It maintains robust discrimination even under strong background interference, substantially reducing false change artifacts. These results collectively verify the robustness and reliability of the proposed framework in multi-temporal change detection, offering a solid technical foundation for accurate change information extraction under complex environmental conditions.

Both the edge-aware Multi-Scale Attention (MSA) module and the Differentiated Transformer Encoding Module (DTEM) contribute essential and complementary performance gains. Theoretical analysis and experimental results consistently confirm their indispensable roles in island building change detection. The MSA module enhances boundary feature representation through multi-scale convolution and integrated channel spatial attention, effectively improving Recall by capturing subtle and irregular boundary variations. In contrast, the DTEM explicitly constructs temporal difference representations and employs global Transformer-based encoding, substantially improving Precision by suppressing background interference and alleviating class imbalance.

Ablation results demonstrate that the removal of either module leads to marked performance degradation. MSA primarily drives improvements in Recall, while DTEM governs Precision enhancement. When both are incorporated, the model attains the highest F1 score and IoU, indicating strong complementarity. Corresponding heatmap analyses

confirm that omitting either component results in missed small-scale changes or scattered false detections. Therefore, MSA and DTEM jointly ensure the robustness and accuracy of the proposed architecture in complex island environments and serve as the core modules underpinning the superior performance of MSDT-Net.

4.4. Comparative Experiments

A dedicated island building change detection dataset was constructed using high-resolution Gaofen-1 (GF-1) satellite imagery to systematically evaluate the overall performance of the proposed MSDT-Net. To further assess its generalization capability and robustness under diverse environmental conditions, additional cross-domain experiments were conducted on two widely used public benchmark datasets, namely the LEVIR-CD and WHU Building datasets.

4.4.1. Comparative Analysis of Change Detection Results on the Island Building Dataset

In this section, seven representative change detection methods were selected for comparison, including five convolutional neural network (CNN)-based models—PC-EC-diff [43], PC-EC-conc [44], FresNet [45], UNet [11], and UNet++ [46]—and two Transformer-based models, ChangeFormer [47] and BIT-CD [16], as well as an attention-driven model, UA-BCD [48]. The models were comprehensively evaluated using multiple quantitative metrics, including F1-score, Recall, IoU, and Boundary Preservation Index, complemented by qualitative visualizations to assess both detection accuracy and boundary integrity.

PC-EC-diff and PC-EC-conc are dual-branch networks employing feature differencing and concatenation, respectively, within an encoder–decoder framework to extract and fuse temporal features. FresNet enhances multi-scale sensitivity by introducing hierarchical feature refinement modules to capture change regions of varying sizes. UNet and UNet++ adopt classical U-shaped architectures for pixel-level change segmentation. ChangeFormer leverages Vision Transformer (ViT) backbones and self-attention mechanisms to capture long-range contextual dependencies, while BIT-CD integrates CNNs and Transformers via dual-temporal feature interaction to improve structural consistency. Collectively, these approaches represent the major paradigms in modern change detection research, encompassing convolutional, multi-scale, U-shaped, and Transformer-based strategies. For fair comparison, all models were trained and evaluated on identical data partitions, using 273 samples from the validation subset. Quantitative evaluations focused on F1-score, Recall, and Boundary Preservation, providing a multi-dimensional assessment of the proposed method's detection precision and completeness in island building change monitoring.

The comparative results summarized in Table 4 clearly demonstrate that MSDT-Net achieves the best overall performance across all evaluation metrics. Specifically, MSDT-Net attains an F1-score of 0.9273 and an IoU of 0.8702, indicating superior accuracy and spatial consistency. Although FresNet achieves competitive Recall (0.8965) and F1-score (0.8765), its Precision (0.8709) remains inferior to that of MSDT-Net, reflecting limited control over false positives. Similarly, PC-EC-diff and PC-EC-conc yield moderate results with F1-scores of 0.8690 and 0.8535, respectively, suggesting insufficient representation of complete change regions. Classical CNN-based architectures such as UNet and UNet++ perform acceptably in terms of Precision (0.8339 and 0.8728, respectively), but their markedly low Recalls (0.7538 and 0.7568) highlight weaknesses in detecting subtle or small-scale changes. Transformer-based models (BIT-CD, ChangeFormer) and attention-driven UA-BCD show suboptimal results, with F1-scores below 0.74 and relatively poor Recall values, indicating difficulty in maintaining boundary consistency and detecting sparse or small-scale changes. In contrast, MSDT-Net achieves both high Recall (0.9250) and Precision (0.9296), reflecting a balanced trade-off between sensitivity and specificity, thereby minimizing both

missed detections and false alarms. These results affirm the robustness and generalization capability of the proposed multi-scale differential Transformer framework in complex island environments.

Table 4. Performance comparison of different methods on the island building dataset.

Method	Recall	F1	IoU	Precision
PC-EC-diff	0.8697	0.8690	0.7755	0.8734
PC-EC-conc	0.8794	0.8535	0.7546	0.8794
FresNet	0.8965	0.8765	0.7862	0.8709
Unet	0.7538	0.6822	0.7918	0.8339
Unet++	0.7568	0.7124	0.8107	0.8728
ChangeFormer	0.8536	0.7307	0.8444	0.8355
BIT_CD	0.8619	0.7128	0.8323	0.8047
UA-BCD	0.7056	0.7865	0.6482	0.8884
MSDT-Net	**0.9250**	**0.9273**	**0.8702**	**0.9296**

To further evaluate boundary delineation performance, representative visual comparisons were conducted using test samples from diverse island scenarios, as shown in Figure 6. The selected samples cover a wide range of geographical and structural conditions—including dense urban areas, scattered rural settlements, and large-scale port regions—characterized by substantial background complexity. The visual comparisons demonstrate that MSDT-Net produces precise and coherent change maps with clearly defined boundaries, successfully suppressing background-induced false alarms and minimizing omission errors.

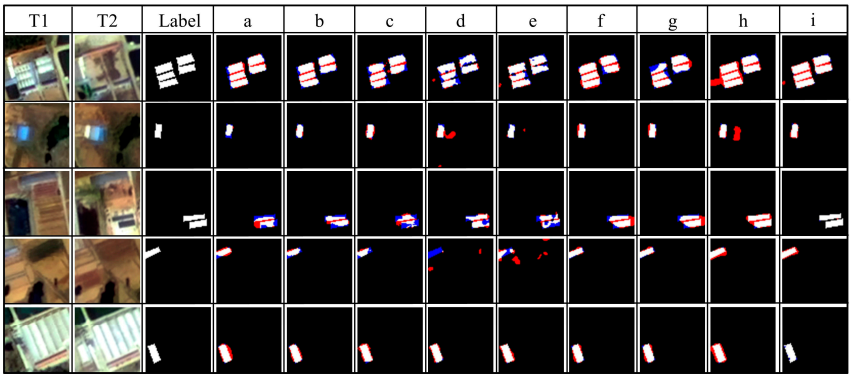


Figure 6. Experimental Comparison Diagram of island building change detection dataset. The rendered colors represent: white for TP, red for FP, and blue for FN. (a–i correspond to PC-EC-diff, PC-EC-conc, FresNet, Unet, Unet++, ChangeFormer, BIT_CD, UA-BCD, MSDT-Net).

The visualization results indicate that, compared with other methods, MSDT-Net demonstrates clear advantages in both change region identification and boundary localization. The proposed method accurately captures true change areas while effectively suppressing false alarms caused by background noise. In contrast, PC-EC-diff, PC-EC-conc, and FresNet exhibit noticeable deficiencies in maintaining the integrity of change region detection. UNet and UNet++ generate a large number of false positives and false negatives, making it difficult to reliably identify actual building changes in island environments. BIT-CD and ChangeFormer tend to miss or misclassify changes along object boundaries, thereby limiting their overall performance. The UA-BCD model demonstrates moderate performance in change detection. Although it successfully captures a considerable number of true positives, it still produces noticeable false positives and false negatives. Compared

with other methods, the detected change boundaries are relatively well defined; however, they still lack the sharpness and accuracy achieved by MSDT-Net.

In terms of boundary clarity, MSDT-Net demonstrates exceptional contour preservation and accurate reconstruction of building outlines, even in cluttered and heterogeneous regions. While UNet++ performs relatively better than other CNN-based models, it still struggles to delineate fine boundaries of small-scale structures. Overall, these results underscore the superior boundary-preserving capability and high detection reliability of MSDT-Net, which effectively reduces both false positives (FP) and false negatives (FN), thereby achieving significant improvements in detection accuracy, completeness, and robustness for multi-temporal building change detection in complex island environments.

4.4.2. Comparative Analysis of Change Detection Results on the LEVIR-CD Dataset

The LEVIR-CD dataset is a representative benchmark for urban building change detection, consisting of 637 pairs of high-resolution (0.5 m) remote sensing images with a size of 1024 × 1024 pixels. It covers various urban areas with building demolition and construction scenes. Owing to its precise annotations and diverse scenarios, this dataset has been widely adopted for evaluating the performance of building change detection models. Since it contains numerous samples with slight geometric distortions, it imposes higher demands on the geometric robustness and feature stability of models, making it particularly suitable for assessing detection capability under complex urban environments.

As shown in Table 5, MSDT-Net achieves the best performance across all evaluation metrics, with an F1-score of 0.9131, an IoU of 0.8401, and a Precision of 0.9380. These results indicate that the proposed method maintains strong feature representation and change recognition capabilities even in the presence of geometric distortions. PC-EC-conc and Unet++ obtain relatively high Recall values (0.8747 and 0.8979, respectively), demonstrating strong sensitivity to change regions. However, their IoU and F1-scores remain lower than those of MSDT-Net, suggesting persistent misdetections or localization deviations under geometric deformation. The traditional Unet exhibits balanced but less competitive performance, reflecting its limited ability to handle geometric inconsistencies. Both ChangeFormer and BIT-CD show comparatively weak results, particularly with poor Recall and IoU scores (e.g., the IoU of BIT-CD is only 0.6902), highlighting their sensitivity to geometric distortions and difficulty in aligning deformed objects within the feature space. Although UA-BCD attains a relatively high Recall of 0.8889—indicating good responsiveness to change regions—its IoU and F1-scores remain inferior to those of MSDT-Net. This suggests that UA-BCD may produce false detections or exhibit lower localization precision under geometric distortions, and that it is less effective than MSDT-Net in balancing true change detection with false alarm suppression. Visualization results are shown in Figure 7, further corroborating these findings.

Table 5. Performance comparison of different methods on LEVIR-CD.

Methods	Recall	Iou	F1	Precision
PC-EC-diff	0.8651	0.7870	0.8806	0.8983
PC-EC-conc	0.8747	0.8164	0.8980	0.9260
FresUNet	0.8750	0.7954	0.8860	0.8984
Unet	0.8509	0.7797	0.8758	0.9040
Unet++	0.8979	0.8112	0.8956	0.8953
ChangeFormer	0.7647	0.6396	0.7802	0.7963
BIT-CD	0.7430	0.6902	0.8167	0.9067
UA-BCD	0.8889	0.7885	0.7865	0.8748
MSDT-Net	**0.8895**	**0.8401**	**0.9131**	**0.9380**

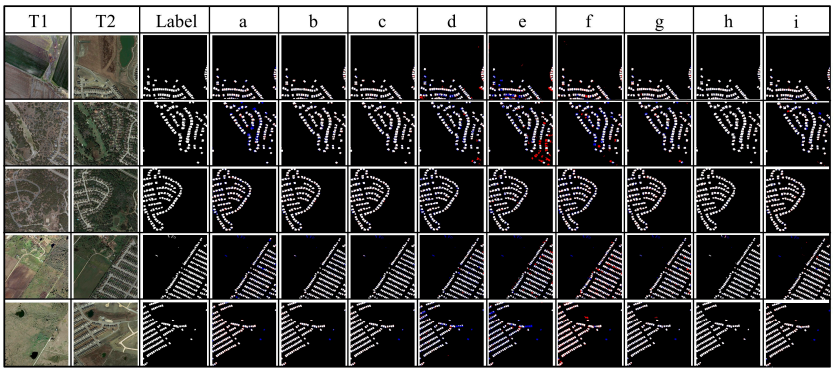


Figure 7. Experimental Comparison Diagram of LEVIR-CD. The rendered colors represent: white for TP, red for FP, and blue for FN. (a–i correspond to PC-EC-diff, PC-EC-conc, FresNet, Unet, Unet++, ChangeFormer, BIT_CD, UA-BCD, MSDT-Net).

4.4.3. Comparative Analysis of Change Detection Results on the WHU-CD Dataset

The WHU-CD dataset originates from high-resolution aerial imagery publicly released by Wuhan University, with a spatial resolution of approximately 0.3 m. It covers multiple representative urban areas and is primarily used for building extraction and change detection research. Although the bi-temporal images are well-registered, the dataset features diverse building types and complex backgrounds, including a large number of spectrally and texturally similar objects (e.g., rooftops and roads). These characteristics pose significant challenges in distinguishing subtle changes from pseudo-changes, making the dataset ideal for evaluating model discrimination capability in high-resolution and complex urban scenes. For example, the spectral and textural similarity between building rooftops and asphalt roads often leads to feature confusion and false alarms, thereby imposing stringent requirements on the discriminative ability of change detection models. Consequently, extracting distinctive building features is crucial for accurately identifying genuine changes.

As presented in Table 6, the PC-EC series achieves excellent Recall performance (e.g., 0.9798 for PC-EC-diff) but at the expense of relatively low Precision. This reflects their strong sensitivity to change regions while also suggesting a tendency toward false detections. Traditional models such as Unet and Unet++ attain higher Precision but lower Recall, indicating conservative predictions that tend to miss actual changes. BIT-CD delivers the lowest overall performance among all compared methods (with an IoU of only 0.6004), underscoring its limitations in capturing subtle feature variations in high-resolution scenes. UA-BCD demonstrates competitive overall performance as its high Recall indicates strong capability in detecting change regions, while its high Precision suggests effective suppression of false alarms. Nevertheless, despite its promising results, MSDT-Net still outperforms UA-BCD in several key metrics, highlighting its superior balance between sensitivity and accuracy in complex high-resolution environments. Visualization results are shown in Figure 8, providing further intuitive evidence for these comparative analyses.

Table 6. Performance comparison of different methods on WHU-CD.

Methods	Recall	Iou	F1	Precision
PC-EC-diff	**0.9798**	0.8240	0.9053	0.8313
PC-EC-conc	0.9734	0.8462	0.9126	0.8721
FresUNet	0.9067	0.8118	0.8799	0.8766
Unet	0.8550	0.8163	0.8856	0.9568
Unet++	0.8616	0.8067	0.8949	0.9376
ChangeFormer	0.8388	0.7482	0.8560	0.8739
BIT-CD	0.8732	0.6004	0.7503	0.6577
UA-BCD	0.9461	**0.9120**	**0.9540**	0.9620
MSDT-Net	0.9098	0.8807	0.9355	**0.9634**

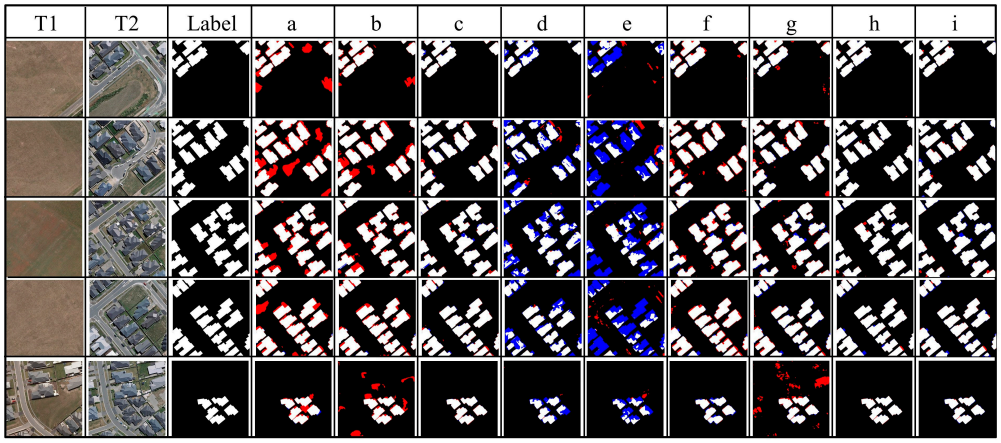


Figure 8. Experimental Comparison Diagram of WHU-CD. The rendered colors represent: white for TP, red for FP, and blue for FN. (a–i correspond to PC-EC-diff, PC-EC-conc, FresNet, Unet, Unet++, ChangeFormer, BIT_CD, UA-BCD, MSDT-Net).

4.5. Comparison Experiments in Different Scenarios

To comprehensively assess the scene adaptability of MSDT-Net in island building change detection and its effectiveness in addressing class imbalance, five representative island scenarios were selected for comparative experiments. As shown in Figure 9, these scenarios encompass diverse spatial patterns and change characteristics, including: (1) dispersed building distribution scenes (252 samples), where sparse target distribution and strong background interference increase detection difficulty; (2) dense settlement scenes (235 samples), characterized by compact building layouts and well-defined boundaries; (3) large-scale port and industrial facility change scenes (216 samples), featuring extensive and structurally complex change regions; (4) suburban expansion and land transformation scenes (245 samples), exhibiting pronounced spatial heterogeneity and dynamic evolution processes; (5) seasonal cycle-driven landscape change scenes (255 samples), where temporal periodicity leads to non-structural and naturally induced changes.

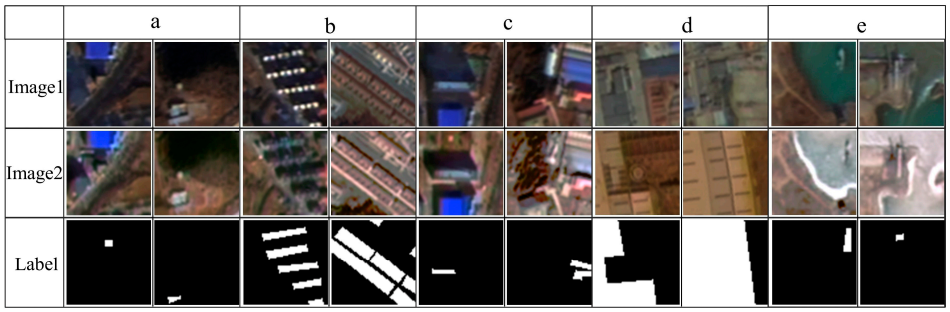


Figure 9. Image examples of different island building scenes. (a–e correspond to Dispersed Building Distribution Scene, Dense Settlement Scene, Large Port Industrial Facility Change Scene, Suburban Expansion and Land Transformation Scene, Seasonal Cycle-Driven Landscape Change Scene).

These scenarios collectively represent the diversity and complexity of island building change patterns while also highlighting different manifestations of class imbalance in practical detection tasks. The model's performance across these five scenarios is summarized in Table 7, providing an empirical foundation for subsequent performance analysis and mechanism interpretation.

Table 7. Comparison of Model Change Detection Performance Metrics Across Different Scenarios.

Scene Types	Recall	F1	IoU	Precision
Discrete Building Distribution Scene	0.8532	0.8894	0.8075	0.9347
Dense Settlement Scene	0.9178	0.9336	0.8825	0.9542
Large Port Area Industrial Facility Change Scene	**0.9312**	**0.9473**	**0.9023**	**0.9634**
Suburban Expansion and Land Transformation Scene	0.8973	0.9192	0.8727	0.9615
Seasonal Cycle Driven Landscape Change Scene	0.9021	0.9296	0.8749	0.9568

MSDT-Net exhibited remarkable performance across all five scenarios, achieving an overall recall rate exceeding 85%, precision above 93%, and an average Intersection over Union (IoU) greater than 86%, demonstrating strong and consistent detection capability. Despite variations in scene characteristics, the model effectively adapted to diverse spatial contexts. In the dispersed building distribution scenario, MSDT-Net achieved a precision of 93.47%, indicating robust suppression of false positives under low target density and high background noise. In dense settlement and port industrial scenes, the model's performance further improved, with all evaluation metrics surpassing 90% in the port scenario, reflecting exceptional detection precision and robustness in complex large-scale change areas. For the dense settlement scenario, recall and precision reached 91.78% and 95.42%, respectively, confirming the model's effectiveness in mitigating recognition bias caused by class imbalance. Even in suburban expansion and seasonal landscape change scenarios—where spatial dynamics and temporal variability are pronounced—MSDT-Net maintained stable performance, with key metrics consistently above 87%. The comprehensive analysis of recall, F1-score, IoU, and Kappa coefficient distributions across different scenes revealed that detection performance is closely associated with the structural complexity and change dynamics of the observed areas, underscoring the model's adaptive response mechanisms under varying environmental conditions. Overall, these experimental results demonstrate that the MSDT-Net, enhanced with an edge-aware smoothing module, not only improves overall detection accuracy but also substantially alleviates class imbalance issues. It exhibits strong generalization ability and robustness across diverse island remote sensing scenarios, providing reliable technical support for high-precision, multi-scenario change detection applications. The metric distribution box plots for specific scenarios are shown in Figure 10, offering further visual validation of the model's performance consistency.

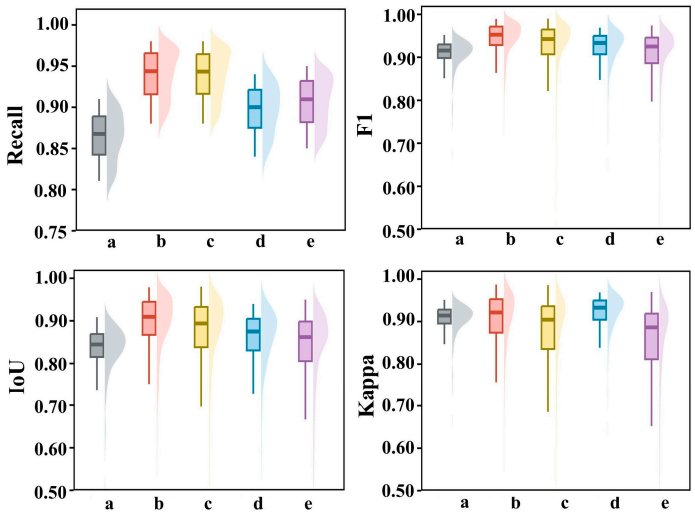


Figure 10. Boxplot and Violin Plot of Metric Distributions. (a–e correspond to Dispersed Building Distribution Scene, Dense Settlement Scene, Large Port Industrial Facility Change Scene, Suburban Expansion and Land Transformation Scene, Seasonal Cycle-Driven Landscape Change Scene).

To further verify the robustness and generalization ability of the proposed method under realistic island conditions, we selected a subset of the island dataset to conduct large-scene inference experiments. In this test, the model was applied in a sliding-window manner to ensure seamless coverage and accurate edge reconstruction across wide areas. As shown in Figure 11, the proposed MSDT-Net successfully detects both small and large building changes under complex coastal backgrounds.

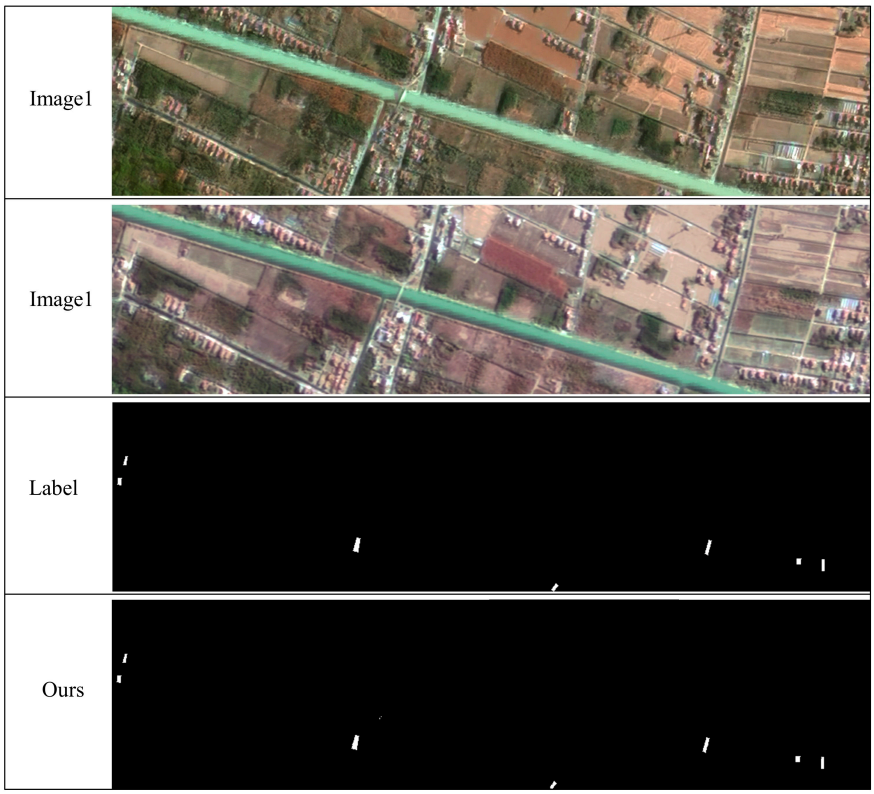


Figure 11. Visualization of full-scene inference results on island areas using MSDT-Net.

5. Discussion

To validate the generalization ability of MSDT-Net across domains, this study directly transfers the model trained on the typical island building change data to a land-based urban building change detection task for testing, as depicted in Figure 12. Island buildings typically have features such as dispersed layouts, large scale variations, and strong background interference, making them visually more challenging. Therefore, the model trained on this type of scenario can learn robust features sensitive to building changes more effectively. When transferred to a land-based scenario with more stable structures and uniform target shapes, it is expected to demonstrate good transfer performance and detection accuracy.

From the visualization results, MSDT-Net demonstrates good cross-domain adaptation ability in land-based urban building change detection. In terms of accuracy, the model can accurately identify building change areas in island building images, with the white detection results highly consistent with the actual annotations, validating its precise detection ability in high-interference backgrounds. In urban land images, despite the presence of various building types and dense distributions, the model is still able to consistently capture the changed targets. This shows that even when the target shape changes, the model maintains high detection performance, indicating that its accuracy does not degrade significantly when transferred to the land-based scenario. In terms of stability, whether facing small-scale targets in island environments or large-scale building changes in land-based

scenarios, the model's output results do not show significant performance fluctuations, reflecting its good robustness. This performance stability is crucial for cross-regional change detection in practical applications, especially for diverse remote sensing data processing scenarios. From the efficiency perspective, the model achieves good performance with no need for retraining on land-based urban buildings, saving a significant amount of data collection, model training time, and computational resources.

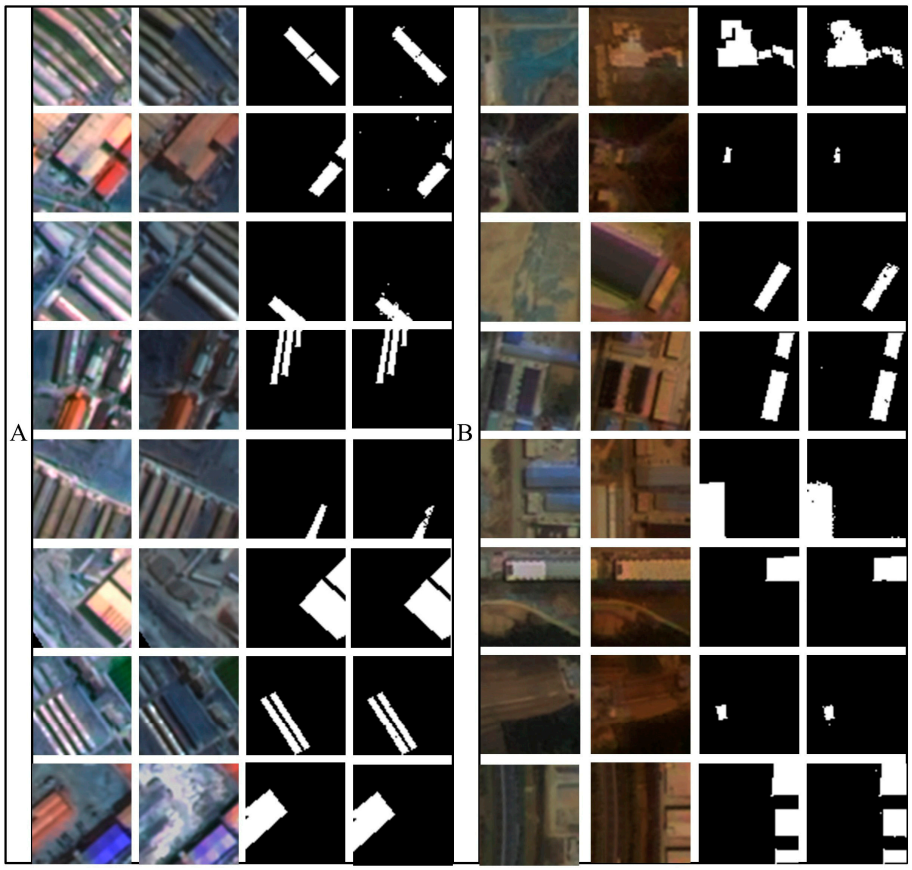


Figure 12. MSDT-Net Building Change Detection Transfer Diagram ((**A**) for land, (**B**) for island).

Although there are differences between island and land buildings in terms of style and environmental background, they share certain geometric structure and boundary characteristics. The model extracts cross-domain representative discriminative features through multi-scale difference modeling and edge-aware mechanisms, enabling effective knowledge transfer from high-complexity scenarios to structurally stable regions.

6. Conclusions

Island building change detection is a key technology for environmental monitoring, disaster early warning, and urban planning. It plays a crucial role in the dynamic monitoring and sustainable development of island resources. However, due to challenges such as class imbalance and boundary blurring, existing methods face significant difficulties in recognizing subtle changes and complex boundaries of island buildings.

This study proposes the MSDT-Net model, which achieves high-precision detection based on a twin ConvNeXt architecture. Its innovations are threefold: the introduction of an edge-aware smoothing module to address recognition accuracy issues caused by boundary blurring; the design of a difference modeling module to enhance detection performance for target and boundary changes; and the introduction of the Focal–Dice–IoU boundary joint

loss function, which effectively alleviates the class imbalance problem and significantly improves the model's sensitivity to sparse change regions.

Experiments show that MSDT-Net significantly improves key metrics on the self-built island dataset, performing excellently in scenarios with minimal change pixels. Multi-scenario tests demonstrate its strong generalization and robustness. Tests across land and island scenes verify its effectiveness in diverse remote sensing data processing scenarios.

In summary, the MSDT-Net model, through innovative architecture design and loss function optimization, successfully addresses the challenges of class imbalance and boundary blurring in island building change detection, providing a reliable technical path for the task. Future work will focus on optimizing the model structure to enhance its adaptability and robustness in complex remote sensing imagery.

Author Contributions: W.M. and X.J. conceived the study; L.Y. and W.M. developed the methodology; Y.C. and Y.Z. performed the software implementation; Z.W. and X.J. validated the results; W.M. conducted formal analysis; L.Y. carried out the investigation; X.J. provided resources; Y.C. curated the data; W.M. and L.Y. prepared the original draft; X.J. and C.Z. reviewed and edited the manuscript; Y.C. created the visualizations; X.J. supervised the study, managed the project, and secured funding. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Open Fund of Nansha Islands Coral Reef Ecosystem National Observation Research Station (grant no. NSICR24201), the Science and Technology Research Project of Jilin Provincial Department of Education (grant no. JJKH20250127KJ), the National Natural Science Foundation of China (grant no. 42301505), and the Key Laboratory of Ocean Geomatics, Ministry of Natural Resources, China (grant no. 2024B15).

Data Availability Statement: The data is sourced from the South China Sea Environmental Monitoring Center of the State Oceanic Administration.

Acknowledgments: The authors gratefully acknowledge the financial support from relevant funding agencies, which greatly facilitated the successful completion of this research.

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

References

1. Güttler, F.N.; Niculescu, S.; Gohin, F. Turbidity retrieval and monitoring of Danube Delta waters using multi-sensor optical remote sensing data: An integrated view from the delta plain lakes to the western-northwestern Black Sea coastal zone. *Remote Sens. Environ.* **2013**, *132*, 86–101. [CrossRef]
2. Khankeshizadeh, E.; Mohammadzadeh, A.; Jamali, S. EDB-HSTEU-Net: Earthquake-damaged building detection using a novel hybrid swin transformer efficient U-Net (HSTEU-Net) and transfer learning techniques from post-event VHR remote sensing data. *J. Build. Eng.* **2025**, *108*, 112889. [CrossRef]
3. Kolokotsa, D.; Lilli, K.; Gobakis, K.; Mavrigiannaki, A.; Haddad, S.; Garshasbi, S.; Mohajer, H.R.H.; Paolini, R.; Vasilakopoulou, K.; Bartesaghi, C.; et al. Analyzing the Impact of Urban Planning and Building Typologies in Urban Heat Island Mitigation. *Buildings* **2022**, *12*, 537. [CrossRef]
4. Begg, S.S.; De Ramon N'Yeurt, A.; Begg, S. Interweaving resource management with indigenous knowledge to build community resilience in the Pacific Islands: Case of the Waimanu Catchment in Viti Levu, Fiji. *Reg. Environ. Chang.* **2023**, *23*, 86. [CrossRef]
5. Du, B.; Shan, L.; Shao, X.; Zhang, D.; Wang, X.; Wu, J. Transform Dual-Branch Attention Net: Efficient semantic segmentation of ultra-high-resolution remote sensing images. *Remote Sens.* **2025**, *17*, 540. [CrossRef]
6. Tasyurek, M. BBD: A new hybrid method for geospatial building boundary detection from huge size satellite imagery. *Multimed. Tools Appl.* **2025**, *84*, 8515–8545. [CrossRef]
7. He, P.; Shi, W.; Zhang, H.; Hao, M. Adaptive superpixel based Markov random field model for unsupervised change detection using remotely sensed images. *Remote Sens. Lett.* **2018**, *9*, 724–732. [CrossRef]
8. Gong, X.; Yao, J.; Li, Q.; Liu, X.; Zhong, P. Feature matching for remote-sensing image registration via neighborhood topological and affine consistency. *Remote Sens.* **2022**, *14*, 2606. [CrossRef]
9. Wang, P.; Zhang, H.; Patel, V.M. An improved change detection approach using tri-temporal logic-verified change vector analysis. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 278–293.

10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
12. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [CrossRef]
13. Xu, S.; Li, H.; Liu, T.; Gao, H. A method for airborne small-target detection with a multimodal fusion framework integrating photometric perception and cross-attention mechanisms. *Remote Sens.* **2025**, *17*, 1118. [CrossRef]
14. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected siamese network for change detection of VHR images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
15. Huang, Y.; Jiao, D.; Huang, X.; Tang, T.; Gui, G. A hybrid CNN-Transformer network for object detection in optical remote sensing images: Integrating local and global feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 241–254. [CrossRef]
16. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
17. Shi, Q.; Liu, M.X.; Li, S.C.; Liu, X.P.; Wang, F.; Zhang, L.P. 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*, 1–6. [CrossRef]
18. Song, K.; Jiang, J. AGCDetNet: An attention-guided network for building change detection in high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4816–4831. [CrossRef]
19. Zhang, Y.; Deng, F.; He, F.; Guo, Y.; Sun, G.; Chen, J. FODA: Building change detection in high-resolution remote sensing images based on feature-output space dual-alignment. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8125–8134. [CrossRef]
20. Dong, W.; Zhao, W.; Wang, S. Multiscale context aggregation network for building change detection using high resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
21. Li, Y.; Zou, S.; Zhao, T.; Su, X. MDFA-Net: Multi-scale differential feature self-attention network for building change detection in remote sensing images. *Remote Sens.* **2024**, *16*, 3466. [CrossRef]
22. Zhao, Y.; Song, X.; Li, J.; Liu, Y. CSCNet: A cross-scale coordination siamese network for building change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 1377–1389. [CrossRef]
23. Yun, K.; Kim, J.; Song, Y.S. Change detection of buildings using siamese neural networks in coastal environments. *J. Coast. Res.* **2023**, *116*, 274–278. [CrossRef]
24. Pan, X.; Zhang, C.; Xu, J.; Zhao, J. Simplified object-based deep neural network for very high resolution remote sensing image classification. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 218–237. [CrossRef]
25. Zheng, H.; Gong, J.; Liu, T.; Jiang, F.; Zhan, T.; Lu, D.; Zhang, M. HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images. *Pattern Recognit.* **2022**, *129*, 108717. [CrossRef]
26. Xiong, J.; Liu, F.; Wang, X.; Yang, C. Siamese transformer-based building change detection in remote sensing images. *Sensors* **2024**, *24*, 1268. [CrossRef]
27. Rao, W.Q.; Gao, L.R.; Qu, Y.; Sun, X.; Zhang, B.; Chanussot, J. Siamese Transformer Network for Hyperspectral Image Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
28. Ren, W.; Wang, Z.; Xia, M.; Lin, H. MFINet: Multi-scale feature interaction network for change detection of high-resolution remote sensing images. *Remote Sens.* **2024**, *16*, 1269. [CrossRef]
29. Zhang, M.; Li, Q.; Yuan, Y.; Wang, Q. Edge Neighborhood Contrastive Learning for Building Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
30. Wang, Q.; Zhang, M.; Ren, J.; Li, Q. Exploring Context Alignment and Structure Perception for Building Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–10. [CrossRef]
31. Alcantarilla, P.F.; Stent, S.; Ros, G.; Arroyo, R.; Gherardi, R. Street-view change detection with deconvolutional networks. *Robot. Auton. Syst.* **2018**, *100*, 187–199. [CrossRef]
32. Wang, P.; Zhang, H.; Patel, V.M. SAR image despeckling using a convolutional neural network. *IEEE Signal Process. Lett.* **2020**, *24*, 1763–1767. [CrossRef]
33. Mou, L.; Bruzzone, L.; Zhu, X.X. Cascade attentive network for remote sensing change detection. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 184–197.
34. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery. *Remote Sens.* **2018**, *10*, 1369.
35. Zhang, Z.; Yang, Y.; Jian, X. MASNet: A novel deep learning approach for enhanced detection of small targets in complex scenarios. *Meas. Sci. Technol.* **2025**, *36*, 045402. [CrossRef]
36. Wang, F.; Zhang, L.; Jiang, T.; Li, Z.; Wu, W.; Kuang, Y. An Improved Segformer for Semantic Segmentation of UAV-Based Mine Restoration Scenes. *Sensors* **2025**, *25*, 3827. [CrossRef]

37. Huo, Y.; Gang, S.; Guan, C. FCIHMRT: Feature cross-layer interaction hybrid method based on Res2Net and transformer for remote sensing scene classification. *Electronics* **2023**, *12*, 4362. [CrossRef]
38. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multi attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4274–4288. [CrossRef]
39. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
40. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
41. Kim, J.; Lee, J.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26–30 June 2016; pp. 1646–1654.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
43. Zhang, M.; Shi, W. A feature difference convolutional neural network-based change detection method. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7232–7246. [CrossRef]
44. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. *IEEE Trans. Med. Imaging* **2018**, *37*, 1342–1355.
47. Bandara, W.G.C.; Patel, V.M. A transformer-based siamese network for change detection. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022.
48. Li, J.P.; He, W.; Li, Z.H.; Guo, Y.J.; Zhang, H.Y. Overcoming the uncertainty challenges in detecting building changes from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 1–17. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.