remote sensing

MDPI

Article

CMSNet: A SAM-Enhanced CNN–Mamba Framework for Damaged Building Change Detection in Remote Sensing Imagery

Jianli Zhang *, Liwei Tao, Wenbo Wei, Pengfei Ma and Mengdi Shi

Army Arms University of PLA, 451 Huangshan Road, Shushan District, Hefei 230031, China; liweitaoo@163.com (L.T.); zhuceyong0113@163.com (W.W.); xiaoma.2002@163.com (P.M.); shismdshi@163.com (M.S.)
* Correspondence: wjxc1982@126.com

Highlights

What are the main findings?

- We propose CMSNet, an end-to-end building damage change detection framework that integrates SAM-derived structural priors with a CNN–Mamba backbone.
- A Pretrained Visual Prior–Guided Fusion Module (PVPF-FM) is introduced to align SAM priors with temporal change cues, enabling sharper boundaries and more accurate local damage extraction.
- We construct RWSBD, a large-scale real-world war-scene dataset comprising 42,732 annotated buildings, offering a fine-grained benchmark for damage analysis.

What is the implication of the main finding?

- The CMSNet significantly improves the accuracy and robustness of remote sensing-based building damage detection in complex war and explosion scenarios.
- The integration of SAM priors and CNN–Mamba allows for better handling of intricate structural variations, enhancing the model's adaptability in real-world disaster assessments.
- The new RWSBD dataset provides a valuable benchmark, advancing both research and practical applications in post-disaster remote sensing.

check for updates

Academic Editor: Javier Marcello

Received: 17 October 2025
Revised: 21 November 2025
Accepted: 28 November 2025
Published: 3 December 2025

Citation: Zhang, J.; Tao, L.; Wei, W.; Ma, P.; Shi, M. CMSNet: A SAM-Enhanced CNN–Mamba Framework for Damaged Building Change Detection in Remote Sensing Imagery. *Remote Sens.* **2025**, *17*, 3913. https://doi.org/10.3390/rs17233913

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/).

Abstract

In war and explosion scenarios, buildings often suffer varying degrees of damage characterized by complex, irregular, and fragmented spatial patterns, posing significant challenges for remote sensing–based change detection. Additionally, the scarcity of high-quality datasets limits the development and generalization of deep learning approaches. To overcome these issues, we propose CMSNet, an end-to-end framework that integrates the structural priors of the Segment Anything Model (SAM) with the efficient temporal modeling and fine-grained representation capabilities of CNN–Mamba. Specifically, CMSNet adopts CNN–Mamba as the backbone to extract multi-scale semantic features from bi-temporal images, while SAM-derived visual priors guide the network to focus on building boundaries and structural variations. A Pre-trained Visual Prior-Guided Feature Fusion Module (PVPF-FM) is introduced to align and fuse these priors with change features, enhancing robustness against local damage, non-rigid deformations, and complex background interference. Furthermore, we construct a new RWSBD (Real-world War Scene Building Damage) dataset based on Gaza war scenes, comprising 42,732 annotated building damage instances across diverse scales, offering a strong benchmark for real-world

scenarios. Extensive experiments on RWSBD and three public datasets (CWBD, WHU-CD, and LEVIR-CD+) demonstrate that CMSNet consistently outperforms eight state-of-the-art methods in both quantitative metrics (F1, IoU, Precision, Recall) and qualitative evaluations, especially in fine-grained boundary preservation, small-scale change detection, and complex scene adaptability. Overall, this work introduces a novel detection framework that combines foundation model priors with efficient change modeling, along with a new large-scale war damage dataset, contributing valuable advances to both research and practical applications in remote sensing change detection. Additionally, the strong generalization ability and efficient architecture of CMSNet highlight its potential for scalable deployment and practical use in large-area post-disaster assessment.

Keywords: remote sensing; damaged buildings; war; foundation models; deep learning

1. Introduction

With the rapid development of remote sensing technology and its widespread applications in land monitoring, urban planning, natural disaster assessment, and battlefield situation awareness, change detection (CD) based on remote sensing imagery has become a key research direction in Earth observation [1,2]. Change detection aims to compare remote sensing data of the same area acquired at different times to identify spatial or attribute changes of ground objects, thereby enabling the monitoring and analysis of the dynamic evolution of surface targets [3,4]. In particular, for tasks such as natural disaster response, emergency management, and damage assessment, change detection can provide governments or military departments with first-hand data support, enabling rapid response and precise deployment.

Among various change detection tasks, building damage detection, as a fine-grained and target-specific form of change detection, has attracted increasing attention in recent years [5,6]. As the primary carriers of urban spatial structure, buildings serve as direct indicators of the intensity of disasters or attacks in a given area. In particular, after war conflicts or major natural disasters, timely and accurate acquisition of building damage information is of great significance for rescue operations, post-disaster reconstruction, and war impact assessment [7,8]. Compared with traditional ground surveys, remote sensing imagery offers advantages such as wide coverage, high timeliness, and repeatable acquisition, making it an ideal tool for war damage monitoring and assessment [9,10].

Although change detection techniques have made initial progress in disaster remote sensing, building damage change detection based on remote sensing imagery still faces multiple challenges [11]. First, unlike macroscopic changes such as land cover transitions or urban expansion, building damage typically manifests as structural destruction, occlusion damage, or partial collapse, which are more fine-grained and irregular. Second, in complex battlefield environments, images are often affected by shadows, smoke, occlusions, and ground disturbances, which severely impair a model's ability to accurately identify true change regions. More critically, there is a lack of publicly available, high-quality datasets of real war-scene building damage. Most existing studies rely on simulated or synthetic data, resulting in limited model generalization and difficulty in meeting the requirements of real-world deployment.

With the rapid advancement of artificial intelligence, particularly the widespread adoption of deep learning in image classification [12], semantic segmentation [13], and object detection [14], remote sensing image change detection has gradually shifted from traditional image differencing methods [15] and machine learning approaches [16] toward

end-to-end strategies based on deep neural networks [17]. The introduction of dual-branch structures (e.g., Siamese Networks) [18], attention mechanisms [19], and Transformer architectures [20] has greatly enhanced the ability of models to represent and robustly detect changes in complex scenarios. In recent years, the rise of foundation models [21]—notably under the pretraining–fine-tuning paradigm—has attracted significant attention in computer vision. Among them, Meta's Segment Anything Model (SAM), as a foundation model for segmentation, demonstrates powerful cross-task transferability and fine-grained boundary perception. It can generate high-quality segmentation masks without task-specific annotations, offering new solutions for weakly supervised or zero-shot remote sensing tasks. However, current foundation vision models like SAM are primarily designed for natural images, and their direct application to remote sensing imagery often leads to semantic shifts and structural misinterpretations [22]. Effectively integrating the prior knowledge of foundation models into task-specific remote sensing networks thus remains an urgent research direction.

At the same time, the development of lightweight neural network models has provided new tools for efficient remote sensing analysis. The recently proposed Mamba architecture [23], which models long-range dependencies through linear recurrent units, achieves a balance between lightweight design and strong temporal modeling capabilities. Its introduction into image modeling tasks has opened new pathways for feature representation beyond CNNs and Transformers, indicating potential advantages for remote sensing change detection in terms of both efficiency and temporal feature extraction.

However, despite these advancements, existing building damage detection methods still face significant limitations that hinder their applicability in real-world complex scenarios. Many CNN- and Transformer-based approaches struggle with structural consistency modeling, making it difficult to maintain the coherence of building geometry under non-rigid deformations, fragmented destruction, or occlusions. Similarly, their adaptability to complex war-zone environments remains insufficient, as fuzzy boundaries, small-scale damage, and cluttered backgrounds often lead to missed or misclassified change regions. While SAM provides strong structural priors and Mamba offers efficient sequence modeling, current approaches still lack an effective mechanism to integrate these complementary strengths.

In addition to these methodological gaps, several critical challenges remain in building damage change detection. First, data scarcity is a major obstacle: there is a lack of publicly available, real, and representative datasets of damaged buildings in remote sensing imagery, particularly from actual war scenarios. This severely limits model training, performance evaluation, and method comparison. Second, even state-of-the-art models struggle to capture fine-grained structural details, often failing to handle fuzzy boundaries and complex damage patterns in real battlefield environments. Third, mechanisms for efficiently incorporating foundation model priors are still immature; existing approaches do not fully leverage the structural cues contained in foundation vision models (e.g., SAM), reducing their potential effectiveness for building damage analysis. Finally, achieving both high accuracy and high efficiency remains challenging: foundation models offer strong representational power but incur high computational cost, whereas lightweight CNN-Mamba architectures improve efficiency but exhibit reduced robustness in highly complex scenes.

To address the above challenges, this paper proposes a novel building damage change detection framework, CMSNet. CMSNet adopts CNN-Mamba as the backbone, leveraging its capability for effective detail and temporal feature modeling to extract multi-scale semantic features from bi-temporal remote sensing imagery. Meanwhile, SAM is introduced as an external structural prior, enhancing intermediate features to guide the model toward building boundaries and damaged regions, thereby improving segmentation accuracy and

structural consistency. To fully exploit the potential of foundation models, we design a Pretrained Visual Prior-Guided Feature Fusion Module (PVPF-FM), which aligns and fuses SAM outputs with CNN-Mamba intermediate features, strengthening the model's ability to detect complex change regions. Moreover, to mitigate the issue of data scarcity, we construct a new high-resolution dataset, RWSBD, based on real war-zone remote sensing imagery from the Gaza Strip. RWSBD covers diverse damage types, large-area scenes, and complex environments, providing a valuable benchmark for building damage change detection in real-world war scenarios.

To comprehensively evaluate the performance of CMSNet, we conducted extensive experiments on four remote sensing change detection datasets: RWSBD, CWBD, WHU-CD, and LEVIR-CD+, and compared it with eight state-of-the-art methods. The results demonstrate that CMSNet consistently outperforms competing approaches across multiple metrics (including F1-score, IoU, Precision, and Recall), showing particular advantages in fine-grained building damage detection with stronger boundary preservation and robustness. In summary, the main contributions of this work are as follows:

(1) We propose CMSNet, a building damage change detection framework that integrates the foundation vision model SAM with CNN-Mamba, combining structural prior enhancement with efficient feature modeling.
(2) We design a Pretrained Visual Prior-Guided Feature Fusion Module (PVPF-FM) to effectively integrate SAM segmentation priors with backbone feature representations.
(3) We construct the RWSBD dataset, providing a high-resolution benchmark of real-world war-zone building damage for remote sensing change detection, supporting future research in this area.
(4) We systematically validate CMSNet on four remote sensing datasets, demonstrating superior accuracy and robustness in complex scenarios.

2. Related Work and Problem Statement

Remote sensing-based building damage assessment has undergone a long evolution, forming two major research paradigms: (1) single-temporal image–based classification, and (2) multi-temporal image–based change detection. Early studies primarily relied on single-temporal images, where damage levels were inferred from post-event optical or SAR imagery using handcrafted features such as texture descriptors, spectral indices, morphological profiles, and rule-based classifiers. Although effective in certain structured environments, these methods could not reliably distinguish true structural damage from naturally dark roofs, shadows, debris, or seasonal variations, due to the absence of explicit temporal comparison. With the availability of multi-temporal remote sensing data, the research paradigm gradually shifted toward change detection, where differences between pre- and post-event images provide more discriminative cues for identifying actual structural destruction. This paradigm has proven more robust for disaster scenarios, especially for recognizing collapse, partial roof failure, or severe structural fragmentation.

Driven by deep learning advancements, remote sensing change detection—one of the core tasks in intelligent image interpretation—has made substantial progress in recent years [24,25]. In the area of building change detection, researchers have extensively explored backbone feature extraction, temporal interaction mechanisms, multi-scale modeling, and structural prior integration [24,25]. Based on backbone modeling strategies, current methods can be broadly divided into four categories: CNN-based methods, Transformer-based methods, Mamba-based methods, and methods integrating foundation models. Despite the improvements contributed by these paradigms, significant challenges remain for detecting damaged buildings in complex war-zone environments characterized by irregular destruction patterns, heterogeneous backgrounds, and fragmented structures.

2.1. CNN-Based Change Detection Methods

Convolutional Neural Networks (CNNs) have long been the mainstream modeling framework for remote sensing change detection tasks [18,26]. Typical methods such as FC-EF, FC-Siam-Conc, and FC-Siam-diff [27] employ end-to-end architectures to extract bi-temporal features, discriminate changes, and generate binary masks. CNNs provide strong local perceptual capabilities and can enhance fine details through multi-scale convolutional designs. However, as inherently local feature extractors, they struggle with long-range spatial dependencies, complex damage patterns, and irregular building boundaries [28]. In real war scenarios, asymmetric destruction, occlusions, and debris-induced texture anomalies are difficult to model with local kernels alone [29]. CNNs also lack long-term temporal modeling capacities, limiting their ability to distinguish subtle structural damage from background noise. Thus, although simple and stable, CNN-based methods often fail to capture global relationships, making them insufficient for modeling discontinuous and irregular war-zone building changes.

2.2. Transformer-Based Change Detection Methods

Transformers [30], widely applied in NLP [31] and image recognition [32], introduce global attention mechanisms to remote sensing change detection tasks [20]. Models such as ChangeFormer [33], BIT [20], and MATNet [34] use self-attention to capture long-range contextual cues, alleviating the receptive field limitations of CNNs. This global modeling ability supports the detection of large-scale changes such as massive collapses or structural disintegration [35]. However, Transformers entail heavy computation, parameter redundancy, and high memory consumption. They also tend to generate blurred change boundaries and exhibit sensitivity to training data volume [36]. In war-zone imagery with limited high-quality samples, Transformers may overfit or fail to converge effectively [37]. Although strong in global modeling, their limitations in boundary precision, efficiency, and small-sample robustness hinder their applicability in real war-damage detection tasks.

2.3. Mamba-Based Change Detection Methods

The Mamba architecture [23] replaces conventional attention with a linear state-space model, enabling efficient long-range dependency modeling with low computation overhead. This makes Mamba an emerging alternative to CNNs and Transformers in vision modeling. Mamba-based change detection approaches such as Change-Mamba [38], CDMamba [39], and M-CD [40] have shown significant advantages in accuracy and efficiency, drawing increasing attention within the remote sensing community. However, Mamba's adaptation to remote sensing imagery remains limited [41]. Current methods often neglect shallow edges and fine texture details, leading to inaccuracies in scenarios involving irregular building boundaries. Most architectures rely on shallow temporal difference modules, lacking deeper bi-temporal interaction strategies. For building damage detection—where changes can be fragmented, subtle, or boundary-focused—the ability of Mamba to model evolving spatiotemporal features remains insufficiently validated. Existing designs are largely generic and not optimized for paired bi-temporal remote sensing images. While theoretically promising, Mamba-based change detection is still in its early stages and lacks task-specific structural optimization.

2.4. Foundation Model-Based Change Detection Methods

Foundation vision models such as SAM [21], CLIP [42], and DINO [43] have demonstrated remarkable generalization ability. Among them, SAM shows strong competence in boundary perception, occlusion handling, and structural delineation, making it particularly attractive for remote sensing segmentation tasks. Existing attempts—such as

RemoteSAM [22], SAM3D [44], and SCD-SAM [45]—have explored using SAM in building extraction or semantic interpretation tasks. However, effectively integrating SAM's structural priors into change detection remains challenging. SAM lacks temporal awareness and may introduce noise if its structural priors are fused improperly. Additionally, SAM's natural-image pretraining limits its adaptation to scale variations and complex backgrounds in remote sensing [22,46]. Key open problems include robust fusion of SAM priors with bi-temporal features, handling semantic shifts, and adapting SAM to irregular building damage structures. While foundation models introduce new opportunities for structural prior integration, substantial modeling challenges remain.

To address these challenges, this paper proposes a novel method—CMSNet—which combines the efficient feature modeling capability of CNN-Mamba with the structural awareness of the SAM foundation model. We design a Pretrained Visual Prior-Guided Feature Fusion Module (PVPF-FM) to enhance structural representation in damage regions. We also construct RWSBD, a real-world war-zone building damage dataset, providing a critical benchmark for the field. CMSNet achieves breakthroughs in structural perception, boundary precision, temporal modeling, and foundation-model-guided fusion, leveraging complementary strengths across modeling paradigms and exploring new strategies for remote sensing building damage change detection.

3. The Dataset of Real War Scenes

To support in-depth research on building damage detection in real war-zone remote sensing imagery, this paper constructs a remote sensing change detection dataset for real war scenarios—RWSBD.

The dataset is based on high-resolution optical imagery and a rigorous manual annotation process, covering typical damage patterns in real war-affected areas, making it highly practical and valuable for research (see Figure 1). RWSBD consists of two high-resolution images of the Gaza Strip acquired by the GF-2 optical remote sensing satellite, captured on 24 September 2022 (pre-conflict) and 3 May 2024 (post-conflict), with a spatial resolution of 0.8 m. The geographic coverage spans the entire Gaza Strip, including urban, suburban, and densely built-up areas. To ensure alignment, the pre- and post-conflict images were carefully registered using feature-based image registration techniques, with key points matched and refined through local affine transformations. Accuracy control was conducted via visual inspection and cross-verification with high-resolution reference data to ensure spatial consistency and annotation reliability. These images effectively capture building structural details and local damage characteristics, providing a high-quality visual foundation for subsequent building damage recognition and change analysis.

During dataset construction, a manual, building-by-building annotation approach was employed to precisely label damaged buildings in the post-conflict images. Annotations were performed by remote sensing professionals with interpretation experience, focusing on severe war damage features such as roof collapse and total building collapse. In total, 42,723 damaged building instances were annotated, covering diverse scenarios including urban and suburban areas, fully reflecting the variety and complexity of war-zone building damage. To meet the training requirements of deep learning models, the pre- and post-conflict images along with their corresponding annotations were uniformly cropped into 256 $\times$ 256-pixel patches. Through cropping and filtering, 28,210 image pairs with aligned and accurately labeled masks were obtained. These samples are representative and discriminative, serving as a high-quality training set for high-resolution war-zone building damage change detection. The construction of the RWSBD dataset not only fills the gap of high-resolution remote sensing datasets for building damage detection in real

war scenarios but also provides a solid benchmark for model training, evaluation, and transfer learning in subsequent research.

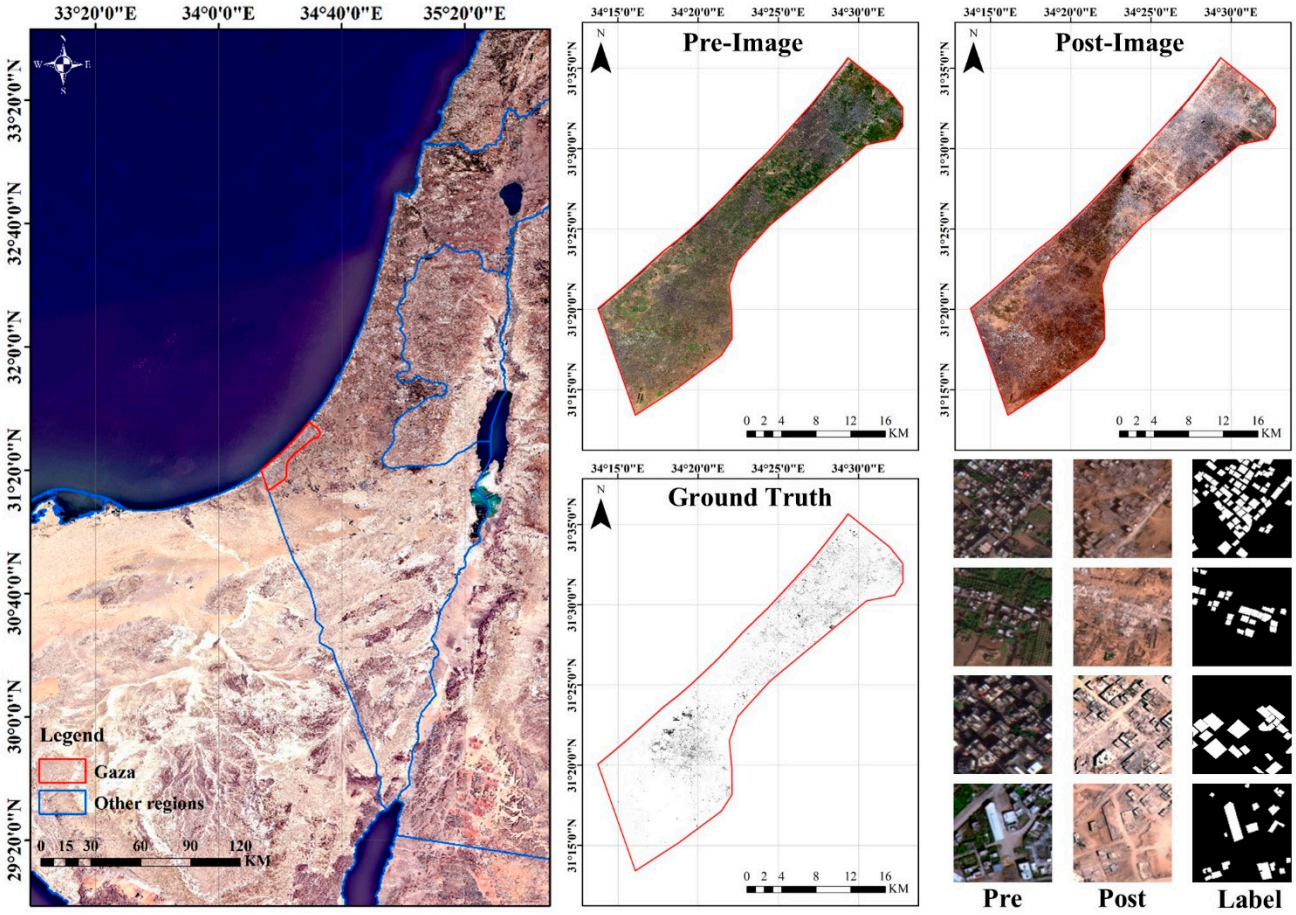


Figure 1. Overview of the RWSBD dataset.

During annotation, partially damaged buildings were labeled by marking only the visually confirmed damaged portions rather than the entire building footprint, ensuring precise localization of destruction. Since this study focuses on binary damaged/undamaged identification, we did not assign multi-level damage categories, as doing so requires additional standardized criteria and expert consensus; such severity-level annotations will be considered in future extensions of the dataset. Moreover, buildings that were naturally demolished, renovated, or newly constructed between the two temporal images were not included, in order to avoid semantic ambiguity and ensure that the dataset strictly reflects war-induced structural damage.

4. Methodology

4.1. CMSNet Overall Architecture

The CMSNet architecture adopts an encode–fuse–decode design, aiming to simultaneously enhance structural awareness of damaged targets and the modeling of change features. The main network is built with a CNN-Mamba hybrid structure, enabling synergistic modeling of local texture details and global temporal dependencies. In addition, a structural prior awareness path is incorporated: building structure masks extracted by the pretrained SAM foundation model serve as guidance, further enhancing the network's ability to capture building boundaries and morphological changes, enabling precise extraction of fine-grained damaged building regions (see Figure 2). CMSNet employs a dual-stream parameter-sharing structure to process pre- and post-conflict images. Each

stream consists of a shallow CNN module and a deep Mamba module. The CNN extracts low-level local features, such as edges and textures, while Mamba—a structured state-space model—captures long-range dependencies and temporal differences, effectively modeling global semantic changes across time. This combination preserves edge details while providing awareness of large-scale semantic variations. To further enhance structural understanding, CMSNet introduces the SAM structural prior path alongside the backbone. Pre- and post-conflict images are input into a frozen SAM to extract multi-level structural mask features, which are then deeply integrated with backbone semantic features via the Pretrained Visual Prior-Guided Feature Fusion Module (PVPF-FM) during intermediate fusion. This module employs channel and spatial attention mechanisms to dynamically adjust the fusion weights between SAM masks and CNN-Mamba features, improving discrimination in cases of blurred boundaries, minor damage, or dense destruction. SAM remains frozen during training, serving solely as a source of structural priors to provide building integrity constraints. The decoding process uses a multi-scale decoder fusion structure that progressively upsamples and integrates features from both the encoder and SAM branch. A boundary refinement channel further enhances the recovery quality of building edges. The final output is a pixel-level damage mask from the change prediction head, ensuring fine boundaries, clear regional structures, and minimal local errors. Throughout decoding, spatial details from CNN, semantic changes from Mamba, and structural guidance from SAM are synergistically maintained. To jointly optimize pixel-level accuracy, structural consistency, and boundary fidelity, CMSNet employs a multi-component loss system. The base loss uses binary cross-entropy (BCE) to supervise overall damage segmentation, combined with Dice loss to address class imbalance and instability in small target regions. A Structure Alignment Loss is introduced to constrain consistency between SAM-guided features and actual change features, improving structural prediction stability. The total loss is a weighted combination of these components, ensuring synergistic coordination among feature guidance, boundary restoration, and change perception during training.

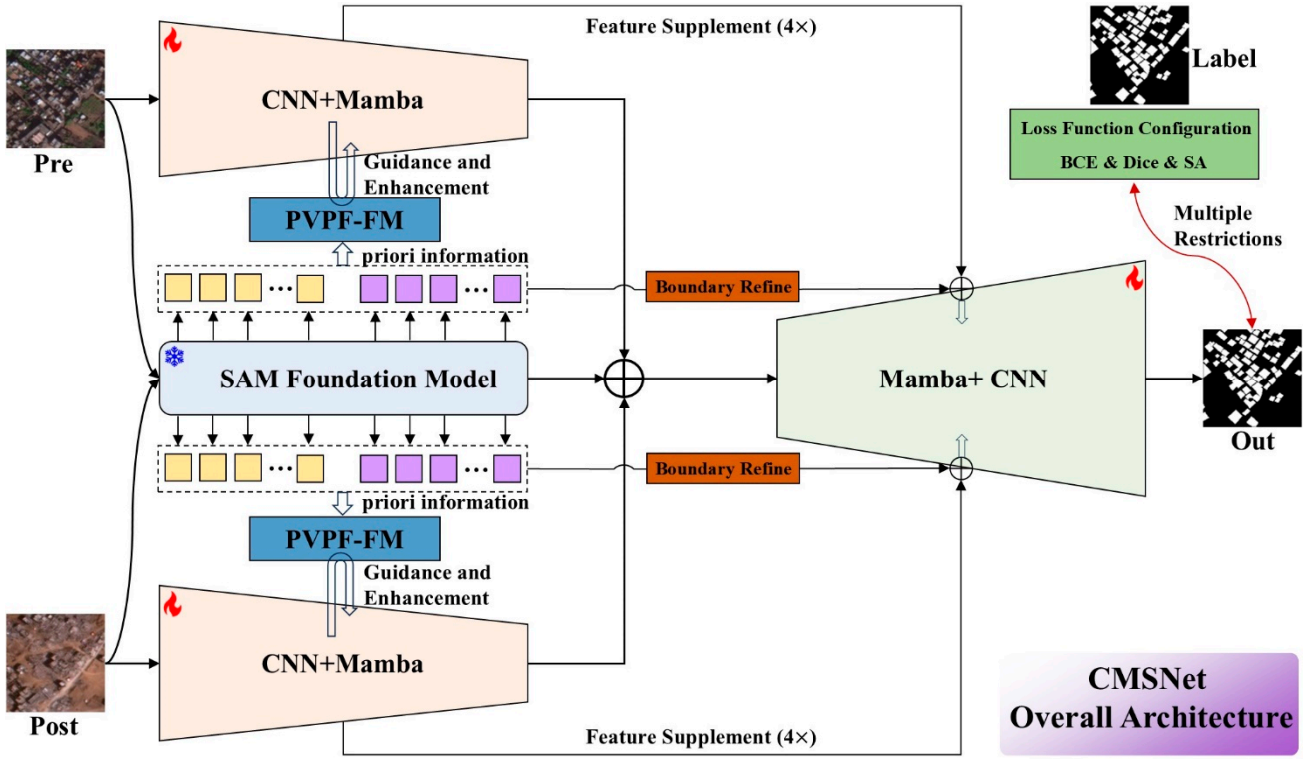


Figure 2. Details and Workflow of CMSNet Network.

4.2. Overview of the SAM Foundational Model

SAM, released by Meta AI in 2023, is a pretrained foundation model with broad visual perception and segmentation capabilities. It provides a general-purpose image segmentation framework with strong zero-shot segmentation ability, enabling mask prediction for any object or region in an image without relying on task-specific downstream training. The core idea of SAM is to unify image segmentation as a "prompt-to-mask" problem, where different types of input prompts—points, boxes, or text—are used to predict high-quality masks for the corresponding regions in the image. This makes SAM widely adaptable across multiple domains, including natural images, medical imaging, and remote sensing, and allows it to serve as a structural prior model embedded within more complex recognition systems. SAM's overall architecture consists of three key modules: an image encoder, a prompt encoder, and a mask decoder. The image encoder uses a Vision Transformer (ViT) to embed the input image into high-dimensional representations, forming a global visual feature tensor. The prompt encoder embeds user-provided prompts—points, boxes, or semantic vectors—into the same feature space. The mask decoder employs a lightweight Transformer architecture to fuse the image features with the prompt embeddings, and uses a multi-head attention mechanism to predict the mask regions corresponding to the prompts. The basic processing workflow can be formally described as follows:

Image embedding generation step: Generate a high-dimensional feature representation based on ViT from an input image I of size $H \times W \times C$:

$$F = ViT_{img}(I),\ F \in R^{H \times W \times C}, \tag{1}$$

where I denotes the input image, $ViT_{img}(x)$ represents the image encoder, and F denotes the resulting high-dimensional feature.

In the prompt vector embedding step, embedding vectors are generated based on the provided prompt inputs:

$$P = Encoder_{prompt}(p),\ p \in \{point, box, text\}, \tag{2}$$

where p represents the input prompt, which can be in the form of points, bounding boxes, or text, and P denotes the output embedding vector in the same feature space as the image features.

In the mask prediction step, the mask output is decoded based on the high-dimensional image features F and the prompt embedding vectors P:

$$\hat{M} = Decoder_{mask}(F, P), \tag{3}$$

where $Decoder_{mask}(x, y)$ represents the mask decoder, $\hat{M}$ denotes the predicted mask, and $\hat{M} \in [0, 1]^{H \times W}$.

Through the above process, the SAM acquires the ability to parse structural regions from arbitrary visual prompts. Its mask outputs typically exhibit high boundary quality and structural integrity, making them particularly suitable as structural prior inputs for tasks in the remote sensing domain, such as building recognition and damage detection. In this study, we freeze the SAM and use only the building structure masks generated from pre- and post-event remote sensing images as prior features to guide the backbone network, enabling more accurate focus on damaged regions and improving the completeness of damage boundaries and semantic consistency (see Figure 3).

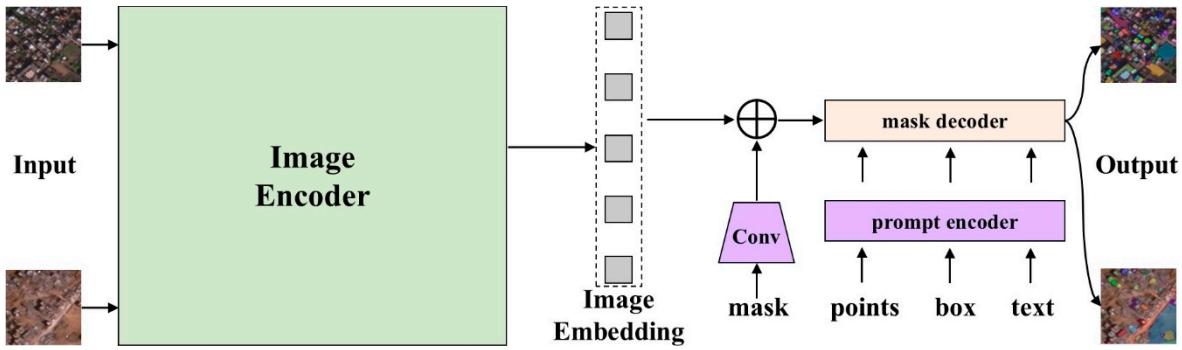


Figure 3. Schematic diagram of SAM principle.

In our implementation, the SAM encoder is kept frozen during CMSNet training. The pretrained SAM extracts high-level structural mask features from the input images, which serve as external visual priors to guide the network's focus on building boundaries and potential structural changes. No fine-tuning is applied to SAM's weights, ensuring that the foundation model's structural knowledge is preserved and consistently integrated into the PVPF-FM module. Only the downstream CNN-Mamba backbone and fusion layers are updated during training, which maintains experimental reproducibility and stability across different datasets. This design allows CMSNet to leverage SAM's global structural awareness without incurring additional computational overhead or risking overfitting to the training data.

4.3. CNN-Mamba Encoder–Decoder Architecture

In remote sensing building damage change detection, building structures are often complex and diverse, with changes ranging from severe structural collapse to roof damage. These changes typically involve multi-scale spatial information, long-range dependencies, temporal semantic shifts, and blurred edge details. Traditional CNN-based networks perform well in extracting local features but are limited by their receptive field and lack of global modeling capability, making it difficult to capture cross-scale and large-area semantic changes. Transformer-based architectures, while providing global perception, incur high computational costs and are less sensitive to boundaries, rendering them less suitable for large-scale, high-resolution remote sensing scenarios. To address these challenges, this study proposes an effective integration of CNN with the state-space modeling architecture Mamba, forming a powerful CNN-Mamba encoder–decoder backbone network that enables precise perception and reconstruction of multi-scale, multi-temporal building changes.

The backbone of CMSNet follows a symmetrical dual-stream encoder–decoder architecture, processing input pairs of pre- and post-event remote sensing images (I_{pre}, I_{post}). Each stream is composed of four stacked CNN-Mamba blocks, forming multi-scale semantic representations. The final change mask is reconstructed through cross-stream difference modeling and feature fusion. In this architecture, CNN-extracted edge details are combined with Mamba-learned global temporal change patterns, enabling local-global collaborative perception: shallow layers focus on structural details, while deeper layers capture semantic and spatial dependencies. The decoder integrates multi-scale features from the encoding stage to reconstruct high-precision damage masks. The encoder employs a hierarchical design, with each layer containing a CNN-Mamba block that includes convolutional feature extraction, downsampling, state-space modeling, and residual connections. Within each block, standard convolution is first applied to extract local spatial features, preserving fine details such as boundary shapes, building contours, and texture variations:

$$F_{CNN}^{(l)} = Conv_{3\times3}\left(BN\left(ReLU\left(F^{(l-1)}\right)\right)\right), \tag{4}$$

where $F^{(l-1)}$ denotes the output from the previous layer, and $F^{(l)}_{CNN}$ represents the output of the convolutional branch at the current layer. In this work, standard ResNet modules are employed to construct the shallow feature representations.

The Mamba module unfolds spatial features along rows or columns and employs a state-space modeling mechanism to capture long-range contextual dependencies and global temporal dynamics. Unlike Transformers, Mamba has linear time complexity $O(n)$, making it suitable for large-scale image inputs. Its fundamental unit can be expressed as:

$$s_t = A \cdot s_{t-1} + B \cdot x_t, \tag{5}$$

$$y_t = C \cdot s_t + D \cdot x_t, \tag{6}$$

where x_t denotes the sequence of convolutional feature vectors, s_t represents the hidden state, y_t is the output feature representation, and A, B, C, and D are learnable parameters.

As shown in Figure 4, at the image level, this can be understood as splitting the 2D feature tensor F_{CNN} along the H and W dimensions into sequences, processing them through the Mamba module, and then reassembling them into a 2D feature map F_{Mamba}.

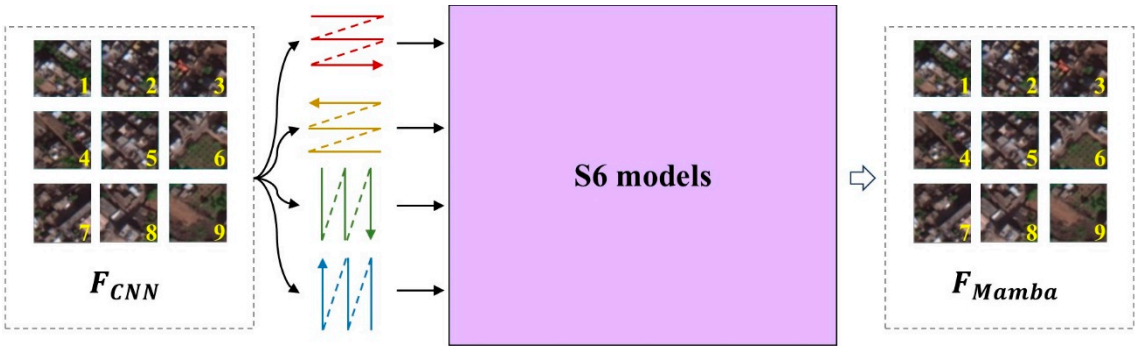


Figure 4. The Image Scanning Mechanism of Mamba.

Finally, the outputs of the CNN and Mamba branches are fused. In this step, Mamba efficiently models spatiotemporal change patterns through multi-dimensional temporal dependency modeling, showing strong responsiveness to long-term changes such as roof collapses or complete building destruction. The CNN branch excels at preserving boundary clarity, particularly for detecting edges of medium-scale damage in remote sensing images, maintaining fine-grained details. By combining the two, global–local joint optimization is achieved, and residual fusion ensures semantic consistency while avoiding distortion in global modeling. Moreover, Mamba's linear computational complexity and parallelizable structure allow the model to maintain high efficiency without sacrificing performance. The feature fusion process between CNN and Mamba is computed as follows:

$$F^{(l)} = F^{(l)}_{CNN} + \alpha \cdot F^{(l)}_{Mamba}, \tag{7}$$

where α is a learnable weighting coefficient used to balance the contributions of the CNN and Mamba features. The fused features retain fine spatial details while incorporating global modeling capabilities.

In CMSNet, the decoder restores spatial resolution in a top-down, multi-scale manner, integrating outputs from each encoder layer via skip connections. Corresponding SAM feature masks are also incorporated at each decoding stage, fusing with existing features to improve boundary mapping accuracy during mask reconstruction. The decoder employs a structure combining VSS modules with learnable transposed convolutions. This design not only restores spatial information from compressed features but also preserves long-range

dependencies and semantic consistency, making it especially suitable for reconstructing complex, edge-blurred damaged regions in remote sensing images. Each decoding unit performs progressive upsampling, aggregation, and reconstruction of change regions, which can be formally expressed as:

$$\hat{F}^{(l-1)} = D_l(TConv_l(V_l(\hat{F}^{(l)}))), \tag{8}$$

where V_l denotes the VSS module at layer l, $TConv_l$ represents the transposed convolution upsampling operation, and D_l refers to the skip-connection fusion and subsequent processing at that layer.

The VSS module in the decoding stage focuses on modeling structural consistency and semantic alignment across multi-scale change regions during feature reconstruction. Its processing flow involves unfolding the current feature map into sequences, applying state-space modeling to capture structural dependencies across pixels, and reconstructing the image. This module enhances the preservation of spatial structures during decoding, ensuring that building change regions remain coherent when upsampled, particularly for fractured edges or collapsed areas.

It is important to clarify that SAM is not employed as a feature extraction backbone or a temporal modeling module in our method. Instead, SAM is used only once to generate structural priors that provide coarse building masks and boundary cues. These priors are lightweight to obtain, do not introduce Transformer-based computational burdens during training or inference, and are fused into the CNN–Mamba feature space solely for structural enhancement. Therefore, the integration of SAM does not contradict the limitations of Transformer architectures discussed earlier; rather, it complements the proposed model by supplying high-quality structural guidance without affecting the overall efficiency and boundary modeling capability of the CNN–Mamba backbone.

4.4. Feature Enhancement Mechanisms of Foundation Model

This section focuses on the structural principles and workflow of the PVPF-FM module, a feature enhancement component in CMSNet built on top of foundation vision models. Serving as a bridge between a pretrained foundation model and the backbone change detection network, this module fully leverages the structural and boundary priors encoded in foundation models to improve the perception and alignment of building structures and damaged regions. Although the CNN-Mamba backbone network exhibits strong local-global feature fusion capabilities, it may still encounter challenges such as false positives, missed detections, or blurred boundaries when processing remote sensing images with dense buildings, complex change patterns, and indistinct edges. To address this, we introduce structural prior information from a universal foundation model such as SAM to guide the backbone network in focusing on critical regions and enhancing its structural awareness of building targets.

The core idea of the PVPF-FM module is to incorporate the structural mask features generated by SAM into the change detection process, thereby guiding the backbone network to enhance the structural consistency, boundary integrity, and semantic clarity of damaged regions during the mid-to-high-level feature representation stage. As shown in Figure 5, the module consists of four subcomponents: the mask embedding encoder, which transforms the structural masks into high-dimensional guidance features; the temporal guidance fusion module, which performs guided fusion separately for pre- and post-event features; the feature difference enhancement module, which strengthens responses in change regions by leveraging differences in SAM masks; and the feature fusion output module, which outputs structure-aware enhanced feature maps for further decoding. Specifically, F_{pre} and F_{post} denote the pre- and post-event image features extracted by the backbone encoder,

while M_{SAM}^{pre} and M_{SAM}^{post} represent the building structure masks generated by SAM for the corresponding images. The binary SAM mask $M_{SAM} \in \{0,1\}^{H \times W}$ is first passed through multiple convolutional layers to obtain a guidance feature map $E_{SAM} \in R^{H \times W \times C}$, which encodes spatial boundary and structural contour information of buildings.

$$E_{SAM} = \varnothing_{conv}(M_{SAM}), \tag{9}$$

where $\varnothing_{conv}$ denotes the convolutional feature transformation block, which consists of a standard convolution (Conv), batch normalization (BN), and a ReLU activation function.

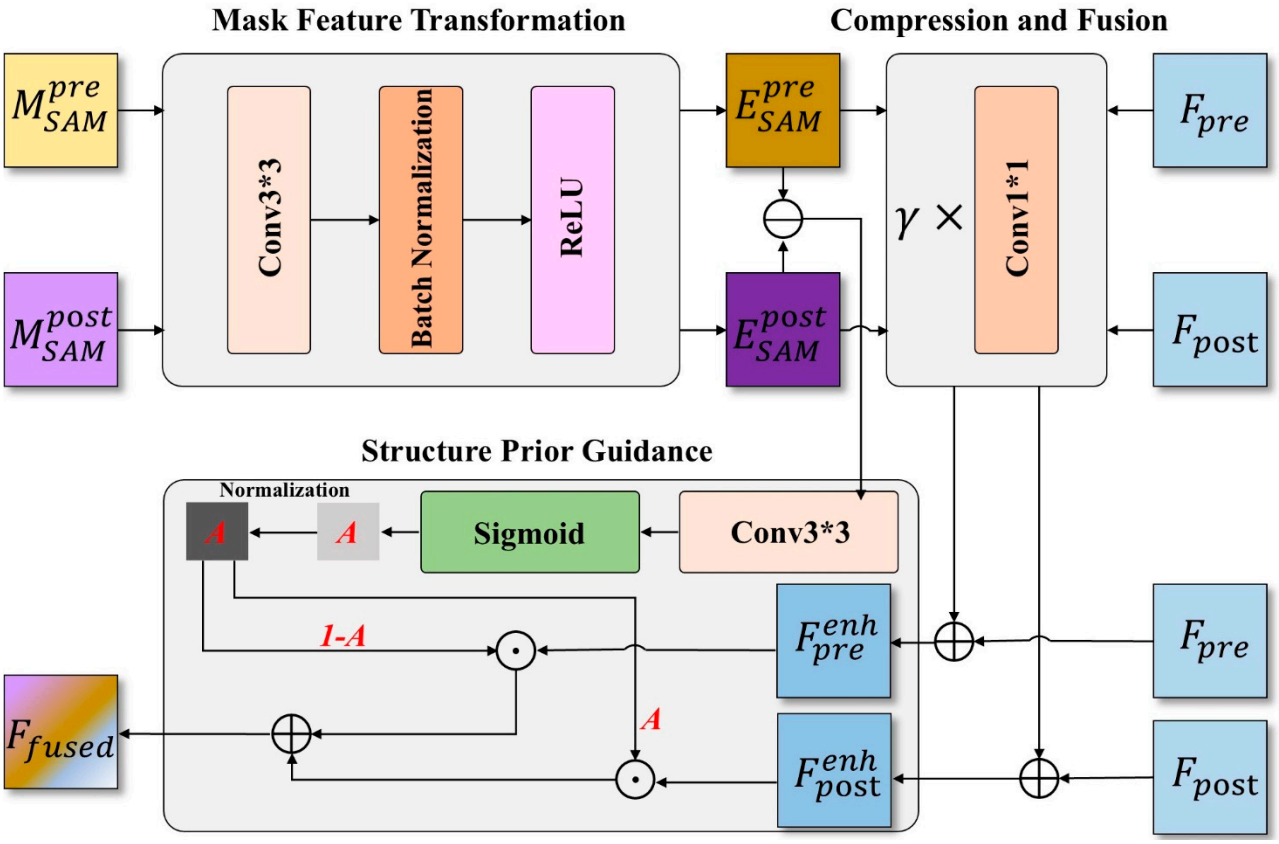


Figure 5. Schematic diagram of the principle of PVPF-FM.

Based on E_{SAM}, feature fusion is performed separately with F_{pre} and F_{post} to enable the features from the foundation visual model to be deeply integrated with the backbone features. In this work, a residual-enhanced fusion strategy is adopted, and the computation can be expressed as follows:

$$F_{pre}^{enh} = F_{pre} + \gamma \cdot \varphi\left(F_{pre}, E_{SAM}^{pre}\right), \tag{10}$$

$$F_{post}^{enh} = F_{post} + \gamma \cdot \varphi\left(F_{post}, E_{SAM}^{post}\right), \tag{11}$$

where φ denotes the operation of dimension reduction and attention-based fusion applied after feature concatenation using a 1×1 convolution, while γ represents a learnable coefficient.

After obtaining the features enhanced by the foundation visual model, the difference between the pre- and post-event structural masks is computed to guide the network's focus toward the damaged regions. The difference feature is calculated as follows:

$$\Delta M = \left| M_{SAM}^{post} - M_{SAM}^{pre} \right|, \tag{12}$$

Finally, the difference mask guidance map is embedded into the feature space and applied to the enhanced backbone features, thereby combining the "structural location prior" with "temporal semantic differences" to improve the model's ability to focus on key regions (damaged building areas) for change recognition. The feature embedding process is defined as follows:

$$F_{fused} = Attention_{fuse}\left(F_{pre}^{enh}, F_{post}^{enh}, \Delta M\right), \tag{13}$$

The output of PVPF-FM is a structure-aware enhanced bi-temporal fused feature, whose dimensionality is consistent with that of the backbone features and can be directly fed into the decoder for change region reconstruction. In this process, strong structural priors, guided by SAM mask information, direct the model to focus on building areas, thereby improving the recognition of small-scale damage. Meanwhile, boundary enhancement helps address issues of blurred boundaries and misclassification in change regions. In addition, PVPF-FM can be flexibly inserted as a module into feature layers of different scales, demonstrating good compatibility. Furthermore, knowledge transfer from the SAM foundation model enhances the model's stability under small- and medium-scale sample conditions.

4.5. Loss Function Configuration

To effectively improve the detection accuracy of building damage areas by CMSNet in real remote sensing war-damage scenarios, we design a multi-objective optimization strategy composed of three complementary loss functions, which jointly consider pixel-level classification accuracy, regional integrity, and structural prior consistency. In this strategy, the basic loss supervises the accuracy of mask prediction, while the structure-guided loss further enhances the model's focus on building regions. The overall loss function is defined as follows:

$$L = \mu_1 L_{BCE} + \mu_2 L_{Dice} + \mu_3 L_{SAM}, \tag{14}$$

where μ_1, μ_2, and μ_3 denote the weighting coefficients of the respective loss functions. In this work, we set $\mu_1 = 1$, $\mu_2 = 1$ and $\mu_3 = 0.5$.

Among them, the BCE loss is the most commonly used pixel-wise binary classification loss function, which measures the per-pixel discrepancy between the predicted mask and the ground truth damage labels, and is defined as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)], \tag{15}$$

where $\hat{y}_i$ represents the predicted damage probability for the i-th pixel, y_i is the corresponding ground truth label, and N denotes the total number of pixels.

Dice serving as a complementary region-level supervision, helps improve the detection of small targets and sparse change areas. It emphasizes the overlap between the predicted region and the ground truth, making it suitable for evaluating segmentation quality. It is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_i \hat{y}_i y_i + \delta}{\sum_i \hat{y}_i + \sum_i y_i + \delta}, \tag{16}$$

where δ denotes a small constant term.

Considering that CMSNet incorporates structural masks provided by SAM as auxiliary supervision, we design a Structure-Guided Loss to direct the model's attention toward actual building regions, thereby improving the spatial plausibility and structural alignment of the predicted masks. It is defined as follows:

$$L_{SAM} = BCE(\hat{y}, M_{SAM}), \tag{17}$$

where M_{SAM} denotes the building region mask extracted by SAM, serving as a weak supervisory signal during training, particularly useful in cases of incomplete annotations or severe occlusions. This loss encourages the model to focus on the main building areas during change detection, thereby reducing the risk of false positives in the background.

5. Experiments and Analysis

5.1. Dataset Details

To comprehensively evaluate the performance of CMSNet in multi-scenario building damage detection tasks, experiments were conducted on four representative datasets, covering various types such as real-world war zones and typical urban changes. These four datasets are RWSBD, CWBD, WHU-CD, and LEVIR-CD+, with the following specific details:

RWSBD is a high-resolution remote sensing building damage change detection dataset independently constructed in this study, specifically designed for real-world war scenarios, with strong practical applicability and research value. The dataset is based on two GF-2 satellite optical images with a spatial resolution of 0.8 m, acquired on 24 September 2022, and 3 May 2024, respectively, covering the entire Gaza Strip. Building damage regions in the bi-temporal images were manually annotated, resulting in a total of 42,723 damaged building instances. During preprocessing, all images and labels were cropped into 256 $\times$ 256-pixel patches, forming 28,210 paired image samples along with their corresponding mask labels.

The CWBD dataset [6] focuses on building damage scenarios caused by war, explosions, and conflicts. Its remote sensing images are sourced from multiple times, locations, and sensors, featuring a higher spatial resolution of 0.3 m, which allows for capturing finer damage details. The CWBD dataset contains a total of 4065 pairs of pre- and post-event image samples, each consisting of a 256 $\times$ 256-pixel bi-temporal image pair and the corresponding damage mask. The dataset exhibits significant diversity in damaged building regions, posing greater challenges to the accuracy and robustness of change detection algorithms.

WHU-CD [47] is a high-quality remote sensing dataset focused on post-earthquake building reconstruction and change detection. The dataset covers an area affected by a 6.3-magnitude earthquake in February 2011, along with the reconstruction progress from 2012 to 2016. It contains high-resolution aerial imagery with significant building changes. After cropping, the images are divided into 256 $\times$ 256-pixel patches, with a spatial resolution of approximately 0.2 m. WHU-CD provides high-quality bi-temporal remote sensing images and precise change annotations, making it widely used for developing and evaluating various change detection methods, particularly suitable for binary building change detection tasks.

The LEVIR-CD+ dataset [48] is specifically designed for urban-scale remote sensing building change detection tasks. It covers multiple urban scenarios and includes various types of building changes, such as expansions, new constructions, and demolitions. The dataset contains 985 pairs of high-resolution image samples, each with a size of 1024 $\times$ 1024 pixels and a spatial resolution of 0.5 m. Each sample pair includes a pre-change image, a post-change image, and a change annotation map. Overall, the dataset encompasses approximately 80,000 building change instances, offering high sample density and diversity. With improvements in image resolution, change density, and annotation precision, LEVIR-CD+ has become one of the widely used benchmark datasets in the field of remote sensing building change detection.

5.2. Experimental Implementations

All experiments in this study were conducted on a high-performance server equipped with an NVIDIA GeForce RTX 3090 GPU (NVIDIA Corporation, Santa Clara, CA, USA) with 24 GB of memory, running Ubuntu 20.04. The deep learning framework used was PyTorch (version 12.2). During model training, the batch size was set to 6, with a total of 100 training epochs. The Adam optimizer was employed, initialized with a learning rate of 1×10^{-4} and dynamically adjusted based on validation performance. The loss function followed the combined strategy defined in Section 4.5 (BCE, Dice, and SAM-guided loss) to enhance pixel-wise classification accuracy, region integrity, and structural consistency. All input images were cropped to 256×256 pixels, and standard data augmentation techniques, including random flipping, slight rotation, and color perturbation, were applied during training to improve model generalization and robustness. This consistent experimental setup provided a reliable foundation for the stable training of CMSNet and for performance comparisons across multiple remote sensing change detection datasets.

5.3. Evaluation Methods and Metrics

To comprehensively evaluate the performance of CMSNet in detecting building damage changes from remote sensing images, this study selected eight representative mainstream change detection methods as baselines and employed four commonly used quantitative metrics for performance comparison and analysis. These baseline methods span different network architecture paradigms, including CNN, Transformer, and Mamba, as well as diverse encoding strategies, allowing a thorough assessment of CMSNet's effectiveness and robustness from multiple perspectives. Specifically, the eight methods cover mainstream CNN architectures, Siamese network structures, Transformer-based models, and the latest Mamba state-space modeling paradigm, enabling comparison of CMSNet's change modeling capabilities across different modeling levels. The details are described as follows:

ChangeMamba [38]: A change detection method based on spatiotemporal state-space modeling, which uses the Mamba structure to model pre- and post-event images and capture long-range dependencies. It is one of the recently proposed efficient change detection models.

RS-Mamba [49]: An improved Mamba network optimized for remote sensing scenarios, enhancing feature extraction capability for high-resolution images and providing better spatial modeling performance.

ChangeFormer [33]: A typical Transformer-based change detection method that introduces cross-temporal feature interaction and global modeling capability, suitable for detecting complex structures and long-range changes.

SNUNet: A symmetric multi-scale nested U-Net network that fuses semantic features across scales to improve change region detection accuracy, demonstrating strong local detail modeling capability.

IFN [49]: A CNN-based network incorporating pre- and post-event interaction mechanisms, enhancing the model's responsiveness to change features through explicit information exchange.

FC-Siam-diff [50]: Similar to FC-Siam-conc but models feature differences explicitly, emphasizing changes between pre- and post-event features and focusing on fine-grained differences.

FC-Siam-conc [50]: A classical fully convolutional Siamese network that fuses pre- and post-event features through concatenation, directly modeling differences between the two images.

FC-EF [50]: An early-fusion change detection network that merges pre- and post-event images at the input stage, focusing on extracting overall temporal difference features.

To objectively evaluate the performance of these methods in building damage change detection, this study employs four standard metrics: F1-score (F1), Intersection over Union (IoU), Precision (P), and Recall (R). All metrics are computed based on pixel-level predictions against ground truth labels. The calculation formulas are as follows:

$$Recall = \frac{TP}{TP + FN}, \tag{18}$$

$$Precision = \frac{TP}{TP + FP}, \tag{19}$$

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

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{21}$$

where TP represents true positives, FP represents false positives, and FN represents false negatives. P indicates the proportion of predicted change regions that are actually true changes; a high precision means fewer false alarms and misclassified areas. R reflects the proportion of actual change regions successfully detected by the model; a high recall indicates fewer missed detections and more comprehensive identification of true changes. IoU measures the overlap between predicted and ground-truth change regions, with higher values indicating closer alignment between the predicted and true masks. F1-score is the harmonic mean of precision and recall, representing a balanced evaluation of the model's overall performance; higher F1 values indicate better trade-off between precision and recall.

5.4. Benchmark Comparison

In this section, we conduct a systematic comparison between the proposed method and several baseline models on the RWSBD, CWBD, WHU-CD, and LEVIR-CD+ datasets. To provide an intuitive understanding of the prediction performance, Figures 6–9 present visualization results of representative examples, with three typical instances selected from each dataset for a clear comparison of predictions across different methods. Meanwhile, Tables 1–4 summarize the quantitative evaluation results, covering P, R, F1, and IoU. These results together enable a comprehensive analysis of the strengths and weaknesses of different approaches across diverse datasets.

Before presenting the comparative results, we briefly summarize the key characteristics of the eight state-of-the-art methods evaluated in this study. FC-EF, FC-Siam-conc, and FC-Siam-diff are classical CNN-based Siamese networks, focusing on local feature extraction and multi-scale convolutional representation. IFN and SNUNet are advanced CNN-based architectures with enhanced feature fusion and attention mechanisms, aiming to improve change detection accuracy in complex scenarios. ChangeFormer is a Transformer-based model that leverages global attention to capture long-range dependencies and contextual information. RS-Mamba and ChangeMamba are Mamba-based methods that employ linear state-space modeling to efficiently capture temporal dependencies and structural changes in bi-temporal images. These methods cover a wide range of paradigms, from purely CNN-based to Transformer- and Mamba-based architectures, providing a comprehensive benchmark for evaluating the performance and generalization capabilities of CMSNet.

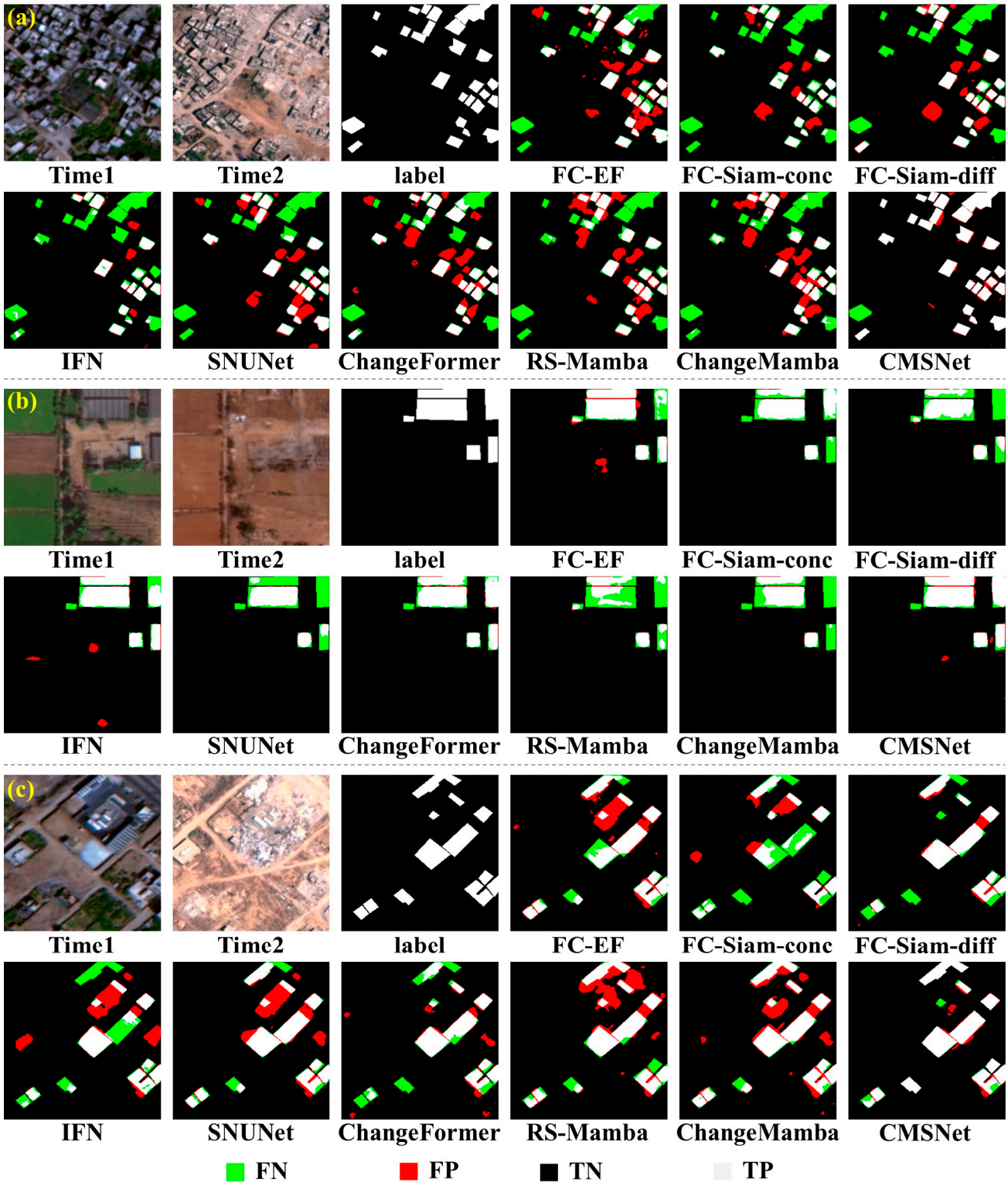


Figure 6. The general comparison of all models on the RWSBD dataset.

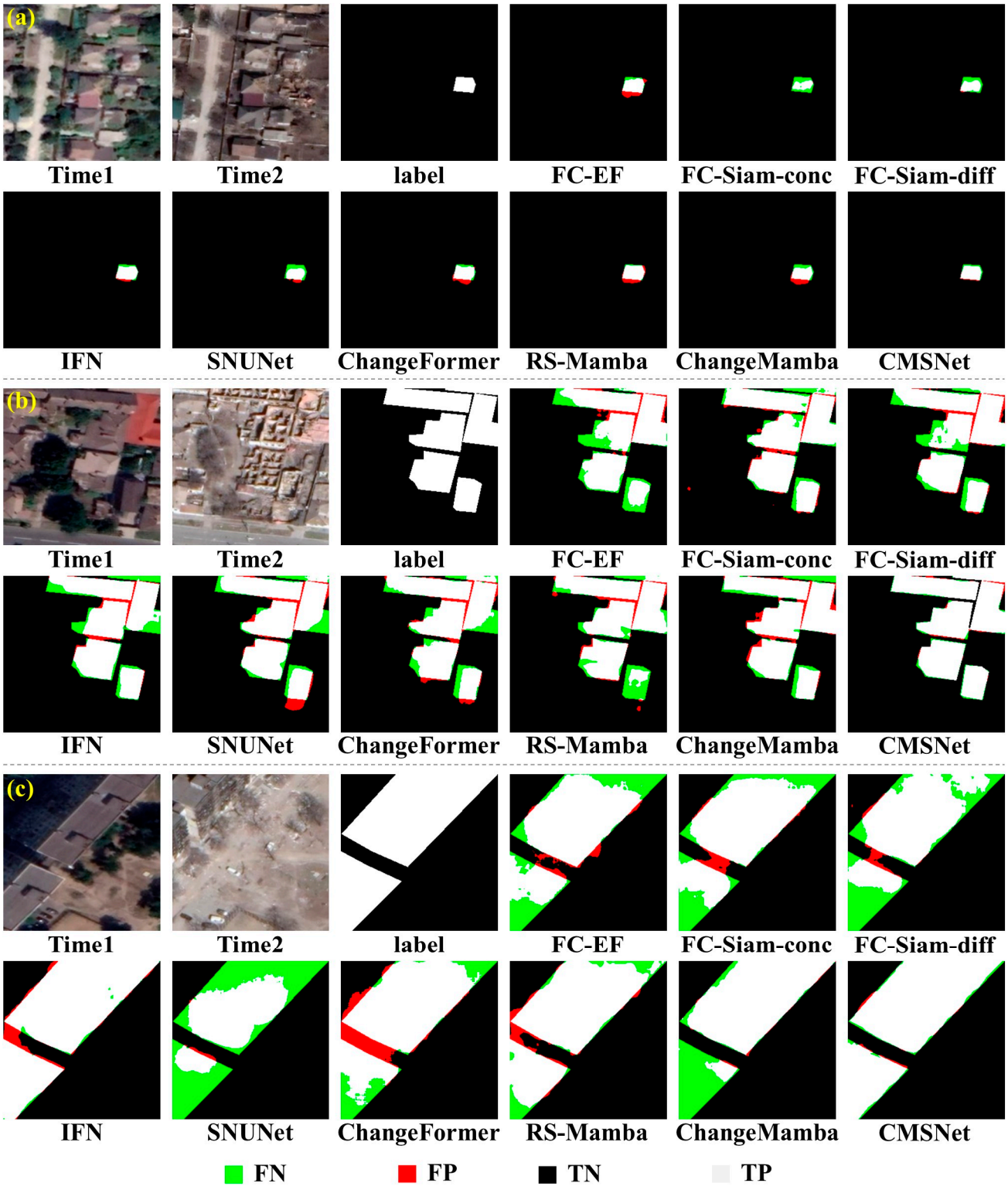


Figure 7. The general comparison of all models on the CWBD dataset.

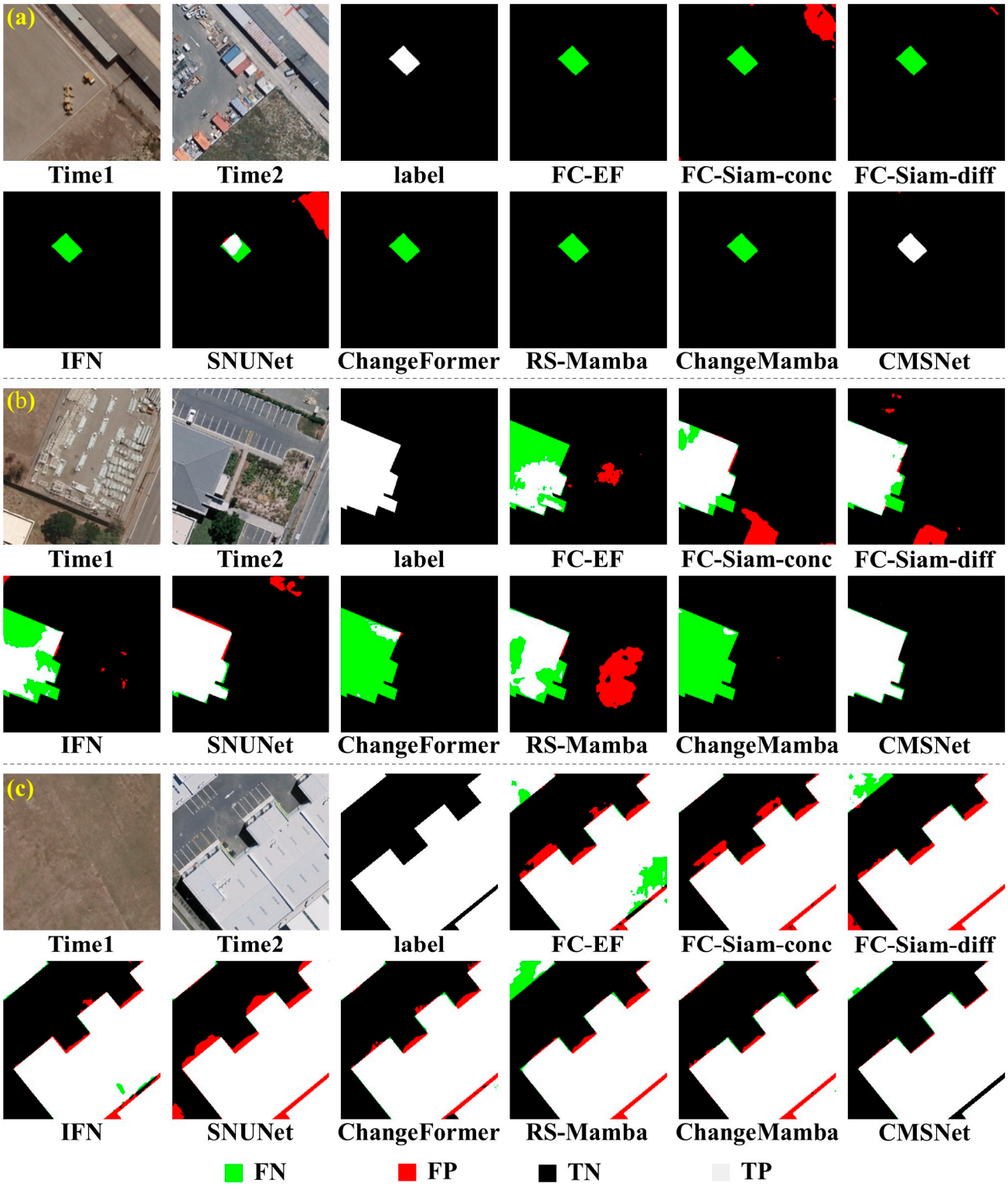


Figure 8. The general comparison of all models on the WHU-CD dataset.

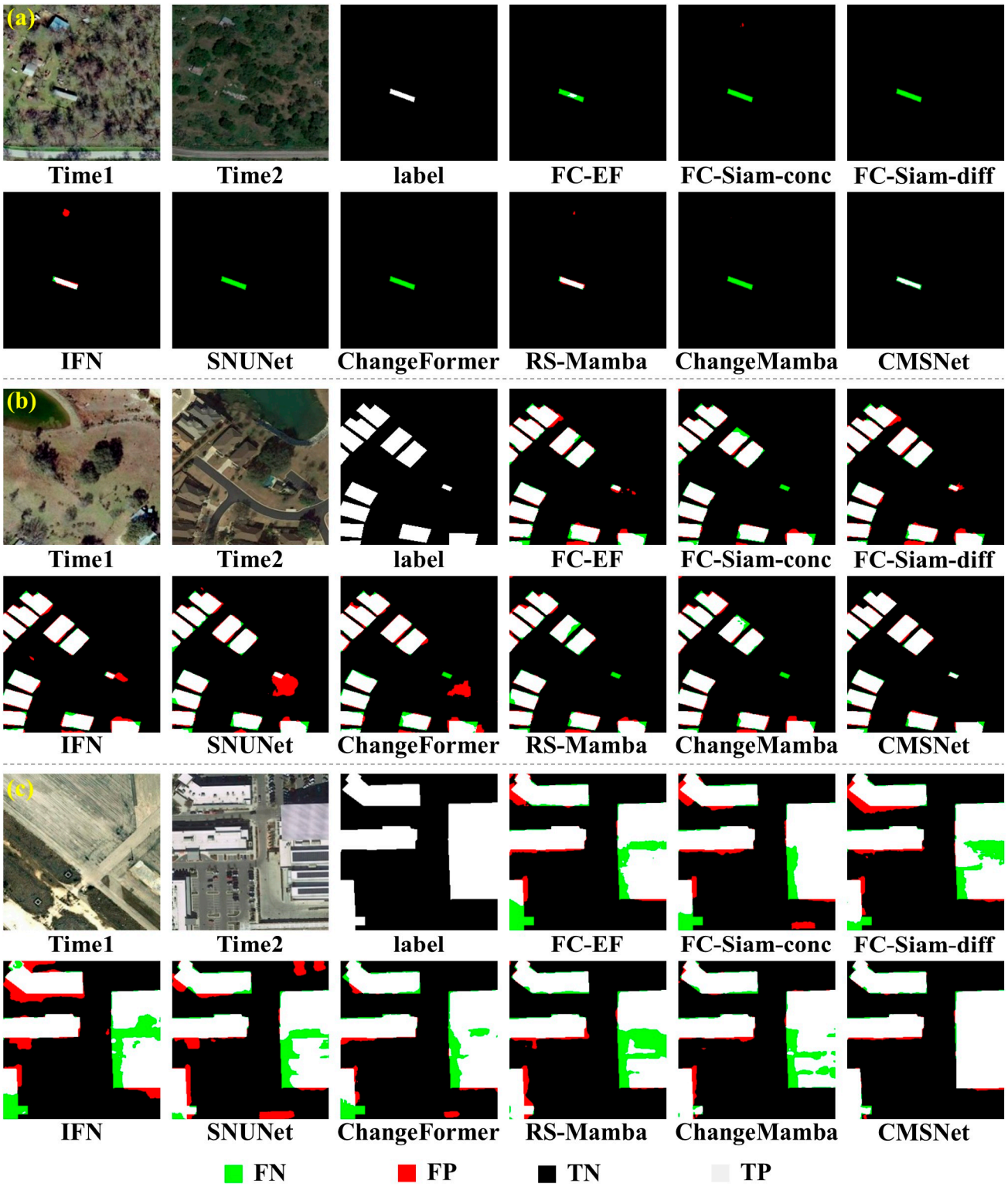


Figure 9. The general comparison of all models on the LEVIR-CD+ dataset.

Table 1. Comparison of accuracy of all models on the RWSBD dataset (%).

Methods	P	R	F1	IoU
FC-EF	78.58	82.68	80.59	70.87
FC-Siam-conc	83.60	77.62	80.49	70.81
FC-Siam-diff	83.19	77.87	80.45	70.78
IFN	80.98	81.03	81.01	71.54

Table 1. *Cont.*

Methods	P	R	F1	IoU
SNUNet	80.39	80.24	80.31	70.77
ChangeFormer	81.35	80.77	81.06	71.60
RS-Mamba	79.80	81.20	80.49	70.93
ChangeMamba	80.91	80.51	80.71	71.21
CMSNet	83.95	81.28	82.60	73.13

Table 2. Comparison of accuracy of all models on the CWBD dataset (%).

Methods	P	R	F1	IoU
FC-EF	84.64	82.83	83.72	74.08
FC-Siam-conc	85.55	84.52	85.03	75.78
FC-Siam-diff	83.92	86.22	85.00	75.71
IFN	84.86	83.85	84.35	74.89
SNUNet	85.26	82.26	83.75	74.06
ChangeFormer	87.23	84.12	85.65	76.55
RS-Mamba	88.31	86.23	87.26	78.77
ChangeMamba	86.60	85.76	86.17	77.30
CMSNet	87.04	88.06	87.54	79.04

Table 3. Comparison of accuracy of all models on the WHU-CD dataset (%).

Methods	P	R	F1	IoU
FC-EF	95.80	94.19	94.99	90.62
FC-Siam-conc	88.91	93.79	91.28	84.12
FC-Siam-diff	92.79	94.64	93.80	88.39
IFN	96.65	95.22	95.93	92.28
SNUNet	89.74	93.47	91.59	84.72
ChangeFormer	94.59	93.22	88.76	93.90
RS-Mamba	96.02	94.90	95.45	91.45
ChangeMamba	95.92	95.51	95.71	91.92
CMSNet	95.75	97.27	96.50	90.17

Table 4. Comparison of accuracy of all models on the LEVIR-CD+ dataset (%).

Methods	P	R	F1	IoU
FC-EF	94.02	93.45	93.74	88.61
FC-Siam-conc	94.86	93.99	89.76	94.43
FC-Siam-diff	95.00	93.72	94.36	89.64
IFN	94.22	94.11	94.17	89.33
SNUNet	92.78	92.74	92.76	87.02
ChangeFormer	94.97	93.35	94.15	89.28
RS-Mamba	95.17	93.78	94.47	89.83
ChangeMamba	95.18	94.10	94.64	90.12
CMSNet	94.65	94.79	94.72	90.27

From the prediction results shown in Figures 6–9, it is evident that the detection accuracy and robustness of different models vary significantly across the four datasets. The earlier FC-series models (FC-EF, FC-Siam-conc, FC-Siam-diff) often suffer from blurred boundaries and misdetections in complex scenarios, performing poorly especially when building structures are intricate or exhibit large scale variations. In contrast, Transformer- and Mamba-based models (e.g., ChangeFormer, RS-Mamba, ChangeMamba) achieve superior performance in maintaining boundary continuity and suppressing false changes, with

predictions that are closer to the ground truth. Notably, CMSNet produces more complete change regions with sharper boundaries across multiple datasets, demonstrating strong cross-scene generalization capability. In the sample visualizations, CMSNet and ChangeMamba deliver the most outstanding results: both are able to capture small-scale targets while effectively distinguishing adjacent building clusters, thereby reducing missed detections and over-merged predictions. RS-Mamba also shows relatively stable performance, though slight boundary breaks are still observed in certain areas. By comparison, SNUNet generates coarser predictions, struggling to delineate fine structural details. These results indicate that network architectures incorporating sequence modeling and long-range dependency mechanisms achieve stronger representational power. Such enhancements also improve the models' adaptability to complex change scenarios.

From the quantitative results shown in Tables 1–4, it can be observed that the traditional FC-series models generally lag behind the more recent approaches across all four datasets. For example, on the RWSBD dataset, the F1 scores of FC-EF, FC-Siam-conc, and FC-Siam-diff remain around 80%, while IFN, ChangeFormer, and ChangeMamba achieve a better balance between IoU and F1. Notably, CMSNet achieves an F1 score of 82.60% and an IoU of 73.13% on RWSBD, ranking as the best among all models. On the CWBD dataset, RS-Mamba and CMSNet demonstrate particularly outstanding performance, with F1 scores of 87.26% and 87.54% and IoU values of 78.77% and 79.04%, respectively, clearly surpassing other models. On the large-scale WHU-CD and LEVIR-CD+ datasets, the performance gap becomes even more pronounced. Taking WHU-CD as an example, CMSNet achieves an F1 score of 96.50%, significantly outperforming other models and highlighting its robustness and accuracy in large-scale scenarios; ChangeMamba also delivers nearly optimal performance, with an F1 score of 95.71%. On LEVIR-CD+, both CMSNet and ChangeMamba remain leading, with F1 scores of 94.72% and 94.64%, respectively, and IoU values exceeding 90%. Overall, the Mamba series and CMSNet consistently achieve stable and superior performance across different datasets, demonstrating their strong generalization ability and application potential in building change detection.

From the experimental results, it can be observed that CMSNet exhibits higher sensitivity to damaged buildings and can effectively distinguish between destroyed and intact areas, which is particularly evident in complex backgrounds and multi-scale scenarios. The performance gains mainly stem from the network architecture and the synergistic effect of its key components. On one hand, the SAM highlights salient features of damaged regions and suppresses background interference, thereby enhancing the model's local discriminative power. On the other hand, the PVPF-FM module enables deep interaction with backbone semantic features, effectively integrating multi-level information and improving the delineation of building boundaries and subtle damage details. In addition, the overall network design ensures sufficient multi-scale feature representation, allowing CMSNet to adapt well to both small-scale damage instances and large-scale destruction scenarios. Collectively, the modules complement each other at different levels, jointly enhancing the model's sensitivity and discriminative capacity for damaged buildings, thus providing strong support for final detection performance. At the architectural level, CMSNet adopts a hierarchical feature extraction and fusion strategy, balancing local and global information. This not only guarantees accurate detection of small-scale damage but also maintains high adaptability to widespread destruction. Meanwhile, the structure is designed to balance efficiency and representational power, enabling high-quality modeling of complex disaster scenarios under relatively low computational complexity. In summary, the modular architecture and hierarchical design of CMSNet act in concert to establish its significant advantages in building damage detection tasks.

A quantitative analysis of relative performance improvements highlights the advantages of CMSNet over existing methods. On the RWSBD dataset, CMSNet achieves an F1 score of 82.60%, outperforming the best baseline (ChangeMamba, 80.71%) by 1.89%, and an IoU of 73.13%, which is 1.92% higher than the best baseline (ChangeMamba, 71.21%). On the CWBD dataset, CMSNet improves the F1 score to 87.54%, surpassing the best baseline (RS-Mamba, 87.26%) by 0.28%, and the IoU to 79.04%, exceeding the best baseline (RS-Mamba, 78.77%) by 0.27%. For the WHU-CD dataset, CMSNet achieves the highest F1 score of 96.50%, which is 0.79% higher than the best baseline (IFN, 95.71%), although its IoU is slightly lower than the top baseline (IFN, 92.28%). On the LEVIR-CD+ dataset, CMSNet attains an F1 of 94.72%, surpassing the best baseline (ChangeMamba, 94.64%) by 0.08%, and an IoU of 90.27%, slightly exceeding the best baseline (ChangeMamba, 90.12%) by 0.15%. These results clearly demonstrate that CMSNet consistently improves both F1 and IoU across multiple datasets, confirming its effectiveness in capturing fine-grained building damage and maintaining boundary consistency, as well as its generalization ability across diverse war-zone and urban scenarios.

5.5. Ablation Study

To further verify the effectiveness of the PVPF-FM module in enhancing the overall performance of the model, ablation experiments were conducted in this section. To ensure a comprehensive and fair evaluation of the proposed network design, all ablation models were constructed based on the full CMSNet architecture and were trained and tested on all four datasets used in this study, including RWSBD, CWBD, WHU-CD, and LEVIR-CD+. Since the PVPF-FM module functions in both the encoder and decoder stages of the network, we designed three ablation settings: removing the PVPF-FM module from the encoder (del-e), removing it from the decoder (del-d), and removing it from both the encoder and decoder (del-e&d). These three settings correspond to three different ablation models. To ensure fairness, all models were retrained under the same configurations and datasets as the baseline experiments, followed by a comprehensive comparative evaluation.

From the visual results shown in Figure 10, it can be observed that removing the PVPF-FM module significantly weakens the model's performance in boundary delineation and fine-grained region detection. Specifically, the del-e model, which removes the encoder interaction branch, tends to produce incomplete target regions in complex backgrounds, while the del-d model, which removes the decoder interaction branch, yields relatively blurry predictions along boundaries. The del-e&d model, with both branches removed, performs the worst, exhibiting a noticeable increase in missed detections and false alarms in damaged regions. In contrast, CMSNet with the complete PVPF-FM module achieves more accurate and consistent results across all regions.

The bar charts in Figure 11 quantitatively validate these observations. Across all four evaluation metrics (Precision, Recall, F1, and IoU), removing the PVPF-FM module leads to performance degradation. While del-e and del-d show moderate declines, del-e&d results in a substantial drop, particularly in IoU and F1. This pattern reflects the varying roles of encoder and decoder interactions: encoder-side interaction establishes cross-scale semantic associations at shallow layers, strongly influencing global feature coherence and completeness; decoder-side interaction mainly contributes to detail refinement and structural recovery. Removing both interactions simultaneously impairs global modeling and local detail enhancement, resulting in the poorest performance. Overall, these results highlight the critical importance of dual-stage interaction in CMSNet and demonstrate the essential contribution of the PVPF-FM module to network robustness and detection accuracy.

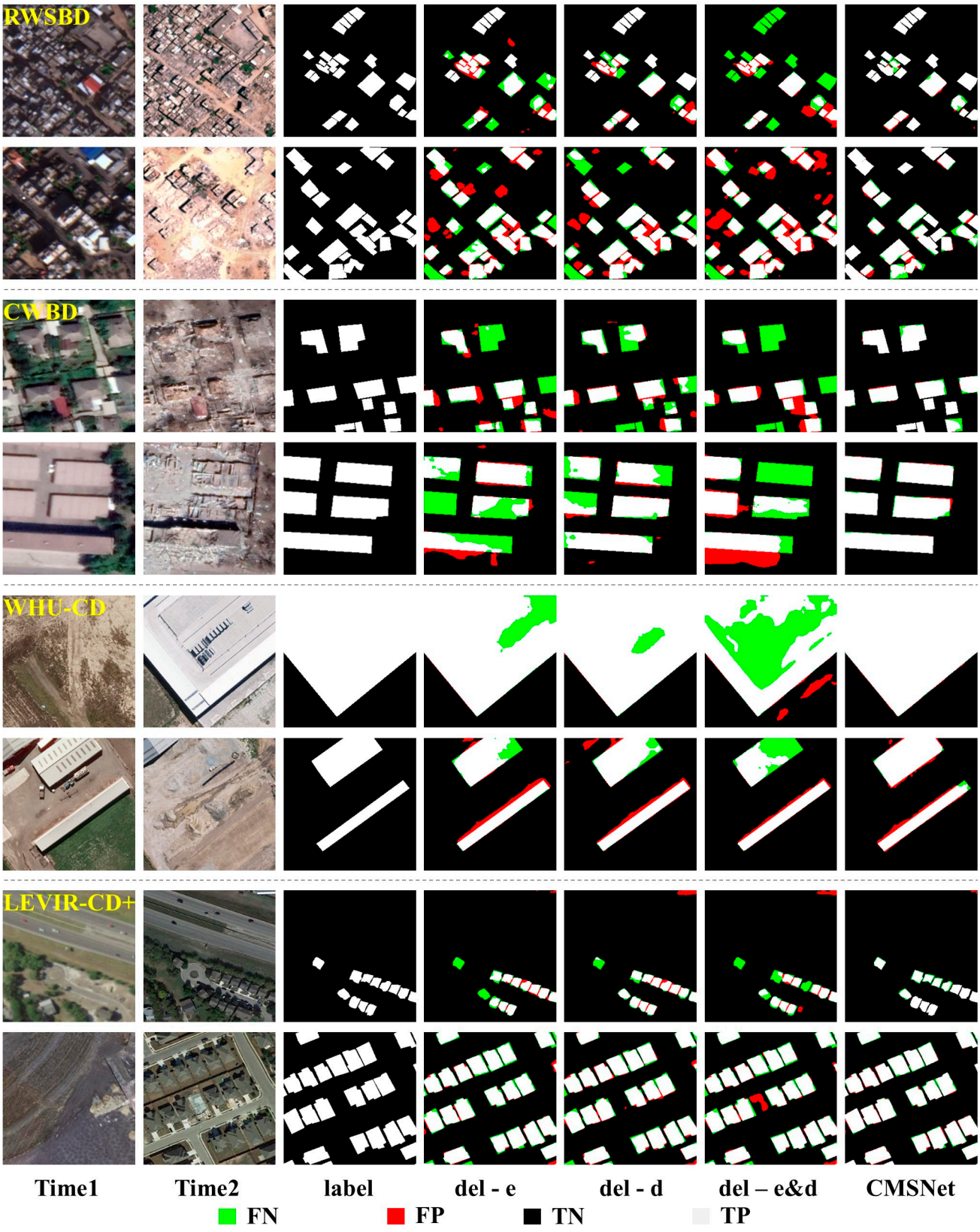


Figure 10. Qualitative Analysis after Ablation of the PVPF-FM Module.

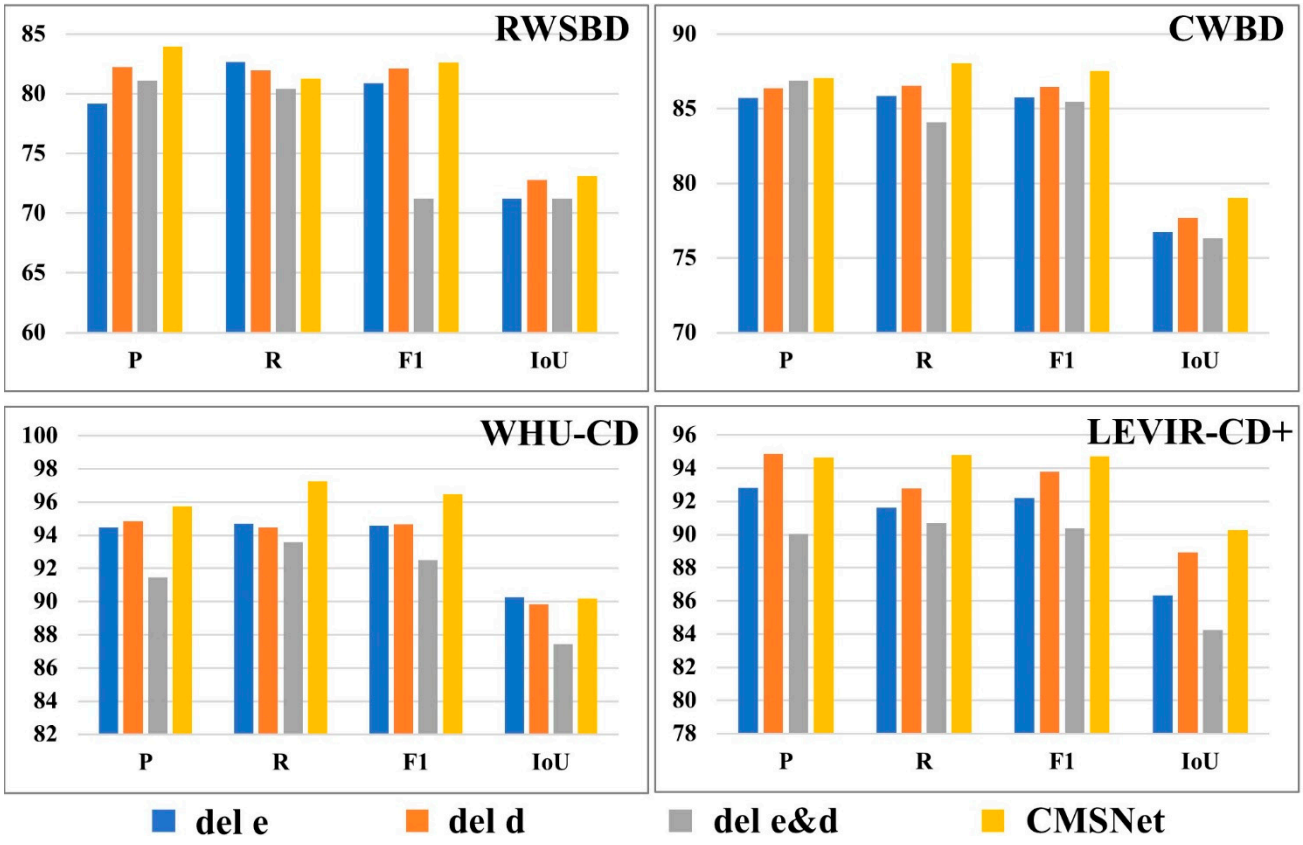


Figure 11. Accuracy analysis of ablation experiments.

6. Discussion

6.1. Module Ablation Analysis

To further validate the effectiveness of the proposed network's key components and loss function design, a series of comparative experiments were conducted on the basis of the complete model. These experiments included: (1) degrading the network into a standard Siamese encoder–decoder structure by completely removing the SAM module; (2) three ablation variants of the PVPF-FM module; and (3) three ablation variants of the loss function, where BCE loss, Dice loss, and SAM loss were individually removed. Together with the complete model, a total of seven heatmap results were generated. To provide an intuitive illustration of how different designs affect feature learning and change-region representation, three categories of comparative heatmaps were produced: Figure 12 presents the functional analysis of SAM based on feature maps, Figure 13 shows the functional analysis of PVPF-FM, and Figure 14 illustrates the functional analysis of the loss functions.

No dataset was excluded. The quantitative comparison results reported in Table 5, as well as the visual analyses in Figures 12–14, are therefore derived from a complete set of experiments conducted across all datasets. This setting guarantees that the impact of each module and loss function is consistently validated under diverse imaging conditions and scene complexities.

As shown in Figure 12, when the SAM foundation model is completely removed, the discriminability of the feature maps decreases significantly. This is especially evident in complex backgrounds and boundary transition regions, where the representation of change targets is notably weakened, manifesting as blurred heatmap responses and partial loss of local targets. This occurs because SAM provides global perception and semantic constraints during feature alignment and semantic modeling, substantially enhancing

the representation of change features. Without SAM, the model degenerates into purely local alignment and convolutional modeling, failing to fully exploit the complementary information in the images, which results in suboptimal heatmap performance in boundary details and challenging regions. These observations confirm the critical role of SAM within the overall network architecture.

Table 5. Comparison of accuracy of all ablation models on the four datasets (F1/%).

Methods	RWSBD	CWBD	WHU-CD	LEVIR-CD+
del-SAM	70.12	70.12	70.12	70.12
del-e	80.88	85.77	94.59	92.21
del-d	82.1	86.46	94.66	93.81
del-e&d	71.24	85.45	92.51	90.38
del-BCE	81.45	81.45	81.45	81.45
del-Dice	82.25	82.25	82.25	82.25
del-SAM-f	82.40	82.40	82.40	82.40
CMSNet	82.60	87.54	96.50	94.72

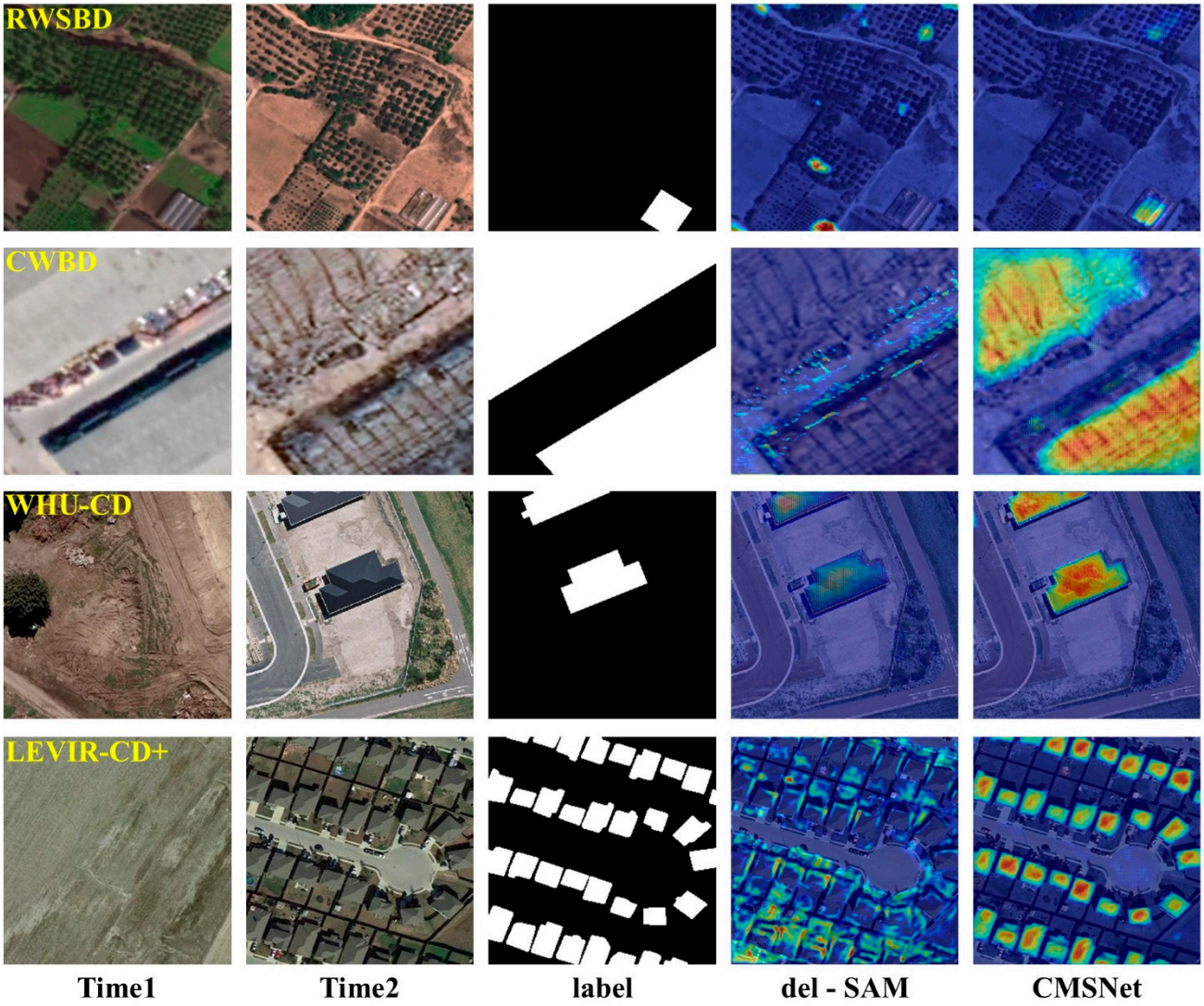


Figure 12. Functional Analysis of SAM Based on Feature Maps.

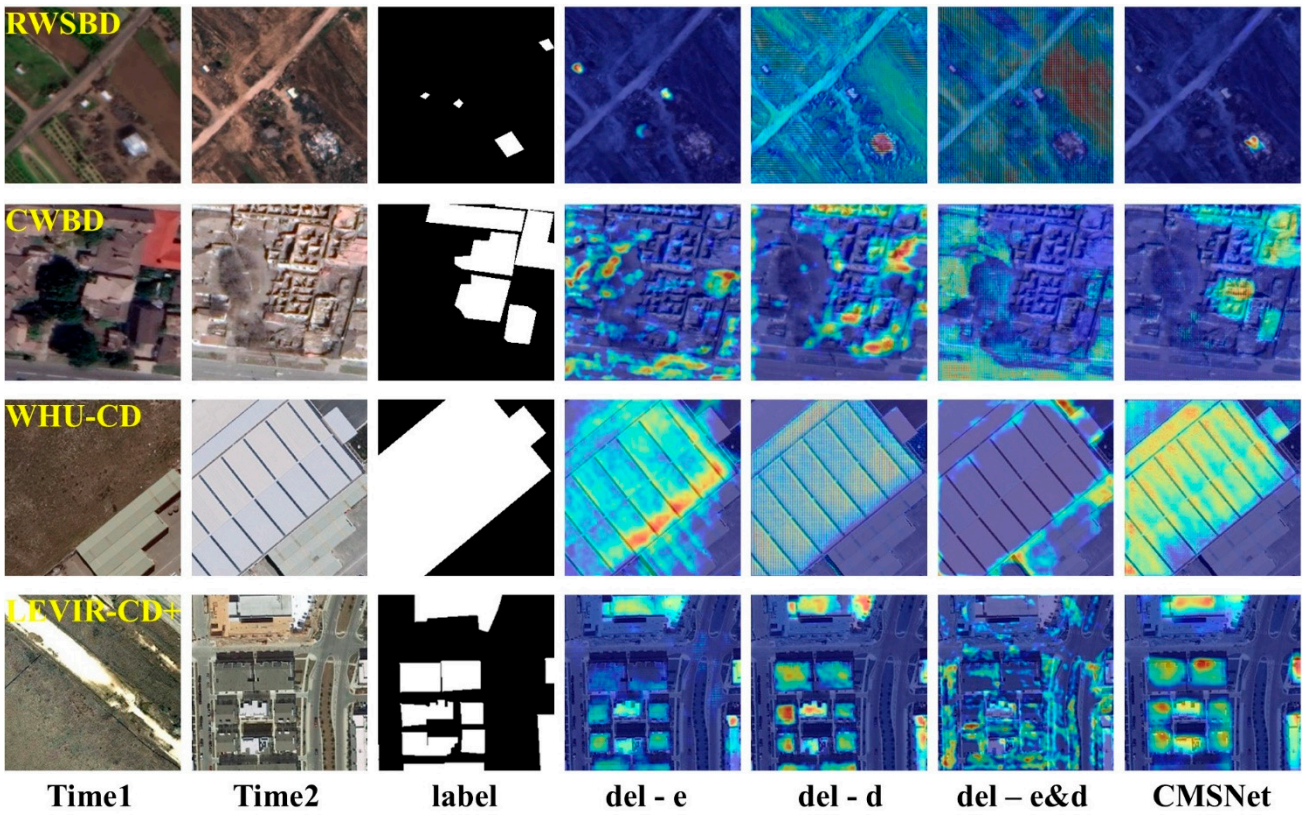


Figure 13. Functional Analysis of PVPF-FM Based on Feature Maps.

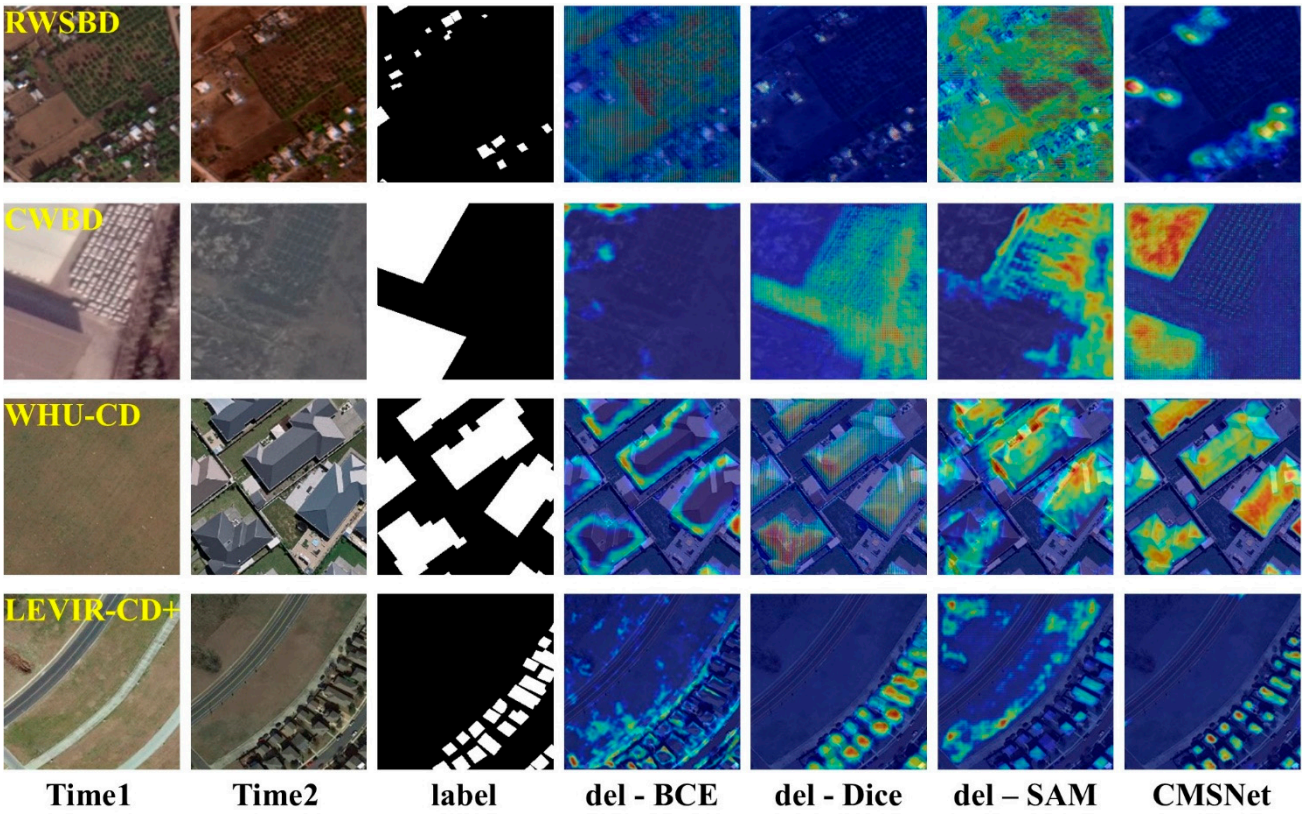


Figure 14. Functional Analysis of loss function Based on Feature Maps.

Figure 13 presents the comparison results of the three ablation experiments on the PVPF-FM module. Among them, the del-d model is able to retain the main change regions

effectively, with relatively clear boundaries, indicating that the differential feature branch plays a significant role in enhancing cross-temporal change perception. The del-e model maintains a certain level of discriminability in complex backgrounds but shows deficiencies in capturing small-scale details. In contrast, the del-e&d model performs the worst, with scattered heatmap responses and blurred target boundaries, demonstrating that the dual-branch collaborative modeling is crucial for suppressing background interference and strengthening discriminative features. These results indicate that the PVPF-FM module not only improves the overall feature discriminability but also enables finer-grained change perception at the detail level.

Figure 14 compares the performance of three models with the BCE, Dice, and SAM losses removed, respectively. The results show that when the BCE loss is omitted, the model's overall classification capability declines, leading to missed detections in some change regions. Removing the Dice loss results in insufficient representation of small targets and fine boundaries, manifesting as blurred target edges. When the SAM loss is removed, the feature alignment ability significantly decreases, causing an increase in false detections in complex backgrounds. Among these, the omission of the BCE loss leads to the most severe performance degradation, as the BCE loss plays a crucial role in supervising the model during training. In contrast, the complete loss function provides complementary benefits in classification, boundary refinement, and feature alignment, ensuring strong performance in both global and local feature modeling, thereby validating the rationality and necessity of the multi-loss joint optimization strategy.

The quantitative results reported in Table 5 are consistent with the visual observations in Figures 12–14 and further validate the effectiveness of the key components in CMSNet. When the SAM module is removed, the del-SAM exhibits the most severe performance degradation across all datasets, with F1 scores dropping by more than 5% compared with the complete model. This decline aligns with the blurred boundaries and weakened structural representations observed in the heatmaps, confirming the crucial role of SAM-derived structural priors in enhancing global perception and stabilizing boundary responses. For the PVPF-FM module, the three ablation variants present a clear accuracy hierarchy: del-d > del-e > del-e&d, which matches their respective feature-map behaviors. Among them, del-d achieves moderate performance due to retaining differential cues, whereas del-e struggles with small-scale details, and del-e&d performs the worst due to the absence of both embedding and differential branches. These results quantitatively demonstrate the complementary benefits of the dual-branch fusion strategy.

Similarly, the three loss-function ablations show consistent trends with the qualitative analyses. Removing BCE loss leads to the largest F1 decline among the loss variants, highlighting its essential role in maintaining classification stability. The del-Dice model shows weakened boundary and small-object performance, while the del-SAM-f variant exhibits increased false alarms due to reduced alignment capability. Across all datasets, each ablation results in a measurable performance drop, whereas CMSNet achieves the highest F1 scores by jointly optimizing BCE, Dice, and SAM losses. These findings confirm that the complete loss combination provides complementary supervision for global classification, fine-grained localization, and structural alignment, thereby ensuring superior detection robustness.

In addition, the PVPF-FM module, which aligns and fuses structural priors with bi-temporal image features, is potentially transferable to other change detection tasks beyond war-damage scenarios. Its dual-branch feature enhancement and prior-guided fusion strategy can be adapted to various applications where structural consistency and fine-grained boundary preservation are critical, such as urban expansion monitoring, disaster assessment, or environmental change analysis.

6.2. Analysis of Weight Coefficients in Loss Functions

To further evaluate the impact of different loss function weights, we conducted ablation experiments by adjusting the coefficients of BCE (μ_1), Dice (μ_2), and SAM (μ_3) losses. Each configuration was used to retrain CMSNet on all four datasets, and the resulting F1 scores are summarized in Table 6.

Table 6. Quantitative evaluation of different loss function coefficients (F1/%).

$[\mu_1, \mu_2, \mu_3]$	RWSBD	CWBD	WHU-CD	LEVIR-CD+
[1, 1, 1]	82.72	82.72	82.72	82.72
[1, 0.5, 1]	82.60	87.60	96.50	94.72
[0.5, 1, 1]	82.50	87.42	96.52	94.70
[1, 0.5, 0.5]	82.58	87.50	96.48	94.65
[0.5, 1, 0.5]	82.40	87.35	96.45	94.60
[0.5, 0.5, 1]	82.45	87.38	96.47	94.62
[1, 1, 0.5]	82.60	87.54	96.50	94.72

The results show that while the F1 scores fluctuate slightly across different weight settings, the model maintains consistently high performance, indicating the robustness of the composite loss design. No single combination dominates in all datasets, suggesting that the proposed multi-loss strategy provides complementary benefits in classification, boundary refinement, and structural feature alignment. The original configuration [1, 1, 0.5] remains effective, but alternative weightings also yield competitive results, validating the practicality and flexibility of the composite loss function.

6.3. Analysis of Model Size and Efficiency

To provide a comprehensive comparison of the computational characteristics of all evaluated models, the parameter count, FLOPs, training time, and inference time are illustrated in Figure 15, where four histograms jointly present the performance of representative change detection methods. This visualization enables a clear and intuitive understanding of architectural differences and their impact on computational overhead, offering a holistic view of model complexity and runtime behavior across the benchmarked methods. It is pointed out that the training time and inference time were tested on the RWSBD dataset.

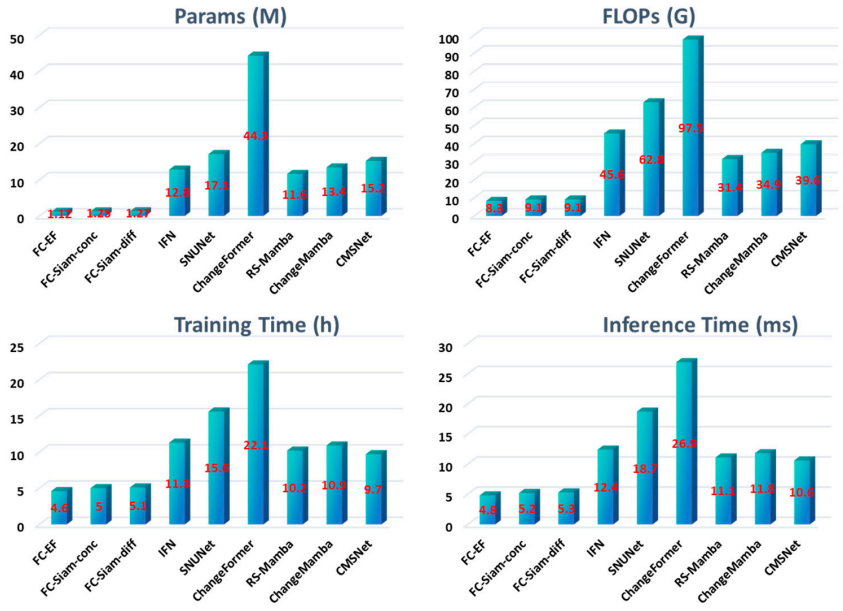


Figure 15. Computational comparison of all models.

As illustrated in Figure 15, the models exhibit noticeable variation in computational demands due to differences in network depth, attention mechanisms, and multi-branch feature processing. Architectures incorporating heavy transformer blocks and long-range dependency modeling (e.g., ChangeFormer, RS-Mamba, and ChangeMamba) tend to show substantially higher parameter counts and FLOPs. In contrast, lightweight convolution-based models yield smaller complexity but often at the cost of representational power. CMSNet demonstrates a balanced complexity profile: although it contains multi-scale spectral–spatial decomposition and cross-modal fusion modules, its parameter count and FLOPs remain moderate and distinctly lower than those of transformer-based and Mamba-based networks. This confirms that CMSNet effectively avoids redundant computation while retaining strong feature modeling capability.

The runtime statistics presented in Figure 15 further reveal that models with larger FLOPs generally require longer training and inference times, whereas more compact designs provide faster execution. CMSNet achieves competitive training efficiency and notably shorter inference latency compared with computation-intensive architectures, demonstrating its practicality for large-scale or time-sensitive remote sensing applications. Moreover, to enhance the interpretability of CMSNet, Table 7 provides a detailed breakdown of parameter quantities across all CMSNet modules, including the spectral–spatial decomposition heads, cross-modal fusion blocks, transformer-based refinement units, and the final prediction module. This module-level analysis shows that no single component dominates the parameter distribution, confirming that the overall computational cost of CMSNet is well-balanced and structurally optimized. Together, Figure 15 and Table 7 demonstrate that CMSNet achieves a favorable trade-off among accuracy, complexity, and efficiency.

Table 7. Detailed parameter distribution of each module in the proposed CMSNet.

Module	Description	Params (M)
CNN Encoder	Shared dual-stream CNN encoder (4 stages)	3.8
Mamba Temporal Modeling Block	4-layer Mamba state-space sequence modeling	4.1
PVPF-FM (SAM Prior Fusion)	Feature enhancement with dual-branch (encoder/decoder) differential fusion	5.2
Decoder	Two-level upsampling with skip connections	1.6
Output Head	Convolution + Sigmoid	0.5
CMSNet	/	15.2

6.4. Performance Analysis of the Model in Building Damage Scenarios

To further evaluate the practicality and generalization capability of the proposed method in real post-disaster scenarios, supplementary experiments were conducted on the xBD dataset, which provides multi-class building damage annotations with diverse disaster categories and complex structural variations. All comparison models were retrained on the xBD training set under uniform experimental settings, ensuring fair comparison across methods. Figure 16 presents the visualization of three representative samples, while Table 8 reports the quantitative evaluation results, including Precision, Recall, F1-score, and IoU.

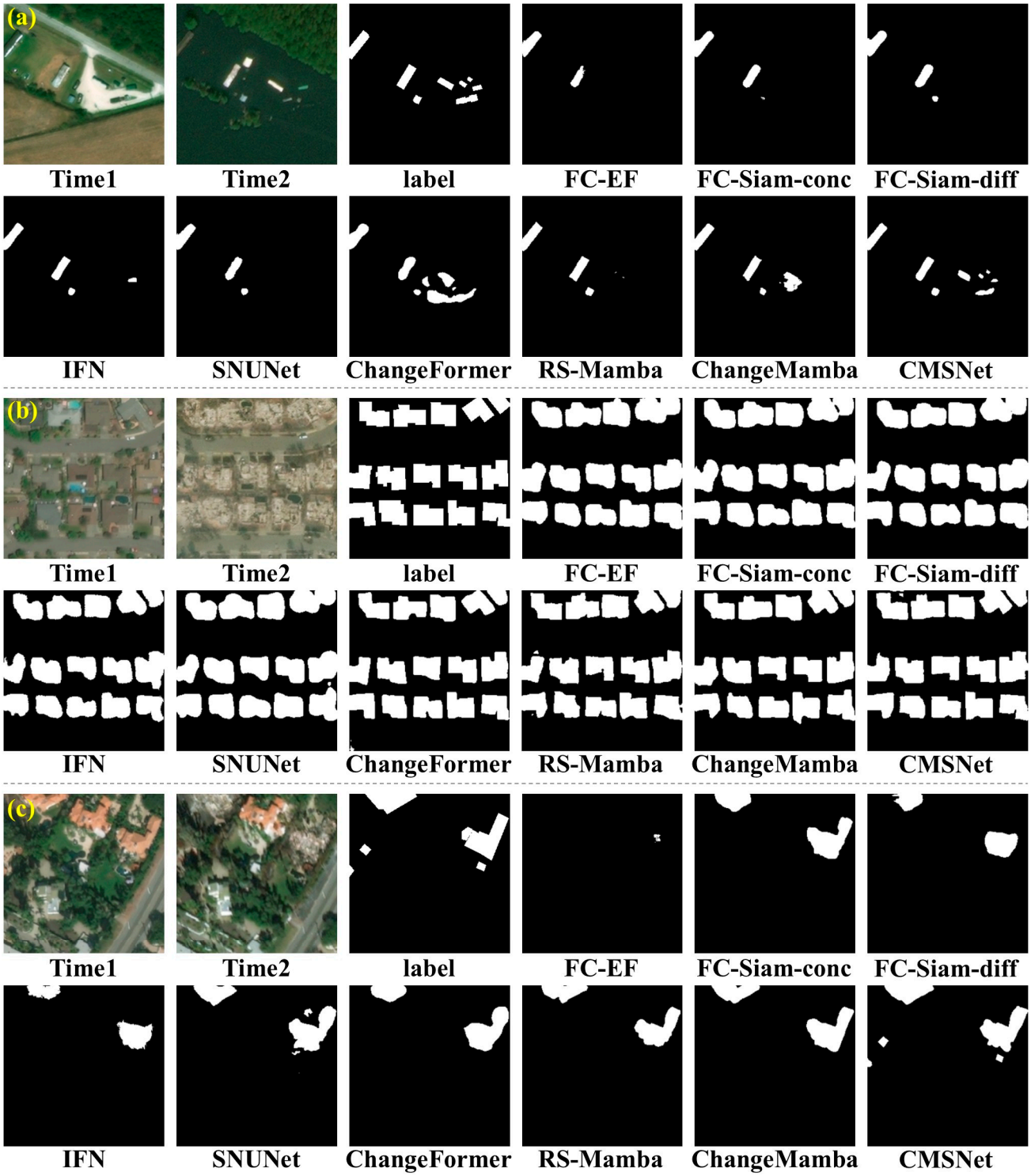


Figure 16. The general comparison of all models on the xBD dataset.

As shown in Table 8, traditional FC-based methods (FC-EF, FC-Siam-conc, and FC-Siam-diff) exhibit relatively low detection accuracy, with F1-scores below 75%. These architectures rely primarily on local convolutional features and therefore struggle to accurately capture irregular building contours and heterogeneous damage patterns present in xBD. Middle-tier models such as IFN and SNUNet achieve notable performance gains, benefiting from improved multi-scale feature aggregation; however, their F1-scores remain below 81%, indicating limited capability in handling large-scale destruction and complex debris regions.

Table 8. Comparison of accuracy of all models on the xBD dataset (%).

Methods	P	R	F1	Igou
FC-EF	71.82	68.94	70.35	54.23
FC-Siam-conc	74.51	70.26	72.33	56.69
FC-Siam-diff	76.94	72.81	74.82	59.73
IFN	82.13	76.92	79.45	65.55
SNUNet	83.27	77.31	80.18	66.90
ChangeFormer	84.62	78.85	81.65	68.75
RS-Mamba	85.13	79.42	82.17	69.42
ChangeMamba	85.56	79.68	82.54	69.98
CMSNet	87.40	80.83	83.91	71.89

More advanced methods achieve progressively better results. ChangeFormer demonstrates a higher F1-score (81.65%), verifying the advantages of transformer-based global modeling in large-disaster scenes. RS-Mamba and ChangeMamba further improve performance, reflecting the effectiveness of state-space modeling in capturing long-range temporal dependencies. Nevertheless, both models still exhibit certain weaknesses in boundary precision and small-object detection.

CMSNet achieves the best performance among all methods, reaching an F1-score of 83.91% and an IoU of 71.89%, which represents a substantial improvement over existing CNN-, Transformer-, and Mamba-based approaches. These results confirm that the hybrid design of CMSNet—combining CNN-based local perception, Mamba-based temporal modeling, and SAM-guided structural priors—effectively enhances the representation of complex damage structures. The visual comparisons in Figure 16 also show that CMSNet produces more accurate delineation of damaged regions, sharper boundaries, and fewer false detections in cluttered environments. Overall, the experimental results on the xBD dataset validate the robustness, generalization capability, and practical applicability of the proposed method in real-world building damage assessment tasks.

7. Conclusions

This paper proposes a novel change detection method for damaged buildings in re-al-world war-zone remote sensing imagery—CMSNet. The method synergistically combines the structural prior awareness of the SAM foundation model with the temporal modeling and fine-detail representation capabilities of CNN-Mamba, effectively addressing challenges such as non-rigid deformations, localized fragmentation, and blurred boundaries of buildings in complex warzone environments. By incorporating the PVPF-FM module, high-level structural guidance from SAM is deeply fused with bi-temporal optical image features, enabling CMSNet to enhance the modeling of structural consistency, semantic integrity, and boundary continuity in the mid-to-high-level feature representation stage. Additionally, this work introduces RWSBD, a large-scale, high-resolution real-world war-damaged building dataset from the Gaza region, providing a solid data foundation for war damage change detection. Extensive experimental results demonstrate that CMSNet significantly outperforms state-of-the-art methods across RWSBD and three other public datasets, achieving leading performance in IoU, F1-score, Precision, and Recall. The model exhibits particularly strong robustness and generalization in identifying fine-grained damaged regions and adapting to complex scenarios. This study not only innovatively integrates foundational visual models with efficient temporal modeling structures at the methodological level but also provides a valuable data benchmark for real-world war-damage building detection, offering substantial academic and practical significance.

Although CMSNet demonstrates high accuracy and robustness in detecting damaged buildings in real-world war-zone scenarios, certain limitations remain. First, this study

relies solely on optical remote sensing imagery and does not incorporate cross-modal data such as SAR, which limits model stability under extreme conditions like cloud cover or optical image failure. Second, while the CNN-Mamba architecture achieves a balance be-tween feature modeling capability and computational efficiency, there remains room for improvement in large-scale rapid inference and lightweight deployment. Future work will focus on exploring cross-modal fusion techniques and more efficient network architecture to enhance the method's applicability and scalability for multi-source remote sensing data and rapid post-disaster response over large areas.

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

Funding: This research received no external funding.

Data Availability Statement: RWSBD and CWBD will be announced soon. If any researcher needs them, they can contact us via email to obtain the right to use the data. WHU-CD is downloaded from http://gpcv.whu.edu.cn/data/building_dataset.html (accessed on 29 August 2025). LEVIR-CD+ is downloaded from https://www.kaggle.com/datasets/mdrifaturrahman33/levir-cd-change-detection/data (accessed on 29 August 2025).

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

References

1. Zhang, Z.; Jiang, H.; Pang, S.; Hu, X. Review and prospect in change detection of multi-temporal remote sensing images. *Acta Geod. Cartogr. Sin.* **2022**, *51*, 1091–1107.
2. Ding, L.; Hong, D.; Zhao, M.; Chen, H.; Li, C.; Deng, J.; Yokoya, N.; Bruzzone, L.; Chanussot, J.J.A. A Survey of Sample-Efficient Deep Learning for Change Detection in Remote Sensing: Tasks, Strategies, and Challenges. *arXiv* **2025**, arXiv:2502.02835. [CrossRef]
3. Cheng, G.; Huang, Y.; Li, X.; Lyu, S.; Xu, Z.; Zhao, H.; Zhao, Q.; Xiang, S. Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review. *Remote Sens.* **2024**, *16*, 2355. [CrossRef]
4. Wu, C.; Zhang, L.; Du, B.; Chen, H.; Wang, J.; Zhong, H. UNet-Like Remote Sensing Change Detection: A review of current models and research directions. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 305–334. [CrossRef]
5. Holail, S.; Saleh, T.; Xiao, X.; Zahran, M.; Xia, G.-S.; Li, D. Edge-CVT: Edge-informed CNN and vision transformer for building change detection in satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 48–68. [CrossRef]
6. Zhang, H.; Ma, G.; Fan, H.; Gong, H.; Wang, D.; Zhang, Y. SDCINet: A novel cross-task integration network for segmentation and detection of damaged/changed building targets with optical remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 422–446. [CrossRef]
7. Zhang, H.; Zhang, Y.; Wang, D.; Ma, G. Damaged Building Object Detection From Bitemporal Remote Sensing Imagery: A Cross-Task Integration Network and Five Datasets. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5648827. [CrossRef]
8. Li, J.; He, W.; Li, Z.; Guo, Y.; Zhang, H. Overcoming the uncertainty challenges in detecting building changes from remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *220*, 1–17. [CrossRef]
9. Fyleris, T.; Kriščiūnas, A.; Gružauskas, V.; Čalnerytė, D.; Barauskas, R. Urban Change Detection from Aerial Images Using Convolutional Neural Networks and Transfer Learning. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 246. [CrossRef]
10. Oubara, A.; Wu, F.; Maleki, R.; Ma, B.; Amamra, A.; Yang, G. Enhancing Adversarial Learning-Based Change Detection in Imbalanced Datasets Using Artificial Image Generation and Attention Mechanism. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 125. [CrossRef]
11. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sens. Environ.* **2021**, *265*, 112636. [CrossRef]
12. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. RSMamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [CrossRef]
13. Zhang, S.; Liu, J.; Qian, T.; Li, X. Prompt-Guided Dual-Path UNet with Mamba for Medical Image Segmentation. *arXiv* **2025**, arXiv:2503.19589.

14. Zhang, R.; Xu, C.; Xu, F.; Yang, W.; He, G.; Yu, H.; Xia, G.-S. S_3OD: Size-unbiased semi-supervised object detection in aerial images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 179–192. [CrossRef]
15. Singh, A. Review Article Digital change detection techniques using remotely-sensed data. *Int. J. Remote Sens.* **1989**, *10*, 989–1003. [CrossRef]
16. Bovolo, F.; Bruzzone, L.; Marconcini, M. A Novel Approach to Unsupervised Change Detection Based on a Semisupervised SVM and a Similarity Measure. *Geosci. Remote Sens. IEEE Trans.* **2008**, *46*, 2070–2082. [CrossRef]
17. Daudt, R.C.; Saux, B.L.; 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.
18. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [CrossRef]
19. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.-S. CBAM: Convolutional Block Attention Module. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
20. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 3095166. [CrossRef]
21. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
22. Yao, L.; Liu, F.; Chen, D.; Zhang, C.; Wang, Y.; Chen, Z.; Xu, W.; Di, S.; Zheng, Y.J.A. RemoteSAM: Towards Segment Anything for Earth Observation. *arXiv* **2025**, arXiv:2505.18022. [CrossRef]
23. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
24. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
25. Shi, W.; Zhang, M.; Zhang, R.; Chen, S.; Zhan, Z. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sens.* **2020**, *12*, 1688. [CrossRef]
26. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
27. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
28. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 548–558.
29. Liu, C.; Sui, H.; Wang, J.; Ni, Z.; Ge, L.J.R.S. Real-time ground-level building damage detection based on lightweight and accurate YOLOv5 using terrestrial images. *Remote Sens.* **2022**, *14*, 2763. [CrossRef]
30. Vaswani, A. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
31. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K.J.A. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805. [CrossRef]
32. Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; Sun, Q.J.A. Feature Pyramid Transformer. *arXiv* **2020**, arXiv:2007.09451. [CrossRef]
33. Bandara, W.G.C.; Patel, V.M.J.A. A Transformer-Based Siamese Network for Change Detection. *arXiv* **2022**, arXiv:2201.01293. [CrossRef]
34. Zhang, Z.; Liu, S.; Qin, Y.; Wang, H. MATNet: Multilevel attention-based transformers for change detection in remote sensing images. *Image Vis. Comput.* **2024**, *151*, 105294. [CrossRef]
35. Zhang, H.; Ma, G.; Zhang, Y.; Wang, B.; Li, H.; Fan, L. MCHA-Net: A multi-end composite higher-order attention network guided with hierarchical supervised signal for high-resolution remote sensing image change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 40–68. [CrossRef]
36. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. PVT v2: Improved baselines with Pyramid Vision Transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
37. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 3144894. [CrossRef]
38. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. ChangeMamba: Remote Sensing Change Detection with Spatiotemporal State Space Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3417253. [CrossRef]
39. Zhang, H.; Chen, K.; Liu, C.; Chen, H.; Zou, Z.; Shi, Z. CDMamba: Incorporating Local Clues Into Mamba for Remote Sensing Image Binary Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 3545012. [CrossRef]

40. Paranjape, J.N.; Melo, C.d.; Patel, V.M. A Mamba-based Siamese Network for Remote Sensing Change Detection. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 1186–1196. [CrossRef]
41. Shen, Y.; Yao, S.; Qiang, Z.; Pei, G. SD-Mamba: A lightweight synthetic-decompression network for cross-modal flood change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104409. [CrossRef]
42. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
43. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]
44. Zhang, D.; Liang, D.; Yang, H.; Zou, Z.; Ye, X.; Liu, Z.; Bai, X.J.A. SAM3D: Zero-Shot 3D Object Detection via Segment Anything Model. *arXiv* **2023**, arXiv:2306.02245. [CrossRef]
45. Mei, L.; Ye, Z.; Xu, C.; Wang, H.; Wang, Y.; Lei, C.; Yang, W.; Li, Y. SCD-SAM: Adapting Segment Anything Model for Semantic Change Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 3407884. [CrossRef]
46. Li, S.; Wang, S.-Y.; Sun, Z.; Xiao, J. Semantic Localization Guiding Segment Anything Model For Reference Remote Sensing Image Segmentation. *arXiv* **2025**, arXiv:2506.10503. [CrossRef]
47. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
48. 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]
49. 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]
50. Rodrigo, C.D.; Bertrand, L.S.; Alexandre, B. Fully Convolutional Siamese Networks for Change Detection. *arXiv* **2018**, arXiv:1810.08462. [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.