

CRD-Net: Frequency-Adaptive Feature Injection and Change Decoupling for Building Damage Assessment

Yao Zheng¹, Yuanxin Ye¹[0000–0001–6843–6722]*, Tan Shu¹, and Liwei Cai¹

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, Sichuan, China

yeyuanxin@home.swjtu.edu.cn

Abstract. Building damage assessment is crucial for post-disaster relief and reconstruction. However, current assessment methods utilizing high-resolution remote sensing imagery struggle to represent fine-grained details, as subtle damage signals are easily overwhelmed by complex backgrounds. To address this, we propose a Frequency-Adaptive Feature Injection and Change Decoupling framework. For feature encoding, we fine-tune a DINOv3 encoder with a frequency-adaptive guidance strategy to acutely capture subtle damage. For decoding, we introduce the Change Representation Decoupling and Discriminative Decoder (CRD²). CRD² leverages the Mamba model to efficiently fuse multi-scale features for spatiotemporal dynamics. To prevent minor damage from being submerged, it utilizes orthogonal subspace projection to isolate pure change features and applies class-prototype contrastive learning for precise damage-level identification. Extensive experiments show our CRD-Net achieves state-of-the-art Joint F1 scores on the xBD (81.26%) and EBD (81.88%) datasets.

Keywords: Building damage assessment · DINOv3 · State space model · Change Decoupling

1 Introduction

The intensifying effects of global climate change have led to a significant increase in the frequency and severity of natural disasters, posing profound threats to human society and the environment [37]. Consequently, rapidly and accurately assessing building damage post-disaster is crucial for emergency rescue operations, reconstruction planning, and risk assessment [20].

Post-disaster building damage assessment is critical for classifying structural damage and guiding recovery efforts [21]. Early methods relied on manual field inspections, utilizing engineering experience to visually assess physical deformations. However, despite their interpretability, these manual approaches are inefficient, highly subjective, and hazardous to personnel, making them fundamentally unsuitable for large-scale, rapid disaster evaluation [28].

* Corresponding author.

The rapid advancement of remote sensing technology—characterized by non-contact operation, broad coverage, and rapid response—has provided critical data for quantitative building damage analysis [25]. Concurrently, deep learning has driven the development of various damage assessment methods [30]. Early approaches predominantly relied on Convolutional Neural Networks (CNNs), which excel at extracting local spatial textures but struggle to capture long-range contextual dependencies due to their limited receptive fields [1, 33]. To address this, Vision Transformers (ViTs) were widely adopted, leveraging self-attention mechanisms to establish global representations, albeit often at the cost of high computational overhead [3, 11, 17]. More recently, state space models like Mamba have emerged as a promising alternative, achieving linear-time sequence modeling to efficiently balance global context and local feature extraction [4, 5]. Building on this architectural evolution, Visual Foundation Models (VFM) such as DINO [2], CLIP [24], and SAM [18] have recently opened new avenues for remote sensing image understanding. Benefiting from robust generalization derived from large-scale pre-training, these models have been successfully adapted to complex scenes using Parameter-Efficient Fine-Tuning (PEFT) techniques, such as LoRA and Adapters [22, 31].

Despite these successes, mainstream VFMs predominantly rely on the standard Vision Transformer (ViT) architecture, which poses inherent challenges for fine-grained tasks [35]. Limited by large-stride initial patch embedding and single-scale representations, these models struggle to construct hierarchical multi-scale pyramids. This early-stage loss of spatial resolution impedes the extraction of high-frequency local details, severely degrading performance on subtle damage patterns like roof cracks and edge deformations. Several recent studies have attempted to mitigate this issue. For instance, ChangeViT [40] employs DINOv2 and injects fine-grained spatial features into high-level semantics to enhance detail perception; however, direct spatial-domain injection frequently introduces substantial feature redundancy. Alternatively, the UGFF-VLM model [27] explores fusion in the frequency domain by decomposing features into amplitude and phase components. While this method selectively enhances high-frequency boundaries for farmland extraction, it lacks the explicit fine-grained spatial injection required for building damage assessment, which relies heavily on microscopic details.

To tackle the aforementioned challenges, we propose a comprehensive framework designed to achieve stable and high-precision building damage assessment. This framework operates on three key strategies: explicitly injecting frequency-domain representations to compensate for detail loss, leveraging Mamba to efficiently fuse multi-scale spatio-temporal dynamics, and thoroughly isolating change information from the background via feature decoupling. Built upon a fine-tuned DINOv3 backbone, our approach is operationalized through two core modules:

- 1) A Spatio-Frequency Adaptive Feature Enhancement (SFAFE) module that leverages learnable Gaussian band-pass filters to decouple multi-scale features into distinct frequency sub-bands. By performing cross-branch interaction in

both the spatial and frequency domains, SFAFE adaptively injects local high-frequency details from a CNN branch into the global low-frequency semantics of the ViT backbone, effectively compensating for the inherent detail loss.

2) We propose a Mamba-based Change Representation Decoupling and Discrimination (CRD²) decoder. To overcome the limitation that minor damage signals are easily overwhelmed by the background during simple feature differencing, this decoder first leverages Mamba’s superior long-sequence modeling capability to fuse hierarchical features, accurately capturing fine-grained spatio-temporal semantic differences. Subsequently, it thoroughly decouples the change information from the background via orthogonal subspaces. Coupled with prototype-guided contrastive learning, it ultimately achieves high-precision discrimination of fine-grained damage levels.

2 Related Work

2.1 DINOv3 in Remote Sensing

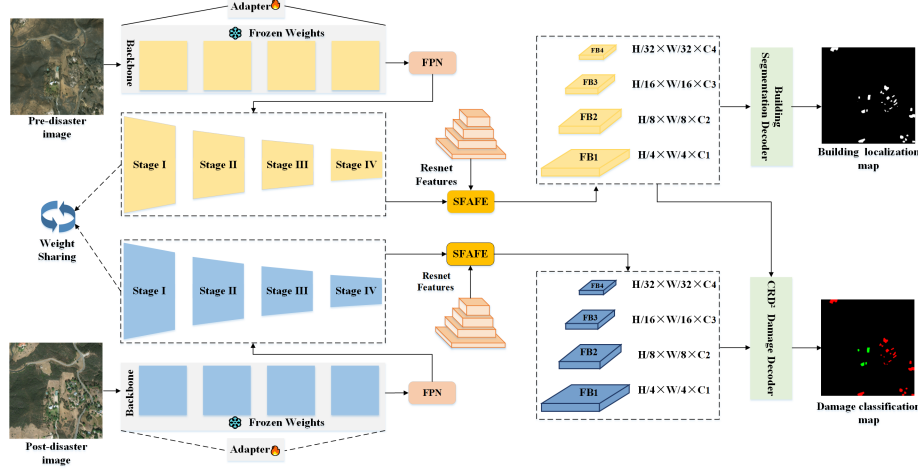
DINOv3 is an advanced self-supervised Vision Transformer (ViT) that learns highly discriminative, semantically consistent features from unlabeled data via an improved self-distillation strategy [26, 39]. Benefiting from its strong generalization and multi-scale modeling capabilities, it has been widely applied in medical imaging [8, 36], object detection [15, 19], and remote sensing [6, 7, 38]. For remote sensing change detection, recent methods adapt DINOv3 by integrating parameter-efficient fine-tuning (PEFT) [7], constructing feature pyramids with lightweight backbones [6], and designing multi-scale cross-attention difference modules [38]. However, despite exploiting DINOv3’s representational capacity, these approaches still exhibit limitations in fine-grained difference modeling and effective multi-scale feature enhancement. Consequently, they struggle to meet the high-precision requirements for structural and semantic change analysis in complex scenarios, such as building damage assessment.

2.2 Parameter-Efficient Fine-Tuning Methods

Vision foundation models such as DINOv3, which are pre-trained on large-scale natural image datasets, exhibit strong feature representation and cross-task transfer capabilities [16]. However, due to significant differences in imaging mechanisms and spatial distributions between remote sensing images and natural images, directly applying these models to remote sensing tasks often leads to performance bottlenecks [14, 23]. To address this issue, parameter-efficient fine-tuning (PEFT) methods have been proposed, which achieve performance comparable to full-parameter fine-tuning while substantially reducing computational costs by optimizing only a small number of additional parameters. Representative approaches include LoRA [13] and Adapter [34]. As presented in Table 1, we comparatively evaluate the performance of LoRA and Adapter fine-tuning methods applied to the DINOv3 architecture using the xBD dataset. The experimental results demonstrate that LoRA exhibits a superior capacity for detecting

Table 1: Comparison of Fine-Tuning Methods on the xBD Dataset

Method	F_1^{loc}	F_1^{cls}	F_1^{overall}	Damage F_1 Score per Class			
				No damage	Minor damage	Major damage	Destroyed
DINOv3-Adapter	85.85	74.88	78.17	93.85	55.76	75.87	85.87
DINOv3-LoRA	86.74	76.46	79.55	94.40	59.54	75.42	85.73

**Fig. 1:** Overall Architecture of the Proposed Network.

minor damages compared to the Adapter method, thereby proving more suitable for building damage assessment tasks.

3 Methodology

Fig. 1 presents the overall architecture of CRD-Net for the building damage assessment task. Operating under an end-to-end learning paradigm, the framework is constructed around three pivotal stages: feature extraction, feature enhancement, and feature refinement. During feature extraction, a LoRA-finetuned DINOv3 serves as the backbone. Following this, the SFAFE module is introduced to enhance fine-grained feature representations. In the final refinement stage, the decoding process is decoupled into two distinct branches. The building decoder directly leverages pre-disaster imagery to extract spatial location information of buildings. Concurrently, the damage classification decoder employs a Change Representation Decoupling and Discriminative Decoder for feature refinement. This architectural design substantially elevates the extraction precision of subtle damage features, effectively mitigating the suppression of damage signals by background noise or intact building characteristics.

3.1 Spatio-Frequency Adaptive Feature Enhancement

DINOv3 operates at a fixed 1/16 resolution, lacking multi-scale representation. Inspired by MTP [29], we insert lightweight neck modules after the 5th, 11th, 17th, and 23rd layers, mapping channels to $\{96, 192, 384, 768\}$ via standard and depthwise separable convolutions, producing a four-level feature pyramid (1/4, 1/8, 1/16, 1/32). To recover high-frequency details lost in patch embedding, we introduce an auxiliary ResNet branch and propose a Spatio-Frequency Adaptive Feature Enhancement module (SFAFE), as illustrated in Fig. 2. SFAFE decomposes features into frequency sub-bands via learnable Gaussian band-pass filters and performs adaptive cross-branch interaction, achieving collaborative CNN–ViT enhancement in both spatial and frequency domains. Let the CNN features and DINO features at the current scale be denoted as $\mathbf{F}_c \in \mathbb{R}^{B \times C \times H \times W}$ and $\mathbf{F}_v \in \mathbb{R}^{B \times C \times H \times W}$, respectively. First, the CNN and DINO features are transformed from the spatial domain to the frequency domain via the fast Fourier transform (FFT):

$$\mathcal{F}_c = \text{FFTshift}(\text{FFT2}(\mathbf{F}_c)), \quad (1)$$

$$\mathcal{F}_v = \text{FFTshift}(\text{FFT2}(\mathbf{F}_v)). \quad (2)$$

Next, a set of learnable Gaussian band-pass filters is constructed to decompose the frequency-domain features into multiple frequency subbands, including low-frequency, mid-frequency, and high-frequency components. Specifically, the radial distance grid is defined as

$$r(u, v) = \sqrt{u^2 + v^2}, \quad (3)$$

where u and v denote the horizontal and vertical frequency coordinates in the two-dimensional Fourier spectrum, and $r(u, v)$ represents the radial distance from the frequency point (u, v) to the spectrum center $(0, 0)$.

The k th Gaussian band-pass filter is defined as

$$B_k(u, v) = \exp\left(-\frac{(r(u, v) - \mu_k)^2}{2\sigma_k^2}\right) \alpha_k, \quad k = 0, \dots, K, \quad (4)$$

where μ_k , σ_k , and α_k are trainable parameters controlling the center frequency, bandwidth, and gain of the k th subband, respectively.

For each frequency subband, the corresponding features are extracted as

$$\mathcal{F}_c^k = \mathcal{F}_c \odot B_k, \quad \mathcal{F}_v^k = \mathcal{F}_v \odot B_k, \quad (5)$$

where $\odot$ denotes element-wise multiplication.

The inverse Fourier transform is then applied to map the features back to the spatial domain, and only the real component is retained:

$$\mathbf{X}_c^k = \Re(\text{IFFT2}(\mathcal{F}_c^k)), \quad \mathbf{X}_v^k = \Re(\text{IFFT2}(\mathcal{F}_v^k)). \quad (6)$$

We apply different fusion strategies based on frequency characteristics. Low and high frequencies ($k = 1, 3$) use Cross-Attention (Queries from ViT, Keys/Values

from CNN) to accurately inject local details. Mid-frequencies ($k = 2$) use a gated fusion controlled by MLP-derived coefficients (β_1, β_2) for bidirectional interaction:

The subband features are processed using different fusion strategies according to their frequency characteristics. Specifically, the output features of the k th subband are defined as

$$\mathbf{Y}_k = \begin{cases} \text{CrossAttn}(\mathbf{X}_v^k, \mathbf{X}_c^k), & k = 1, \\ \beta_1 \mathbf{X}_c^k + \beta_2 \mathbf{X}_v^k, & k = 2, \\ \text{CrossAttn}(\mathbf{X}_v^k, \mathbf{X}_c^k), & k = 3, \end{cases} \quad (7)$$

where $\text{CrossAttn}(\cdot)$ denotes the cross-attention operation.

The fused results of all subbands are then transformed back into the frequency domain through the forward Fourier transform and accumulated into the global spectrum $\mathcal{F}_{\text{rea}}$. Subsequently, frequency-domain aggregation and inverse Fourier transformation are performed to obtain the frequency-fused features $\mathbf{F}_{\text{rea}}$:

$$\mathbf{F}_{\text{rea}} = |\text{IFFT2}(\text{FFTshift}(\mathcal{F}_{\text{rea}}))|. \quad (8)$$

To mitigate numerical errors from FFT operations and enhance channel-wise semantics, we introduce a complementary lightweight spatial fusion branch. We generate a preliminary fused feature $\mathbf{F}_d$:

$$\mathbf{F}_d = \text{Conv}_{1 \times 1}(\text{Concat}(\mathbf{F}_c, \mathbf{F}_v)). \quad (9)$$

A spatial gating map $\mathbf{G} \in [0, 1]^{H \times W}$ is then computed using the frequency-fused features as prior knowledge to adaptively emphasize key structural regions and suppress background noise:

The spatial gating map is formulated as

$$\mathbf{G} = \sigma(\text{Conv}_{1 \times 1}(\text{Concat}(\mathbf{F}_c, \mathbf{F}_d, \mathbf{F}_{\text{rea}}))), \quad (10)$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function. The gating map $\mathbf{G} \in \mathbb{R}^{H \times W}$ takes values in the range $[0, 1]$, where larger values indicate stronger discriminative responses at the corresponding spatial locations. During training, this gating map automatically learns to emphasize key structural regions, such as building edges and texture patterns, while suppressing irrelevant background clutter.

The final fused feature representation is computed as

$$\mathbf{F}_{\text{fused}} = \text{Conv}_{3 \times 3}(\text{Concat}(\mathbf{G} \odot \mathbf{F}_{\text{rea}}, (1 - \mathbf{G}) \odot \mathbf{F}_d)). \quad (11)$$

where $\odot$ denotes element-wise multiplication.

All Gaussian filter parameters μ_k , σ_k , and α_k , attention modules, and gating networks are jointly optimized in an end-to-end manner together with the backbone network.

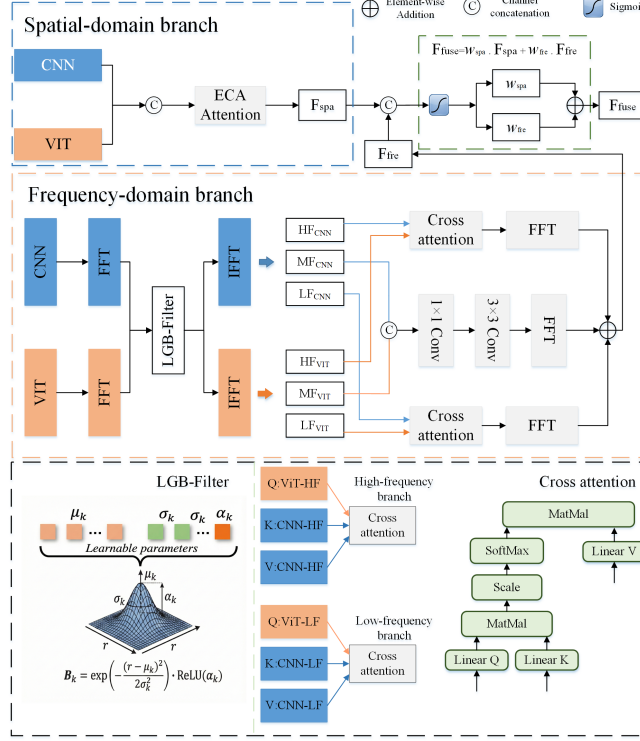


Fig. 2: Spatio-Frequency Adaptive Feature Enhancement Module.

3.2 Change Representation Decoupling and Discrimination (CRD^2) Decoder

We propose the Change Representation Decoupling and Discrimination (CRD^2) Decoder, built upon the decoder architecture of ChangeMamba [4]. It consists of two branches: the localization branch uses only pre-disaster features with progressive upsampling to produce building masks; the damage assessment branch further fuses multi-scale features from three spatio-temporal branches via skip connections at each level. Within the damage assessment branch, CRD^2 introduces two successive components to separate change signals from the background and enhance inter-class discriminability:

Orthogonal Subspace Decoupling (OSD) At each decoder level l , the spatiotemporal interaction module produces a fused feature $\hat{f}_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ by aggregating bi-temporal representations through multiple interaction strategies (i.e., channel-wise concatenation, column-wise interleaving, and horizontal concatenation) followed by state space modeling. This fused feature encapsulates both change-related and change-irrelevant information in an entangled manner, which may hinder precise damage assessment. To disentangle them, we apply

two independent projections $\mathcal{P}_c^l$ and $\mathcal{P}_i^l$ to decompose $\tilde{f}_l$ into a *change subspace* and an *invariant subspace*:

$$z_c^l = \mathcal{P}_c^l(\tilde{f}_l), \quad z_i^l = \mathcal{P}_i^l(\tilde{f}_l) \in \mathbb{R}^{D \times H_l \times W_l} \quad (12)$$

where z_c^l captures disaster-induced structural changes and z_i^l encodes temporally stable scene context. To ensure mutual independence, we flatten and L_2 -normalize both representations, then minimize their cross-correlation:

$$\mathcal{L}_{orth} = \frac{1}{N} \|\bar{Z}_c \bar{Z}_i^\top\|_F^2 \quad (13)$$

where $\bar{Z}_c, \bar{Z}_i \in \mathbb{R}^{D \times N}$ are the normalized feature matrices with $N = H_l \times W_l$.

Furthermore, to eliminate unchanged background residuals (e.g., artifacts caused by illumination) that might remain in z_c , we utilize the invariant subspace features as an inverse prior to design a Change Gating mechanism. By evaluating the response intensity of the invariant features, spatial gating weights are generated to modulate the change features:

$$g = \sigma(W_g * [z_c; z_i]) \quad (14)$$

$$z'_c = z_c \odot g \quad (15)$$

where σ denotes the Sigmoid activation function, and $\odot$ represents element-wise multiplication. This mechanism effectively exploits the high confidence of invariant regions to suppress spurious background changes.

Prototype-Guided Hard-sample Contrastive Learning (PHCL) Although OSD achieves orthogonal decoupling of features, in disaster assessment tasks, “minor damage” and “intact building” are visually highly similar, and decoupling alone struggles to guarantee sufficient inter-class discriminability. Therefore, we propose the PHCL module to explicitly enlarge the feature distance between confusing classes by constructing three prototype centers and employing adaptive temperature contrastive learning.

We maintain three momentum-updated feature prototype centers: μ_d (damage), μ_b (intact building), and μ_g (background). To avoid the burden of additional building annotations, we utilize the energy response of the invariant subspace z_i to approximate the distinction between buildings and the background. Since buildings possess rich structural textures, their energy response in z_i is significantly higher than that of flat background regions. Defining the spatial energy as $E = \|z_i\|_2$, the confidence probability of a building is:

$$P_{bldg} = \sigma\left(\frac{5(E - \mu_E - 0.2\sigma_E)}{\sigma_E}\right) \quad (16)$$

where μ_E and σ_E are the spatial mean and standard deviation of the feature map energy, respectively.

In pixel-level contrastive learning, we prioritize sampling hard examples near change boundaries. Let $D(x)$ be the normalized distance from pixel x to the ground-truth change boundary. We design a semantic-adaptive temperature mechanism: when using a damage feature as the anchor, if the negative sample is an intact building, we adopt a higher temperature τ_{high} to tolerate their inevitable intrinsic structural similarity; if the negative sample is the background, we apply a lower temperature τ_{low} to impose a stronger discrimination penalty. The specific sample temperature is calculated as:

$$\tau(x, y) = \begin{cases} \tau_{high} + \tau_{\alpha} e^{-3D(x)}, & \text{if } y \text{ is building} \\ \tau_{low} + \tau_{\alpha} e^{-3D(x)}, & \text{if } y \text{ is background} \end{cases} \quad (17)$$

Combining distance weighting and semantic weighting, the pixel-level InfoNCE loss is defined as:

$$\mathcal{L}_{pixel} = -\frac{1}{|A|} \sum_{i \in A} \log \frac{\exp(s_i^+ / \tau_{pos}^i)}{\exp(s_i^+ / \tau_{pos}^i) + \sum_{k \in K} w_k \exp(s_{i,k}^- / \tau_{neg}^{i,k})} \quad (18)$$

Simultaneously, we constrain the mean of the anchor features $\bar{z}_a$ to move towards its corresponding ground-truth class prototype center, forming the prototype center loss:

$$\mathcal{L}_{center} = -\log \frac{\exp(\bar{z}_a \cdot \mu_{gt} / \tau)}{\sum_c \exp(\bar{z}_a \cdot \mu_c / \tau)} \quad (19)$$

Finally, the total loss of the PHCL module is:

$$\mathcal{L}_{PHCL} = \mathcal{L}_{pixel} + \lambda \mathcal{L}_{center} \quad (20)$$

where λ balances the pixel-level and center-aware contrastive losses. Combining the OSD decoupling constraint across all scales and the PHCL contrastive constraint, the auxiliary loss term is expressed as:

$$\mathcal{L}_{aux} = \alpha \mathcal{L}_{orth} + \mathcal{L}_{PHCL} \quad (21)$$

where α controls the strength of the orthogonality regularization. We empirically set $\lambda = 1$ and $\alpha = 0.5$ throughout all experiments.

3.3 Loss Function

The objective of building localization is to predict a binary mask of building locations. Considering that accurate building regions are crucial for subsequent damage classification, we combine the binary cross-entropy loss (BCE) with the Lovász-softmax loss to obtain a more robust building localization performance under class-imbalanced datasets. These losses are formulated as follows:

$$\mathcal{L}_{lovász} = \frac{1}{NC} \sum_{n=1}^N \sum_{c \in C} \Delta_{J_c}(m_i^n(c)) \quad (22)$$

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (23)$$

Here, $\mathcal{L}_{\text{lovász}}$ and $\mathcal{L}_{\text{bce}}$ denote the Lovász-softmax loss and binary cross-entropy loss, respectively. N and C represent the number of samples and classes. $m_i^n(c)$ is the error vector for class c in the n -th image of the batch, computed based on the network output and ground-truth labels. For the i -th pixel, if $y_i = c$, then $m_i^n(c) = 1 - p_i$; otherwise, $m_i^n(c) = p_i$. Here, p_i and y_i denote the predicted probability and the ground-truth label of pixel i , respectively.

Δ_{J_c} is the Lovász extension of the Jaccard (IoU) loss for class c , which serves as a differentiable convex surrogate facilitating gradient-based optimization.

Accordingly, the final localization loss is defined as:

$$\mathcal{L}_{\text{loc}} = \mathcal{L}_{\text{bce}} + \alpha_1 \cdot \mathcal{L}_{\text{lovász}} \quad (24)$$

where α_1 is a hyperparameter that balances the two loss terms.

The damage classification sub-task aims to determine the damage level of each building. Considering that the damage levels in the xBD dataset possess an ordinal nature, we jointly employ the Lovász-softmax loss and the multi-class cross-entropy loss (CE) to supervise this sub-task. Accordingly, the damage classification loss is formulated as:

$$\mathcal{L}_{\text{dmg}} = \mathcal{L}_{\text{ce}} + \alpha_2 \cdot \mathcal{L}_{\text{lovász}} \quad (25)$$

where α_2 is a hyperparameter that balances the Lovász-softmax loss with the other loss term. The weighting coefficients α_1 and α_2 are set to 0.5 and 0.75, respectively, following the configuration of ChangeMamba [4] to ensure fair comparison. The localization loss $\mathcal{L}_{\text{loc}}$, the damage classification loss $\mathcal{L}_{\text{dmg}}$ and the auxiliary loss $\mathcal{L}_{\text{aux}}$ are jointly optimized to improve both building localization accuracy and damage classification performance. The overall loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{dmg}} + \mathcal{L}_{\text{aux}} \quad (26)$$

4 Experimental Results

4.1 Datasets

xBD Dataset The xBD dataset [10] is a large-scale building damage assessment (BDA) benchmark using high-resolution (< 0.8 m/pixel) pre- and post-disaster satellite imagery. Spanning diverse disaster types across $\approx 45,362$ km², it provides four-level damage annotations (*no damage*, *minor*, *major*, and *destroyed*) for over 850,000 buildings, facilitating robust evaluation of BDA methods in complex real-world scenarios.

EBD Dataset The EBD dataset [32] provides valuable supplementary data for post-disaster building damage assessment (BDA) and emergency mapping. It contains $> 18,000$ high-resolution pre- and post-disaster image pairs across 12 events, featuring four-level damage annotations for over 175,000 buildings. Since no official split is provided, we randomly divide the dataset into training, validation, and test sets at an 8:1:1 ratio.

4.2 Evaluation Metrics

To facilitate fair comparison with existing methods, this paper adopts the official evaluation metrics of the xBD dataset. Building localization performance is evaluated using the F_1 score, while damage classification performance is assessed by the harmonic mean of the F_1 scores across all damage categories.

The overall F_1 score is computed as a weighted sum of the two, with a higher weight assigned to the classification term to emphasize damage classification accuracy:

$$F_{1_b} = \frac{2TP}{2TP + FP + FN}, \quad (27)$$

$$F_{1_d} = \frac{m}{\sum_{n=1}^m \frac{1}{F_{1_{c_n}}}}, \quad (28)$$

$$F_1^{over} = 0.3 \times F_{1_b}^b + 0.7 \times F_{1_d}^d, \quad (29)$$

where TP , FP , and FN denote the number of true positive, false positive, and false negative pixels, respectively. $F_{1_{c_n}}$ is the per-category F_1 score and m is the total number of damage categories. As the harmonic mean of precision and recall, the F_1 score penalizes models that trade off one for the other, making it particularly suitable for evaluating building localization and damage classification under severe class imbalance.

4.3 Experimental Details

All models are implemented in PyTorch and trained on a single NVIDIA A6000 GPU (48 GB). We use the SGD optimizer with an initial learning rate of 1×10^{-4} , momentum of 0.9, weight decay of 5×10^{-4} , and a batch size of 8. Training runs for 100 epochs with per-epoch evaluation.

Data Preprocessing: Original 1024×1024 images from the xBD dataset were randomly cropped into 512×512 patches. Data augmentations (random rotation, random erasing, and Gaussian noise) were applied to mitigate class imbalance.

Baseline Comparisons: The proposed CRD-Net was systematically compared against several state-of-the-art models: Siamese-UNet [12], ChangeOS [37] (employing a ResNet-101 backbone), as well as Transformer-based models, including DamFormer [3] and DAHiTra [17]. In addition, Mamba-based approaches were evaluated using ChangeMamba [4] and FlowMamba [9].

Table 2: Quantitative Evaluation Results of Comparative Experiments on the xBD Dataset (%)

Method	Overall Performance			Damage F_1 Score per Class			
	F_1^{loc}	F_1^{cls}	F_1^{overall}	No damage	Minor damage	Major damage	Destroyed
Siamese-UNet	86.22	66.13	72.16	87.34	51.22	64.33	71.58
ChangeOS	85.49	71.13	75.44	89.11	53.11	72.45	80.79
DamFormer	86.29	73.29	77.19	94.96	54.50	72.67	83.75
DAHITra	85.69	72.76	76.64	93.98	54.22	72.83	82.25
ChangeMamba	85.91	76.68	79.45	94.90	59.10	76.12	86.44
FlowMamba	87.66	76.89	80.12	95.06	59.30	76.36	86.63
CRD-Net	87.10	78.76	81.26	95.73	62.68	77.81	86.70

4.4 Quantitative Analysis

As shown in Tab. 2, the proposed CRD-Net achieves state-of-the-art results on the xBD dataset, recording the highest overall F_1 score and damage classification F_1 score. Notably, CRD-Net outperforms the second-best approach by 1.14% in the overall metric.

Regarding the building localization task, FlowMamba attains the top extraction accuracy—leveraging multi-directional scanning with Mamba features—and edges out CRD-Net by 0.56% in the localization F_1 score. However, in the more critical task of damage classification, CRD-Net demonstrates a distinct superiority, surpassing the next best method by 1.87%.

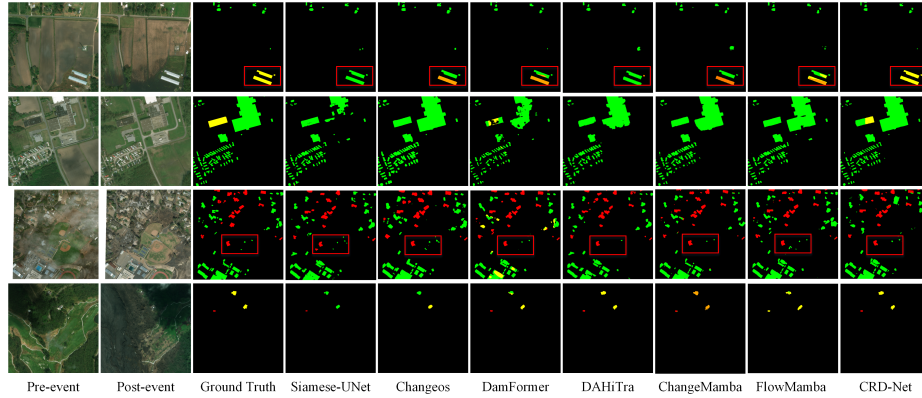
Furthermore, CRD-Net achieves the best scores across all four damage levels, with particularly notable improvements in the no-damage, minor-damage, and major-damage categories. For instance, CRD-Net outperforms the second-best method by 0.67%, 3.38%, and 1.45% in these three categories, respectively. Even in the destroyed category, CRD-Net maintains a slight edge, confirming its robust capability to model complex, multi-level building damage patterns.

As shown in Tab. 3, the proposed CRD-Net achieves state-of-the-art results on the EBD dataset, recording the best performance across all three key metrics: the overall, building localization, and damage classification F_1 scores. Notably, CRD-Net outperforms the second-best method by 1.33% in the overall F_1 score.

Notably, FlowMamba ranks second overall with an F_1 score of 80.55%, owing to its class-distance-aware loss function, which enhances discrimination among damage severity levels by explicitly penalizing confusion between adjacent damage classes. However, noticeable performance gaps remain between FlowMamba and CRD-Net in the minor-damage and major-damage categories, where CRD-Net outperforms FlowMamba by 2.54% and 2.32%, respectively. It is worth noting that in the destroyed category, DAHiTra achieves the highest score of 64.50%, surpassing CRD-Net by 2.92%. This suggests that while CRD-Net excels at identifying moderate damage levels, further improvement is needed for extreme destruction scenarios.

Table 3: Quantitative Evaluation Results of Comparative Experiments on the EBD Dataset (%)

Method	Overall Performance			Damage F ₁ Score per Class			
	F ₁ ^{loc}	F ₁ ^{cls}	F ₁ ^{overall}	No damage	Minor damage	Major damage	Destroyed
Siamese-UNet	86.72	63.02	70.13	84.21	74.98	60.25	46.16
ChangeOS	84.55	71.48	75.40	87.12	81.28	71.26	55.11
DamFormer	87.82	72.15	76.85	86.55	80.84	72.72	56.28
DAHiTra	85.68	74.99	78.20	85.96	80.75	72.36	64.50
ChangeMamba	89.09	74.99	79.22	89.71	84.99	70.95	61.23
FlowMamba	91.20	75.98	80.55	89.26	85.58	75.14	60.79
CRD-Net	92.00	77.54	81.88	90.18	88.12	77.46	61.58

**Fig. 3:** Building Damage Assessment Results on the xBD Dataset.

Based on the visual evidence in Fig. 3, the segmentation maps of the xBD dataset underscore the robust performance of CRD-Net. When evaluated alongside several state-of-the-art models, CRD-Net exhibits a superior spatial awareness, effectively isolating target features from high-interference environments.

As depicted in Fig. 4, the EBD dataset visualization results further demonstrate the effectiveness of CRD-Net. Compared with multiple competing algorithms, CRD-Net demonstrates stronger discriminative capability in distinguishing building damage categories and can more accurately characterize the differences among various damage levels. Whether in complex backgrounds, building areas with similar damage severity, or scenarios involving small targets and connected buildings, CRD-Net is able to produce more consistent and finer segmentation results, showing a clear overall visual superiority over other comparative methods.

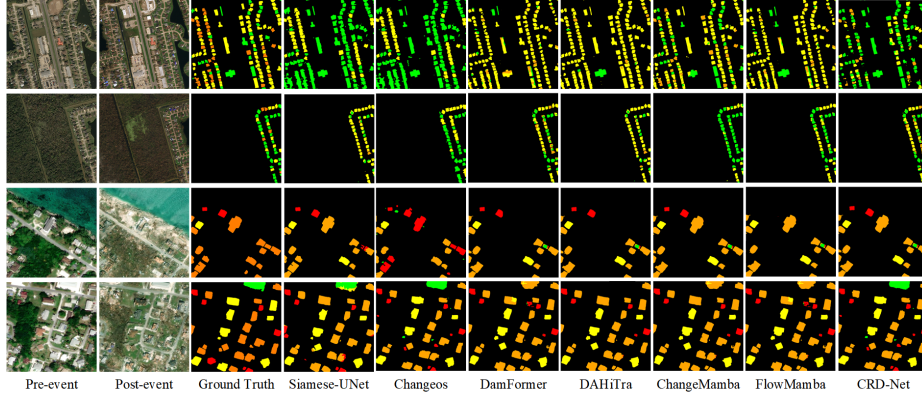


Fig. 4: Building Damage Assessment Results on the EBD Dataset.

Table 4: Ablation Study on the xBD Dataset. All configurations use the DINOv3-LoRA backbone.

Configuration	F_1^{loc}	F_1^{cls}	F_1^{overall}
DINOv3-Frozen	85.74	74.06	77.56
DINOv3-LoRA (baseline)	86.74	76.46	79.55
+ SFAFE	87.46	77.98	80.83
+ CRD ²	86.22	78.33	80.70
+ OSD only	87.33	77.08	80.16
+ PHCL only	87.25	77.60	80.50
+ SFAFE + CRD ² (Full)	87.10	78.76	81.26

4.5 Ablation Experiments

As shown in Table 4, the ablation study on the xBD dataset validates the step-wise effectiveness of the proposed modules. Except for the frozen baseline, all configurations share the same DINOv3-LoRA backbone, and “+” denotes modules added on top of this baseline (row 2), ensuring every comparison is controlled.

Compared to the frozen baseline (F_1^{overall} 77.56), introducing LoRA increases F_1^{overall} to 79.55, demonstrating the importance of downstream task adaptation. SFAFE and CRD² individually contribute +1.28 and +1.15 to Joint F_1 , with complementary strengths: SFAFE strengthens localization (+0.72 F_1^{loc}), while CRD² boosts classification (+1.87 F_1^{cls}). Within CRD², combining OSD and PHCL trades a small amount of F_1^{loc} for a substantial F_1^{cls} gain, as PHCL prioritizes inter-class separation over boundary precision. In the full model, SFAFE’s high-frequency cues complement CRD², achieving the best F_1^{cls} (78.76, +0.78 over SFAFE-only) at the cost of only a 0.36 drop in F_1^{loc} —a favourable trade-off that lifts Joint F_1 to its peak (81.26).

Table 5: Comparison of Model Complexity and Inference Efficiency.

Method	Params (M)	FLOPs (G)
Siamese-UNet	28.62	69.42
ChangeOS	56.25	170.23
DamFormer	47.32	71.69
DAHiTra	25.18	48.80
ChangeMamba	52.33	152.35
FlowMamba	68.52	130.80
CRD-Net	86.22	108.33

4.6 Computational Efficiency

Table 5 compares the model complexity of CRD-Net with several representative methods. CRD-Net has the largest number of parameters (86.22 M), mainly due to its dual-encoder design. However, its computational cost is not the highest: CRD-Net requires 108.33 G FLOPs, which is lower than ChangeOS (170.23 G), ChangeMamba (152.35 G), and FlowMamba (130.80 G) by 36.4%, 28.9%, and 17.2%, respectively. Although lighter models such as DAHiTra and Siamese-UNet are less computationally demanding, they trail CRD-Net by 4.62 and 9.10 percentage points in Joint F_1 on xBD, respectively. This suggests that CRD-Net achieves a favorable trade-off between model capacity, computational cost, and accuracy.

5 Conclusion

We propose CRD-Net, an efficient building damage assessment framework based on the DINOv3 model. To adapt DINOv3 to remote sensing scenarios, we employ parameter-efficient LoRA fine-tuning. We also introduce the SSAFE module to capture fine-grained local details and the novel CRD² Decoder—integrating Orthogonal Subspace Decoupling and Prototype-Guided Hard-sample Contrastive Learning—to tackle blurred boundaries, background noise, and imbalanced damage levels. Experiments on the xBD and EBD datasets show that CRD-Net achieves state-of-the-art (SOTA) performance in post-disaster evaluation.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2024YFC3015404). The authors acknowledge the use of large language models solely for English language editing and proofreading. All scientific content, analyses, and conclusions were developed and verified by the authors.

References

1. Braik, A.M., Koliou, M.: Automated building damage assessment and large-scale mapping by integrating satellite imagery, gis, and deep learning. *Computer-Aided Civil and Infrastructure Engineering* **39**(15), 2389–2404 (2024)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9650–9660 (2021)
3. Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L.: Dual-tasks siamese transformer framework for building damage assessment. In: *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. pp. 1600–1603. IEEE (2022)
4. Chen, H., Song, J., Han, C., Xia, J., Yokoya, N.: Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing* **62**, 1–20 (2024)
5. Chen, K., Chen, B., Liu, C., Li, W., Zou, Z., Shi, Z.: Rsmamba: Remote sensing image classification with state space model. *IEEE Geoscience and Remote Sensing Letters* **21**, 1–5 (2024)
6. Cheng, C.H., Hsu, C.C.: Changedino: Dinov3-driven building change detection in optical remote sensing imagery. *arXiv preprint arXiv:2511.16322* (2025)
7. Dong, S., Hu, Y., Wang, L., Chen, G., Meng, X.: Pefcd: Leveraging vision foundation models with parameter-efficient fine-tuning for remote sensing change detection. *arXiv preprint arXiv:2509.09572* (2025)
8. Gao, Y., Li, H., Yuan, F., Wang, X., Gao, X.: Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909* (2025)
9. Guo, X., Kang, X., Wang, Z., Wu, Q., Duan, P., Yang, B., Li, S.: Flowmamba: Building damage assessment via optics flow-based state space model. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
10. Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M.: Creating xbd: A dataset for assessing building damage from satellite imagery. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. pp. 10–17 (2019)
11. Han, Y., Li, J., Qu, Y., Wang, L., Pan, X., Huang, X.: Hfnet: semantic and differential heterogeneous fusion network for remote sensing image change detection. *Journal of Geovisualization and Spatial Analysis* **9**(1), 1 (2025)
12. Hao, H., Baireddy, S., Bartusiak, E.R., Konz, L., LaTourette, K., Gribbons, M., Chan, M., Delp, E.J., Comer, M.L.: An attention-based system for damage assessment using satellite imagery. In: *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. pp. 4396–4399. IEEE (2021)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
14. Hu, L., Lu, W., Yu, H., Yin, D., Sun, X., Fu, K.: Tea: A training-efficient adapting framework for tuning foundation models in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
15. Huang, S., Hou, Y., Liu, L., Yu, X., Shen, X.: Real-time object detection meets dinov3. *arXiv preprint arXiv:2509.20787* (2025)
16. Jiang, W., Sun, Y., Lei, L., Kuang, G., Ji, K.: Adaptvfms-rscd: Advancing remote sensing change detection from binary to semantic with sam and clip. *ISPRS Journal of Photogrammetry and Remote Sensing* **230**, 304–317 (2025)

17. Kaur, N., Lee, C.C., Mostafavi, A., Mahdavi-Amiri, A.: Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *Computer-Aided Civil and Infrastructure Engineering* **38**(15), 2072–2091 (2023)
18. 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. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4015–4026 (2023)
19. Liao, Z., Zhao, Y., Shan, X., Yan, Y., Liu, C., Lu, L., Ji, X., Chen, J.: Rt-detr4: Painlessly furthering real-time object detection with vision foundation models. *arXiv preprint arXiv:2510.25257* (2025)
20. Lin, C., Li, Y., Liu, Y., Wang, X., Geng, S.: Building damage assessment from post-hurricane imageries using unsupervised domain adaptation with enhanced feature discrimination. *IEEE Transactions on Geoscience and Remote Sensing* **60**, 1–10 (2021)
21. Lu, W., Wei, L., Nguyen, M.: Bitemporal attention transformer for building change detection and building damage assessment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**, 4917–4935 (2024)
22. Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.O., Huang, B.: Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
23. Ma, X., Zhang, X., Pun, M.O., Huang, B.: A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* (2025)
24. 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: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
25. Sha, X., Guo, Z., Sang, X., Wang, S., Zhao, Y.: Ges: A new building damage data augmentation and detection method based on extremely imbalanced data and unique spatial distribution of satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024)
26. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. *arXiv preprint arXiv:2508.10104* (2025)
27. Tan, K., Wu, Y., Yang, H., Ma, X.: Ugff-vlm: Uncertainty-guided and frequency-fused vision-language model for remote sensing farmland segmentation. *Remote Sensing* **18**(2), 282 (2026)
28. Vetrivel, A., Gerke, M., Kerle, N., Nex, F., Vosselman, G.: Disaster damage detection through synergistic use of deep learning and 3d point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS journal of photogrammetry and remote sensing* **140**, 45–59 (2018)
29. Wang, D., Zhang, J., Xu, M., Liu, L., Wang, D., Gao, E., Han, C., Guo, H., Du, B., Tao, D., et al.: Mtp: Advancing remote sensing foundation model via multitask pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **17**, 11632–11654 (2024)
30. Wang, L., Wu, J., Yang, Y., Tang, R., Ya, R.: Deep learning models for hazard-damaged building detection using remote sensing datasets: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024)

31. Wang, M., Zhou, L., Zhang, K., Li, X., Hao, M., Ye, Y.: Esam-cd: Fine-tuned efficient sam network with lora for weakly supervised remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
32. Wang, Z., Wu, C., Zhang, F., Xia, J.: Constructing an extensible building damage dataset via semi-supervised fine-tuning across 12 natural disasters. *Journal of Remote Sensing* **5**, 0733 (2025)
33. Wu, C., Zhang, F., Xia, J., Xu, Y., Li, G., Xie, J., Du, Z., Liu, R.: Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets. *Remote Sensing* **13**(5), 905 (2021)
34. Wu, J., Wang, Z., Hong, M., Ji, W., Fu, H., Xu, Y., Xu, M., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis* **102**, 103547 (2025)
35. Xu, J., Yang, W., Kong, L., Liu, Y., Zhou, Q., Zhang, R., Li, Z., Chen, W.M., Fei, B.: Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems* (2025)
36. Yang, S., Wang, H., Xing, Z., Chen, S., Zhu, L.: Segdino: An efficient design for medical and natural image segmentation with dino-v3. *arXiv preprint arXiv:2509.00833* (2025)
37. 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 Sensing of Environment* **265**, 112636 (2021)
38. Zhenga, K., Wu, Z., Wei, F., Zhou, M., Lie, K., Guo, H., Ding, L., Zhang, W., Dong, H.C.: Csd: Change semantic detection with only semantic change masks for damage assessment in conflict zones. *arXiv e-prints pp. arXiv-2511* (2025)
39. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832* (2021)
40. Zhu, D., Huang, X., Huang, H., Cheng, Q., Huang, Z., Shao, Z.: Changevit: Unleashing plain vision transformers for change detection in remote sensing images. *Pattern Recognition p. 112539* (2025)