

Multi-Stage Building Extraction from Staring Satellite Video under Disaster-Induced Appearance Degradation

Xingbo Huang, Huijiao Qiao* and Yuhe Zhu

College of Geological and Surveying Engineering, Taiyuan University of Technology, Taiyuan, China



INFORMATION

Keywords:
Building extraction
staring satellite video
multi-stage
disaster-induced appearance degradation
local–global feature fusion
prior guidance

DOI: 10.23967/j.rimni.2026.10.87997

SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

Multi-Stage Building Extraction from Staring Satellite Video under Disaster-Induced Appearance Degradation

Xingbo Huang, Huijiao Qiao* and Yuhe Zhu

College of Geological and Surveying Engineering, Taiyuan University of Technology, Taiyuan, China

ABSTRACT

Timely and accurate building extraction from disaster imagery is important for emergency response, damage assessment, and post-disaster recovery. Staring satellite video provides stage-specific observations of the same affected area as a disaster evolves. However, building appearances vary substantially across stages, from intact structures to deformed buildings and fragmented remnants, resulting in pronounced stage-dependent within-class variation. Existing methods are commonly developed using single-date imagery, and their reliability across heterogeneous damage conditions remains insufficiently investigated. To address this problem, we propose PGDGNet, a Prior-Guided Detail–Global Fusion Network for frame-wise multi-stage building extraction. PGDGNet employs a dual-branch CNN–Transformer architecture to capture local structural details and global semantic context. Its Detail–Global Feature Fusion (DGFF) module applies branch-specific attention and spatial gating to reconcile original and attention-enhanced features before local–global integration. A joint color-and-boundary prior-guidance strategy provides complementary appearance and contour cues when structural information is weakened, while boundary-aware supervision facilitates the delineation of incomplete buildings. A patch-level dataset is constructed from frames representing the early, intermediate, and late stages of the 2018 Indonesian tsunami–earthquake event. Experiments show that PGDGNet achieves the highest Building IoU and recall across all three stages. Compared with the strongest baseline, CFFormer, PGDGNet improves Building IoU by 0.59, 0.63, and 0.81 percentage points and recall by 1.45, 0.80, and 1.08 percentage points in the early, intermediate, and late stages, respectively. The results demonstrate robust and reliable performance across disaster stages.

OPEN ACCESS

Received: 26/06/2026

Accepted: 21/08/2026

Published: 21/09/2026

DOI
10.23967/j.rimni.2026.10.87997

Keywords:
Building extraction
staring satellite video
multi-stage
disaster-induced appearance degradation
local–global feature fusion
prior guidance

Abbreviations

DGFF	Detail–Global Feature Fusion
CNN	Convolutional Neural Network
PGDGNet	Prior-Guided Detail–Global Fusion Network
FCN	Fully Convolutional Network

*Correspondence: Huijiao Qiao (qiaohj@tyut.edu.cn) This is an article distributed under the terms of the Creative Commons BY-NC-SA license

DSM	Digital Surface Model
GFB	Global Feature Branch
LDB	Local Detail Branch
WTConv	Wavelet Convolutions
PE	Patch Embedding
PM	Patch Merging
BFBs	BuildFormer Blocks
W-LMHSA	Window-based linear multi-head self-attention
C-MLP	Convolutional Multilayer Perceptron
OA	Overall Accuracy
IoU	Intersection over Union
CA	Coordinate Attention
ECA	Efficient Channel Attention

1 Introduction

Natural disasters such as earthquakes and tsunamis can rapidly disrupt urban built environments, making timely building extraction important for emergency response and post-disaster assessment [1,2]. Staring satellite video-derived observations provide short-interval views of the same affected area, offering information beyond a single-date snapshot [3,4]. In the disaster event examined in this study, buildings exhibit distinct stage-dependent appearances: structures remain largely intact in the early stage, visible deformation emerges in the intermediate stage, and severe damage leaves many buildings fragmented in the late stage. These transformations disrupt building geometry, texture, and boundary continuity, producing substantial within-class appearance variation. As structural damage intensifies, local structural cues become incomplete and recognizable building patterns weaken. The central challenge is therefore to learn building representations that remain discriminative across representative stages of disaster-induced appearance degradation.

Traditional building extraction methods, including edge detection, geometric feature-based approaches, and object-oriented approaches, generally rely on handcrafted spectral, textural, and geometric features together with predefined rules [5,6]. Although these methods can be effective when buildings exhibit clear spectral contrast and regular geometric shapes, their fixed feature design limits their ability to accommodate substantial variations in building appearance.

Deep learning has significantly advanced building extraction by enabling hierarchical feature learning directly from remote sensing imagery. Early studies predominantly adopted fully convolutional and encoder–decoder architectures for pixel-level building prediction. Bittner et al. [7] proposed a multisource FCN architecture that jointly exploits panchromatic, multispectral, and stereo-DSM data, while Yu et al. [8] proposed ConvBNet, a modern pure-convolution architecture based on ConvNeXt, for building footprint extraction from very-high-resolution imagery. These studies demonstrate the capability of CNNs to capture roof textures, edges, and local geometric patterns when such cues are clearly preserved. However, CNN representations still depend strongly on local feature reliability, and long-range spatial relationships are modeled indirectly through stacked local operations [9,10]. To broaden the contextual range of CNN representations, subsequent studies introduced multiscale feature extraction and attention mechanisms. RU-Net [11], BMFR-Net [12] and EU-Net [13] enhance multiscale contextual learning through residual learning, atrous convolution, or dense spatial pyramid pooling. MAP-Net [14], RFA-UNet [15] and AttUNet+ [16] further improve feature discrimination by incorporating attention-based mechanisms. These methods enrich CNN

features under conventional building morphology, where buildings are usually spatially continuous and scale variation is a major challenge. However, they are not specifically designed for disaster scenes in which building regions become spatially discontinuous and structurally fragmented. In such cases, simply broadening the contextual range does not guarantee that fragmented regions can be consistently associated with the building category.

Transformer architectures provide a more explicit mechanism for modeling long-range spatial dependencies [17–19]. However, pure Transformer representations without sufficient local inductive biases may underrepresent the fine-grained structural details required for accurate building delineation. Consequently, CNNs and Transformers offer complementary representation capabilities, motivating dual-branch architectures that jointly exploit local structural information and global contextual semantics.

BuildFormer [20] introduced a CNN–Transformer dual-branch framework for building extraction, combining a CNN-based detail path with a Transformer-based global branch through hierarchical feature fusion. Subsequent studies further improved cross-branch interaction, but many still integrate the two branches through direct aggregation or unified post-fusion enhancement. DSFENet aligns Transformer representations with CNN feature maps and fuses them mainly through concatenation before redistributing the aggregated features to both branches [21]. To enhance cross-branch feature interaction, attention-guided fusion and nonlinear transformation mechanisms were further introduced into CNN–Transformer architectures. DPENet [22] employs concatenation followed by channel attention within a dual-path interactive fusion module, while MDTrans [23] introduces attention-guided fusion to enhance local–global interaction after feature aggregation. DFRTNet [24] further adopts an MLP-based strategy to model nonlinear interactions between local and global representations. Although these methods improve CNN–Transformer collaboration, their fusion mechanisms generally treat CNN-derived local features and Transformer-derived global features as complete streams to be aggregated or reweighted. They do not explicitly refine each branch according to its own representation characteristics before fusion. This limitation becomes more critical in satellite video-derived multi-stage disaster observations, where buildings progressively change from intact structures to deformed buildings and fragmented remnants. In this process, local structural details may become fragmented and spatially unreliable, while global semantic responses may also become ambiguous as regular building shapes are gradually lost. Directly fusing such heterogeneous features may introduce unreliable or redundant responses into the fused representation. These observations motivate the exploration of pre-fusion branch refinement, rather than direct aggregation, for integrating CNN-derived local details and Transformer-derived global context under progressive structural degradation.

Prior knowledge provides a complementary source of information for building extraction. Existing priors can be broadly categorized according to their primary roles into structural and scene-context cues. Structural priors include geometry-aware methods [25,26], which explicitly incorporate building shape information and geometric significance cues to enhance structural completeness and preserve regular building contours during extraction, and edge-aware methods [27,28], such as gradient-based contour information or multiscale DoG-derived edge responses, to facilitate accurate contour localization and building boundary delineation. Scene-context priors exploit relationships between buildings and their surroundings. For example, Ok [29] incorporated shadow evidence into a graph-cut framework to distinguish buildings from surrounding objects in complex urban scenes. However, many existing priors depend on coherent geometry, continuous boundaries, or stable contextual relationships. The reliability of these cues decreases as disaster-induced damage progresses. Even when overall structural regularity is disrupted, color-derived appearance cues and partial boundary

evidence may remain observable in some damaged regions. Their complementary use may support the discrimination of building regions and the delineation of their contours, motivating their incorporation as auxiliary priors for multi-stage building extraction.

Beyond feature representation and prior guidance, existing task settings rarely provide sufficient support for studying building extraction throughout disaster evolution. Conventional building extraction studies are mostly based on single-date imagery, while disaster-oriented change detection and damage assessment studies often use pre- and post-event image pairs [30,31]. These settings capture either one building state or two endpoint states, but provide limited information about the intermediate process in which buildings gradually deform, collapse, and fragment. Staring satellite video-derived observations help fill this gap by providing short-interval observations of the same affected area. This capability motivates the multi-stage building extraction problem studied in this work: how to accurately identify buildings as their appearance changes from intact structures to deformed buildings and fragmented remnants.

To address these task- and representation-level challenges, we propose PGDGNet, a Prior-Guided Detail–Global Fusion Network for multi-stage building extraction from disaster satellite video-derived observations. PGDGNet adopts a dual-branch CNN–Transformer architecture to capture local structural details and global contextual semantics. Its core DGFF module refines heterogeneous local and global features before fusion, reducing the propagation of unreliable responses caused by fragmented structures and ambiguous semantics. A joint color-and-boundary prior-guidance module further provides auxiliary appearance and contour cues when coherent building geometry is disrupted.

The main contributions are summarized as follows:

(1) We investigate a multi-stage building extraction problem from disaster satellite video-derived observations, aiming to accurately extract building-related regions with substantially different appearance and structural forms within the same disaster event. This setting focuses on buildings ranging from intact structures to deformed buildings and fragmented remnants, addressing stage-dependent within-class appearance variation.

(2) We propose a Detail–Global Feature Fusion (DGFF) module to improve local–global feature integration under structural degradation. Instead of directly aggregating CNN-derived local features and Transformer-derived global features, DGFF performs branch-specific refinement and intra-branch gated residual reconciliation before fusion.

(3) We introduce a joint color-and-boundary prior-guidance strategy to provide complementary appearance and contour cues when regular building geometry becomes unreliable. Boundary-aware supervision is further incorporated to improve the delineation of buildings with incomplete or ambiguous boundaries.

2 Methodology

In this section, the proposed methodology is described in detail.

2.1 Overview

PGDGNet adopts a dual-branch architecture consisting of a Global Feature Branch (GFB) and a Local Detail Branch (LDB) to collaboratively model global semantic context and fine-grained local structural details. The RGB image is concatenated with the RBW color guidance map and boundary guidance map to form a five-channel input to the LDB. A Detail-Global Feature Fusion (DGFF) module adaptively integrates local detail and global semantic representations from the two branches

to enhance feature complementarity. The fused features are decoded via an FPN-style decoder to produce the final segmentation map. The overall architecture is illustrated in Fig. 1.

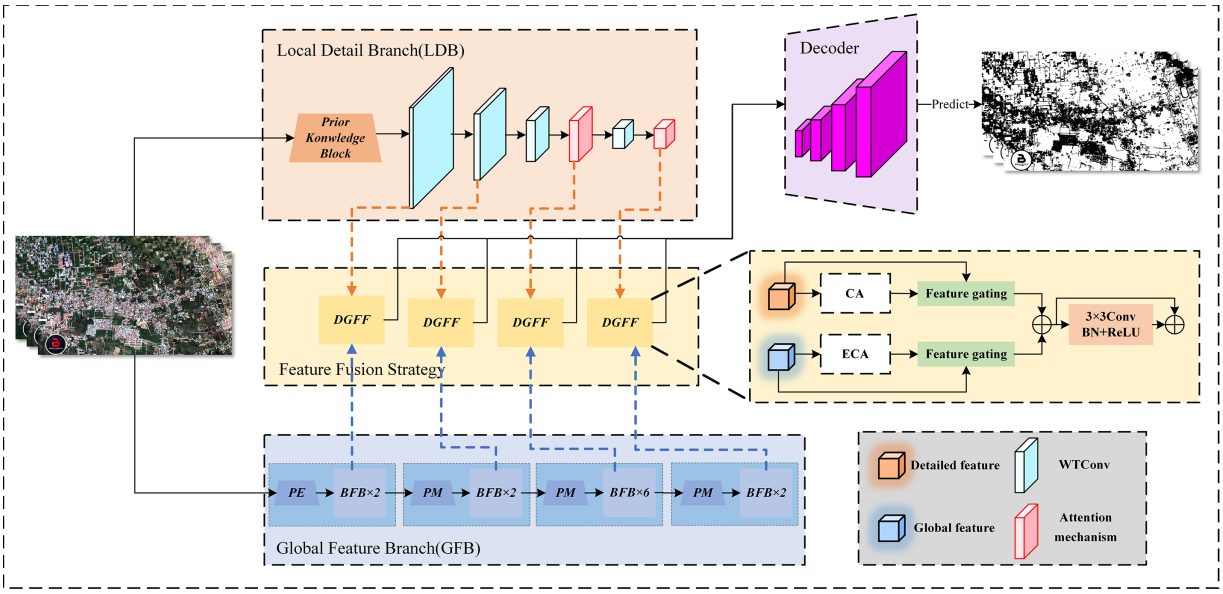


Figure 1: Overall architecture of the proposed PGDGNet

2.2 LDB Backbone

The LDB is built upon a CNN backbone to extract fine-grained spatial features and structural details, comprising three key components: a WTConv-based feature extraction backbone, a prior-guided input enhancement module, and CBAM-based attention recalibration. Standard convolutions in conventional CNN branches are limited in capturing the multi-scale building representations required in disaster imagery, where intact large-scale structures coexist with small-scale fragmented remnants across the appearance degradation spectrum. WTConv is therefore introduced to enhance multi-scale feature representation through wavelet-based decomposition, while CBAM is incorporated after the third and fourth convolutional layers to adaptively recalibrate channel and spatial attention responses, suppressing background interference under complex disaster conditions.

2.2.1 WTConv Block

WTConv performs discrete wavelet decomposition on input features to obtain multi-frequency sub-bands, applies depthwise convolution in the wavelet domain, and reconstructs features via inverse wavelet transform. The structure of the WTConv block is illustrated in Fig. 2.

Given an input feature $X \in R^{C\times H\times W}$, WTConv first performs discrete wavelet decomposition to obtain four frequency sub-bands:

$$X_s = X * f_s, \{s \in LL, LH, HL, HH\} \tag{1}$$

where f_{LL} denotes the low-pass filter and the remaining filters correspond to high-frequency components.

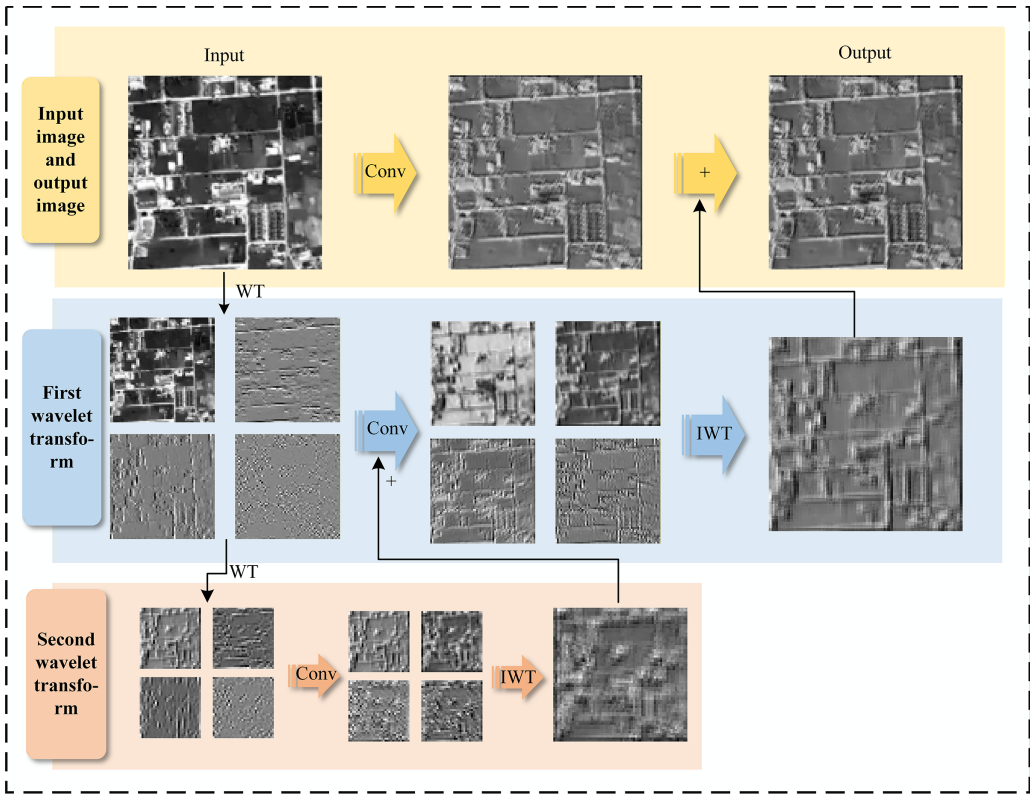


Figure 2: Feature extraction with the WTConv module

At each level i, the sub-bands are concatenated along the channel dimension and processed by depthwise convolution:

$$X^{(i)} = Concat_c\left[X_{LL}^{(i)}, X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)}\right] \tag{2}$$

$$Y^{(i)} = Conv_{dw}\left(W^{(i)}, X^{(i)}\right) \tag{3}$$

The final output fuses spatial and frequency domain features through residual connection:

$$Y = Conv_{dw}(X) + Z^{(0)} \tag{4}$$

Through wavelet decomposition, WTConv separates low- and high-frequency components, facilitating the extraction of coarse structural information and fine-grained boundary details while enlarging the effective receptive field at low computational cost.

2.2.2 Prior-Guided Module

In disaster imagery, structural damage progressively eliminates the textural and geometric regularity that conventional priors rely upon. As established in Section 1, two types of low-level cues remain useful as complementary information under severe damage conditions in the study area. First, building rooftops exhibit relatively consistent color distributions, predominantly white, red, and blue, that remain distinguishable from surrounding backgrounds across all disaster stages. Second, although complete building boundaries are progressively lost as structural damage advances, building

fragmentation may produce locally concentrated and geometrically irregular boundary responses in building-related regions, providing complementary structural cues for identifying damaged building remnants. providing complementary structural guidance that helps distinguish building remnants from surrounding low-gradient ground surfaces. These cues are incorporated as auxiliary guidance channels rather than explicit predictions, providing coarse spatial constraints to encourage the network to focus on building-sensitive regions during feature learning. An illustration of the prior knowledge is shown in Fig. 3.

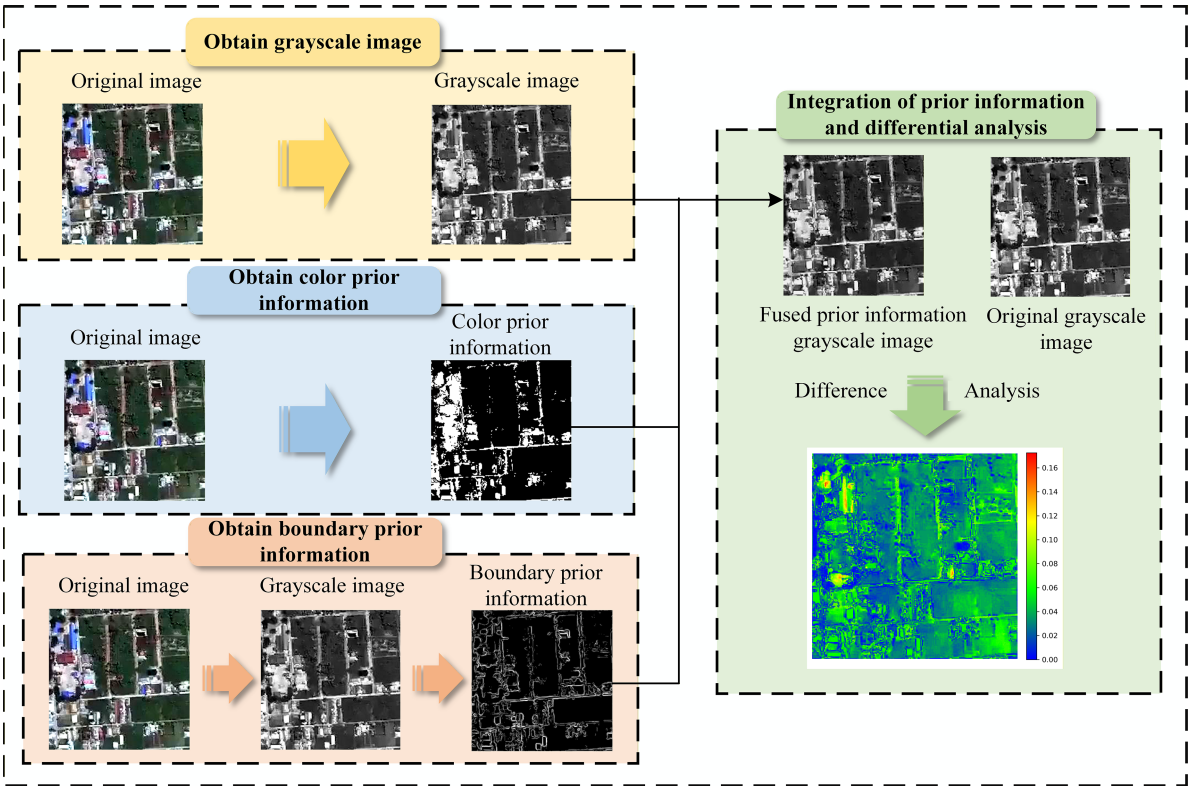


Figure 3: Color and boundary prior extraction and fusion

1. Color prior knowledge

Building rooftops in the Petobo study area show relatively distinct color characteristics, with many white, red, and blue roofs distinguishable from the surrounding background in the selected disaster frames. Based on observations from the training frames, the HSV thresholds used in this study were empirically determined to construct a coarse red–blue–white (RBW) color prior. The testing frames were not used for threshold adjustment. The predefined thresholds are not used as decision rules for final building extraction; instead, they generate an RBW color guidance map that provides complementary color cues for network feature learning. The RBW guidance map is generated in the HSV color space using the following threshold-based masks:

$$M_{red} = [(h < 0.05) \vee (h > 0.90)] \wedge (s > 0.3) \wedge (v > 0.2) \tag{5}$$

$$M_{blue} = (h > 0.55) \wedge (h < 0.75) \wedge (s > 0.3) \wedge (v > 0.2) \tag{6}$$

$$M_{white} = (s < 0.25) \wedge (v > 0.7) \tag{7}$$

where h denotes Hue, s denotes Saturation, and v denotes Value. The resulting guidance map provides coarse spatial constraints for building-sensitive regions under the scene-specific appearance characteristics of the study area.

The three-color masks are further combined to obtain a unified RBW color guidance map highlighting candidate building-sensitive regions in the image. The formulation is defined as follows:

$$M_{RBW} = \text{clip}_{[0,1]}\left(M_{red} + M_{blue} + M_{white}\right) \tag{8}$$

2. Boundary prior knowledge

Progressive building fragmentation in the intermediate and late disaster stages produces densely distributed and geometrically irregular boundaries, increasing the difficulty of learning reliable structural representations from RGB imagery alone. Boundary-aware cues are extracted using the Sobel operator applied to the grayscale image:

$$E_{\text{raw}} = \sqrt{\nabla_x^2 + \nabla_y^2 + \epsilon}\ \left(\epsilon = 10^{-6}\right) \tag{9}$$

Boundary responses are normalized and filtered to suppress background clutter:

$$E_{th}(p) = \begin{cases} E_{norm}(p), & E_{norm}(p) > \tau \\ 0, & \text{otherwise} \end{cases}, \tau = 0.2 \tag{10}$$

$$F_{edge} = \frac{E_{th}}{max\left(E_{th}\right) + \epsilon} \tag{11}$$

Finally, the boundary guidance map is concatenated with the RGB image and the RBW color guidance map to form the five-channel input to the LDB:

$$X_{\text{input}} = \text{Concat}\left[X_{\text{RGB}}, F_{\text{edge}}, M_{\text{RBW}}\right] \tag{12}$$

where $X_{RGB} \in R^{H\times W\times 3}$ denotes the RGB image, $F_{\text{edge}} \in R^{H\times W\times 1}$ represents the one-channel boundary guidance map, and $M_{RBW} \in R^{H\times W\times 1}$ denotes the one-channel RBW color guidance map. These features are concatenated along the channel dimension to construct the fused representation $X_{\text{input}} \in R^{H\times W\times 5}$, providing complementary structural and appearance-aware cues for enhanced local feature representation.

2.3 GFB Backbone

The GFB serves as the global semantic extraction pathway and collaborates with the LDB to form a global–local feature learning framework. Considering the fragmented structures and complex background interference commonly observed in disaster imagery, effective modeling of long-range semantic dependencies is crucial for robust building extraction. Therefore, the GFB is constructed based on the global Transformer branch of BuildFormer, which enables hierarchical global context modeling. Specifically, the GFB consists of a Patch Embedding (PE) layer, multiple BuildFormer Blocks (BFBs), and Patch Merging (PM) layers. The PE layer generates overlapping patches to retain local structural information, while PM layers progressively enlarge the receptive field and construct hierarchical semantic representations. Each BFB integrates window-based linear multi-head self-attention (W-LMHSA) and a convolutional multilayer perceptron (C-MLP), which jointly capture contextual dependencies and facilitate semantic interaction across spatial regions. As a result, the GFB provides robust global semantic representations for subsequent feature fusion.

2.4 DGFF Module

As illustrated in Fig. 4, DGFF is designed to fuse heterogeneous features from the Local Detail Branch (LDB) and Global Feature Branch (GFB).

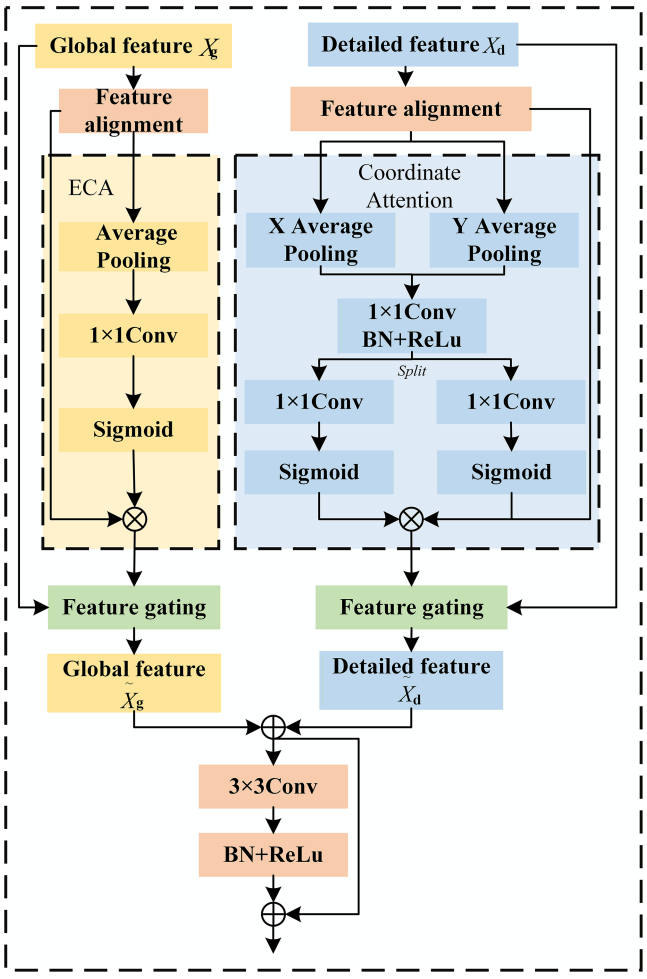


Figure 4: Architecture of the DGFF module

Unlike conventional fusion strategies that mainly rely on direct aggregation or post-aggregation enhancement, DGFF refines each branch before local–global feature interaction. This design is motivated by the different representation characteristics of CNN and Transformer features under disaster-induced appearance degradation. CNN features mainly encode fine-grained structural details, boundary cues, and positional information, which are sensitive to spatial discontinuities caused by building fragmentation. Transformer features provide high-level semantic representations through long-range contextual modeling, where informative semantic channels are more important for global scene understanding.

Based on these characteristics, Coordinate Attention is adopted in the CNN branch to enhance local structural responses while preserving positional information, whereas Efficient Channel Attention (ECA) is employed in the Transformer branch to recalibrate informative semantic channels with low computational cost. Given the input global feature and local feature, both are first projected into

a unified feature space with the same channel dimension and spatial resolution. ECA and Coordinate Attention are then applied to the global feature X_g and local feature X_l, respectively, producing attention-enhanced branch representations $X_g^{'}$ and $X_l^{'}$.

After attention refinement, DGFF introduces an intra-branch gated residual refinement unit. The purpose of this gate is not to directly assign weights between the CNN and Transformer branches, but to adaptively reconcile the original feature and its attention-enhanced counterpart within each branch. This design preserves reliable original responses while selectively incorporating enhanced information, reducing the propagation of over-enhanced or unreliable responses before cross-branch fusion. The channel-aware spatial gating maps are generated from the attention-enhanced local feature $X_l^{'}$ and global feature $X_g^{'}$ and used to blend them with their corresponding original branch representations, as formulated below:

$$G_l = \sigma\left(W_l * X_l^{'}\right) \tag{13}$$

$$G_g = \sigma\left(W_g * X_g^{'}\right) \tag{14}$$

where W_1 and W_g denote learnable 1×1 convolutional weights, and $\sigma(\cdot)$ represents the Sigmoid activation function. The generated gating maps G_l and G_g, both with dimensions of $B \times C_{\text{mid}} \times H \times W$, have the same dimensionality as their corresponding attention-enhanced feature maps $X_l^{'}$ and $X_g^{'}$, respectively. Each element of a gating map represents an adaptive residual weighting coefficient for a specific channel and spatial location. The gating values are constrained to the range [0, 1], where larger values assign greater confidence to the attention-enhanced features, while smaller values preserve a larger proportion of the original branch representations. Consequently, the gating maps dynamically balance the contributions of enhanced and original features through channel-aware spatially adaptive residual blending. The gated fusion process is formulated as follows:

$$\tilde{X}_l = G_l \otimes X_l^{'} + (1 - G_l) \otimes \text{Interpolate}\,(X_l) \tag{15}$$

$$\tilde{X}_g = G_g \otimes X_g^{'} + \left(1 - G_g\right) \otimes \text{Interpolate}\left(X_g\right) \tag{16}$$

where $\otimes$ denotes element-wise multiplication and Interpolate $(\cdot)$ represents bilinear interpolation for spatial alignment. Through adaptive residual refinement, the gating mechanism effectively suppresses feature dominance caused by heterogeneous representations while preserving complementary semantic and structural information from both branches.

Finally, the gated global and local features $\tilde{X}_g$ and $\tilde{X}_l$ are fused through element-wise summation. To further refine the integrated representation after cross-branch fusion, a convolutional refinement block consisting of a 3×3 convolution, batch normalization, and ReLU activation is applied after feature summation. A residual connection is subsequently introduced to combine the refined representation with the original fused feature, facilitating feature preservation and stable information propagation during multi-branch integration.

2.5 Decoder

To obtain the final semantic segmentation results, the fused features output from the DGFF module are further decoded. In this study, a decoding strategy similar to that of Feature Pyramid Networks (FPN) is adopted for feature fusion. The detailed architecture is illustrated in Fig. 5. Four 1×1 convolution layers project encoder features to a unified channel dimension, after which higher-level features are progressively upsampled and aggregated with lower-level features via element-wise summation. The aggregated features are fed into the segmentation head for final prediction.

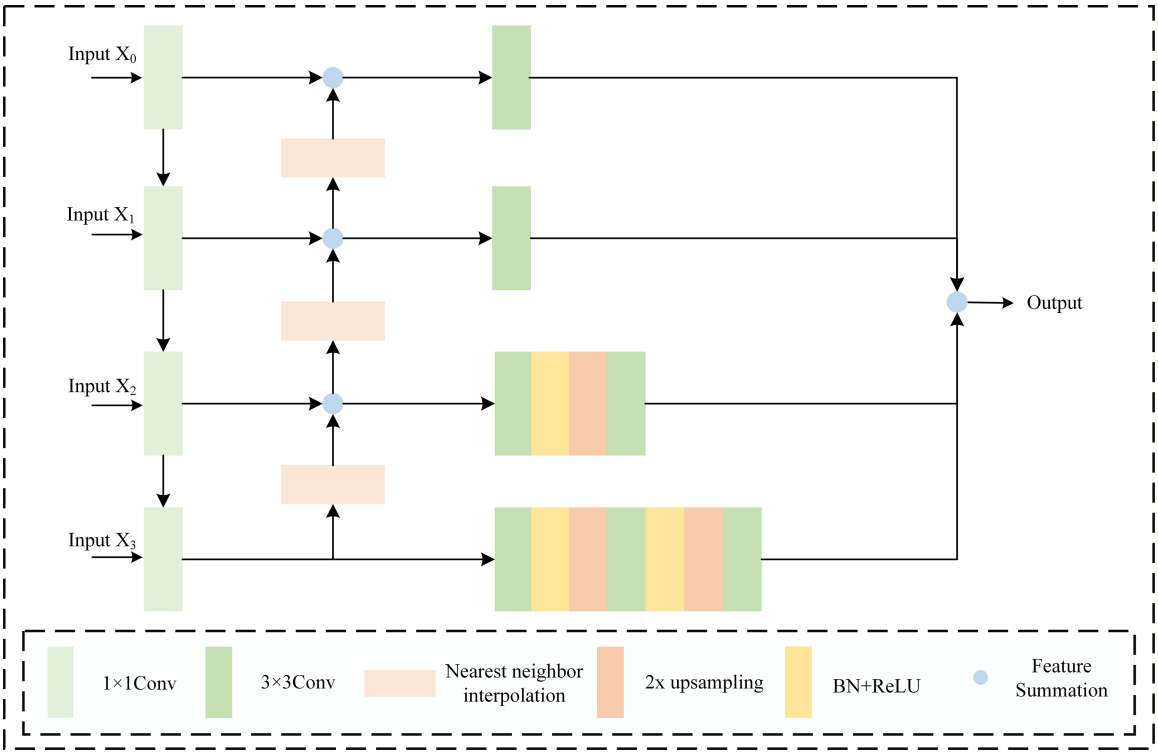


Figure 5: Architecture of the decoder

2.6 Loss Function

The model is optimized using a joint loss combining a main segmentation loss and a boundary-aware loss:

$$L = L_{main} + \lambda_{edge} \cdot L_{edge} \tag{17}$$

$$L_{\text{main}} = L_{\text{focal}} + L_{\text{Dice}} \tag{18}$$

Focal loss addresses class imbalance by emphasizing hard samples, while Dice loss improves region-level overlap for small and fragmented building regions. The boundary loss is defined as:

$$L_{edge} = \text{BCE}\left(Boundary_{pred}, Boundary_{gt}\right) \tag{19}$$

where boundary maps are extracted using a 3×3 Laplacian operator. For the predicted segmentation, the network logits are first converted into class probabilities using the softmax operation, and the Laplacian operator is then applied to the continuous building-class probability map. The absolute Laplacian response is subsequently mapped to a soft predicted boundary map using the hyperbolic tangent function. The ground-truth boundary map is extracted using the same Laplacian operator.

A fixed boundary loss weight is suboptimal as large weights impede early convergence while small weights limit boundary refinement in later stages. A dynamic weighting factor λ_{edge} is therefore introduced:

$$\lambda_{edge} = \min\left(5 + 10 \cdot \frac{epoch}{T} + 10\left(1 - Dice_{edge}\right), \text{max factor}, \frac{L_{main}}{L_{edge} + \varepsilon}\text{max factor}\right) \tag{20}$$

where T denotes the total number of training epochs. The dynamic weighting factor consists of three components. First, a constant base weight of 5 maintains boundary supervision throughout training. Second, a linearly increasing term gradually increases the contribution of the boundary loss as training proceeds, encouraging the optimization to shift from coarse region segmentation to finer boundary refinement when segmentation predictions become more reliable. Third, an adaptive term adjusts the weight according to the current boundary Dice score: lower boundary accuracy leads to a larger boundary-loss weight, whereas improved boundary prediction reduces the adaptive contribution to avoid overemphasizing the auxiliary task. The scaling coefficients were heuristically selected through preliminary experiments to balance convergence stability and boundary refinement. The maximum factor was set to 15 with reference to BuildFormer, where the boundary loss uses a fixed weight of 10. Since the proposed weighting function dynamically increases boundary supervision according to training progress and boundary quality, a slightly larger upper bound allows stronger boundary refinement when needed, while still preventing the auxiliary boundary loss from dominating the overall optimization.

The boundary quality is measured by:

$$Dice_{edge} = \frac{2|Boundary_{pred}^{hard} \cap Boundary_{gt}|}{|Boundary_{pred}^{hard}| + |Boundary_{gt}|} \tag{21}$$

where $Boundary_{pred}^{hard}$ denotes the binary predicted boundary map extracted from the discrete segmentation obtained by argmax, and Dice denotes the standard Dice coefficient computed between the binary predicted and ground-truth boundary maps.

During training, the dynamic weight is recalculated for every mini-batch according to the current training epoch, the boundary Dice score, and the relative magnitudes of the main segmentation loss and the boundary loss. Since the weighting factor is computed from detached quantities, it is treated as a non-trainable scaling coefficient rather than a learnable parameter. Consequently, gradients are propagated through both L_{main} and L_{edge}, while the dynamically updated weighting factor itself is excluded from gradient computation.

3 Result and Discussion

This section presents the datasets, experimental settings, comparative experiments, and ablation studies.

3.1 Experimental Data and Details

The experimental data were derived from a publicly available satellite video of the Petobo region, Indonesia, during the Mw 7.5 earthquake and tsunami event on 28 September 2018. The video was released as a synthesized satellite video product generated from WorldView/DigitalGlobe imagery and was converted into sequential image frames for analysis. Invalid segments unrelated to the study area were removed, resulting in 301 valid frames. Since the publicly released video does not provide original sensor-level metadata, georeferenced spatial resolution, or satellite temporal resolution, all experiments in this study were conducted at the frame-pixel level. The disaster process was divided into early, intermediate, and late stages according to visual interpretation of building damage. The early stage mainly contains buildings with limited deformation, the intermediate stage contains buildings with visible deformation or partial collapse, and the late stage is characterized by severely damaged buildings, debris accumulation, and fragmented building remnants. Two representative frames were selected from each stage: Frames 160 and 165 for the early stage, Frames 290 and 295 for the

intermediate stage, and Frames 420 and 425 for the late stage. Building-related regions were annotated as a single semantic category, including intact buildings, partially damaged structures, and identifiable fragmented remnants spatially attributable to original buildings. The ground-truth labels were first generated using an SVM-assisted procedure based on color and texture features and then refined through manual visual interpretation. Across the six selected frames, the average proportions of building and background pixels were 34.67% and 65.33%, respectively. The selected frames and their corresponding labels were cropped into 256×256 patches using a sliding-window strategy with a stride of 64 pixels, producing 75% overlap between adjacent patches. For each disaster stage, 3000 image–label pairs were generated, resulting in 9000 pairs in total. Frames 160, 290, and 420 were used to generate training samples, while Frames 165, 295, and 425 were used for testing. The training samples were further divided into training and validation subsets at a ratio of 8:2.

The experimental settings used in this study are summarized in Table 1. The input resolution was set to 256×256. The Adam optimizer was adopted with an initial learning rate of 0.0001. The batch size was set to 10, and each model was trained for 100 epochs.

Table 1: Experimental setup and hardware/software configuration

Type	Version
CPU	Intel(R) Core (TM) i7-14700KF
GPU	NVIDIA GeForce RTX 4090D
Memory	64 GB
Operating system	Windows 11
Framework	PyTorch 2.5.0
Python version	3.9.20

3.2 Comparative Experiments

To validate the effectiveness of the proposed method across different disaster stages, comparative experiments were conducted with five representative segmentation models: U-Net and DeepLabv3+ as classical CNN-based baselines, Swin-UNet as a pure Transformer-based architecture, UNetFormer as a CNN-Transformer hybrid architecture, and CFFormer as a dual-branch CNN-Transformer hybrid architecture. Notably, CFFormer adopts a dual-branch design, in contrast to the single-branch architectures of UNetFormer and Swin-UNet.

3.2.1 Quantitative Results

To quantitatively evaluate the proposed method, we report building-specific metrics, including Building IoU, F1-score, Precision, and Recall, together with overall metrics, including OA, Kappa, and mIoU. Boundary F1-score is further used to assess boundary delineation accuracy. To evaluate experimental reliability, all methods were independently trained and tested using three random seeds, and the results in Tables 2–4 are reported as Mean $\pm$ Std.

Table 2: Comparison of building extraction results from different methods for early-stage disaster imagery

Method	Building-specific metrics (%)			Overall evaluation metrics (%)				Boundary F1 (%)
	IoU	F1-Score	Precision	Recall	OA	Kappa	mIoU	
U-Net	88.00 ± 0.89	93.62 ± 0.51	96.49 ± 0.47	90.91 ± 0.64	95.45 ± 0.36	90.08 ± 0.78	90.58 ± 0.70	90.67 ± 1.13
DeepLab v3+	86.83 ± 0.35	92.95 ± 0.20	96.55 ± 0.46	89.61 ± 0.50	95.01 ± 0.14	89.09 ± 0.30	89.69 ± 0.27	90.84 ± 0.33
Swin-Unet	87.46 ± 0.77	93.31 ± 0.44	95.97 ± 0.31	90.79 ± 0.61	95.22 ± 0.31	89.59 ± 0.67	90.15 ± 0.61	90.61 ± 0.91
UNetFormer	85.51 ± 1.13	92.19 ± 0.66	95.35 ± 0.53	89.22 ± 0.77	94.44 ± 0.46	87.80 ± 1.01	88.62 ± 0.89	89.11 ± 1.55
CFFormer	88.96 ± 0.06	94.16 ± 0.04	96.89 ± 0.14	91.57 ± 0.09	95.83 ± 0.03	90.92 ± 0.06	91.34 ± 0.06	92.01 ± 0.11
Ours	89.55 ± 0.20	94.49 ± 0.11	96.02 ± 0.98	93.02 ± 0.96	96.02 ± 0.08	91.37 ± 0.17	91.75 ± 0.15	92.47 ± 0.31

Table 3: Comparison of building extraction results from different methods for intermediate-stage disaster imagery

Method	Building-specific metrics (%)			Overall evaluation metrics (%)				Boundary F1 (%)
	IoU	F1-Score	Precision	Recall	OA	Kappa	mIoU	
U-Net	90.52 ± 0.37	95.02 ± 0.20	98.83 ± 0.15	91.49 ± 0.26	96.10 ± 0.16	91.83 ± 0.33	92.16 ± 0.30	92.15 ± 0.50
DeepLab v3+	88.73 ± 0.51	94.03 ± 0.28	98.50 ± 0.14	89.94 ± 0.62	95.35 ± 0.21	90.24 ± 0.44	90.70 ± 0.40	91.06 ± 0.61
Swin-Unet	90.25 ± 0.52	94.88 ± 0.29	98.80 ± 0.10	91.25 ± 0.48	95.99 ± 0.22	91.60 ± 0.46	91.94 ± 0.42	92.24 ± 0.70
UNetFormer	87.85 ± 0.60	93.53 ± 0.34	97.92 ± 0.12	89.53 ± 0.63	94.97 ± 0.25	89.42 ± 0.54	89.97 ± 0.48	89.66 ± 0.86
CFFormer	91.57 ± 0.76	95.60 ± 0.42	98.98 ± 0.27	92.44 ± 0.69	96.54 ± 0.32	92.75 ± 0.67	93.01 ± 0.62	93.59 ± 0.69
Ours	92.20 ± 0.42	95.94 ± 0.22	98.81 ± 0.37	93.24 ± 0.60	96.79 ± 0.16	93.30 ± 0.36	93.52 ± 0.34	94.30 ± 0.37

Table 4: Comparison of building extraction results from different methods for late-stage disaster imagery

Method	Building-specific metrics (%)			Overall evaluation metrics (%)				Boundary F1 (%)
	IoU	F1-Score	Precision	Recall	OA	Kappa	mIoU	
U-Net	88.55 ± 0.24	93.93 ± 0.14	98.80 ± 0.12	89.51 ± 0.15	96.47 ± 0.08	91.44 ± 0.19	91.84 ± 0.17	88.83 ± 0.34
DeepLab v3+	86.95 ± 0.50	93.02 ± 0.29	98.59 ± 0.06	88.05 ± 0.55	95.97 ± 0.15	90.20 ± 0.39	90.72 ± 0.35	87.42 ± 0.66
Swin-Unet	88.21 ± 0.57	93.74 ± 0.33	98.96 ± 0.10	89.03 ± 0.51	96.37 ± 0.18	91.19 ± 0.45	91.62 ± 0.40	88.54 ± 0.82
UNetFormer	86.15 ± 0.32	92.56 ± 0.19	98.08 ± 0.02	87.63 ± 0.33	95.70 ± 0.10	89.55 ± 0.25	90.14 ± 0.22	85.95 ± 0.42
CFFormer	88.60 ± 0.18	93.96 ± 0.10	98.98 ± 0.19	89.42 ± 0.05	96.49 ± 0.06	91.49 ± 0.15	91.88 ± 0.14	89.13 ± 0.23
Ours	89.41 ± 0.45	94.41 ± 0.25	98.67 ± 0.31	90.50 ± 0.71	96.73 ± 0.13	92.11 ± 0.34	92.45 ± 0.30	90.20 ± 0.60

In the early stage, buildings remain largely intact, but complex backgrounds still cause building–background confusion. As shown in Table 2, CFFormer is the strongest baseline, with a Building IoU of 88.96%, mIoU of 91.34%, and Boundary F1-score of 92.01%. PGDGNet improves these values to 89.55%, 91.75%, and 92.47%, respectively, indicating improved region extraction and boundary delineation under complex backgrounds. In the intermediate stage, most methods achieve higher accuracy than in the early stage. This may be because many buildings are partially affected but still retain recognizable structural and textural cues, while their contrast with the surrounding background

becomes more distinguishable. As shown in Table 3, PGDGNet ranks first in most evaluation metrics. In particular, its Boundary F1-score reaches 94.30%, outperforming CFFormer and indicating more accurate boundary localization. In the late stage, severe collapse, structural fragmentation, and irregular appearances increase extraction difficulty. Nevertheless, PGDGNet maintains clear advantages, improving Building IoU and mIoU over CFFormer by 0.81 and 0.57 percentage points, respectively. Paired t-tests were conducted between PGDGNet and CFFormer for Building IoU and mIoU over three independent runs. PGDGNet achieves higher means across all stages, with statistically significant improvements in the late stage ($p = 0.0337$ for Building IoU; $p = 0.0283$ for mIoU) and a near-significant trend in the early stage. The intermediate-stage differences are not significant due to the close performance of the two methods. Overall, PGDGNet shows the most consistent advantage under severe structural degradation.

Overall, PGDGNet achieves the strongest comprehensive performance across the three disaster stages, with small standard deviations over repeated runs. Although its Precision is slightly lower than that of CFFormer in the intermediate and late stages, PGDGNet consistently achieves higher Recall, Building IoU, mIoU, and Boundary F1-score. This indicates that the proposed method recovers more building regions while reducing omission errors.

3.2.2 Qualitative Results

The visual results in Figs. 6–8 are consistent with the quantitative findings. In the early stage, the proposed method accurately extracts complete building regions with clear boundaries. In the intermediate stage, it preserves structural continuity despite partial collapse and debris interference. In the late stage, it more effectively captures residual building structures under severe fragmentation.

To provide a more detailed comparison of the experimental results, several representative regions were selected from the early, intermediate, and late stages of the disaster imagery for local detailed visualization. Fig. 9 presents zoomed-in views of imagery at different disaster stages along with the corresponding building extraction results. In the early stage, most methods produce reasonable results on intact building regions. The proposed method shows clearer advantages in boundary precision and false positive suppression in cluttered areas, as highlighted in Fig. 9. U-Net and DeepLabv3+ exhibit boundary misclassification, while UNetFormer produces blurred delineation despite reasonable localization.

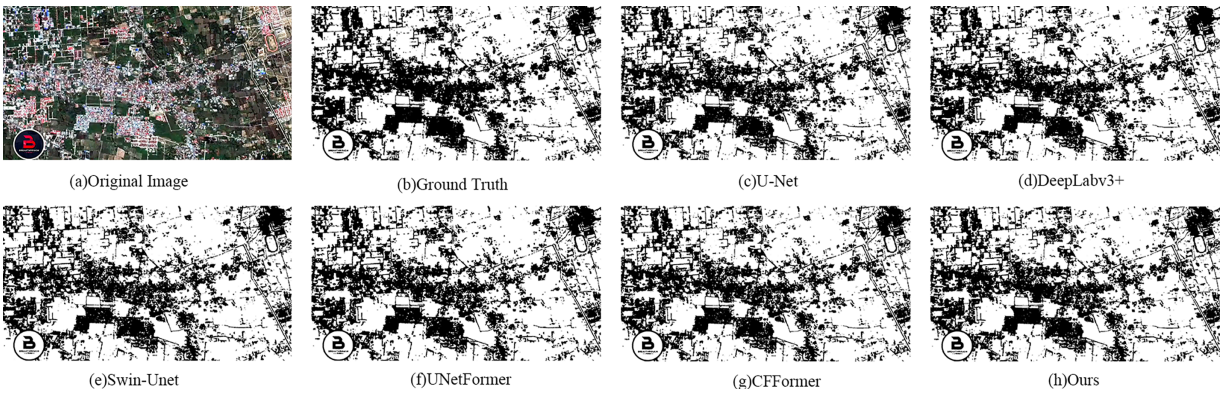


Figure 6: Comparative visualization of building semantic segmentation results from different models for early-stage disaster remote sensing imagery (Frame 165): (**a**) original image; (**b**) ground truth; (**c**) U-Net; (**d**) DeepLabv3+; (**e**) Swin-Unet; (**f**) UNetFormer; (**g**) CFFormer; and (**h**) ours (PGDGNet)

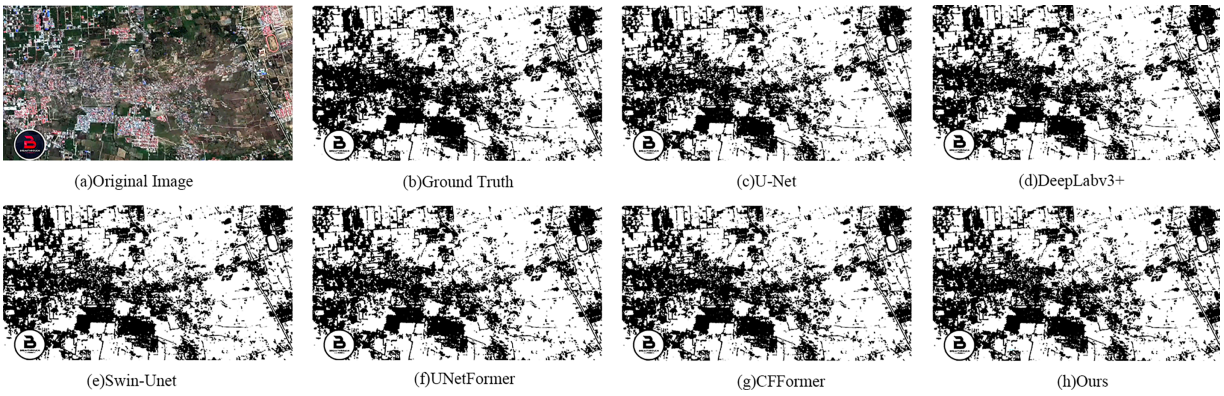


Figure 7: Comparative visualization of building semantic segmentation results from different models for intermediate-stage disaster remote sensing imagery (Frame 295): (**a**) original image; (**b**) ground truth; (**c**) U-Net; (**d**) DeepLabv3+; (**e**) Swin-Unet; (**f**) UNetFormer; (**g**) CFFormer; and (**h**) ours (PGDGNet)

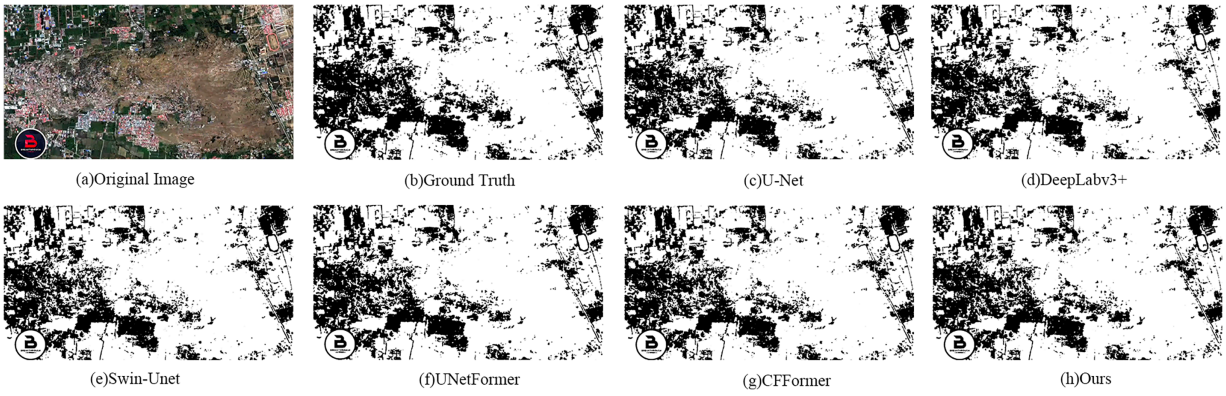


Figure 8: Comparative visualization of building semantic segmentation results from different models for late-stage disaster remote sensing imagery (Frame 425): (**a**) original image; (**b**) ground truth; (**c**) U-Net; (**d**) DeepLabv3+; (**e**) Swin-Unet; (**f**) UNetFormer; (**g**) CFFormer; and (**h**) ours (PGDGNet)

As disaster severity increases, the advantage of the proposed method becomes more pronounced. In the intermediate stage, U-Net and DeepLabv3+ show considerable omission errors in partially collapsed regions, while the proposed method maintains better structural continuity through collaborative local-global feature modeling. In the late stage, where building structures are severely fragmented and visually similar to surrounding debris, UNetFormer and DeepLabv3+ produce discontinuous and incomplete predictions, whereas the proposed method preserves the overall distribution and continuity of residual building structures more reliably, benefiting from the complementary local–global feature representation and prior-guided input design.

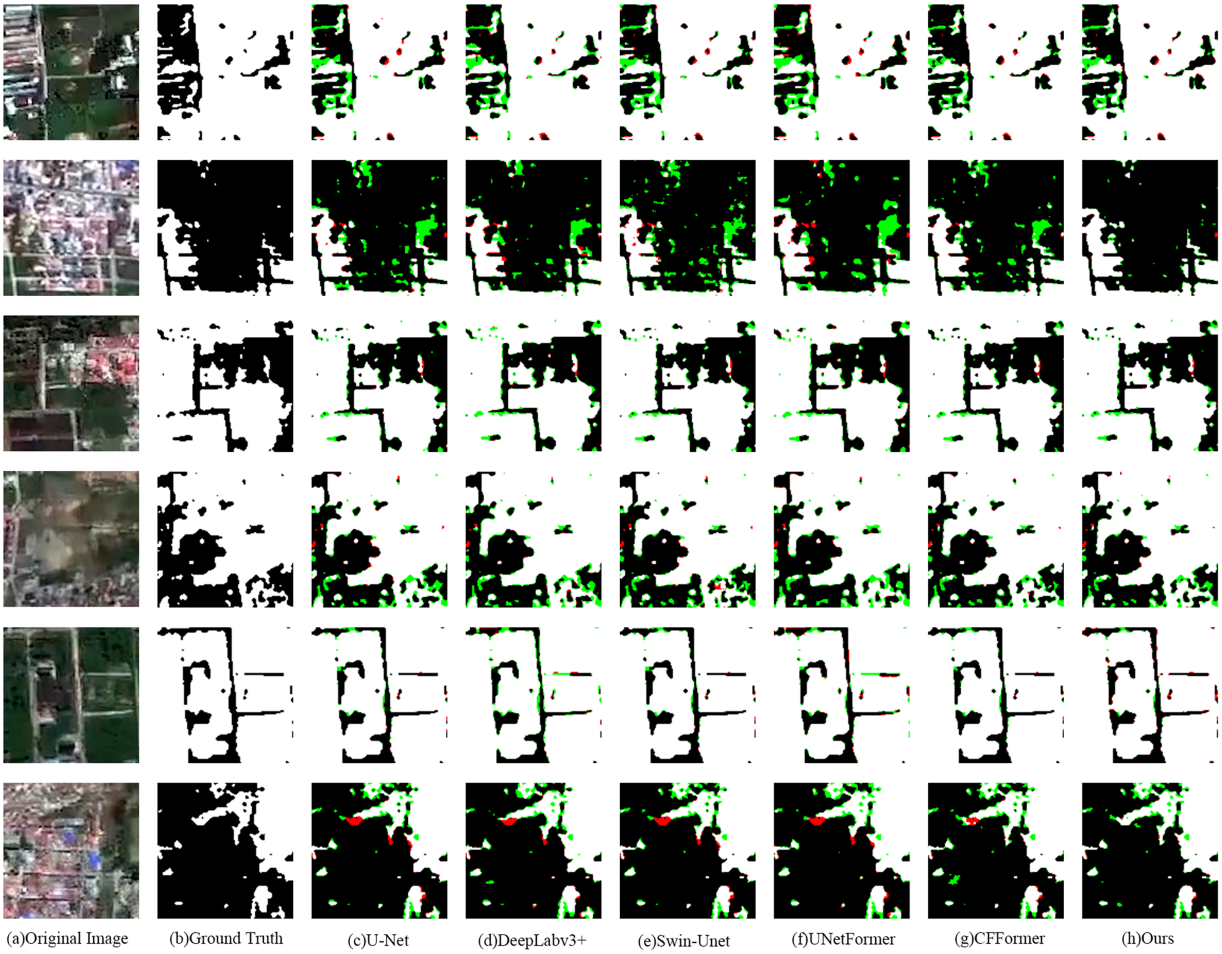


Figure 9: Detailed comparative visualization of building semantic segmentation results from different models across disaster stages: (**a**) original image; (**b**) ground truth; (**c**) U-Net; (**d**) DeepLabv3+; (**e**) Swin-Unet; (**f**) UNetFormer; (**g**) CFFormer; and (**h**) ours (PGDGNet). Rows 1–2 show local details in the early stage (Frame 165), rows 3–4 show local details in the intermediate stage (Frame 295), and rows 5–6 show local details in the late stage (Frame 425). In the error maps, black indicates correctly segmented buildings, green indicates false negatives, and red indicates false positives

3.2.3 Performance Analysis across Disaster Stages

To evaluate the adaptability of different models throughout disaster evolution, we analyze the results from two perspectives. First, a model-level comparison is conducted using Building IoU, Precision, and Recall to examine extraction performance across the early, intermediate, and late stages. Second, a stage-wise dynamic analysis is performed to quantify performance fluctuations between disaster stages, thereby assessing the stability of each model under changing building appearances.

Fig. 10 presents the model-level comparison based on the mean values from repeated experiments. As shown in Fig. 10a, most methods achieve the highest Building IoU in the intermediate stage. This indicates that building extraction difficulty is not determined solely by disaster severity. In the early stage, buildings are largely intact, but complex background objects increase building–background

confusion and reduce Precision. In the intermediate stage, many buildings remain structurally recognizable while becoming more distinguishable from the background, leading to a better balance between Precision and Recall. In the late stage, severe fragmentation and incomplete boundaries reduce Recall, although simplified backgrounds and exposed ground help maintain relatively high Precision. Compared with CFFormer, the strongest baseline, PGDGNet achieves higher Building IoU and Recall across all three stages. Building IoU increases by 0.59, 0.63, and 0.81 percentage points, while Recall improves by 1.45, 0.80, and 1.08 percentage points in the early, intermediate, and late stages, respectively. Although CFFormer shows slightly higher Precision in the intermediate and late stages, PGDGNet obtains higher Recall and Building IoU, indicating that it recovers more building regions with only a limited Precision trade-off. These results demonstrate the effectiveness of PGDGNet for multi-stage building extraction under different levels of disaster-induced appearance degradation.

To evaluate stage-wise stability under multi-stage appearance degradation, the absolute Building IoU differences between disaster stages were analyzed. Adjacent-stage variation was measured using the early–intermediate and intermediate–late differences, while the early–late difference was used to quantify the overall variation across the largest stage interval. A smaller absolute difference indicates less performance fluctuation between stages.

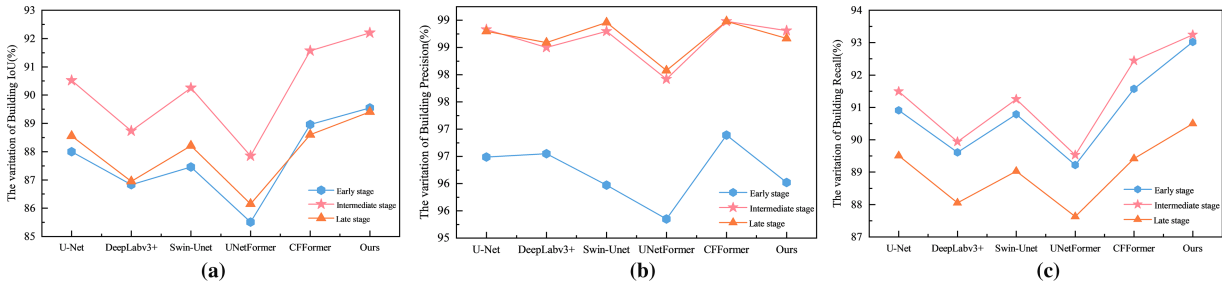


Figure 10: Multi-stage comparison of building-specific metrics for different methods across disaster stages. (**a**–**c**) show the variations in building IoU, precision, and recall across the early, intermediate, and late stages, respectively

As shown in Fig. 11, most methods achieve the highest Building IoU in the intermediate stage. This pattern may be attributed to the fact that many buildings in this stage are partially damaged but still retain recognizable structural and textural cues, while their contrast with the surrounding background becomes more distinguishable. In contrast, early-stage scenes contain more complex background interference, and late-stage scenes suffer from severe fragmentation and boundary discontinuity. The stability results show that PGDGNet does not have the smallest variation in every stage comparison. However, it achieves the highest Building IoU across all three stages while maintaining smaller overall stage-wise variation than the strongest baseline, CFFormer. Therefore, PGDGNet provides a better accuracy–stability trade-off among high-performing methods. These results suggest that PGDGNet maintains reliable extraction performance across stage-dependent appearance changes while preserving high building extraction accuracy.

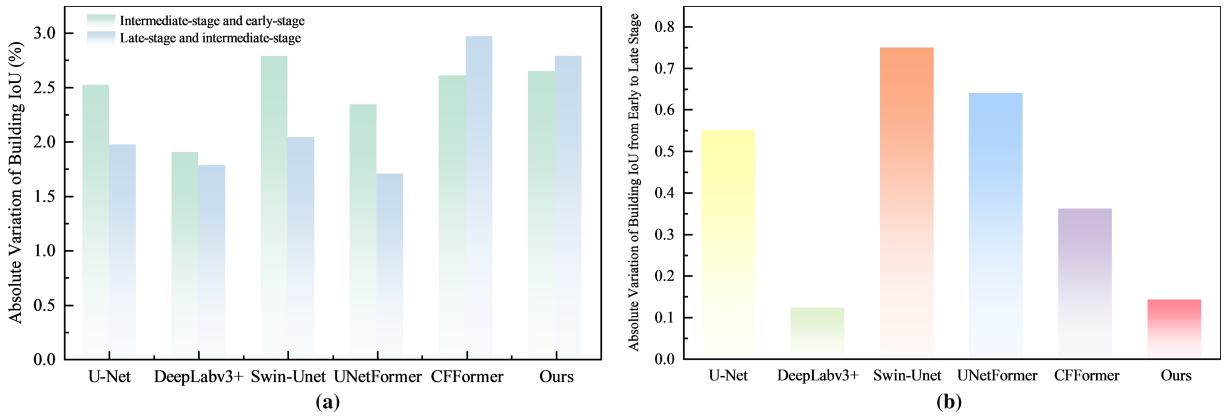


Figure 11: Absolute differences in building IoU across disaster stages: (**a**) early–intermediate and intermediate–late comparisons; (**b**) early–late comparison

3.3 Ablation Experiments

To evaluate the contribution of each component, ablation experiments were conducted by removing or replacing individual modules from the complete framework. The averaged results over the three disaster stages are summarized in Table 5. The complete model achieves the best overall performance, indicating that the main components of the framework are effective.

Table 5: Ablation study of individual network modules

DGFF	Prior knowledge	CBAM	WTConv	IoU (%)		F1-Score (%)		Evaluation metrics (%)	
				Building	Others	Building	Others	OA	Kappa
✓	✓	✓	✓	90.58	94.89	95.05	97.38	96.59	92.43
✗	✓	✓	✓	88.72	93.85	94.02	96.82	95.88	90.85
✓	✗	✓	✓	89.92	94.52	94.69	97.18	96.35	91.87
✓	✓	✗	✓	90.36	94.67	94.93	97.31	96.51	92.24
✓	✓	✓	✗	89.99	94.55	94.73	97.20	96.37	91.93
✗	✗	✗	✗	87.53	93.09	93.35	96.42	95.38	89.78

As shown in Table 5, DGFF contributes the largest performance gain. Removing DGFF reduces Building IoU from 90.58% to 88.72% and Kappa from 92.43% to 90.85%, demonstrating its importance for local–global feature integration. Removing the prior-guidance module decreases Building IoU to 89.92%, indicating that auxiliary color and boundary cues improve building localization and delineation. Replacing WTConv with a standard 3×3 convolution reduces Building IoU to 89.99%, while removing CBAM decreases it to 90.36%, showing that both local feature extraction and attention enhancement contribute to feature discrimination.

Table 6 further evaluates the proposed fusion strategy, prior components, and dynamic boundary loss. Replacing DGFF with simple feature addition reduces Building IoU from 90.58% to 89.21%, confirming that the proposed pre-fusion refinement is more effective than direct feature addition.

For the prior-guidance module, removing the color prior decreases Building IoU to 90.07%, while removing the boundary prior reduces it to 90.35%. This indicates that color and boundary cues provide complementary appearance and contour information. Removing the dynamic boundary loss decreases Building IoU to 89.44% and Kappa to 91.49%, confirming its contribution to boundary-aware optimization and contour delineation.

Table 6: Ablation study of key components in the proposed method

DGFF	Color prior	Boundary prior	DB loss	IoU (%)		F1-Score (%)		Evaluation metrics (%)	
				Building	Others	Building	Others	OA	Kappa
✓	✓	✓	✓	90.58	94.89	95.05	97.38	96.59	92.43
✗	✓	✓	✓	89.21	94.10	94.29	96.82	96.05	91.20
✓	✗	✓	✓	90.07	94.61	94.77	97.23	96.41	92.01
✓	✓	✗	✓	90.35	94.71	94.92	96.91	96.49	92.21
✓	✓	✓	✗	89.44	94.29	94.42	97.06	96.18	91.49

Overall, the ablation results show that DGFF provides the most substantial improvement, while WTConv, CBAM, color and boundary priors, and dynamic boundary loss further enhance the final performance. These results support the effectiveness of the proposed framework design.

3.4 Computational Complexity Analysis

Table 7 reports the model complexity, inference efficiency, and Building IoU of all compared methods. UNetFormer is the most lightweight model, with 11.72M parameters and 58.44 GFLOPs, and achieves 210.79 FPS, but it obtains the lowest Building IoU, suggesting a trade-off between lightweight design and representation capacity. CFFormer achieves competitive accuracy but requires the largest model size and computational cost, with 99.57M parameters and 835.52 GFLOPs. PGDGNet contains 40.27M trainable parameters and 393.23 GFLOPs, which is moderate compared with the evaluated methods, while achieving the highest Building IoU of 90.58%. Although its multi-branch architecture and feature refinement operations increase latency and peak GPU memory consumption, PGDGNet still reaches 46.66 FPS, exceeding the commonly used 30 FPS reference for real-time processing. These results indicate that PGDGNet provides a favorable balance between extraction accuracy and computational efficiency.

Table 7: Comparison of model complexity among different segmentation networks

Method	Params (M)	FLOPs (G)	Latency (ms)	FPS	Peak GPU Memory (MB)	Building IoU
U-Net	31.04	964.80	3.54	282.60	232.50	88.48
DeepLabv3+	44.54	238.60	4.06	246.07	328.17	87.16
Swin-Unet	41.38	270.02	11.85	84.42	519.33	87.93
UNetFormer	11.72	58.44	4.74	210.79	523.31	85.72

(Continued)

Table 7 (continued)

Method	Params (M)	FLOPs (G)	Latency (ms)	FPS	Peak GPU Memory (MB)	Building IoU
CFFormer	99.57	835.52	16.75	59.83	976.00	90.03
PGDGNet	40.27	393.23	21.43	46.66	1313.64	90.58

4 Conclusion

This study proposed PGDGNet for frame-wise multi-stage building extraction from staring satellite video-derived observations under disaster-induced appearance degradation. By combining branch-specific local–global feature refinement with joint color and boundary prior guidance, PGDGNet helps address the difficulty of maintaining discriminative building representations when structural cues become incomplete, fragmented, or inconsistent across disaster stages.

Experiments on representative observations from the early, intermediate, and late stages of the 2018 earthquake–tsunami event in Indonesia demonstrate that PGDGNet achieves the highest Building IoU and Recall across all disaster stages. Compared with CFFormer, PGDGNet improves Building IoU by 0.59, 0.63, and 0.81 percentage points, and Recall by 1.45, 0.80, and 1.08 percentage points in the three stages, respectively. The relatively larger gains in the early and late stages indicate that PGDGNet is particularly effective in challenging cases where complex backgrounds, severe appearance variation, or structural fragmentation reduce the reliability of conventional building cues. These results support the effectiveness of combining local–global feature refinement with prior-guided learning for robust building extraction across multi-stage disaster observations.

Several limitations remain. First, the experiments are based on a single tsunami–earthquake event in the Petobo region, Indonesia. Different disasters may produce distinct building degradation patterns, such as inundation, roof loss, burning traces, debris burial, or collapse, and may introduce different background interference. Therefore, multi-event and multi-region validation is needed to assess the generalization capability of the proposed framework. Second, the color prior is derived from the dominant roof-color characteristics of the current study area, and its transferability may be limited under different roof materials, illumination conditions, sensors, or regional building styles. Finally, this study focuses on building-region extraction rather than explicit damage-level classification or temporal sequence modeling. Future work will address these limitations through multi-event, multi-region, and multi-sensor validation, adaptive prior-guidance strategies, and joint building extraction and damage assessment.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the National Natural Science Foundation of China (Youth Science Fund Project), grant number 42201427; and the Research Project Supported by Shanxi Scholarship Council of China, grant number 2025-090.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Xingbo Huang and Huijiao Qiao; methodology, Xingbo Huang; software, Xingbo Huang; validation, Xingbo Huang and Huijiao Qiao; formal analysis, Xingbo Huang; investigation, Xingbo Huang; resources, Huijiao Qiao; data curation, Xingbo Huang; writing—original draft preparation, Xingbo

Huang; writing—review and editing, Xingbo Huang, Yuhe Zhu and Huijiao Qiao; visualization, Xingbo Huang; supervision, Huijiao Qiao; project administration, Huijiao Qiao; funding acquisition, Huijiao Qiao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Huijiao Qiao, upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Chen M, Wu J, Mao T, Du R, Zhao B, Lin J, et al. An improved method for rapid un-collapsed building extraction from post-disaster high-resolution remote sensing imagery based on multi-scale feature alignment. Int J Digit Earth. 2024;17(1):2344599. doi:10.1080/17538947.2024.2344599.
2. Ji Y, Sumantyo JTS, Chua MY, Waqar MM. Earthquake/tsunami damage assessment for urban areas using post-event PolSAR data. Remote Sens. 2018;10(7):1088. doi:10.3390/rs10071088.
3. Song C, Fan C, Wang M. Image-based adaptive staring attitude control for multiple ground targets using a miniaturized video satellite. Remote Sens. 2022;14(16):3974. doi:10.3390/rs14163974.
4. Masafu C, Williams R. Satellite video remote sensing for flood model validation. Water Resour Res. 2024;60(1):e2023WR034545. doi:10.1029/2023WR034545.
5. Li Q, Mou L, Sun Y, Hua Y, Shi Y, Zhu XX. A review of building extraction from remote sensing imagery: geometrical structures and semantic attributes. IEEE T Geosci Remote. 2024;62:1–15. doi:10.1109/TGRS.2024.3369723.
6. Zhou J, Li Z. Spatial relation-aided method for object-oriented extraction of buildings from high resolution image. J Appl Sci. 2012;30(5):511–16. doi:10.3969/j.issn.0255-8297.2012.05.012.
7. Bittner K, Cui S, Reinartz P. Building extraction from remote sensing data using fully convolutional networks. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2017;XLII-1/W1:481–6. doi:10.5194/isprs-archives-XLII-1-W1-481-2017.
8. Yu T, Tang P, Zhao B, Bai S, Gou P, Liao J, et al. ConvBNet: a convolutional network for building footprint extraction. IEEE Geosci Remote Sens Lett. 2023;20:1–5. doi:10.1109/LGRS.2023.3250091.
9. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain. Red Hook, NY, USA: Curran Associates Inc.; 2016. p. 4905–13.
10. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. Piscataway, NJ, USA: IEEE; 2021. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
11. Wang H, Miao F. Building extraction from remote sensing images using deep residual U-Net. Eur J Remote Sens. 2022;55(1):71–85. doi:10.1080/22797254.2021.2018944.
12. Ran S, Gao X, Yang Y, Li S, Zhang G, Wang P. Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images. Remote Sens. 2021;13(14):2794. doi:10.3390/rs13142794.
13. Kang W, Xiang Y, Wang F, You H. EU-Net: an efficient fully convolutional network for building extraction from optical remote sensing images. Remote Sens. 2019;11(23):2813. doi:10.3390/rs11232813.

14. Zhu Q, Liao C, Hu H, Mei X, Li H. MAP-Net: multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Trans Geosci Remote Sens. 2021;59(7):6169–81. doi:10.1109/TGRS.2020.3026051.
15. Ye Z, Fu Y, Gan M, Deng J, Comber A, Wang K. Building extraction from very high resolution aerial imagery using joint attention deep neural network. Remote Sens. 2019;11(24):2970. doi:10.3390/rs11242970.
16. Saadati D, Hao H. AttUNet+: towards high-fidelity building footprints. Remote Sens Appl Soc Environ. 2025;39:101715. doi:10.1016/j.rsase.2025.101715.
17. Chen K, Zou Z, Shi Z. Building extraction from remote sensing images with sparse token transformers. Remote Sens. 2021;13(21):4441. doi:10.3390/rs13214441.
18. Strudel R, Garcia R, Laptev I, Schmid C. Segmenter: Transformer for semantic segmentation. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. Piscataway, NJ, USA: IEEE; 2021. p. 7242–52. doi:10.1109/ICCV48922.2021.00717.
19. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6–14; Virtual. Red Hook, NY, USA: Curran Associates Inc.; 2021. p. 12077–90.
20. Wang L, Fang S, Meng X, Li R. Building extraction with vision transformer. IEEE Trans Geosci Remote Sens. 2022;60:1–11. doi:10.1109/TGRS.2022.3186634.
21. Xia L, Mi S, Zhang J, Luo J, Shen Z, Cheng Y. Dual-stream feature extraction network based on CNN and transformer for building extraction. Remote Sens. 2023;15(10):2689. doi:10.3390/rs15102689.
22. Chen Z, Luo Y, Wang J, Li J, Wang C, Li D. DPENet: dual-path extraction network based on CNN and transformer for accurate building and road extraction. Int J Appl Earth Obs. 2023;124:103510. doi:10.1016/j.jag.2023.103510.
23. Diao K, Zhu J, Liu G, Li M. MDTrans: multi-scale and dual-branch feature fusion network based on swin transformer for building extraction in remote sensing images. IET Image Process. 2024;18(11):2930–42. doi:10.1049/ipr2.13145.
24. Wang Y, Wang S, Dou A. A dual-branch fusion network based on reconstructed transformer for building extraction in remote sensing imagery. Sensors. 2024;24(2):365. doi:10.3390/s24020365.
25. Wang C, Wu B, Wang H. Geometric parameters extraction of building based on shape prior. J Electron Inf Technol. 2017;39(8):1848–56. doi:10.11999/JEIT161224.
26. Hao M, Lin H, Chen S, Luo W, Zhang H, Zheng N. A geometric significance-aware deep mutual learning network for building extraction from aerial images. Drones. 2024;8(10):593. doi:10.3390/drones8100593.
27. Chen Y, Wu Y, Xu L, Wong A. Quantization in relative gradient angle domain for building polygon estimation. In: Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR); 2021 Jan 10–15; Milan, Italy. Piscataway, NJ, USA: IEEE; 2021. p. 8360–67. doi:10.1109/ICPR48806.2021.9413115.
28. Quan Y, Yu A, Cao X, Qiu C, Zhang X, Liu B, et al. Building extraction from remote sensing images with DoG as prior constraint. IEEE J Sel Top Appl Earth Obs Remote Sens. 2022;15:6559–70. doi:10.1109/JSTARS.2022.3195808.
29. Ok AO. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS J Photogramm Remote Sens. 2013;86:21–40. doi:10.1016/j.isprsjprs.2013.09.004.
30. Turker M, Cetinkaya B. Automatic detection of earthquake-damaged buildings using DEMs created from pre- and post-earthquake stereo aerial photographs. Int J Remote Sens. 2005;26(4):823–32. doi:10.1080/01431160512331316810.
31. 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. doi:10.1016/j.rse.2021.112636.