

Original Article

Multi-Scale Spatial-Channel Reconstruction Transformer for Automatically Assessing Building Damage in Post-Disaster Scenarios

K. S. Raghavendra Reddy¹, B. Sudhakar², K. Venkata Ramanaiah³

^{1,2}Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, Annamalai University, Chidambaram, Tamil Nadu, India.

³Department of Electronics and Communication Engineering, Y.S.R. Engineering College of Yogi Vemana University, Proddatur, Andhra Pradesh, India.

¹Corresponding Author : ks.raghavendrareddy@gmail.com

Received: 12 June 2026

Revised: 10 September 2026

Accepted: 18 September 2026

Published: 27 October 2026

Abstract - Surveying building damage in post-disaster image segmentation is a difficult problem due to noise, low contrast, and ambiguous boundaries in real-life clinical images. The existing transformer models face problems such as feature redundancy, semantic inconsistencies, and poor boundary reconstructions. To overcome such drawbacks, this study introduces a new model named Reconstruction Spatial-Pixel Fusion Multi-Scale Transformer (RSPF-MST). Specifically, the proposed architecture has three main novelties: (i) Reconstruction Spatial-Pixel Convolution (RSPC) is designed to remove redundant features with the help of spatial and channel-wise reconstructions; (ii) Channel Interaction Fusion (CIF) is designed to strengthen multi-scale semantic consistency with cross-channel attention; and (iii) Pixel-Aware Up-sampling Guidance (PAUG) aims at improving boundary accuracy by introducing pixel-level dynamic weights. Moreover, a Dynamic Receptive Multi-Scale Context (DRMSC) module is added in the decoder for capturing context information at various levels. A pruning strategy is adopted for efficiency. Experimental analysis demonstrates that the proposed model significantly improves segmentation accuracy, achieving 98.43% and 98.53%, and IoU scores of 96.13% and 96.12% on the Ida-BD and RescueNet datasets, respectively, while maintaining robustness, particularly for low-quality post-disaster images.

Keywords - Post-disaster image, Building damage, Transformer, RSPF-MST, PAUG, and Segmentation.

1. Introduction

Remote Sensing (RS) [1] is a high-growth area with many new technologies being developed or improved upon for applications like land management, city planning, assessing natural disasters, and understanding what is happening in the field environment, leading to a significant new direction in research and development in earth observation: Change Detection (CD) using RS images [2]. CD aims to compare two or more different times of RS data for a common region to identify spatial and attribute changes of ground objects that will help to monitor and study continuous changes of surface targets [3]. CD is particularly helpful in situations like responding to natural disasters, managing an emergency, and assessing damage, where it provides governments or military organizations with real-time data that allows for quick response and accurate deployment [4, 5]. Building damage detection is one of several fine-grained CD tasks that has attracted increasing attention over recent years [6]. Because buildings are the main carriers of urban spatial structures, they can directly indicate how much space is destroyed after a

disaster or terrorist attack in an area [7]. It is vital to have the most current and accurate information regarding building damage as soon as possible after a conflict or natural disaster in order to properly carry out rescue operations, reconstruct the area, and assess the war damage [8]. The demand for rapid and large-scale assessments of damage is increasing, so there is a growing need to develop automated approaches using RS data [9]. Nowadays, construction damage assessments can be divided into three categories: (i) Manual inspection, (ii) Machine Learning (ML) approaches, and (iii) Deep Learning (DL) approaches [10]. Manual inspections are very accurate but time-intensive (the evaluator needs to be an expert). ML algorithms such as Support Vector Machines (SVM) [11] and Random Forests (RF) [12] are used to automate the assessment process; however, as they rely on handcrafted features, there is limited ability to adapt to various disasters and rely solely on feature engineering to develop a model. This means traditional ML approaches are expensive and do not have the flexibility needed to be effectively utilized across different types of disasters [13-15]. Damage assessments



using deep learning [16-18] have made significant advancements compared to previous methodologies by enabling the extraction of better features and providing more accurate classifications compared to CNN-based algorithms in identifying localized structural damage (e.g., building collapse). Conversely, CNNs are not ideal for identifying minor structural damage, such as small cracks, because CNNs utilize a limited receptive field. CNNs also do not generalize well across numerous damage scenarios. To address these issues, new models based on attention were suggested to enhance global dependencies and adaptive weighting attributes.

The introduction of transformers has allowed for the creation of a self-attention mechanism and has resulted in building damage assessments that are globally contextual in nature [19, 20]. Redundancies exist in spatial and channel activations in densely packed post-disaster images using existing CNN-based frameworks. Although the use of transformers improves long dependency modeling, it brings about additional computations that do not preserve boundaries in detail. Current skip connections cannot capture semantic relations, and conventional upsampling techniques cause boundary blurring. Most multi-scale strategies lack efficient integration of receptive fields and suffer from redundancies and complexities. This indicates an existing research gap towards a unified segmentation framework that is able to reduce redundant features, capture semantic interactions, reconstruct boundaries, and be computationally efficient. The proposed framework incorporates redundancy-aware feature reconstruction, cross-channel semantic fusion, pixel-level boundary-aware up-sampling, multi-scale contextual modeling, and adaptive pruning in a U-shaped transformer architecture. This strategy improves on the weaknesses of existing frameworks by suppressing redundancies, improving semantic consistency, refining boundary details, and reducing computational redundancies, highlighting a more precise representation of complex post-disaster buildings while ensuring efficiency.

1.1. Motivation

At the moment, there are three classifications of damage assessment techniques: 1. Manual Inspection. It can be done with great precision, but tends to take a long period of time to complete, and will usually require the use of a qualified expert assessor. 2. ML Methods. Traditional ML methods require significant feature engineering costs and, therefore, are not very flexible for assessing damage from heterogeneous disasters, even though they have a very high precision for building damage assessments. 3. DL Methods. Many advances have been made in this area due to deeper capabilities in extracting and classifying features. Current Convolutional Neural Network (CNN)-based methods have shown that they can effectively extract local structural features of buildings (ex., complete collapses) but generally lack the ability to capture finer details of damage, such as minor

cracking, due to the small receptive field sizes of the CNN architecture. To address some of the issues associated with extracting features from CNN architectures, some researchers have developed attention-based DL methods that improved the ability to model global dependencies and provide for adaptive feature weighting. Self-attention-based models (Transformers) have completely changed the way that building damage assessment is accomplished by providing a model that can capture and consider the complete context of the building damage via global context modeling. While the field of building damage assessment has advanced significantly in the past few years, the approaches and methodologies that currently exist still have 2 major issues associated with the development of the methods for building damage assessment. One of the issues is that the available damage assessment methodologies do not properly utilize multi-scalar features and, therefore, are unable to identify severe or minor damage; and the other issue is that many models do not include sufficient temporal and spatial contextual data in the assessment methodologies; yet, these two types of data are required for a correct analysis of building damage following a disaster or emergency. Third, the current methods often have trouble generalizing to different kinds of disasters and environments, thus limiting their practical use. To address these limitations, a Reconstruction Spatial-Pixel Fusion multi-scale Fusion Transformer is proposed for post-disaster building damage assessment.

1.2. Objectives

- A Reconstruction Spatial-Pixel Fusion multi-scale Fusion Transformer is proposed for post-disaster building damage assessment.
- A Channel Interaction Fusion (CIF) module is designed based on a Vision Transformer architecture to effectively perform, thereby reducing the semantic discrepancy between low & high-level representations.
- A Pixel Attention Up-sample Guidance (PAUG) module is designed to enhance the model's ability to precisely localize target regions.
- The Dynamic Residual Multi-Scale Convolution Module (DRMSC) uses different features while capturing high-level contextual semantic data.
- In this study, model pruning was employed to decrease computational complexity and model size while highlighting efficient performance results.

The research contributes to the organization as follows: Section 2 presents a recent literature review analysis, Section 3 presents the overview of the proposed study, Section 4 illustrates the experimental setup-based results analysis, and Section 5 provides the conclusion & Future work.

2. Related Works

Xiao et al. [21] developed DamageCAT, an advanced damage classification system that processes post- and pre-disaster satellite images equally. This study used post- and

pre-disaster satellite images, and a hierarchical U-Net-based transformer method was introduced. Results show that the overall system performs well compared to other methods. Nevertheless, the model faced poor multi-scale feature extraction, struggled with real-time inference, and had generalization challenges.

To localize buildings and damage classification, Yu et al. [22] presented a GMSF model to record complex spatial relationships and contextual attributes. GMSF consists of three main novelties: (1) Adaptive Attention (AA), which prioritizes critical regions in the image, (2) the GMFFN approach to improve robustness by highlighting significant damage features, and (3) MWF, which applies wavelets to extract high-resolution details from images. Evaluation results indicate GMSF's excellent results compared to baseline models, with the suggested greater accuracy in minor damage detection. Multiple fusion and attention modules lead to great performance, but generalization, reliability issues, high computational cost, and struggles with multimodality limitations were faced in this study.

To assess hyperlocal on-ground damages, Yang et al. [23] used bi-temporal street-view images and advanced pre-trained vision models. Swin Transformer and ConvNeXt were used for damage categorization. Results indicate the accuracy of damage evaluation increases by over 10% when comparing the Swin Transformer model at 66.14% accuracy vs the dual-channel Feature-Fusion ConvNeXt model at 77.11% accuracy. Grad-CAM results indicate improved ability of the pre-trained vision model to recognize areas of significant change between pre- and post-disaster images by using the previous street view images, thereby improving the model's performance. However, the study faced some limitations, such as scene complexity in disaster-affected areas, data scalability issues, and a high burden of manual labelling. Zhang et al. [24] introduced an end-to-end system called CMSNet that combines the structural priors from the SAM with the efficient

temporal modeling and detailed representation abilities of CNN-Mamba. It also used visual information from SAM to help the network focus on identifying object boundaries and structural changes.

To make the system more robust against local damage, non-rigid movements, and complicated background distractions, a PVPF-FM model was introduced. This module helps align and combine the visual priors with the change features. However, the attachment of the model increased complexity and dependence on additional priors (SAM), which causes concerns about the scalability and poor fusion strategy.

Cha et al. [25] presented a DL-based system to identify and quantify road damage using pre- and post-disaster images. In particular, different techniques such as the overlay method, difference-based method, and Siamese-based method for change detection were validated under a strictly measured validation setting to detect the most effective approach.

To solve problems such as the lack of good multi-scale characteristics, difficulties with real-time inference, generalization problems, scalability constraints of the data, and the high cost of labeling in disaster scenes with surface cracks and smaller debris, where it becomes difficult to detect, resulting in data loss, it is becoming highly important to devise hybrid solutions that incorporate spatial-channel based feature improvement, multi-scale fusion using attention, and efficient optimization methods.

Consequently, in the proposed solution, these shortcomings have been addressed through the use of reconstruction-based convolutions, interactions between channels, upsampling of pixels, multi-scale context, and dynamic pruning to ensure reliable multimodal processing and reduce computational complexity. Table 1 presents the recent literature research performance and challenges.

Table 1. Recent literature research performance and challenges

Authors & ref	Models	Dataset	Performance	Cons
Xiao et al. [21]	DamageCAT-hierarchical U-Net	BD-TypoSAT dataset	MIoU-0.737 and F1-score 0.846	Poor multi-scale features, struggle with real-time inference, and generalization challenges.
Yu et al. [22]	GMSF	XFBD and Ida-BD dataset	F1-score-80.4% and Kappa-77.9%	Generalization, reliability issues, and struggle with multimodality.
Yang et al. [23]	Swin Transformer and ConvNeXt	Collected dataset	Precision-73.13%, Recall-73.15%, and F1-score-71.19%	Complexity in disaster-affected areas, data scalability issues, and a high burden of manual labeling.
Zhang et al. [24]	CMSNet	RWSBD dataset	Precision-83.95, Recall-81.28, and MIoU-73.13	Poor fusion strategy, scalability, and complexity issues.
Cha et al. [25]	Triple UNet	xBD dataset	Precision-0.886, Recall-0.864, and MIoU-0.769	Surface cracks and small debris may become undetectable and result in high levels of data loss.

2.1. Problem Statement

Disaster-induced structure detection is an important process that necessitates the accurate recognition of complex structures in difficult environments characterized by high noise levels, low contrast, and undefined boundaries. Convolutional neural networks, especially U-Net-based networks, have been found to work well with respect to local feature extraction, but they tend to be plagued by feature redundancy, poor global context learning capacity, and poor boundary recovery. Though transformer architectures have helped boost global semantic feature extraction, they face computational complexities and issues regarding multi-scale feature inconsistency.

To overcome these problems, it is increasingly necessary to create hybrid systems capable of integrating spatial-channel feature refinement, attention-based multi-scale fusion, and effective optimization techniques.

Subsequently, the proposed system employs reconstruction-based convolutions, cross-channel feature interaction, pixel-wise upsampling, multi-scale contextualization, and dynamic pruning.

3. Proposed Methodology

Post-disaster building damage assessments present substantial challenges, especially due to the trade-off between assessment speed and accuracy. Poor image quality in debris-filled situations, difficulty distinguishing structural from non-structural damage, and automated models' inability to generalize across multiple building forms and hazard types are significant challenges. To address these challenges, a multi-scale spatial-channel reconstruction transformer is proposed for automatically assessing building damage in post-disaster scenarios. Figure 1 denotes the architecture of the proposed multi-scale spatial-channel reconstruction transformer.

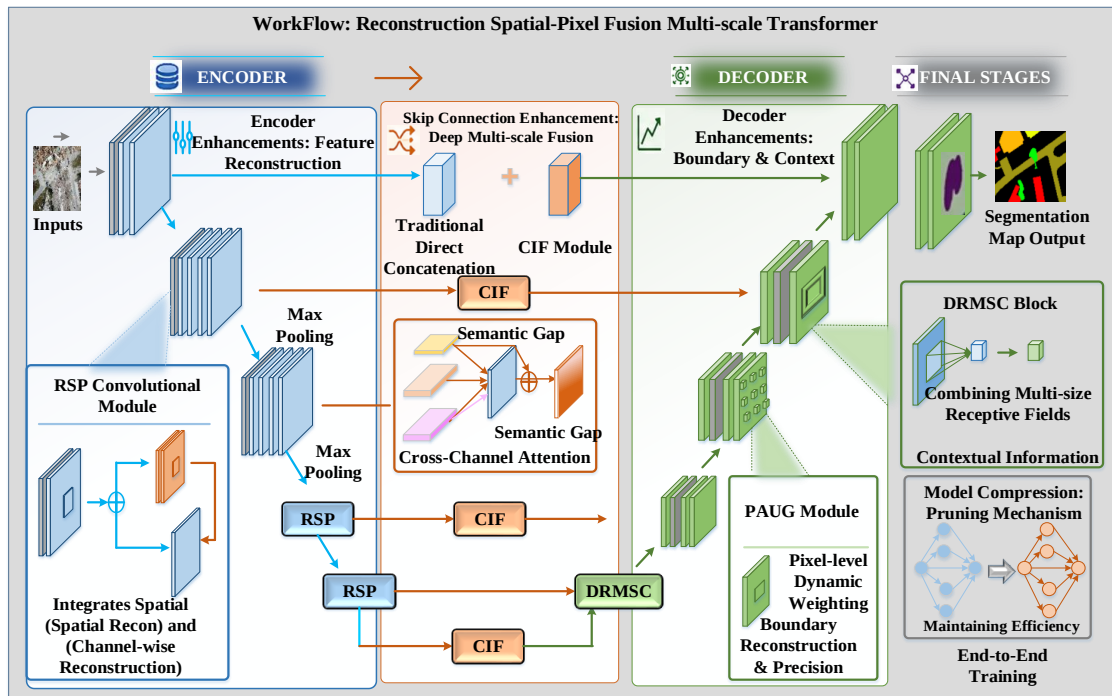


Fig. 1 Architecture of the proposed multi-scale spatial-channel reconstruction transformer

The Reconstruction Spatial-Pixel Fusion multi-scale Fusion transformer improves the standard transformer framework by fixing some of its problems, like too much repeated information, mismatched meanings, and poor details at the edges. In the encoder, regular convolutional (Conv) blocks are replaced with an RSPC module. This module uses both spatial and channel-based reconstruction to reduce unnecessary details. Afterward, in the skip connection path, instead of just putting features together like in UNet, they use a CIF module. This module helps combine features from different scales using attention across channels, making the meaning between the encoder and decoder parts match better. Subsequently, in the decoder, a PAUG module replaces

traditional upsampling layers. This module uses dynamic weighting at the pixel level to improve edge details and make segmentation more accurate. The model keeps the same U-shaped structure and allows end-to-end training. It focuses more on rebuilding features and understanding meaning, which makes it better at handling images that are noisy, low contrast, or have unclear edges. In the decoder phase, DRMSC integrates different features by combining receptive fields of different sizes, capturing richer contextual semantic information. A pruning mechanism was employed to reduce computational complexity and model size while maintaining efficient performance results.

3.1. System Model

The developed model highlights a U-Shaped encoder-decoder framework, improved with transformer-based learning and specialized modules for enhanced segmentation. In the encoder, the ROSC module reduces redundancy by refining spatial and channel features. Skip connections are made robust using the CIF module, which enables cross-channel attention to align multi-scale semantics. In the decoder, the PAUG module enhances boundary precision through dynamic pixel-level weighting, while the DRMSC modules capture rich contextual information across different scales. Furthermore, a pruning mechanism reduces model complexity, ensuring efficient performance without compromising accuracy.

3.1.1. Reconstruction Spatial-Pixel Convolutional Module (Encoder)

Initially, encoder-standard Conv blocks are substituted with an RPSC module, as shown in Figure 2, developed to explicitly suppress feature redundancy through joint spatial and channel reconstruction. Given a feature map $M \in \mathbb{R}^{N \times C \times H \times W}$, the module initially enables a pre-convolution and splits the feature into two branches, where one branch preserves residual information while the other undergoes structured refinement. The spatial redundancy is reduced using the Spatial Reconstruction Unit (SRU), as depicted in Figure 3, where feature normalization is expressed as:

$$\gamma = GN(M) = \frac{M - \mu}{\sqrt{\sigma^2 + \varepsilon}} \cdot \gamma + \beta \quad (1)$$

Where μ indicates the mean of feature values computed across spatial dimensions, σ is the standard deviation of the feature distribution, ε denotes the small constant, γ, β represents the learnable scaling and shifting parameter in GN, and $\bar{M}$ is the feature map after GN. Adaptive spatial weights are attained through a gating mechanism.

$$W = \text{Sigmoid}(DN(M)), M_w = W \otimes M \quad (2)$$

Subsequently, by cross-reconstruction operation to reallocate informative regions:

$$M_{w1} = W_1 \otimes M, M_{w2} = W_2 \otimes M, M_w = M_{w1} \cup M_{w2} \quad (3)$$

Here W is the spatial weight map generated using the sigmoid function, W_1 where W_2 denotes gating masks that separate informative and redundant features, $\otimes$ is the element-wise multiplication, $\cup$ indicates the channel-wise concatenation, M_w and is the spatially refined feature map. Following, Figure 4 presents the CRU, which mitigates channel redundancy by splitting features using a ratio α , transforming them through grouped and point-wise convolutions, and fusing them using channel attention:

$$Y_1 = X_G M_{up} + X_{P1} M_{up}, Y_2 = X_{P2} M_{low} \cup M_{low} \quad (4)$$

Where X_G denotes the weight matrix of grouped convolution (GWC), X_{P1}, X_{P2} is the weight matrices of point-wise convolutions, and Y_1, Y_2 indicate transformed feature outcomes from the upper and lower channel branches.

$$\beta_1 = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}, \beta_2 = \frac{e^{s_2}}{e^{s_1} + e^{s_2}}, Y = \beta_1 Y_1 + \beta_2 Y_2 \quad (5)$$

Here, global terms s_1, s_2 are derived using global average pooling and β indicate the normalized attention weights. This fused spatial-channel reconstruction process applies the encoder to emphasize informative structures while suppressing irrelevant activations, leading to more robust feature extraction, particularly in challenging scenarios such as post-disaster damage assessment, where noise, occlusions, and ambiguous boundaries are prevalent.

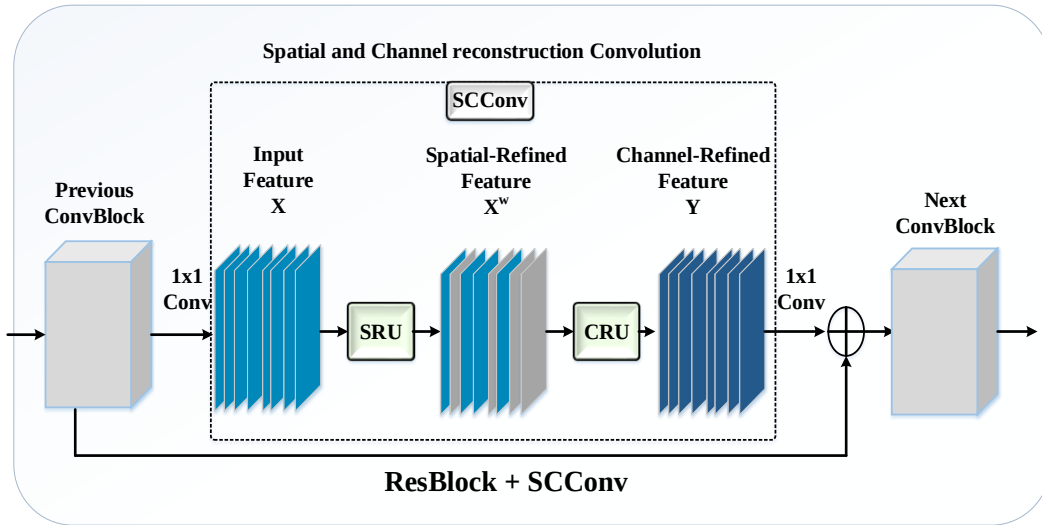


Fig. 2 Spatial-Pixel convolutional module

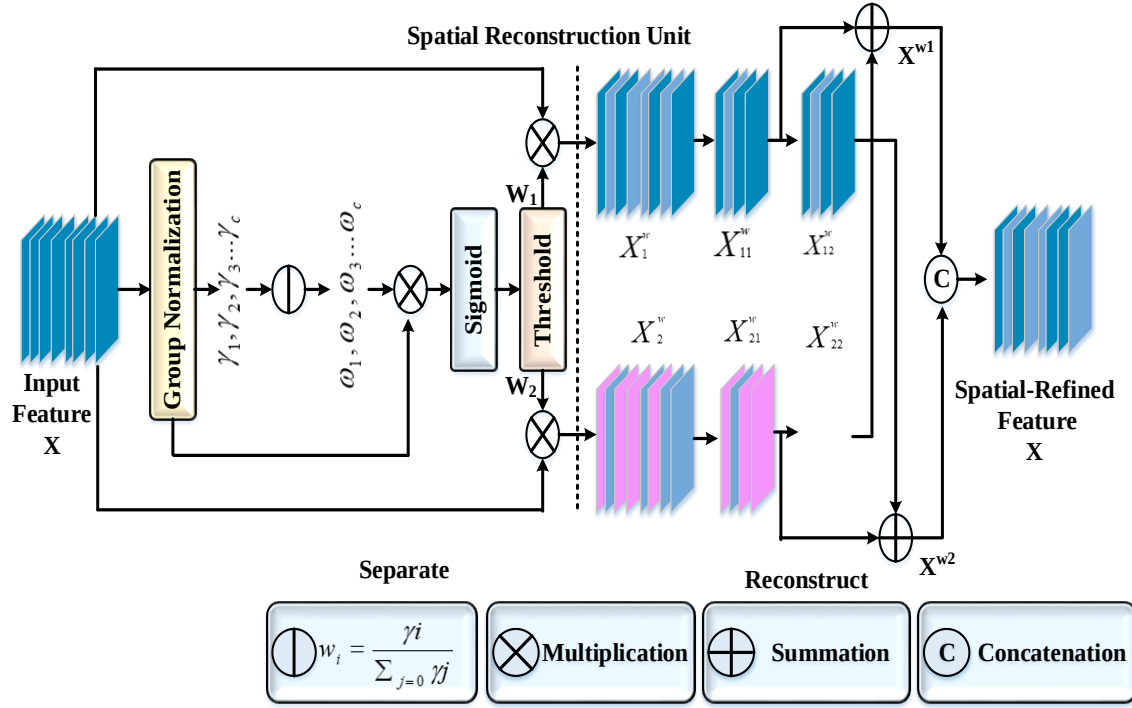


Fig. 3 SRU framework

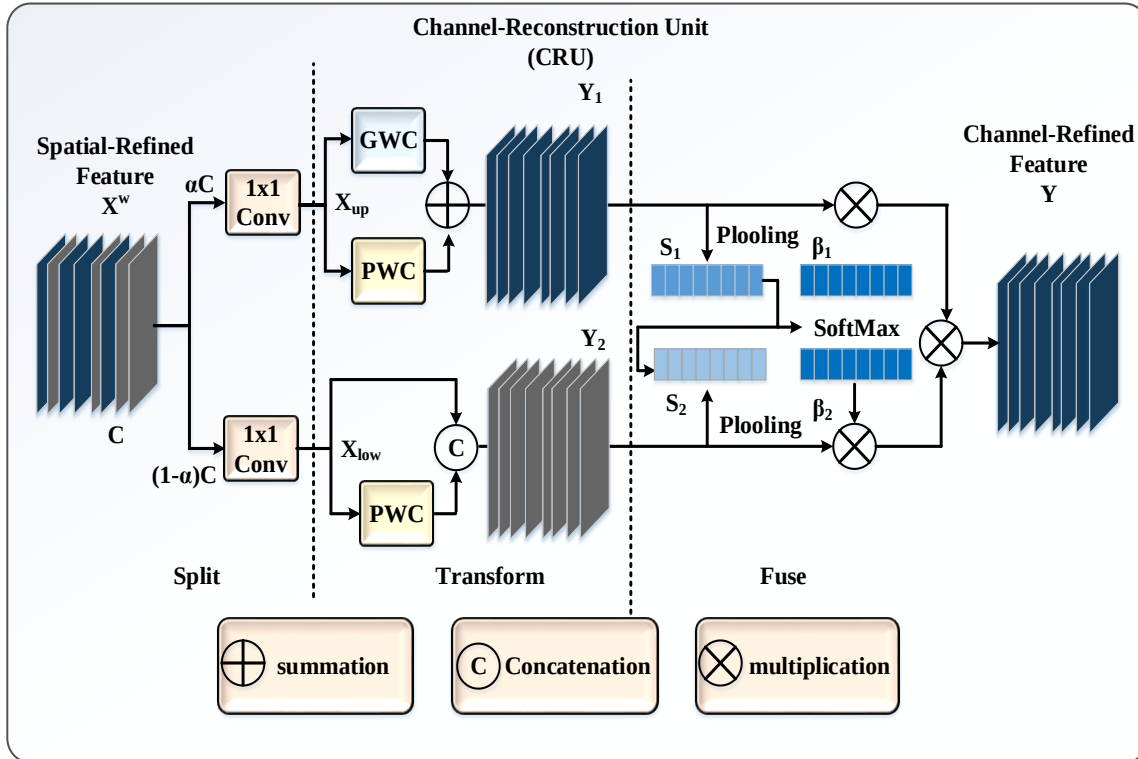


Fig. 4 CRU framework

3.1.2. Channel Interaction Fusion Module (CIF)

After completing the initial stage, in the skip connection pathway, a CIF-based vision transformer architecture is suggested instead of the conventional direct concatenation

used in U-Net, enabling deeper semantic interaction between encoder and decoder features. Given multi-scale encoder features E_i , query, key, and value representations are first expressed as:

$$Q_i = E_i W_Q, K = E_\Sigma W_K, V = E_\Sigma W_V \quad (6)$$

Where E_i is the encoder feature at scale i , E_Σ indicates aggregated multi-scale feature representation, W_Q, W_K, W_V are learnable projections metrics for query, key, and value. Then, cross-channel attention is used to compute feature similarity and aggregation:

$$CA_i = \text{Soft max} \left(\frac{Q_i K^T}{\sqrt{c}} \right) V \quad (7)$$

Where CA_i denotes the cross-attention output for scale i , Q_i is the query representation for feature E_i , K is the key representation, and C represents the channel dimension used for scaling attention. To enhance stability and representation capacity, multi-head channel attention is employed:

$$MC_i = \frac{1}{N} \sum_{n=1}^N CA_i^{(n)} \quad (8)$$

Let N is the amount of attention heads, and MC_i represents the multi-channel attention outcome. Afterward, the final combined representation is attained using residual learning and a non-linear transformation:

$$O_i = MC_i + \text{MLP}(Q_i + MC_i) \quad (9)$$

Where O_i denotes the final fused feature map after attention and MLP, $\text{MLP}(\cdot)$ is the multi-layer perceptron for non-linear transformation, and $Q_i + MC_i$ indicates the residual connection ensures stable learning. This attention-driven fusion module effectively aligns semantic information across different scales, reduces the semantic gap between encoding and decoding attributes, and significantly improves contextual understanding. As a result, the CIF module is highly effective in complex environments such as post-disaster building assessment, where structural variations and ambiguous regions require robust multi-scale semantic integration.

3.1.3. Pixel Attention Upsample Guidance (PAUG) Module (Decoder)

Subsequently, in the decoding part, a new framework, the PAUG module, is presented to substitute the traditional upsampling approaches like transposed convolution and bilinear interpolation that can only upscale the spatial resolution but lack improvement on semantic quality. Different from the aforementioned methods, the PAUG module advances the deep features D_i and semantically enhanced features S_i , generated by the fusion mechanism based on transformers for smart upsampling. Specifically, deep features D_i are interpolated using bilinear interpolation and aligned with the corresponding S_i . Then, both feature paths pass through the same process of 1×1 convolution and normalization; the module learns pixel-wise semantic correlation by two feature mapping functions $f_S(\cdot)$, and $f_D(\cdot)$ generates a dynamic weighting factor:

$$\sigma = \text{Sigmoid}(f_D(\vec{v}_D) \cdot f_S(\vec{v}_S)) \quad (10)$$

Here σ denotes the probability that a feature from both branches belongs to the same object. The adaptive confidence of the last upsampled feature is given as:

$$\text{Output}_{PAUG} = \sigma \vec{v} + (1 - \sigma) \vec{v}_S \quad (11)$$

The dynamic weighting strategy at the pixel level makes it possible for the decoder to concentrate on reconstructing features that contain more accurate information. Afterward, the model maintains the U-shaped architecture symmetrically while ensuring end-to-end training; this model is primarily concerned with reconstructing features and conducting semantic modeling. It has high tolerance for cases such as building damage evaluation after disaster imaging segmentation with noisy, poor-quality images.

3.1.4. Dynamic Residual Multi-Scale Convolution Module (DRMSC)

The DRMSC module is suggested to integrate features from different receptive field sizes, enabling richer contextual semantic understanding. Otherwise, conventional single-scale convolutions are limited; DRMSC integrates dilated convolutions with varying dilation rates to find both well-grained contextual details and wider global search without reducing spatial resolution.

Given an input feature map F , a multi-scale feature is extracted by a set of parallel branches using 3×3 convolutions with various dilated rates $d \in \{1, 2, 5, 8\}$:

$$F_d = \text{Conv}_{3 \times 3}^{(d)}(F) \quad (12)$$

In order to avoid the grid effect that is common in dilated convolutions, the initial branch consists of stacked convolutions with varying dilation rates, whereas further branches gradually reduce the number of dilations for the purpose of achieving variation in their receptive fields. The output from all these branches is concatenated:

$$F_{ms} = \text{Concaten}(F_1, F_2, F_5, F_8) \quad (13)$$

where F_{ms} is the mitigates the feature sequentially, Concaten denotes the concatenation, subsequently, through a 1×1 convolution for channel compression and feature alignment:

$$F_{\text{output}} = \text{Conv}_{1 \times 1}(F_{ms}) \quad (14)$$

To enhance feature propagation and stabilize training, residual connections are integrated within each branch:

$$F_d^{\text{res}} = F_d + F \quad (15)$$

In addition, Dropout is used during training to improve the network's capacity to generalize. The suggested DRMSC block improves segmentation outcomes by combining multi-scale receptive fields with residual learning, which enables the block to collect both local and global structural dependence

information. Evaluation of building damage following disasters using low-quality pictures is one example of a challenging application where a model can be very helpful. Figure 5 shows the framework of the DRMSC module.

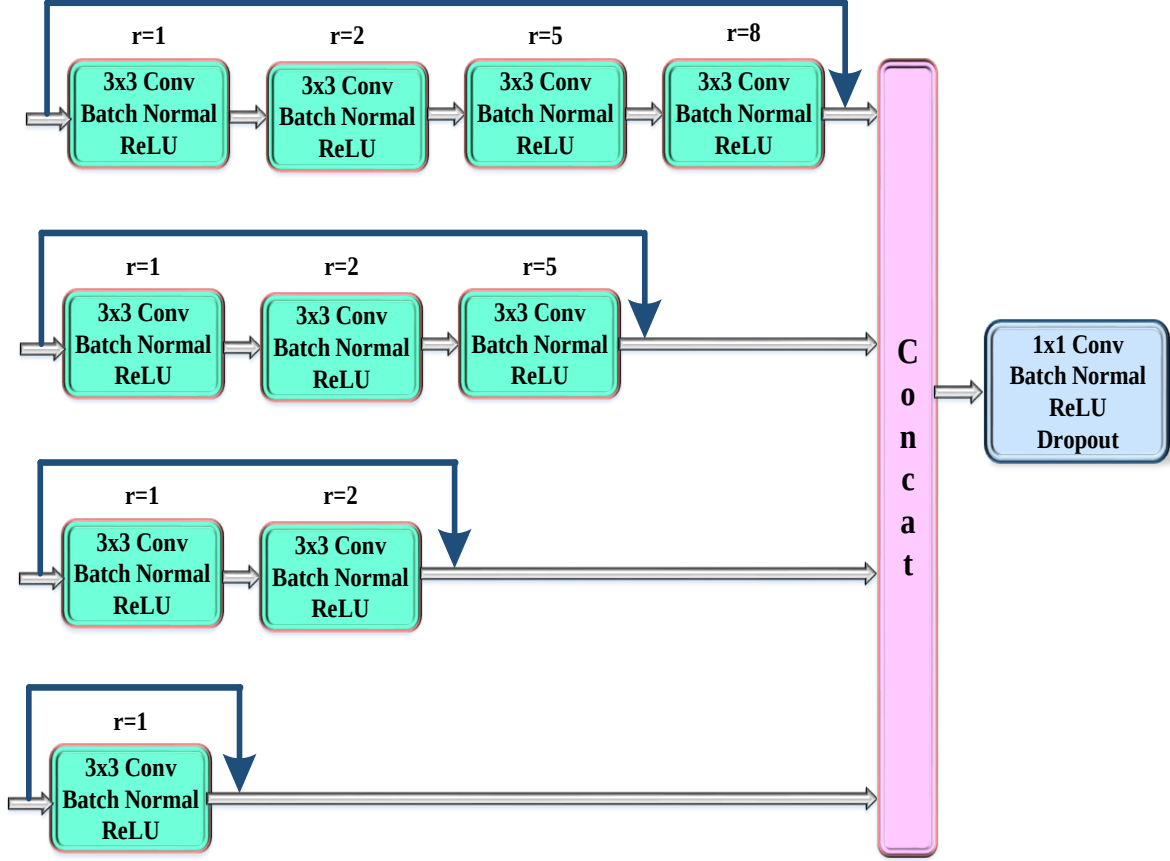


Fig. 5 Framework of DRMSC module

3.1.5. Dynamic Structured Pruning Mechanism

Lastly, a dynamic structured pruning mechanism is utilized to reduce computational complexity and model size while highlighting efficient performance.

Rather than using a fixed threshold, the developed method adaptively learns a pruning threshold that evolves during training, allowing gradual sparsification of the framework. Given a DNN $\theta^{(L)}$ with L layers, the dynamic pruning threshold at iteration t is computed as:

$$\psi_t = \sum_i^L \Omega_i^{(l)} \cdot p_t \quad (16)$$

Here $\Omega_i^{(l)}$ denotes the significance score of i^{th} neuron in level, and p_t indicates the sparsity factor at the iteration t , denoted as:

$$p_t = p_f + (p_i - p_f) \left(1 - \frac{t - t_0}{N \Delta t}\right)^3 \quad (17)$$

Where p_i and p_f denote the initial and target sparsity levels, t_0 is the starting epoch of pruning, and N is the total number of pruning steps. Based on this threshold, neurons are selectively retained or removed using binary masks:

$$m_i^{(l)} = \begin{cases} 1, & \text{if } |m_i^{(l)} \theta_i^{(l)}| \geq \psi_t \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

Through the dynamic nature of this technique, the pruning process is made such that unimportant neurons are trimmed in a gradual manner without compromising important feature representations.

In addition, this method is efficient in achieving the best combination of accuracy and efficiency by gradually varying the level of sparsity during training, as opposed to static methods, which tend to suffer from sudden drops in accuracy. This technique is therefore ideal for application in disaster recovery imaging applications.

3.2. Loss Function

In this research, Categorical cross-entropy is used for the purpose of defining the loss function that would train the segmentation method. Categorical cross-entropy is traditionally used in multi-class classification and segmentation problems, where pixels are mapped to one of a few predefined classes using a ground-truth label distribution against a predicted probability distribution. Specifically, for a given pixel, the categorical-entropy loss is given as:

$$L_{CE} = -\sum_{c=1}^C y_c \log(\hat{y}_c) \quad (19)$$

Here C denotes the total number of classes, y_c indicates the ground truth label, and $\hat{y}_c$ represents the predicted probability for the class obtained after the softmax activation. The loss emphasizes incorrect detections by assigning higher values when the predicted probability deviates from the true class label.

In this research, categorical cross-entropy serves as an effective tool in enforcing accuracy when it comes to pixel-level classifications. Categorical cross-entropy plays an important role in the learning process because it helps the model learn discriminative features from varying semantic classes. It complements the modules that have been proposed, such as RSPC, CIF, PAUG, and DRMSC. This technique is especially helpful for difficult applications, such as building assessments after disasters.

4. Experimental Setup Analysis

Results and discussion describe the evaluation of model performance with other models to prove the RSPF-MST model for automatically assessing building damage in post-disaster scenarios. The proposed model is implemented and validated using a Python tool. Table 2 presents the system settings, and Table 3 indicates the proposed model-based hyperparameter details.

Table 2. System settings

Parameter	Settings
Tool	PYTHON 3.10.14, Anaconda, IDE Spyder
Processor	Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz 3.40 GHz
Installed RAM	16.0GB
System type	64-bit processor system, x64-based processor

Table 3. Proposed model-based hyperparameter details

Parameter	Value	Description
Optimized	Adam	Adaptive gradient optimization
Learning rate	1×10^{-10}	Initial learning rate
Batch size	32	Depends on GPU memory
Epochs	300	Total training iterations
Loss function	Categorical cross-entropy	Pixel-wise classification loss
Weight decay	$1e-5$	Regularization parameter
Dropout rate	0.3	Prevents overfitting
Activation function	Softmax	Non-linear transformation

4.1. Dataset Used

Dataset-1: The data used in the experiment is from the DesignSafe data repository database, which involves very high-resolution pre- and post-disaster satellite pictures. Figure 6 depicts Dataset-2 based on the actual, ground truth & predicted mask outputs.

There is a total of 5,598 samples, split into a training set and test set on an 80:20 split, with 4,478 images in the training set and 1,120 in the testing set. For more dataset details:

<https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ3563/%2Fpublished-data%2FPRJ-3563%2FProject--ida-bd-pre-and-post-disaster-high-resolution-satellite-imagery-for-building-damage-assessment-from-hurricane-ida%2Fdata%2Fmasks>

Dataset-2: Additionally, the paper uses RescueNet as another data benchmark. Figure 7 provides Dataset-2 based on the actual, ground truth & predicted mask outputs. This dataset is a large-scale benchmark used for post-disaster scene analysis with high-resolution UAV imagery.

The main advantage of the benchmark is that it captures urban scenes impacted by natural disasters, which include damaged buildings, debris, road infrastructure, and vegetation, among others.

To balance the test and training process of the model, an 80:20 training and testing distribution is implemented. As a result, the training data set includes 3,595 images, and the testing data set includes 899 images. <https://github.com/BinaLab/RescueNet-A-High-Resolution-Post-Disaster-UAV-Dataset-for-Semantic-Segmentation>

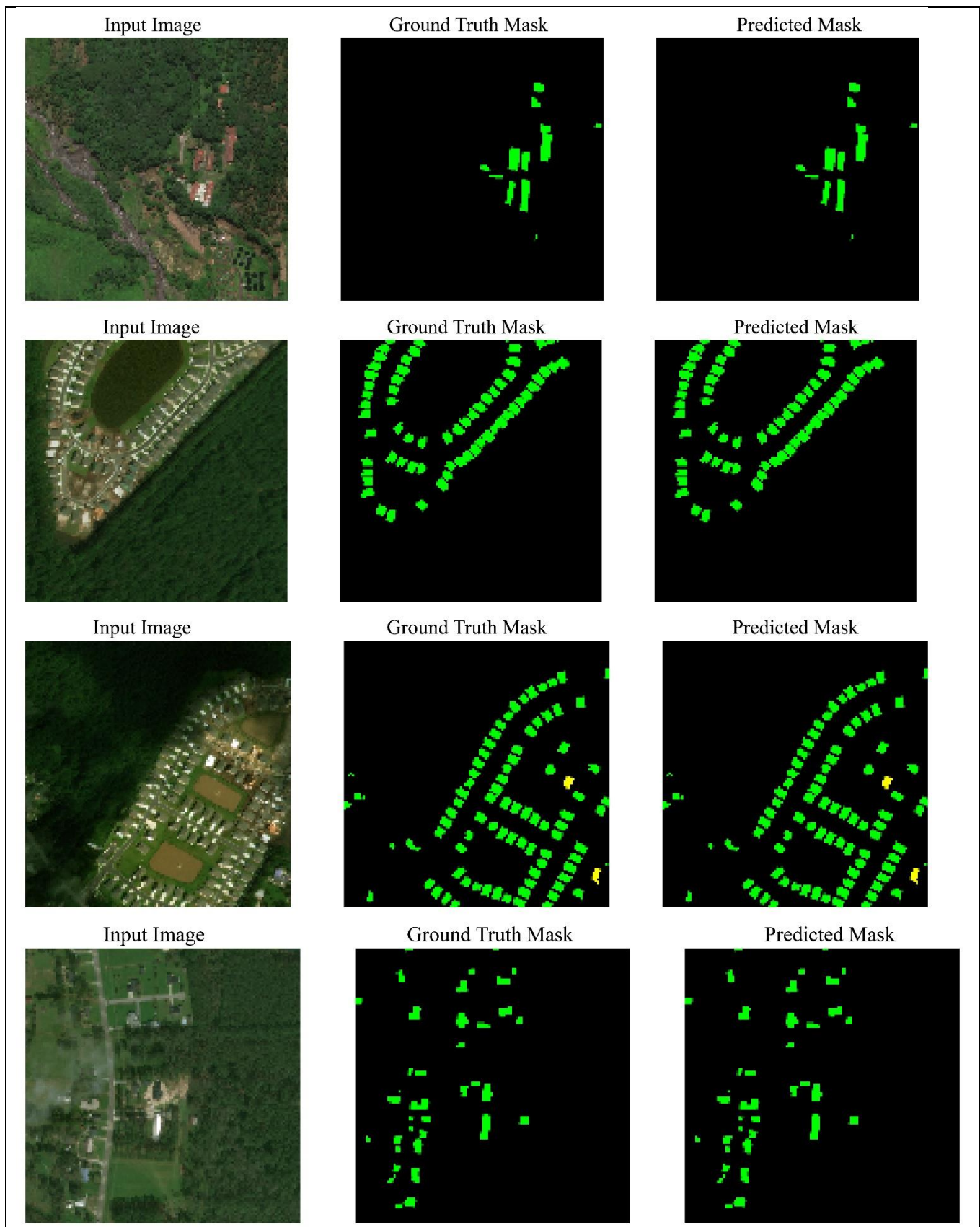


Fig. 6 Dataset-1 based on the actual, ground truth & predicted mask outputs

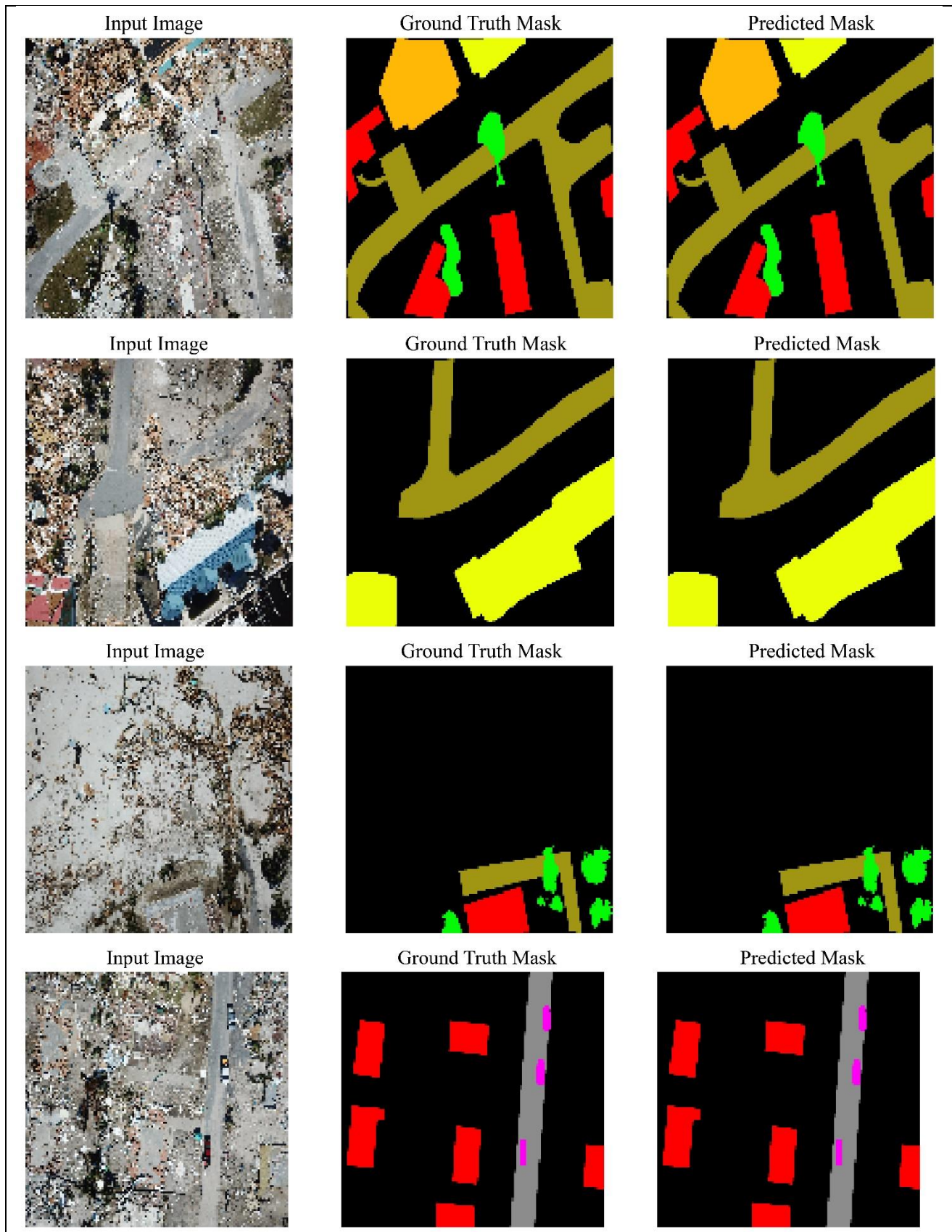


Fig. 7 Dataset- 2 based on the actual, ground truth & predicted mask outputs

It means that training data allow the model to learn diverse spatial relations, structure differences, and features under such complicated conditions as occlusion, lighting changes, and cluttered backgrounds. However, the testing data set allows evaluating the generalization performance of the proposed framework in the case of post-disaster object detection and classification tasks.

4.2. Validation Outcome Metrics

The experimental efficiency of the RSPF-MST method was evaluated using some performance metrics, namely Intersection over Union (IoU), mean Intersection over Union (mIoU), F1-Score, Overall Accuracy (OA), Precision, Recall, and so on. Also, the comparison with state-of-the-art methods is provided.

The precise results refer to the percentage of attributes that are actually progressive among those the model predicts as positive. The following equation can be used to express accuracy in image segmentation:

$$Precision = \frac{TP}{TP+FP} \quad (20)$$

$$Recall = \frac{TP}{TP+FN} \quad (21)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN+FP+TP} \quad (22)$$

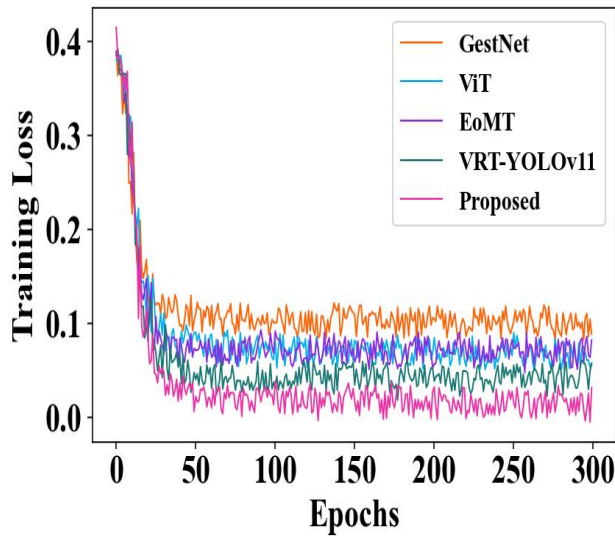
$$Dice = \frac{2TP}{2TP+FN+FP} \quad (23)$$

$$PixelAccuracy = \frac{TP}{TP+FP} \quad (24)$$

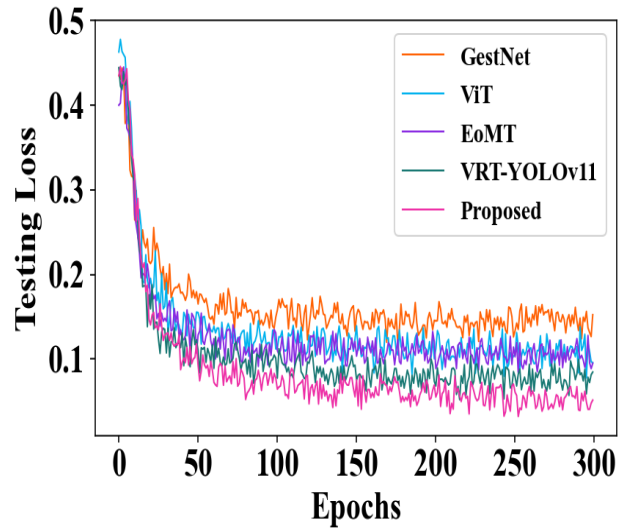
Here, True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

4.3. Results Validation Analysis

Figures 8 (a) and (b) presents the training and validation loss curves over 300 epochs; the RSPF-MST model shows its superiority in terms of convergence and generalization properties. While at the beginning of the training process (epochs 0 to 50) all three models present an exponentially fast decay in training and testing losses, which reflects their quick feature learning capability, the real differences become obvious during the stabilization period of the training.



(a) Training loss



(b) Testing loss

Fig. 8 (a)-(b) Training & testing losses vs. epochs

For the experimental performance in automatic building evaluation after a disaster, it can be seen clearly that the RSPF-MST model performs much better than any baseline architecture concerning both metrics.

Figure 9 (a), the RSPF-MST model attains the top precision rate of 97.04% by achieving the biggest gap with the second top one, VRT-YOLOv11 (94.04%), and greatly surpassing others such as EoMT (91.89%), ViT (90.61%), and GestNet (89.11%). It suggests a high level of reliability in accurately identifying the post-disaster building structures without producing false positives.

The same trend also appears in recall in Figure 9 (b), where the proposed model achieves 96.69%, ensuring nearly all damaged or related buildings are detected. Compared with the second-best model, VRT-YOLOv11 (94.56%), and other models such as EoMT (92.12%), ViT (91.38%), and GestNet (89.6%), they perform relatively worse in detecting buildings with damage.

As a result, both the high scores in precision and recall indicate that the multi-scale fusion model successfully avoids false positives and misses at the same time.

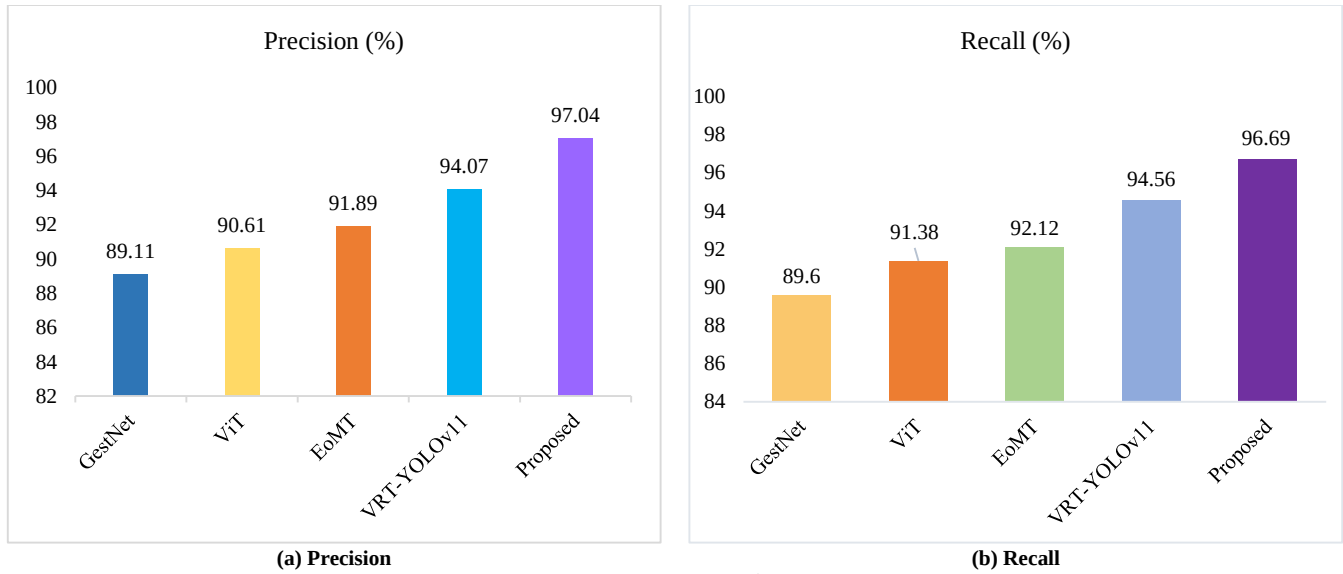


Fig. 9 (a)-(b) Precision & recall performance analysis

Quantitative Analysis of Post-Disaster Automatic Building Assessment shows that the RSPF-MST model performs better than state-of-the-art models in segmentation and localization. Pixel Accuracy Figure 10 (a), the proposed model delivers a high score of 98.43%, which is a small but highly relevant improvement compared to the performance levels achieved by VRT-YOLOv11 (97.33%), EoMT (95.15%), ViT (93.33%), and GestNet (90.55%). The pixel-

based analysis of the proposed model's performance is additionally backed up by the high results delivered by the Jaccard Index, Figure 10 (b), also known as Intersection over Union (IoU), in which the RSPF-MST model scores an amazing 96.13%. The score outperforms the results produced by VRT-YOLOv11 (94.1%) by more than 2% while vastly outscoring those provided by EoMT (93.13%), ViT (91.16%), and GestNet (88.3%).

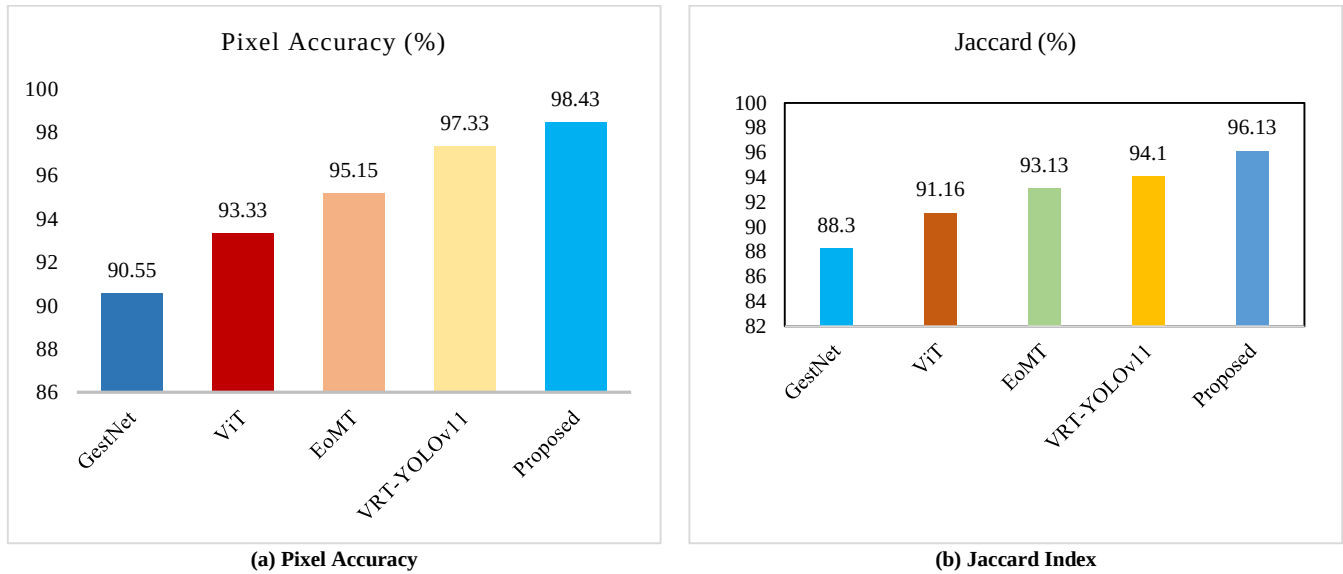


Fig. 10 (a)-(b) Pixel accuracy & Jaccard index analysis

Figures 11 (a) and (b) IoU further proves the capability of the RSPF-MST model in segmenting buildings based on structures. In the case of Dice Score in Figure 11 (a), the proposed model shows excellent overlapping ability with the value of 97.45% and exceeds all other baselines, including VRT-YOLOv11 (95.82%) and EoMT (95.79%), creating an even bigger difference between itself and ViT (92.69%) and

GestNet (89.83%). Similarly, in the case of IoU in Figure 11 (b), the proposed model obtained 96.13%, followed by VRT-YOLOv11 (94.1%), EoMT (93.13%), ViT (91.16%), and GestNet (88.3%). The consistently high values of both Dice and IoU indicate better performance of the multi-scale spatial-pixel fusion technique in boundary detection and region estimation.

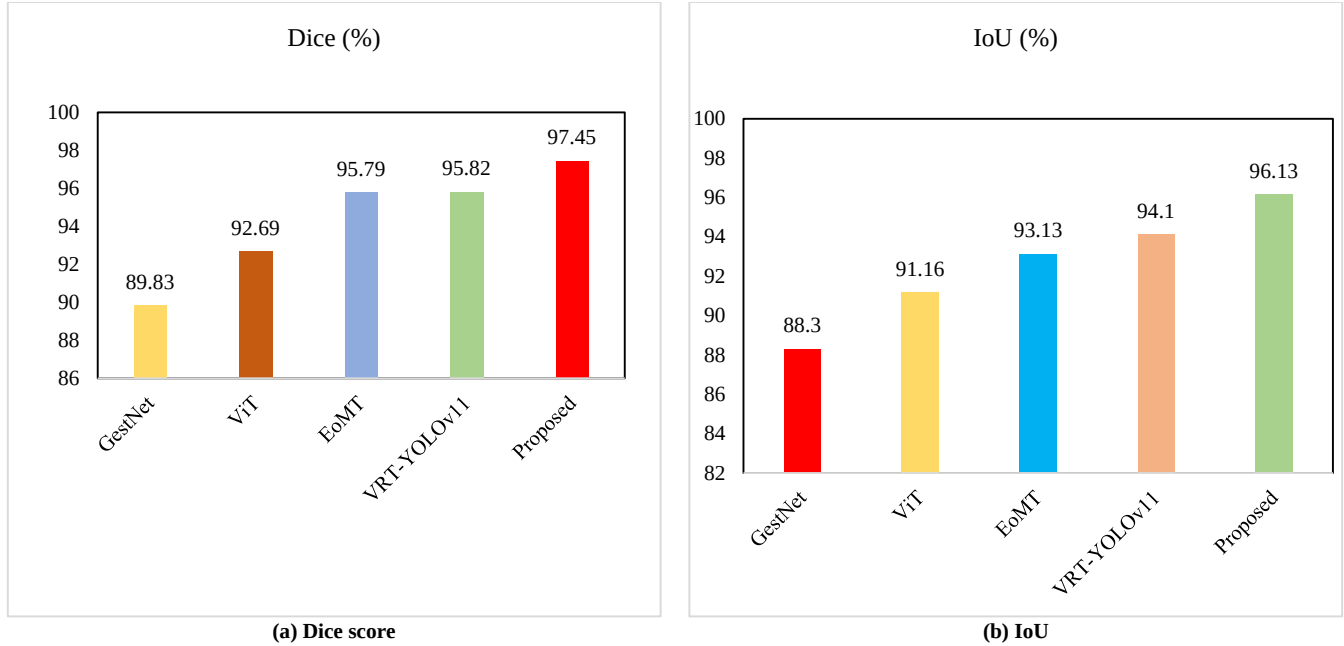


Fig. 11 (a)-(b) Dice score & IoU analysis

Figures 12 (a) and (b) illustrates the Hausdorff Distance (HD (mm)) and FLOP analysis of the precision of the RSPF-MST model, when compared to other segmentation models. As for the metric measuring boundary delineation accuracy through the HD, the smaller its value, the better the model's performance, and the RSPF-MST obtains an outstandingly low value of only 2.08 mm. This considerably surpasses VRT-YOLOv11 (5.46 mm), EoMT (6.31 mm), ViT (9.32 mm), and GestNet (10.61 mm), thereby implying that the multi-scale spatial-pixel fusion dramatically improves boundary accuracy

and enables precise structural location. On the other hand, considering the metric representing the computational challenges in FLOPs, it turns out that the RSPF-M model value (17.94) is slightly greater compared to those for VRT-YOLOv11 (17.72), EoMT (17.36), ViT (16.98), and GestNet (16.45). However, it may be argued that this slight increase in floating-point operations can be considered completely acceptable in view of the corresponding performance improvement.

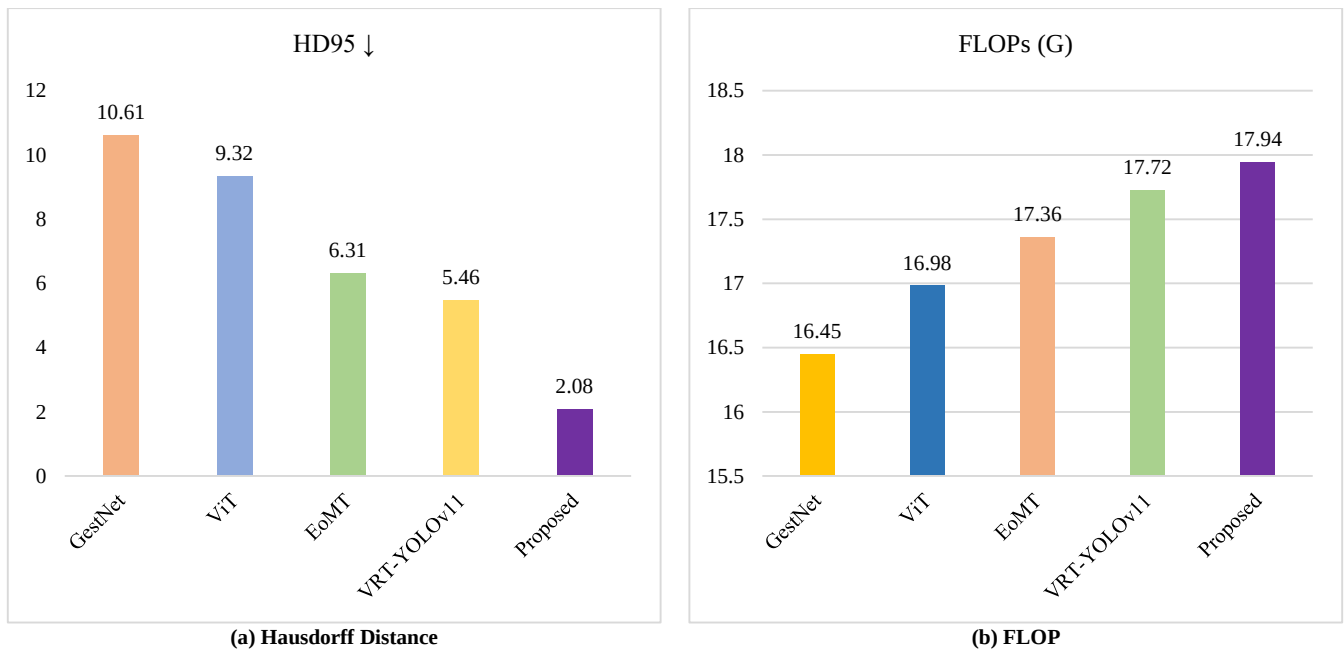


Fig. 12 (a)-(b) HD & FLOP analysis

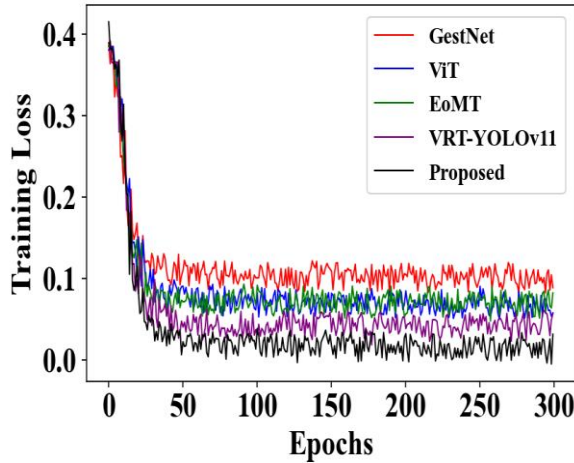
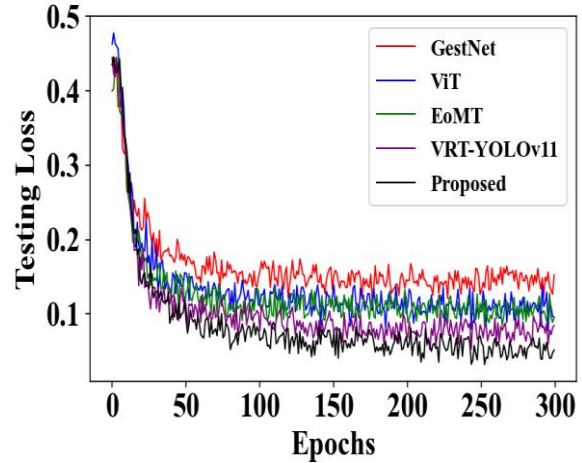
Table 4. Performance validation of the proposed model

Model	Dice (%)	IoU (%)	Jaccard (%)	HD95 ↓	FLOPs (G) ↓	Pixel Accuracy (%)	Precision (%)	Recall (%)
GestNet	89.83	88.3	88.3	10.61	16.45	90.55	89.11	89.6
ViT	92.69	91.16	91.16	9.32	16.98	93.33	90.61	91.38
EoMT	95.79	93.13	93.13	6.31	17.36	95.15	91.89	92.12
VRT-YOLOv11	95.82	94.1	94.1	5.46	17.72	97.33	94.07	94.56
Proposed	97.45	96.13	96.13	2.08	17.94	98.43	97.04	96.69

Dataset-2

Figures 13 (a) and (b) indicates the results in training and test loss; the RSPF-MST model performs better than other models in automatic post-disaster building damage assessment after 300 epochs. All models start with a large reduction in losses but converge to the smallest loss achieved

by the RSPF-MST model. From 50 to 300 epochs, the RSPF-MST model has the smallest and most stable training loss values at almost 0.0 and test loss at 0.05. The proposed model is much superior to VRT-YOLOv11, which shows high loss values compared to GestNet.

**(a) Training loss****(b) Testing loss****Fig. 13 (a)-(b) Training & testing loss analysis**

Figures 14 (a) and (b) represents the precision & Recall performance analysis, highlighting that the RSPF-MST model performs significantly better than other competing models for the task of automated assessment of buildings after disasters.

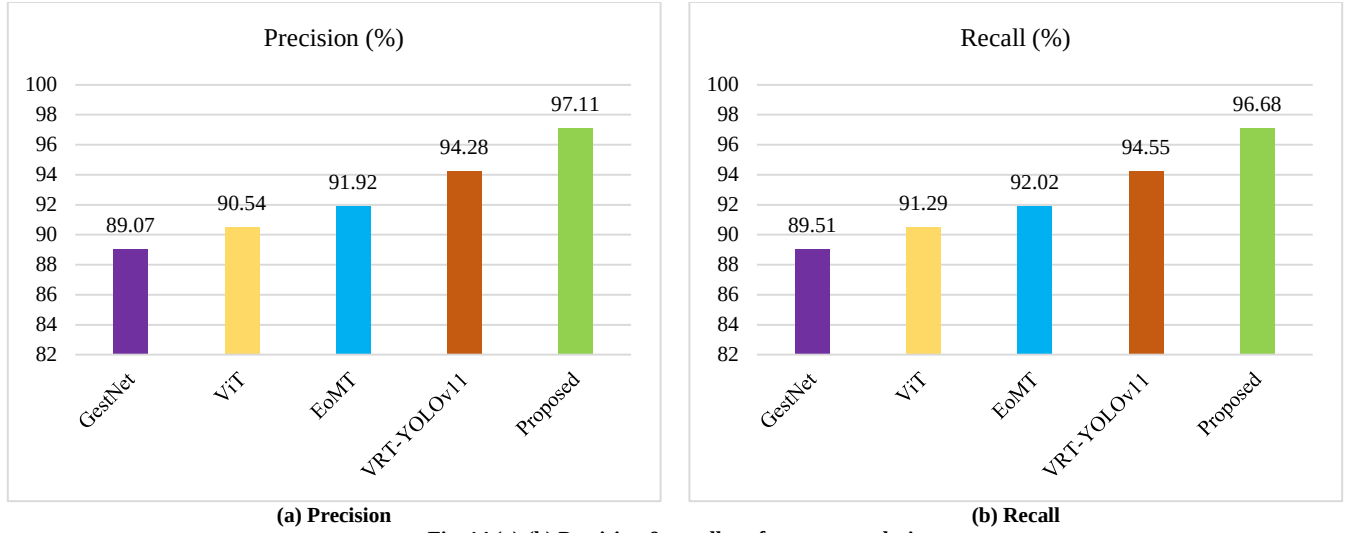
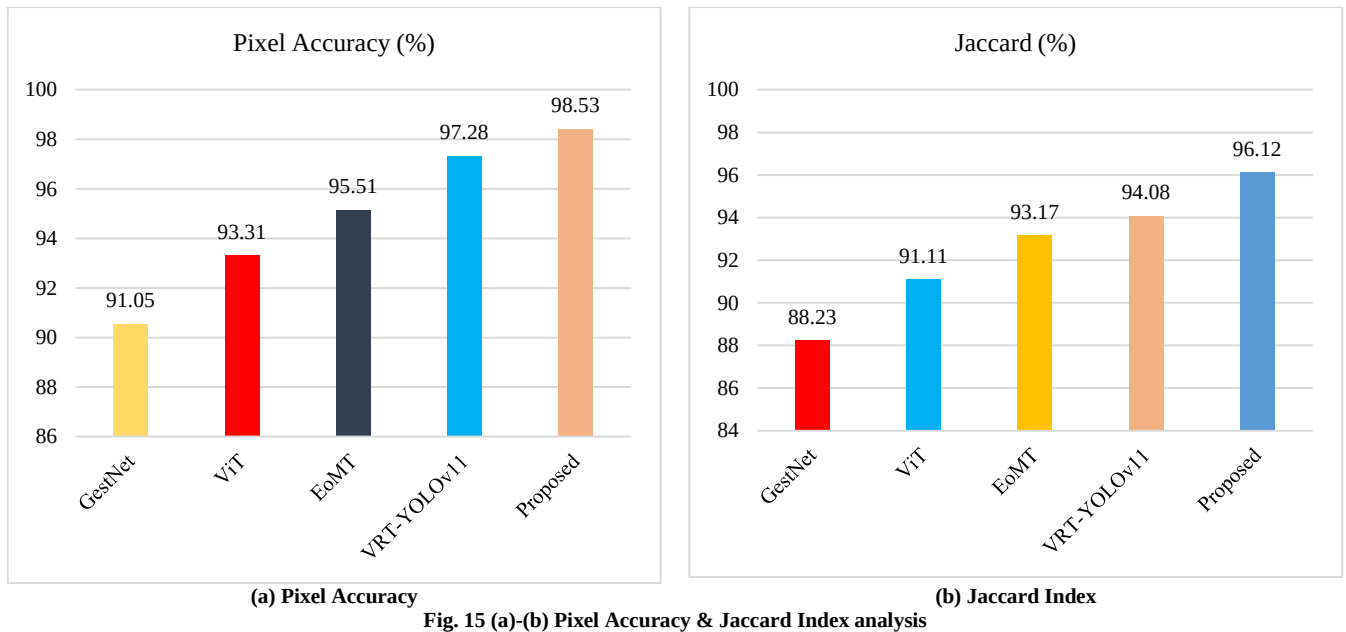
Figure 14 (a) Precision, the proposed model shows the highest efficiency of 97.11% compared to competing methods, including VRT-YOLOv11 (94.28%), EoMT (91.92%), ViT (90.54%), and GestNet (89.07%). Thus, this model is able to accurately classify damaged buildings without causing any false positive detections.

Furthermore, similar outstanding performance can be seen in Figure 14 (b) Recall, where the highest value of 96.68% is recorded for the proposed model. In comparison, VRT-YOLOv11 (94.55%), EoMT (92.02%), ViT (91.29%), and GestNet (89.51%) show relatively lower values.

Figures 15 (a) and (b) displays that the RSPF-MST model segmentation efficacy is superior in the automatic post-disaster building assessments, based on the proposed metrics. First, Figure 15 (a) shows that the Pixel Accuracy based on the RSPF-MST model has the best performance with 98.53%.

In addition, compared to other competing models like VRT-YOLOv11 (97.28%), EoMT (95.51%), ViT (93.31%), and GestNet (91.05%), the RSPF-MST model outperforms all the others.

The accurate localization of the objects is also evident when analyzing the Jaccard Index in Figure 15 (b), where the RSPF-MST model takes the lead with 96.12% against VRT-YOLOv11 (94.08%), EoMT (93.17%), ViT (91.11%), and GestNet (88.23%).



Figures 16 (a) and (b) validates the greater overlapping and segmenting abilities of the RSPF-MST approach for automatic evaluation of buildings post-disaster. From the results of the Dice Score in Figure 16 (a), it can be seen that the RSPF-MST method scores its best performance of 97.46%, surpassing the nearest state-of-the-art methods, VRT-YOLOv11 (95.84%) and EoMT (95.82%), and setting a much larger margin ahead of ViT (92.74%) and GestNet (89.85%). The same trend holds for Figure 16 (b) IoU, in which the RSPF-MST method scores the maximum at 96.12%, beating VRT-YOLOv11 (94.08%), EoMT (93.17%), ViT (91.11%), and GestNet (88.23%). The parallel achievement of the maximum Dice and IoU metrics confirms the effectiveness of the multi-scale spatial-pixel fusion methodology to capture geometric information and alignment with boundaries, and hence minimize false segments.

Figures 17 (a) and (b) depicts the performance measurement criteria of the RSPF-MST model, showing the balance between boundary precision and computation time required for automatic post-disaster building analysis. It should be mentioned that the RSPF-MST model performs better in terms of achieving a lower HD value (2.07 mm), which proves the higher quality of boundary precision in comparison with other base models: VRT-YOLOv11 (5.46 mm), EoMT (6.19 mm), ViT (9.3 mm), and GestNet (10.57 mm). In turn, the RSPF-MST model requires 17.45 FLOPs, which is a slight decrease in computations compared to VRT-YOLOv11 (17.88), EoMT (17.92), ViT (18.17), and GestNet (18.61). Therefore, the increase in computations can be regarded as a fair compromise for significant improvements in geometry edge detection and localization. The performance validation of the proposed model is provided in Table 5.

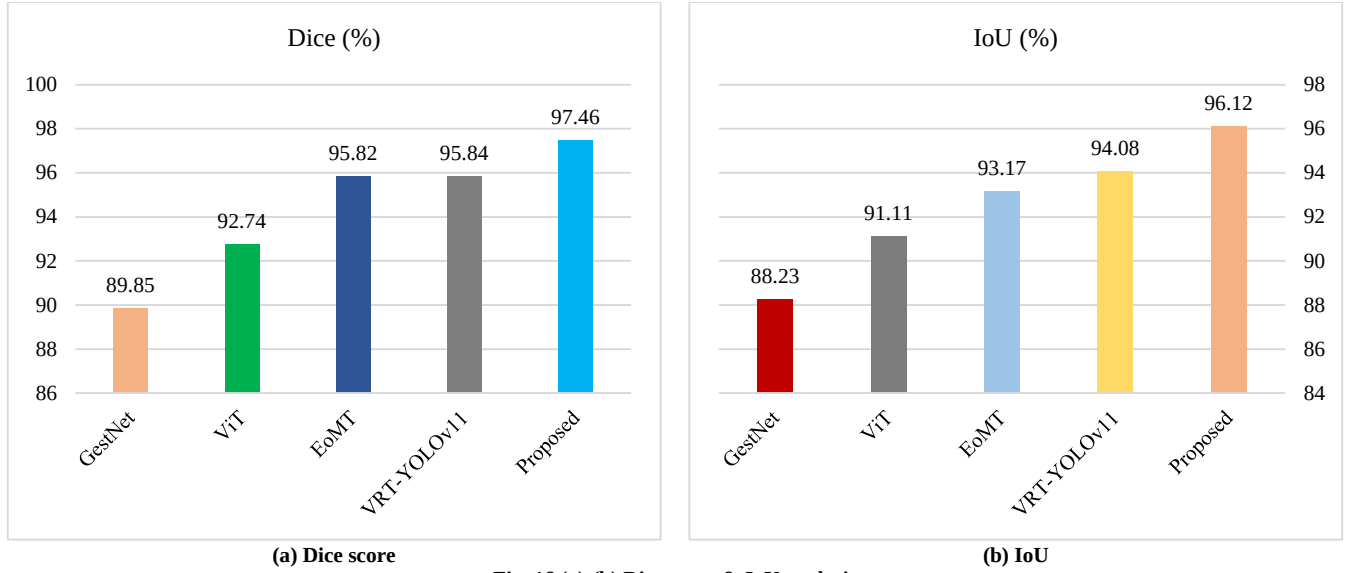


Fig. 16 (a)-(b) Dice score & IoU analysis

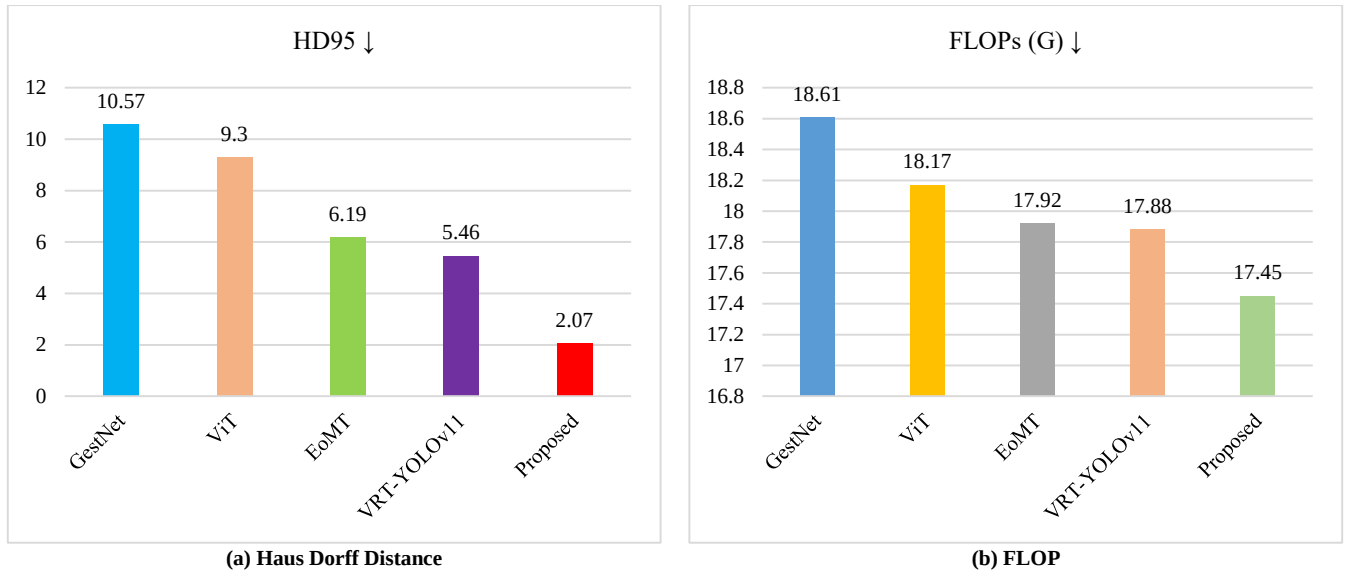


Fig. 17 (a)-(b) HD & FLOP analysis

Table 5. Performance validation of the proposed model

Model	Dice (%) ↑	IoU (%) ↑	Jaccard (%) ↑	HD95 ↓	Pixel Accuracy (%) ↑	Precision (%) ↑	Recall (%) ↑	FLOPs (G) ↓
GestNet	89.85	88.23	88.23	10.57	91.05	89.07	89.51	18.61
ViT	92.74	91.11	91.11	9.3	93.31	90.54	91.29	18.17
EoMT	95.82	93.17	93.17	6.19	95.51	91.92	92.02	17.92
VRT-YOLOv11	95.84	94.08	94.08	5.46	97.28	94.28	94.55	17.88
Proposed	97.46	96.12	96.12	2.07	98.53	97.11	96.68	17.45

4.4. Discussion

Given the combined shortcomings of DamageCAT with hierarchical U-Net, GAMSF, Swin Transformer-ConvNeXt, CMSNet, and Triple U-Net, the suggested framework outperforms these alternatives in terms of segmentation performance. Among DamageCAT's strongest points is its

hierarchical U-Net architecture, which allows for hierarchical feature extraction. However, this approach isn't without its limitations, such as the possibility of redundant features and an inability to adequately capture long-range semantic relationships (caused by convolution-based feature representation). Even though GAMSF undergoes traditional

multi-scale fusion, which causes semantic inconsistency between encoder and decoder representations, it efficiently aggregates features across scales.

On the other hand, Swin Transformer and ConvNeXt have high performance in terms of global and local feature learning. Still, they are more complicated in terms of architecture and might lead to more computation due to their complexity and lack of focus on fine boundary reconstruction. Moreover, GestNet, Vision Transformer (ViT), EoMT, and VRT-YOLOv11 are some existing models that show good performance in segmentation and detection. However, each of them faces certain limitations when used for solving complex problems related to building evaluation after disaster images of poor quality. Specifically, GestNet is dependent on convolution operations for feature extraction, thereby limiting the model's capability to detect long dependency relationships between features and resulting in redundant features.

While ViT addresses the problem of global context modeling, it fails in local feature representations and needs huge amounts of data for proper training. EoMT improves multi-scale feature learning, but it can lead to inconsistencies between the representations of semantic information among

encoder and decoder levels due to the lack of a feature alignment mechanism. To address these limitations, the proposed framework incorporates an RSPC, CIF, PAUG, DRMSC, and dynamic structured pruning.

Specifically, RSPC is aimed at decreasing the redundancy of spatial and channel levels in the process of encoder feature extraction in order to improve the retention of structural features. CIF substitutes traditional skip connections with cross-channel interactions, thus matching the spatial details from shallow levels with the semantic representation from deeper levels. PAUG implies applying the adaptive feature weighting at the pixel level in order to improve boundary preservation and recovery of damaged areas. Lastly, DRMSC extracts the information about local and global context via the use of several receptive fields. The pruning technique gradually eliminates less significant neurons, thereby minimizing the size of the model and the level of computational complexity, but still retaining all significant features. Hence, the method's superior performance stems from the synergistic effects of these components, leading to more accurate segmentation results, especially in complex post-disaster scenarios. Table 6 indicates the RSPF-MST model compared with state-of-the-art models.

Table 6. RSPF-MST model compared with state-of-the-art models

Authors & ref	Models	Dataset	Performance
Liu et al. [26]	TatchGANs	MDS-EB dataset	Accuracy-90%
Bhardwaj et al. [27]	ResNet-UNet	High-resolution satellite image dataset	Accuracy-96.91%
Sundaresan et al. [28]	DeepLabV3+	UAV imagery dataset	Accuracy-0.8620, MIoU-0.6286
Le et al. [29]	CMSNet	3DAeroRelif dataset	Precision-83.95, and MIoU-0.4584
Kaya [30]	MaskR-CNN-DNN	RS, RADARSAT, ENVISAT, COSMO-SkyMed and Sentinel-1	Precision-93.43, F1-score-95.42
Proposed	RSPF-MST model	Ida-BD and RescueNet datasets	Dataset- 1 accuracy of 98.43%, IoU of 96.13%, and Dataset-2- accuracy-98.53% and IoU-96.12%

5. Conclusion & Future Scope

The present work proposes a model known as Reconstruction Spatial-Pixel Fusion Multi-Scale Transformer to facilitate accurate segmentation in tough environments, including post-disaster building damage inspection. Using the RSPC component for redundancy elimination, CIF for efficient multi-scale semantic fusion, PAUG for accurate boundary-aware up-sampling, and DRMSC for extracting contextual information, the proposed model performs better in feature extraction and segmentation than other existing techniques. Moreover, employing dynamic pruning improves the efficiency of the model by decreasing its complexity without sacrificing its performance. Experimental evaluation shows that the model works effectively with noisy, low-contrast, and structurally difficult images. In terms of future work, the proposed approach can be extended to 3D volume

data for sophisticated real-time disaster monitoring systems. The proposed system can be further enhanced by introducing lightweight attention modules for edge computation and using self-supervised or semi-supervised learning to reduce reliance on labeled data. In addition, the use of other modalities, such as LiDAR and thermal data, can improve the performance of the model in challenging situations.

Conflicts of Interest

The author(s) declare that there is no conflict of interest regarding the publication of this paper.

Funding Information

No funding is provided for the preparation of the manuscript

References

- [1] Zhenghang Yuan et al., “RRSIS: Referring Remote Sensing Image Segmentation,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-12, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Chen Wu et al., “UNet-Like Remote Sensing Change Detection: A Review of Current Models and Research Directions,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 12, no. 4, pp. 305-334, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] Guangliang Cheng et al., “Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review,” *Remote Sensing*, vol. 16, no. 13, pp. 1-36, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Dongzhe Han et al., “A Multi-Level Damage Assessment Model based on Change Detection Technology in Remote Sensing Images,” *Natural Hazards*, vol. 121, no. 6, pp. 7367-7388, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Wenfan Qiao et al., “Revolutionizing Building Damage Detection: A Novel Weakly Supervised Approach using High-Resolution Remote Sensing Images,” *International Journal of Digital Earth*, vol. 17, no. 1, pp. 1-26, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Lili Wang et al., “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*, vol. 17, pp. 15301-15318, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] M. Sadman Sakib, and C. Emdad Haque, “Engaging Military with Domestic Disaster Response: A Comparative Study of Institutional Structure,” *International Journal: Canada’s Journal of Global Policy Analysis*, vol. 80, no. 3, pp. 486-512, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Hongruixuan Chen et al., “BRIGHT: A Globally Distributed Multimodal Building Damage Assessment Dataset with Very-High-Resolution for All-Weather Disaster Response,” *Earth System Science Data*, vol. 17, no. 11, pp. 6217-6253, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Aaqib Ali et al., “Applications of Remote Sensing Technologies to Assess Ground Displacements and Seismic Hazards: Precision, Scale, and Insight,” *Arabian Journal for Science and Engineering*, vol. 50, no. 20, pp. 16357-16378, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Sultan Al Shafian, and Da Hu, “Integrating Machine Learning and Remote Sensing in Disaster Management: A Decadal Review of Post-Disaster Building Damage Assessment,” *Buildings*, vol. 14, no. 8, pp. 1-32, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Jiancheng Gu et al., “Advances in Rapid Damage Identification Methods for Post-Disaster Regional Buildings based on Remote Sensing Images: A Survey,” *Buildings*, vol. 14, no. 4, pp. 1-35, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Dilum Rajapaksha et al., “Systematic Mapping of Global Research on Disaster Damage Estimation for Buildings: A Machine Learning-Aided Study,” *Buildings*, vol. 14, no. 6, pp. 1-26, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Tania Islam et al., “A Systematic Review of Urban Flood Susceptibility Mapping: Remote Sensing, Machine Learning, and other Modeling Approaches,” *Remote Sensing*, vol. 17, no. 3, pp. 1-37, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Ronny Francis Ribeiro Junior, and Guilherme Ferreira Gomes, “On the use of Machine Learning for Damage Assessment in Composite Structures: A Review,” *Applied Composite Materials*, vol. 31, no. 1, pp. 1-37, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Abdullah M. Braik, and Maria Koliou, “Automated Building Damage Assessment and Large-Scale Mapping by Integrating Satellite Imagery, GIS, and Deep Learning,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 39, no. 15, pp. 2389-2404, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Lan Nguyen-Ngoc et al., “Deep Neural Network and Evolved Optimization Algorithm for Damage Assessment in a Truss Bridge,” *Mathematics*, vol. 12, no. 15, pp. 1-25, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Shayan Ghazimoghadam, and S.A.A. Hosseinzadeh, “A Novel Unsupervised Deep Learning Approach for Vibration-based Damage Diagnosis using a Multi-Head Self-Attention LSTM Autoencoder,” *Measurement*, vol. 229, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Wen Lu, Lu Wei, and Minh Nguyen, “Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 4917-4935, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Masoomeh Gomroki et al., “UNet- GCViT: A UNet-based Framework with Global Context Vision Transformer Blocks for Building Damage Detection,” *International Journal of Remote Sensing*, vol. 46, no. 6, pp. 2587-2610, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Ehsan Govahi, Mojtaba Salkhordeh, and Reza Karami Mohammadi, “A Strengthened Convolutional Neural Network Algorithm for Identifying the Extent of Seismic Damage in Moment-Frame Buildings,” *Structure and Infrastructure Engineering*, pp. 1-22, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Yiming Xiao, and Ali Mostafavi, “DamageCAT: A Deep Learning Transformer Framework for Typology-based Post-Disaster Building Damage Categorization,” *International Journal of Disaster Risk Reduction*, vol. 128, pp. 1-21, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [22] Bo Yu et al., “Post-Disaster Building Damage Assessment based on Gated Adaptive Multi-Scale Spatial-Frequency Fusion Network,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 141, pp. 1-13, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Yifan Yang et al., “Hyperlocal Disaster Damage Assessment using Bi-Temporal Street-View Imagery and Pre-Trained Vision Models,” *Computers, Environment and Urban Systems*, vol. 121, pp. 1-14, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Jianli Zhang et al., “CMSNet: A SAM-Enhanced CNN-Mamba Framework for Damaged Building Change Detection in Remote Sensing Imagery,” *Remote Sensing*, vol. 17, no. 23, pp. 1-36, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Jungeun Cha, Seunghyeok Lee, and Hoe-Kyoung Kim, “Deep Learning-based Detection and Assessment of Road Damage caused by Disaster with Satellite Imagery,” *Applied Sciences*, vol. 15, no. 14, pp. 1-16, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [26] Chaoxian Liu, Haigang Sui, and Lihong Huang, “Minor Damage Recognition from Postearthquake Buildings with an Improved Generative Adversarial Semantic Segmentation Network,” *Natural Hazards Review*, vol. 26, no. 3, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Diwakar Bhardwaj et al., “Enhanced Satellite Imagery Analysis for Post-Disaster Building Damage Assessment using Integrated Resnet-U-Net Model,” *Multimedia Tools and Applications*, vol. 84, no. 5, pp. 2689-2714, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Akila Agnes Sundaresan, and Appadurai Arun Solomon, “Post-Disaster Flooded Region Segmentation using Deeplabv3+ and Unmanned Aerial System Imagery,” *Natural Hazards Research*, vol. 5, no. 2, pp. 363-371, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Nhut Le, Ehsan Karimi, and Maryam Rahnemoonfar, “3DAeroRelief: The First 3D Benchmark UAV Dataset for Post-Disaster Assessment,” *Scientific Data*, vol. 13, no. 1, pp. 1-12, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Efdal Kaya, “Assessment of Building Extraction and Geometric Corrections from UAV Image using Mask R-CNN Deep Neural Network Algorithm,” *European Journal of Remote Sensing*, vol. 59, no. 1, pp. 1-17, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]