Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.088119

ARTICLE

Semantic Segmentation-Based Identification of Seismic-Induced Building Damage by Using UAV Orthophoto

Haijia Wen[1,2,*], **Wanying Zhang**[1], **Yingqi Zeng**[1], **Fengguang Xiong**[1] and **Yujie Li**[1]

[1]School of Civil Engineering, Chongqing University, Chongqing. No. 83 Shabeijie, Shapingba, Chongqing, China
[2]State Key Laboratory of Safety and Resilience of Civil Engineering in Mountain Area, Chongqing, China

*Corresponding Author: Haijia Wen. Email: jhw@cqu.edu.cn

Received: 29 June 2026; Accepted: 21 September 2026; Published: 21 October 2026

ABSTRACT: Due to transportation disruptions after the earthquake, on-site surveys could not quickly assess house damage. Remote sensing can quickly acquire disaster-related information over large-scale affected areas and support rapid localization and morphological identification of damaged buildings from top-view observations. Nevertheless, on-site inspection by experienced engineers is still the most accurate method for judging real building structural damage, since UAV orthophotos only record surface features and cannot capture internal building damage. Field investigation data in our study served as ground-truth references for UAV image interpretation and labelling. Current remote sensing-based post-earthquake building assessment mostly adopts a binary classification system of collapsed and non-collapsed, which makes it difficult to distinguish between lightly damaged and intact buildings, and the imbalance between positive and negative sample sizes leads to low identification accuracy. This article created the orthophoto images from unmanned aerial vehicles in Shuanghe Town, Changning County, Sichuan Province and established a classification standard for building seismic damage based on visual interpretation. The earthquake-damaged buildings were ultimately classified into three damage levels: completely collapsed, partially damaged, and basically intact. Semantic segmentation can identify different objects in an image. This article compares the performance of different models, sample sizes, and loss functions. This study compares 50% overlapping cropping with non-overlapping cropping, and compares a Focal + Dice hybrid loss with cross-entropy, on UAV orthophoto patches of Shuanghe Town. Reported scores are mean metrics on the test block of this single mosaic. The results show that the MobileNetV3 model with Focal Loss and Dice Loss reported the highest mean mIoU, mPA, and F1 scores of 92.36%, 97.32%, and 96.77%, respectively, on the evaluated test block with 512 × 512 pixels and 50% overlapping sampling.

KEYWORDS: Seismic-induced; building damage identification; semantic segmentation; UAV orthophoto

1 Introduction

An earthquake is a sudden and destructive natural disaster that can cause significant harm to people's lives and property [1]. After an earthquake occurs, it is necessary to promptly determine the distribution of earthquake intensity based on the degree of damage [2]. Researchers need to conduct on-site investigations of the damage to buildings to assess earthquake intensity [3].

Following the 1906 San Francisco earthquake, early attempts at aerial documentation included kite-mounted camera systems used to capture post-earthquake imagery of the city, representing one of the earliest known applications of aerial photography for post-earthquake damage assessment. With the continuous advancement of remote sensing technology [4–6], remote sensing has been used for earthquake damage

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

monitoring since the 1960s, and aerial photographs have been used to document damage patterns in affected areas [7]. At first, interpretation relied mainly on manual observation to obtain information on house collapses, with damage levels assigned from visual differences in collapse appearance. There are subjective differences in traditional manual observation methods. It is difficult to guarantee accuracy with this method and cannot meet the needs of rapid earthquake assessment.

Pixel classification relies on mathematical statistics [8]. This method determines the nature of the ground object class represented by the pixel through the spectral information of the image bands. Traditional pixel-based classification technology has the "salt-and-pepper noise" in image analysis [9]. Therefore, it is impossible to distinguish problems such as "different spectra of the same object" and "same spectra of different objects". Compared with traditional pixel classification, object-oriented classification contains richer semantics and has great improvement in understanding depth, classification accuracy, and classification level [10].

The object-oriented classification method begins by using image segmentation technology to generate polygonal objects. The image target is used as a basic unit to extract information. Then, a classification method is used to complete the automatic extraction of the target. Wen et al. established an object-oriented multi-level classification system for the post-earthquake area of Shuanghe Town using the ESP2 algorithm [11]. Researchers have utilized machine learning methods to make identification results faster and more accurate [12,13]. Günen constructed multi-dimensional stacked feature sets from very-high-resolution imagery and terrain-derived products, compared multiple machine-learning classifiers for building rooftop detection, and achieved competitive detection accuracy via feature-selection procedures [14]. Such multi-feature fusion schemes provide valuable references for building-damage interpretation in disaster-affected areas. It should be noted that UAV orthophotos are purely 2D planar representations and may miss some three-dimensional and internal structural details of buildings. Gautam et al. developed machine-learning and soft-voting ensemble classification frameworks for earthquake-induced bridge damage assessment, and quantified the influence of data homogeneity and heterogeneity on model performance [15]. Such feature-driven machine-learning pipelines provide alternative solutions for seismic damage evaluation when annotated deep-learning training samples are insufficient.

In recent years, the development of deep learning technology has enabled the extraction of some important information from images. Deep learning technology has been widely used in agricultural pest detection [16], farmland crop classification [17], autonomous driving [18–20], and other fields. At present, various deep-learning semantic segmentation models have been used to identify the collapse and damage of houses [21]. Hong et al. [22] proposed a new convolutional neural network, namely the Earthquake Building Damage Classification Network (EBDC-Net), which can classify damaged buildings into three and four categories according to the degree of damage. Wang et al. [23] designed a novel loss function that integrates geometric consistency constraints (GCC) with cross-entropy (CE) loss for post-earthquake building segmentation to reduce misidentification. Xu et al. [24] proposed a computer vision and machine learning-based framework for seismic damage assessment of reinforced concrete (RC) structures to reduce errors.

However, most existing studies simplify earthquake damage assessment of buildings as a binary classification, i.e., whether or not the building has collapsed [25]. The traditional binary classification system (collapsed/not collapsed) cannot distinguish between lightly damaged and intact buildings in earthquake damage assessment. Therefore, there is a risk of misclassification, where buildings that are not collapsed are mistakenly deemed safe. Some convolutional neural network-based models perform poorly in house collapse detection due to class imbalance caused by the significant pixel disparity between houses and background scenes. Under the imbalanced training samples, some deviations will be generated, which will affect the

accuracy of the classification results [26,27]. This results in poor identification accuracy of under-represented buildings. In response to the issues raised, this article has made the following contributions:

1. Systematically evaluate and compare four mainstream semantic segmentation models (UNet, DeepLabv3+, MobileNetV3, and OCRNet) for post-earthquake building damage identification using UAV orthophotos in mountain towns, reporting comparative mean metrics on a single orthophoto mosaic under complex terrain and occlusion conditions.
2. Compare 50% overlapping cropping and a Focal + Dice hybrid loss with non-overlapping crops and cross-entropy on this mountainous-town UAV dataset, and report the resulting mean segmentation metrics.
3. Treat seismic damage assessment of buildings as a three-class classification problem.

2 Study Area and Data

2.1 Study Area

On June 17, 2019, a 6.0-magnitude earthquake struck Changning County, Yibin City, Sichuan Province. This earthquake was the strongest within a 50-km radius of the area since records began. As a shallow earthquake, it caused severe damage to buildings. Changning County is located on the southern edge of Sichuan Province, China, with a longitude ranging from 104°44′22″ E to 105°03′30″ E and a latitude ranging from 28°15′18″ N to 28°47′48″ N. The epicentral area, Shuanghe Town, lies to the south of Changning County, adjacent to Guanxing Town and Meidong Town in the east, Xingwen County in the south, Gongquan Town of Gong County in the west, and Longtou Town in the north. The administrative area of Shuanghe Town covers 135.35 km^2. The dataset used in this study was derived from UAV orthophotos of Shuanghe Town, Changning County, as shown in Fig. 1.

The study area is mainly Shuanghe Town, located in Changning County, Yibin City, Sichuan Province, which is the epicentral area of this earthquake. UAV remote sensing images were acquired on June 18, 2019, with a spatial resolution of 0.6 m and three spectral bands (RGB). The image size is 39,200 × 38,282 pixels. After obtaining multiple aerial photographs, PhotoScan software was used to process the images. The remote sensing images and POS data acquired at each monitoring point were imported into the system, and ground control point files were loaded. Through operations such as aligning photos and building dense point clouds, a Digital Elevation Model (DEM) and a Digital Orthophoto Map (DOM) were generated. The UAV orthophoto covers Shuanghe Town only. Surrounding towns and other affected areas of the Changning earthquake are outside the mosaic and are not evaluated.

Shuanghe Town is a typical mountain town in Southwest China. After the earthquake, it features fragmented terrain, dense buildings, and severe occlusion by trees and mountains, which collectively reflect the main challenges of earthquake damage in mountain towns. It encompasses both dense building clusters and scattered rural houses, with a highly imbalanced distribution of partially collapsed, completely collapsed, and largely intact buildings, along with complex image conditions such as mountain shadows, tree occlusions, and variations in roof materials. These conditions are common in mountain towns of Southwest China, but the present dataset covers only Shuanghe Town in one campaign. Transfer to other towns is not demonstrated.

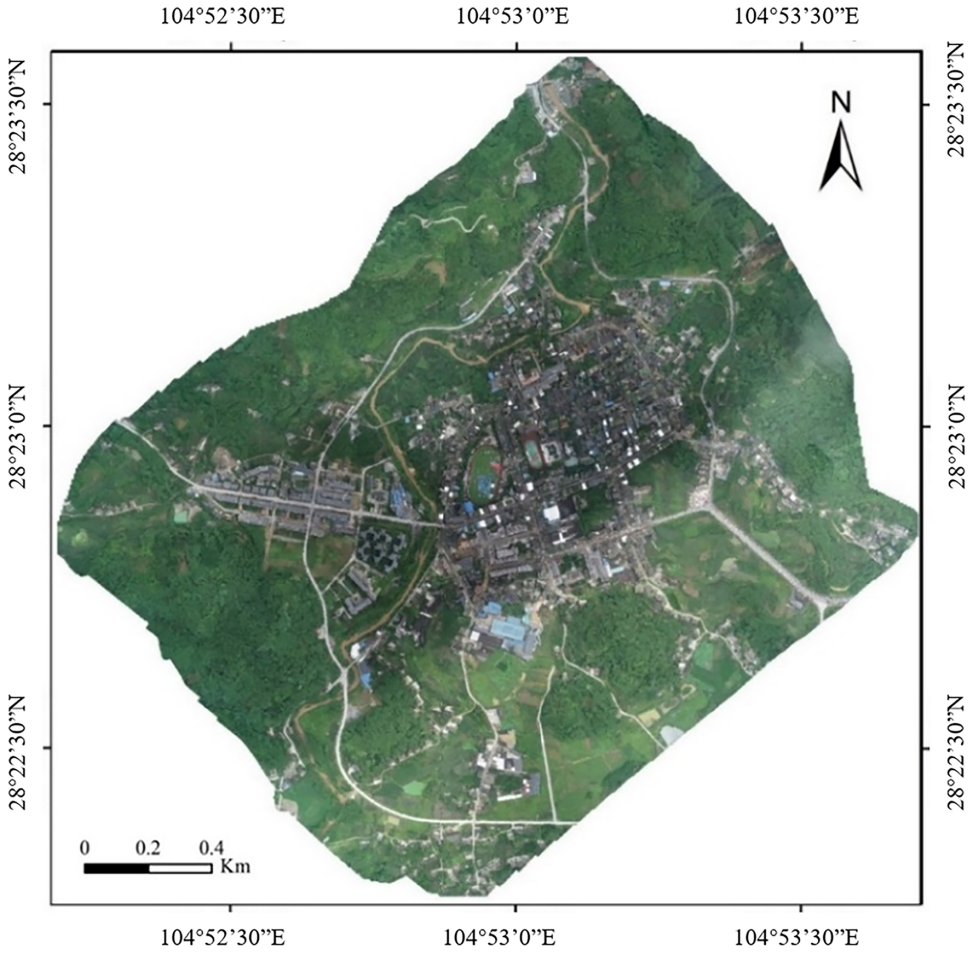


Figure 1: UAV orthophoto of Shuanghe Town.

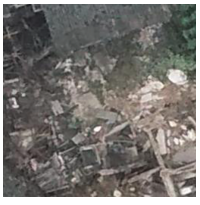
2.2 Data Source

The European Macroseismic Scale 1998 (EMS-98), developed by the European Seismological Commission (ESC), is a widely adopted macroseismic intensity scale for assessing and describing seismic intensity. It places particular emphasis on detailed classification of building damage levels, providing a systematic framework for evaluating earthquake impacts on structures [28]. In an ideal scenario, the progression of damage grades from Grade 1 to Grade 5 should correspond to a linear increase in seismic intensity. However, achieving this in practice is challenging due to the complex relationship between structural damage and ground motion. This difficulty is heavily influenced by the description methodology of damage grades. In practice, damage grade descriptions must prioritize clear differentiation between levels to ensure usability, while acknowledging that seismic intensity scales do not explicitly list all possible combinations of vulnerability categories and damage grades for each intensity value. Instead, they provide representative examples to guide users in interpreting real-world scenarios, balancing comprehensiveness with practical applicability.

Practical obstacles in implementing EMS-98 multi-level seismic-damage classification have been identified since the 2010s. Gautam et al. analysed these real-world limitations for seismic vulnerability evaluation and put forward targeted adjustment suggestions [29]. Consistent with these earlier findings, we also found that EMS-98 grades bring high requirements for training-data quality and annotation consistency, with ambiguous grade boundaries (e.g., crack-width thresholds between Grade 2 and Grade 3) producing

classification uncertainties during field-based annotation. To mitigate such ambiguities, our study consolidates intermediate states by regrouping: Grades 1–2 as "basically intact", Grades 3–4 as "partially damaged", and Grade 5 as "completely collapsed". This regrouping reduces the number of ambiguous grade boundaries in the image labels and is used here for roof-view UAV interpretation. We established a classification system for the degree of building damage, as shown in Table 1. Building damage is classified into three levels: basically intact, partially damaged, and completely collapsed. In the post-disaster images, black, red, green, and yellow represent background, Basically intact, Partially damaged, and Completely collapsed, respectively. Sample images and their annotations are shown in Fig. 2.

Table 1: Building seismic damage classification standards.

Damage Degree	**Basically Intact**		**Partially Damaged**		**Completely Collapsed**
The classification of building damage levels in EMS-98	Grade 1: Negligible to slight damage	Grade 2: Moderate Damage	Grade 3: Substantial to Heavy Damage	Grade 4: Very Heavy Damage	Grade 5: Destruction
Building damage level descriptions in EMS-98	No structural damage. Fine cracks in plaster, falling of loose tiles	Slight structural damage. Thin cracks in walls	Substantial structural damage. Large cracks in walls, partial failure of non-load-bearing elements	Heavy structural damage. Major cracks, failure of some load-bearing elements, partial collapse possible	Total or near-total collapse of the building.
On-site image features	The building structure remains largely intact with clear contours; visible damage is limited to minor features such as fine plaster cracks or small tile losses that do not compromise structural integrity.		The roof has collapsed or severely deformed, and walls are partially destroyed.		The structure is fully collapsed, with blurred boundaries and unrecognizable contours.
Remote sensing images					

(Continued)

Table 1 (continued)

Damage Degree	Basically Intact	Partially Damaged	Completely Collapsed
On-site images			

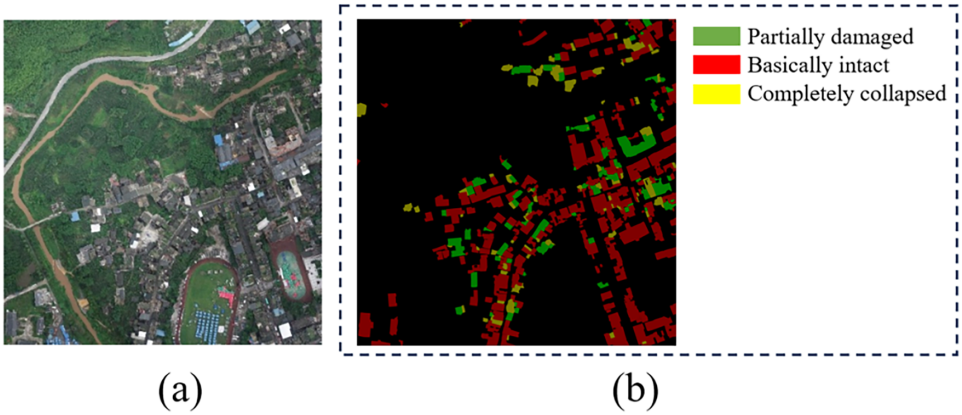


Figure 2: Sample images. (**a**) Original drawing; (**b**) Annotations.

In addition, the reliability of the building damage ground-truth labels in this study is validated by systematic on-site field investigations. On July 9, 2019, our research group conducted field verification and *in-situ* building structural damage assessments in Shuanghe Town. The team was divided into two subgroups to carry out comprehensive building safety surveys, and 855 valid building seismic damage records were collected in total. Image labels were assigned by visual interpretation of the UAV orthophoto and then checked against the 855 field records. Roof-view imagery cannot show internal structural damage; field engineering notes were used to reduce this mismatch, but residual label uncertainty remains for damage that is not expressed on the roof. The fieldwork was conducted in collaboration with Chongqing University Major Engineering Quality Inspection Co., Ltd. as part of the post-earthquake building damage assessment, with our research team specifically responsible for the building seismic-damage survey in Shuanghe Town. Six graduate students majoring in civil engineering participated in the field investigation; before labelling, all annotators were trained on a unified seismic-damage annotation standard. All building survey records were compiled strictly for physical damage evaluation and contain no personal or identifiable human data.

2.3 Data Processing

The UAV orthophoto was partitioned into spatially contiguous training and test blocks at a ratio of 7:3 (Fig. 3). The split was applied to image regions rather than by shuffling individual pixels or building objects. Adjacent blocks on the same orthophoto may still be spatially autocorrelated; this limitation is stated in Section 4.4. Overlapping cropping was then applied only inside the training block; the test block was cropped without overlap and without augmentation. The resulting numbers of training patches were approximately 6394 (128 × 128), 1906 (256 × 256), and 874 (512 × 512). The data set used in this study originates

from the post-earthquake field investigation and UAV survey of Shuanghe Town reported in our earlier publications [11–13], which document the data collection, orthophoto processing, building-level allocation, and ground-truth annotation procedures in full; as the present work is a follow-up study based on that surveyed data, these established procedures are not repeated here.

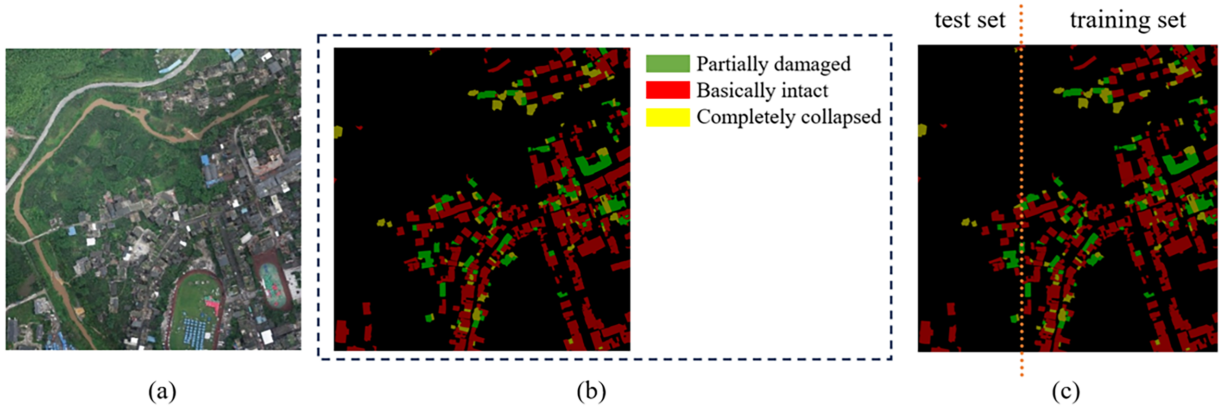


Figure 3: Dataset partitioning of the Shuanghe Town UAV orthophoto. (**a**) Orthophoto; (**b**) Label map; (**c**) Training and test blocks.

Across the complete surveyed scene, 855 building-damage records were available before dataset partitioning, comprising 256 basically intact buildings (29.9%), 522 partially damaged buildings (61.1%), and 77 completely collapsed buildings (9.0%). These object-level counts characterize the source scene. In this study, 50% overlapping cropping was used as a spatial-shift augmentation to increase patch coverage and to reduce truncation of buildings at crop boundaries. Because class-wise pixel and patch counts after cropping were not archived, whether this operation changed the effective class proportions cannot be quantified from the available data. These object-level counts characterize the source scene rather than the post-cropping training set. Building pixels account for only a small proportion of the total image pixels, resulting in a severe foreground-background imbalance.

Oversampling methods typically employ image enhancement and random replication to augment samples of the minority class to address imbalanced classification. In this study, overlapping sampling was adopted. Overlapping sampling is implemented by adjusting the sliding step size of the cropping window to extract patches from the labeled images. This method not only increases the number of samples but also ensures sample diversity. A schematic diagram is shown in Fig. 4. A building remote sensing image cropped by the overlapping sliding window is shown in Fig. 5.

The specific procedure of overlapping sampling is as follows. First, a fixed-size cropping window is defined. The step size is set to a value smaller than the window size. The window is slid horizontally and vertically across the entire orthophoto image and its corresponding label map, and a patch is extracted at each position. Because the step size is smaller than the window size, adjacent patches share overlapping regions. As illustrated in Fig. 4, a building that would generate only one sample under ordinary non-overlapping sampling can be covered by up to nine overlapping patches. Each patch captures the same building from a slightly different spatial context, including different surrounding environments and partial views of the building itself. Adopting 50% overlap ensures that the same building appears intact in multiple adjacent patches with different offsets, effectively mitigating the problems of edge truncation and texture fragmentation caused by splitting buildings across different patches in non-overlapping sampling. If the overlap ratio is further increased, the number of samples grows quadratically, which greatly increases training

time and GPU memory pressure, while the marginal gain in performance diminishes. Therefore, 50% is a commonly adopted compromise that balances the effect of data augmentation with training feasibility.

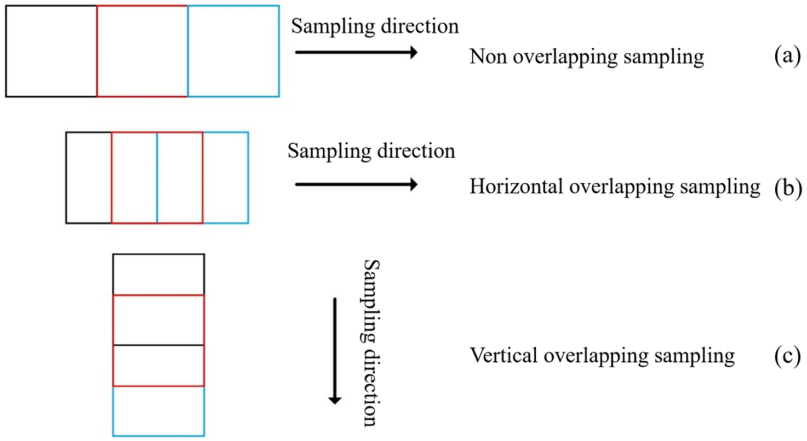


Figure 4: Different sampling methods. (**a**) Non overlapping sampling; (**b**) Horizontal overlapping sampling; (**c**) Vertical overlapping sampling.

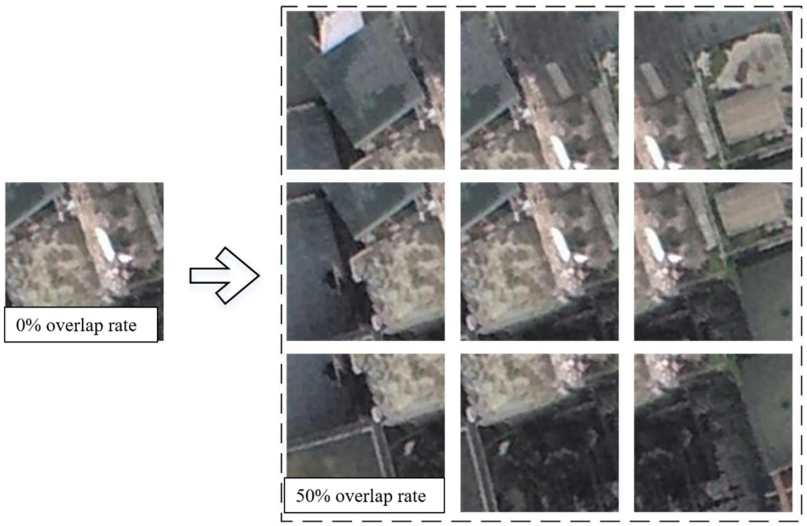


Figure 5: Overlapping sliding window cropped building remote sensing image.

Overlapping sampling offers several key advantages. It increases the total number of training patches. Under a fixed number of epochs this also increases the number of gradient updates, so the comparison with non-overlapping cropping is not matched on training budget. The observed mean-metric gains should therefore be attributed to the combined effects of additional patches, additional optimization steps, reduced boundary truncation, and richer spatial context, rather than to the overlap operator alone. By including each building in several overlapping patches, the model is exposed more frequently to the structural details of buildings, thereby enhancing its ability to distinguish between subtle damage levels such as “partially damaged” and “basically intact”. Furthermore, overlapping sampling enriches local information about building edges and damage fragments, which is critical for accurate semantic segmentation in complex post-earthquake scenes.

In this study, the training set and the test set were strictly separated from the very beginning of the data partitioning stage, ensuring that they do not overlap with each other. During the subsequent data augmentation process, the overlapping sampling operation was applied only to the training set, aiming to increase the number of training samples and enrich their diversity by halving the sliding window stride; the test set, on the other hand, always maintained the original non-overlapping sampling method and was not subjected to any form of duplication or augmentation. As a result, although the model is exposed to more samples from different perspectives of the same region during training, the test patches were not duplicated by overlapping cropping. This reduces the risk of identical train–test patches, but it does not remove spatial autocorrelation within the same orthophoto. Model, patch size, sampling strategy, and loss function were selected using the same test set; no geographically independent validation set is available. This design choice is driven by the limited size of the annotated dataset and the severe class imbalance. The "completely collapsed" category accounts for only 9.0% of building records, so reserving a further validation block would leave too few minority-class samples for stable training and reliable evaluation. A simplified train/test partition without an independent validation set has been adopted in prior post-disaster building damage assessment studies under comparable data constraints [30–32]. The reported scores should therefore be read as estimates on this test block, not as an independent confirmation of generalization.

3 Method and Experiment Schemes

Semantic segmentation is a pixel-level image analysis technique [33,34]. It enables a deep understanding and analysis of image content by classifying each pixel. Currently, various typical network models exist in the field of semantic segmentation [35–37]. This study selects four mainstream networks for comparison. Fig. 6 shows the flowchart for identifying building damage degrees using semantic segmentation.

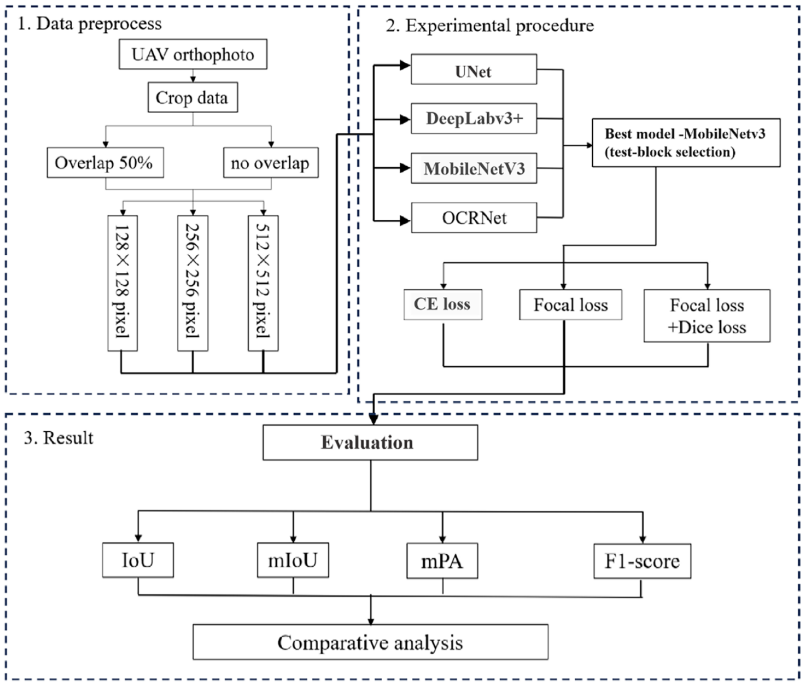


Figure 6: Technical flowchart of building damage identification. Note: the loss-function comparison (CE loss, focal loss, and focal loss + dice loss) was performed on MobileNetV3 only; see Section 4.3.

For the task of classifying building damage levels in post-earthquake orthophotos of mountainous towns using UAV, this study selects four typical semantic segmentation networks for comparison, UNet, DeepLabv3+, MobileNetV3, and OCRNet. As a classic encoder-decoder architecture, UNet effectively fuses shallow detail features with deep semantic information through skip connections, maintaining stable segmentation boundaries under small-sample annotation conditions. This makes it particularly suitable for fine localization of building edges and locally damaged areas. DeepLabv3+ employs atrous convolution and the ASPP module to expand the receptive field without reducing spatial resolution, enabling the capture of multi-scale contextual information. This is crucial for identifying buildings with different sizes and collapse morphologies, such as debris areas from partial collapse and large-scale completely collapsed zones. MobileNetV3, based on depthwise separable convolution and the SE attention mechanism, substantially reduces parameter count and computational cost while maintaining high accuracy, facilitating rapid inference in disaster emergency response. OCRNet, through object region representation and pixel-to-pixel relation modeling, enhances the learning of intra-class consistency and inter-class boundary distinguishability. It is suitable for the fine classification task in complex post-earthquake scenes, where the differences between "basically intact" and "partially damaged" are subtle and intra-class variations are large. In summary, the four networks represent distinct design paradigms, symmetric encoder-decoder, atrous spatial pyramid pooling, lightweight convolution, and context-aware object representation, allowing a comparative evaluation of overlapping sampling and hybrid loss functions across four architectures under the present experimental setting.

3.1 UNet

UNet is a popular Convolutional Neural Network (CNN) architecture [38]. This network is primarily used for image segmentation tasks, especially for biological image segmentation in medical image analysis. In the context of this study, the encoder is trained to extract building damage features from UAV orthophotos, while the decoder fuses multi-scale feature maps through skip connections to enhance the network's pixel-wise classification accuracy for building damage levels. The encoder extracts building damage features from images, while the decoder fuses multi-channel signals with the extracted features to enhance the network's pixel-wise classification accuracy. Fig. 7 depicts the schematic diagram of the UNet structure.

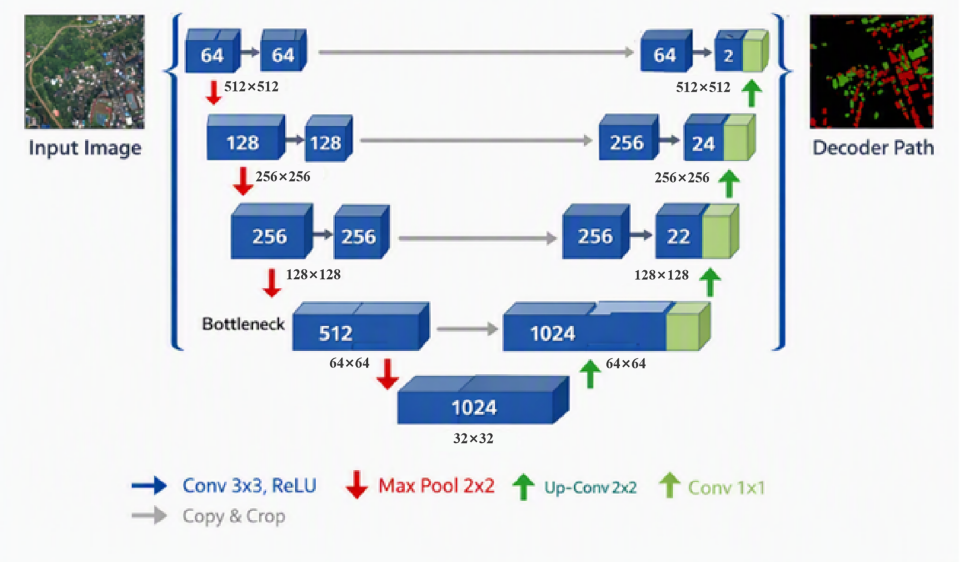


Figure 7: The schematic diagram of the UNet structure.

3.2 DeepLabv3+

DeepLabv3+ is an improvement of the DeepLabv3 model [39]. Fig. 8 depicts the schematic diagram of the DeepLabv3+ structure. DeepLabv3+ employs atrous convolution to expand the receptive field [40], enhancing feature extraction without increasing computational complexity. The Atrous Spatial Pyramid Pooling (ASPP) module incorporates global average pooling [41]. It can capture multi-scale information in images at a lower computational cost. Fig. 9 shows schematic diagrams of different dilation rates.

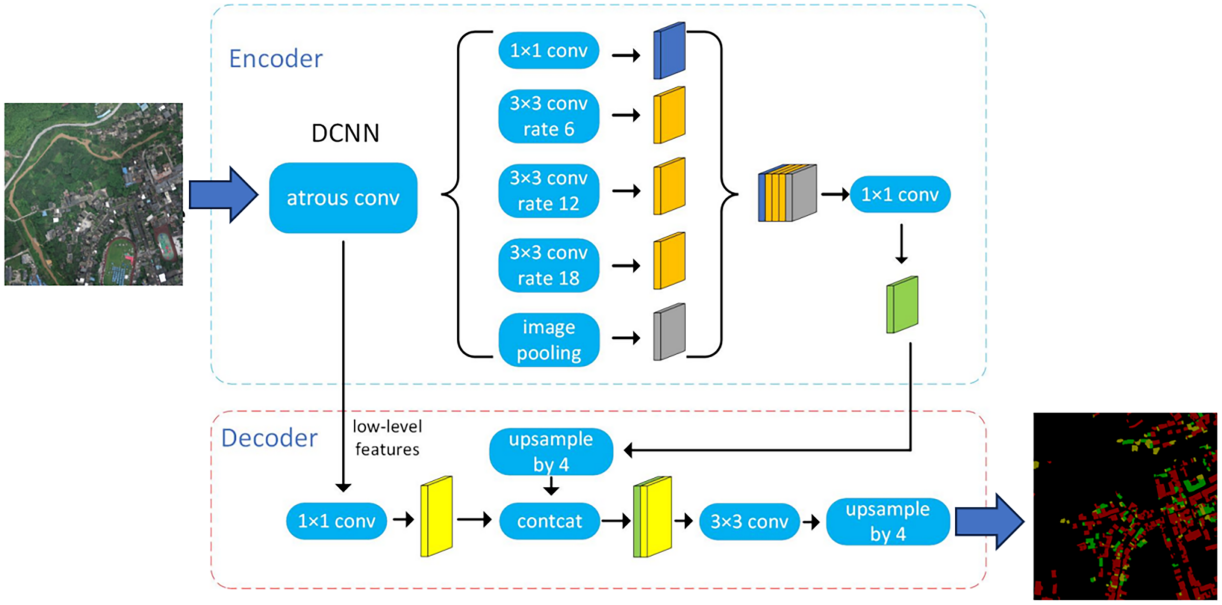


Figure 8: The schematic diagram of the DeepLabv3+ structure.

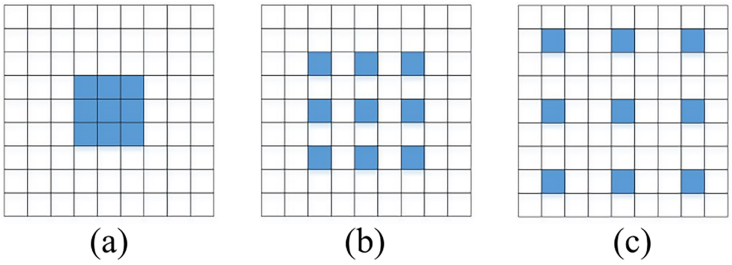


Figure 9: Atrous convolution. (**a**) Dilation rate = 1; (**b**) Dilation rate = 2; (**c**) Dilation rate = 3.

3.3 MobileNet3

The core innovation of MobileNet is depthwise separable convolution, which decomposes standard convolution into two steps: depthwise convolution and pointwise convolution. This substantially reduces model parameters and computational cost while maintaining high accuracy, making it suitable for devices with limited computational resources [42,43]. MobileNetV3 further introduces the Squeeze-and-Excitation (SE) module, which enhances the feature responses of key channels and improves the ability to distinguish between complex backgrounds and small targets in UAV-captured images. In this study, MobileNetV3-Large is utilized as the backbone, combined with the Lite R-ASPP segmentation head to complete pixel-wise semantic segmentation prediction.

3.4 OCRNet

OCRNet (Object-Contextual Representation Network) is a semantic segmentation architecture that enhances pixel representation by exploiting object-region contextual information [44,45]. It mainly comprises two components: a backbone feature extraction network and an object-contextual representation module that models pixel-to-pixel relations within object regions. In this paper, HRNet-W18 is adopted as the backbone for multi-resolution feature extraction, followed by an object-contextual representation module and a 1×1 convolution prediction head for four-class semantic segmentation tasks. It demonstrates a certain degree of robustness against factors such as image noise and lighting variations. The schematic diagram of the OCRNet structure is shown in Fig. 10.

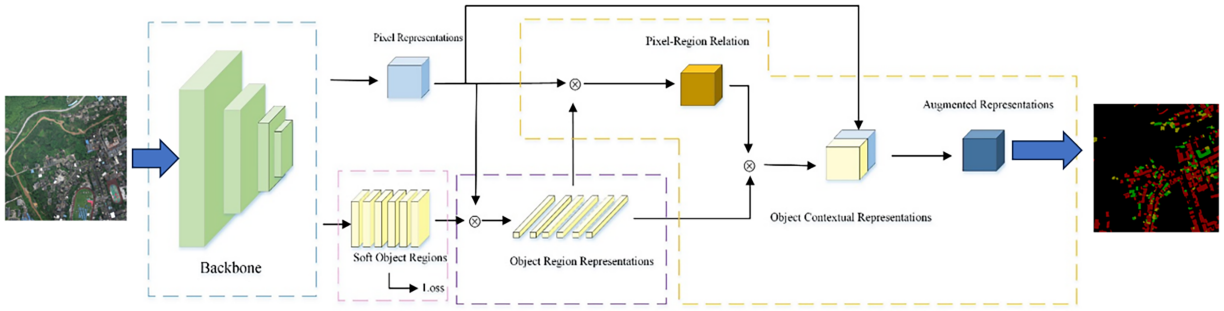


Figure 10: Schematic diagram of OCRNet structure.

3.5 Experimental Schemes

The experiments are conducted on a Linux system using PaddlePaddle 2.4 and PaddleSeg 2.6. The CPU is an Intel Core quad-core, eight-thread processor with 32 GB of RAM. The GPU is an NVIDIA Tesla V100 with 32 GB of memory. The number of training epochs is set to 100, and the batch size is set to 12. The experimental environment and hyperparameter settings are shown in Table 2. To ensure fair and reproducible comparisons among different deep learning models, all experiments are performed under identical hardware, software, and parameter settings. In Table 2, 128-0, 256-0, and 512-0 denote non-overlapping sampling, where the numbers 128, 256, and 512 indicate the input image size. 128-50, 256-50, and 512-50 denote 50% overlapping sampling. For example, 256-0 means input images of 256×256 pixels obtained via non-overlapping sampling, while 256-50 means input images of 256×256 pixels obtained via 50% overlapping sampling.

Table 2: Experimental environment and hyperparameter settings.

Parameter	Parameter Settings
Operating system	Linux system using PaddlePaddle 2.4 and PaddleSeg 2.6
CPU	Intel Core quad-core
Graphics card	NVIDIA Tesla V100 32 GB
Batch size	12
Epoch	100
Learning rate	1e−4
Segmentation model	UNet, DeepLabv3+, MobileNetV3, OCRNet
Optimizer	SGD
Loss function	CE loss, Focal loss, Focal loss + Dice loss
Data set	128-0, 256-0, 512-0, 128-50, 256-50, 512-50

The hybrid loss adopted in this study is a weighted combination of Focal Loss and Dice Loss. Focal Loss is introduced to mitigate severe sample imbalance across different building-damage categories and background, while Dice Loss optimizes region-level overlap between prediction and ground-truth for segmentation tasks.

Focal Loss is defined as:

$$L_{Focal} = -\sum_{c=0}^{C-1} \alpha_c \left(1 - p_C\right)^{\gamma} log(p_C)_-$$

where C is the total number of classes; α_c denotes class weight for category c; γ represents the focusing factor of Focal Loss. The hybrid loss is $L_{hybird} = \lambda \cdot L_{Focal} + (1 - \lambda) L_{Dice}$. Class weights α_c and the focusing factor γ followed the Focal Loss defaults in PaddleSeg 2.6 and were not grid-searched.

Dice Loss is expressed as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=2}^{N} p_i^2 + \sum_{i=2}^{N} g_i^2}$$

where p_i and g_i are the predicted probability and ground-truth label of pixel i, respectively.

The final hybrid loss is formulated as:

$$L_{hybird} = \lambda \cdot L_{Focal} + (1 - \lambda) L_{Dice}$$

where λ is the balance coefficient. In this work, a fixed weight λ =0.5 is adopted for simple weighted combination of two loss components, without further hyper-parameter searching for optimal weighting coefficients.

3.6 Evaluation Metrics

The models output four labels (background and three damage classes). Overall accuracy is computed over all pixels, including background. Per-class IoU, PA, and $F_{1\text{-score}}$ are reported for the three building-damage classes only. Let N_{ij} be the number of pixels whose ground-truth label is i and whose predicted label is j, with $i, j \in \{0, 1, 2, 3\}$ where indices 0, 1, 2, and 3 correspond to background, basically intact, partially damaged, and completely collapsed, respectively. Per-class metrics are computed for the three building-damage classes (i = 1, 2, 3).

$$Accuracy = \frac{\sum_{i=0}^{3} N_{ij}}{\sum_{i=0}^{3}\sum_{j=0}^{3} N_{ij}} \tag{1}$$

$$IoU_i = \frac{N_{ij}}{\sum_{k=1}^{3} N_{ik} + \sum_{k=1}^{3} N_{ki} - N_{ii}} \tag{2}$$

$$mIoU = \frac{1}{3}\sum_{i=1}^{3} IoU_i \tag{3}$$

$$PA_i = Recall_i = \frac{N_{ij}}{\sum_{k=1}^{3} N_{ik}} \tag{4}$$

$$mPA = \frac{1}{3}\sum_{i=1}^{3} PA_i \tag{5}$$

$$F_{1-score_i}=\frac{2\times Precision_i\times Recall_i}{(Precision_i+Recall_i)}\ Precision_i=\frac{N_{ij}}{\sum_{k=1}^{3}N_{ki}} \tag{6}$$

where N_{ij} denotes the number of pixels whose ground-truth label is class i and whose predicted label is class j; indices $i, j \in \{1, 2, 3\}$ correspond to the three building-damage classes.

4 Results and Discussion

This study evaluates semantic segmentation models on three datasets with sizes of 128 × 128, 256 × 256, and 512 × 512, which are hereafter referred to as the 128, 256, and 512 datasets, respectively. We selected the most suitable size and subsequently compared loss functions on the selected setting. All experiments were repeated ten times, and the tables report the arithmetic means of the evaluation metrics. The run-level numerical outputs were not retained. Standard deviations, confidence intervals, and formal significance tests therefore cannot be reconstructed. All between-method comparisons below are descriptive comparisons of mean values and should not be interpreted as evidence of statistically significant performance differences.

4.1 Not Overlapping

Table 3 shows the results of four models on different datasets. On the 128 dataset, the UNet model performs poorly in identifying partially damaged and completely collapsed buildings. DeepLabv3+ achieves the highest mean mIoU, mPA, and F1-score among all models. The accuracy of all four models for identifying non-collapsed buildings is greater than 70%. The minimum IoU values for identifying partially damaged and completely collapsed buildings in the model are 29.71% and 34.78%, respectively. This suggests that damaged buildings are more prone to misclassification. On the 256 dataset, MobileNetV3 achieves the highest mIoU and mPA among all models. On the 512 dataset, MobileNetV3 achieves the highest mIoU, mPA, and F1-score. In contrast, on the 128 dataset, DeepLabv3+ achieves the highest values among the four models across all three metrics. Compared to other models, UNet remains the worst-performing model, which can be attributed to its relatively shallow encoder and simple skip connection architecture that cannot effectively capture the complex fracture textures in post-earthquake scenes. In contrast, MobileNetV3's depthwise separable convolution combined with the squeeze-and-excitation attention mechanism enables efficient multi-scale feature extraction, contributing to its higher mean mIoU, mPA, and F1-score on the 256 and 512 datasets. On the 512 dataset, MobileNetV3 achieves the highest mIoU, mPA, and F1-score. All models achieve mIoU values above 45%, with MobileNetV3 leading at 56.06%. However, per-class IoU for the "partially damaged" and "completely collapsed" categories remains relatively low, with DeepLabv3+ reporting a partially damaged IoU of 38.05%.

Based on the comparison results in Table 3, a detailed analysis of the performance of the four semantic segmentation models under non-overlapping sampling is provided.

From an overall trend, as the input image size increases from 128 × 128 to 512 × 512, most models (UNet, MobileNetV3, and OCRNet) show improved mIoU, mPA, and F1-score. However, DeepLabv3+ exhibits a non-monotonic response, achieving its best performance at 128-0 and declining at larger sizes. This indicates that the optimal input size is model-dependent, and larger inputs do not universally benefit all architectures. However, the responses to size variation differ notably among the models.

UNet performs the worst on all datasets. Its mIoU at 512-0 is 47.01%, only slightly higher than the 46.27% at 256-0. Notably, the IoU for the "completely collapsed" category at the 512 size is 32.00%, which is lower than the 34.94% at the 256 size, showing a non-monotonic change. This reflects that the shallow encoder and simple skip connections of UNet cannot effectively integrate the fracture textures in complex post-earthquake scenes; when the input size becomes too large, redundant noise may be introduced or local details may be lost.

Table 3: Comparison of different models without overlapping.

Models	Dataset	Evaluation Metrics					
		IoU			mIoU	mPA	F1-score
		Basically Intact	Partially Damaged	Completely Collapsed			
UNet	128-0	70.54%	29.71%	34.78%	45.01%	71.32%	61.86%
	256-0	70.88%	33.00%	34.94%	46.27%	71.68%	62.64%
	512-0	72.25%	36.77%	32.00%	47.01%	71.58%	63.50%
DeepLabv3+	128-0	78.55%	44.43%	41.02%	54.67%	73.26%	70.52%
	256-0	73.19%	40.28%	39.78%	51.08%	72.48%	68.49%
	512-0	78.14%	38.05%	41.53%	52.57%	73.24%	68.79%
MobileNetV3	128-0	73.53%	31.03%	39.94%	48.17%	72.17%	65.48%
	256-0	75.72%	41.03%	36.67%	51.14%	72.99%	67.21%
	512-0	79.42%	46.16%	42.60%	56.06%	73.51%	70.96%
OCRNet	128-0	72.26%	32.71%	38.84%	47.94%	71.79%	64.49%
	256-0	74.58%	40.70%	36.16%	50.48%	72.62%	66.78%
	512-0	78.19%	43.96%	43.77%	55.31%	73.34%	69.75%

DeepLabv3+ achieves its best mIoU of 54.67% on the 128-0 dataset, but its performance fluctuates on the 256-0 and 512-0 datasets, with mIoU values of 51.08% and 52.57%, respectively. The IoU for the "completely collapsed" category remains relatively stable, ranging from 39.78% to 41.53%, but the IoU for the "partially damaged" category gradually decreases from 44.43% at 128-0 to 38.05% at 512-0. This suggests that although the atrous convolution and ASPP module of DeepLabv3+ expand the receptive field, the model adapts better to small-sized inputs, whereas excessively large inputs may dilute feature responses in partially damaged areas.

MobileNetV3 exhibits the best scaling ability with respect to image size. From 128-0 to 512-0, its mIoU increases from 48.17% to 56.06%, the largest improvement among all models. In particular, the IoU for the "partially damaged" category jumps from 31.03% to 46.16%, reaching the highest value among all models at all sizes. The IoU for the "completely collapsed" category increases from 39.94% to 42.60%. The mPA (73.51%) and F1-score (70.96%) of this model on the 512-0 dataset also lead the other networks. The lightweight depthwise separable convolution combined with the squeeze-and-excitation attention mechanism enables efficient multi-scale feature extraction with limited parameters, providing stronger discriminative ability for the fragmented textures of collapsed buildings.

OCRNet performs between MobileNetV3 and DeepLabv3+. Its mIoU at 512-0 is 55.31%, second only to that of MobileNetV3. The IoU values for partially damaged and completely collapsed categories are 43.96% and 43.77%, respectively, which are relatively balanced. Through object region representation and pixel-to-pixel relation modelling, OCRNet maintains overall segmentation accuracy while achieving consistent recognition ability for the two damage categories, without a notably lower performance in any single category.

In summary, under non-overlapping sampling, the IoU values for the two key damage categories, "partially damaged" and "completely collapsed", are notably lower than those for the "basically intact" category, with the highest values reaching only 46.16% and 43.77%, respectively. This pattern is consistent with the inherent limitation of fixed-grid discrete cropping: the sampling window boundaries cut off building edges and fracture textures, disrupting the spatial continuity of damaged structures and consequently leading to misclassification. Fig. 11 compares the performance of the four models across the three input image sizes, and Fig. 12 shows representative prediction results under non-overlapping sampling.

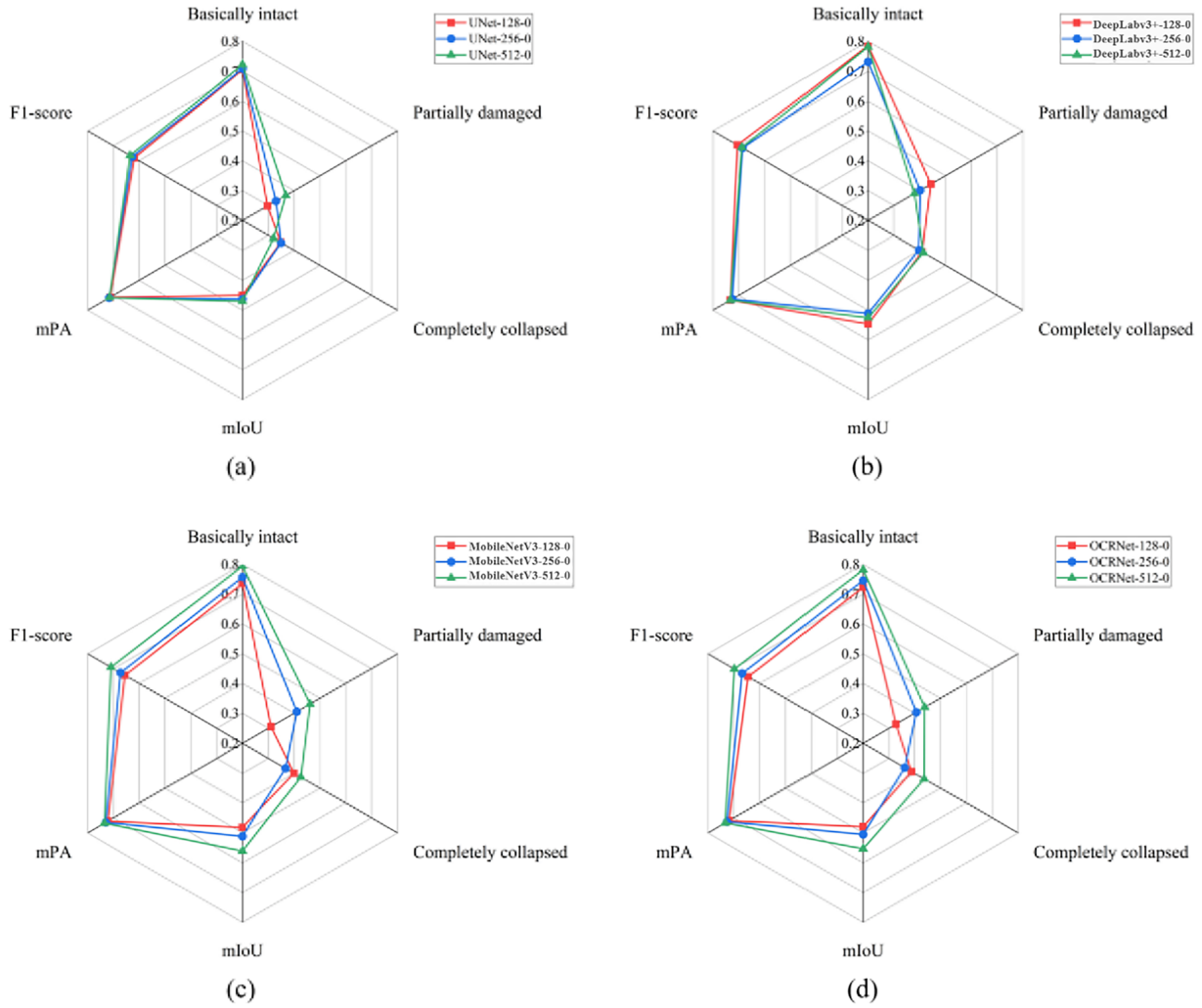


Figure 11: Comparison of input image sizes without overlapping sampling. (**a**) UNet; (**b**) DeepLabv3+; (**c**) MobileNetV3; (**d**) OCRNet.

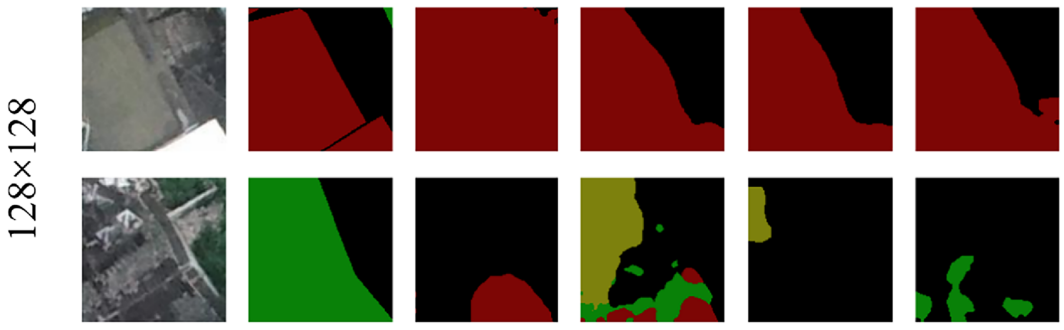


Figure 12: (Continued)

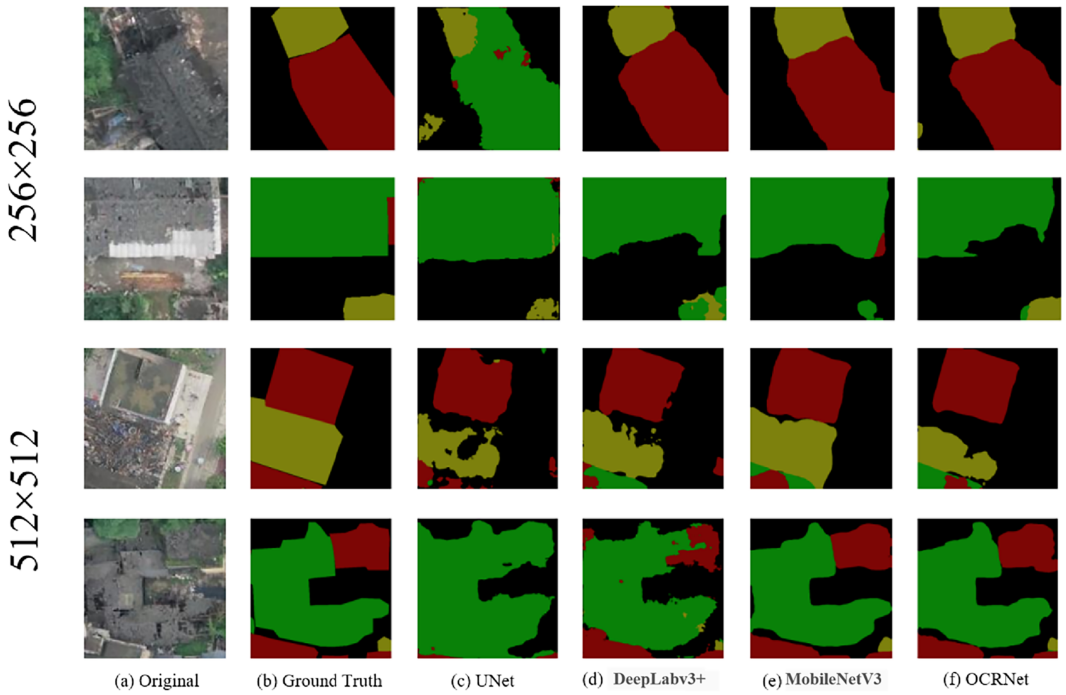


Figure 12: Prediction results for buildings without overlapping.

4.2 Overlapping Sampling

Higher evaluation scores were observed under the overlapping-sampling configuration. The model's performance across different datasets is presented in Table 4. On the 128-50 dataset, DeepLabv3+ reported the highest mean mIoU, mPA, and F1-score among the four models, whereas OCRNet reported the lowest mean metrics. On the 256-50 dataset, both MobileNetV3 and OCRNet showed higher mean mIoU, mPA, and F1-score than on the corresponding non-overlapping settings. On the 512-50 dataset, MobileNetV3 achieved the highest mean mIoU, mPA, and F1-score, followed by OCRNet, while UNet exhibited the lowest performance. On the 512-50 dataset, the IoU values for the two non-collapsed building classes ranged from 87.22% to 94.16%, and most approached or exceeded 90%, indicating high mean IoU for the basically intact class on this test block.

Based on the results with 50% overlapping sampling shown in Table 4, a detailed analysis of the performance of the four semantic segmentation models is provided. Compared with Table 3, the 50% overlapping setting yields higher mean IoU, mPA, and F1-score for all models. Because the number of patches and gradient updates also increases at a fixed epoch count, these differences are not an isolated effect of the overlap ratio, with larger mean-metric increases for the two challenging categories, "partially damaged" and "completely collapsed". Fig. 13 compares the four models across the three input image sizes under overlapping sampling.

From an overall trend, as the input image size increases from 128 × 128 to 512 × 512, the performance metrics of all models generally increase, although some models exhibit fluctuations at the size of 256. Notably, all models achieve their best or near-best mIoU, mPA, and F1-score at the size of 512, indicating that larger input images combined with overlapping sampling can more completely preserve fracture textures and edge information.

Table 4: Comparison of different models with 50% overlapping sampling.

Models	Dataset	Evaluation Metrics					
		IoU			mIoU	mPA	F1-score
		Basically Intact	Partially Damaged	Completely Collapsed			
UNet	128-50	84.47%	69.41%	63.63%	72.50%	95.59%	84.61%
	256-50	84.77%	67.13%	61.77%	71.22%	95.54%	83.74%
	512-50	92.04%	87.22%	81.80%	87.02%	96.86%	93.70%
DeepLabv3+	128-50	93.41%	89.06%	84.06%	88.84%	97.15%	94.72%
	256-50	87.20%	70.14%	54.59%	70.64%	95.74%	82.97%
	512-50	93.91%	89.95%	85.55%	89.80%	97.21%	95.26%
MobileNetV3	128-50	87.52%	80.49%	74.84%	80.95%	96.20%	90.12%
	256-50	93.87%	89.87%	84.93%	89.56%	97.19%	95.13%
	512-50	94.16%	93.70%	87.38%	91.75%	97.26%	96.33%
OCRNet	128-50	77.45%	53.08%	50.40%	60.31%	94.26%	75.52%
	256-50	90.38%	81.83%	75.35%	82.52%	96.55%	91.03%
	512-50	94.14%	91.43%	86.75%	90.77%	97.24%	95.80%

Note: all results were obtained with the CE loss; the loss-function comparison on MobileNetV3 is reported in Table 5. Overlapping cropping was applied only to the training set, and the test block remained non-overlapping. Values are arithmetic means of ten runs.

Table 5: Comparison of different loss functions on MobileNetV3.

Data	Loss Functions	Evaluation Metrics					
		IoU			mIoU	mPA	F1-score
		Basically Intact	Partially Damaged	Completely Collapsed			
512-0	CE loss	79.42%	46.16%	42.60%	56.06%	73.51%	70.96%
	Focal loss	81.22%	45.23%	48.61%	58.35%	74.00%	71.12%
	Focal loss + Dice loss	83.56%	46.44%	51.32%	60.44%	78.50%	72.38%
512-50	CE loss	94.16%	93.70%	87.38%	91.75%	97.26%	96.33%
	Focal loss	94.20%	92.10%	87.41%	91.24%	97.31%	96.50%
	Focal loss + Dice loss	94.22%	93.70%	89.15%	92.36%	97.32%	96.77%

Note: all values are arithmetic means of ten runs; run-level variability was unavailable.

UNet achieves an mIoU of 87.02% on the 512-50 dataset, an increase of 40 percentage points from 47.01% under non-overlapping sampling. Its IoU for "partially damaged" and "completely collapsed" jumps from 36.77% and 32.00% to 87.22% and 81.80%, respectively, representing the largest improvement among all models. The observed improvement is consistent with the hypothesis that the additional patches, optimization steps, and reduced boundary truncation associated with overlapping sampling may provide richer contextual information for UNet.

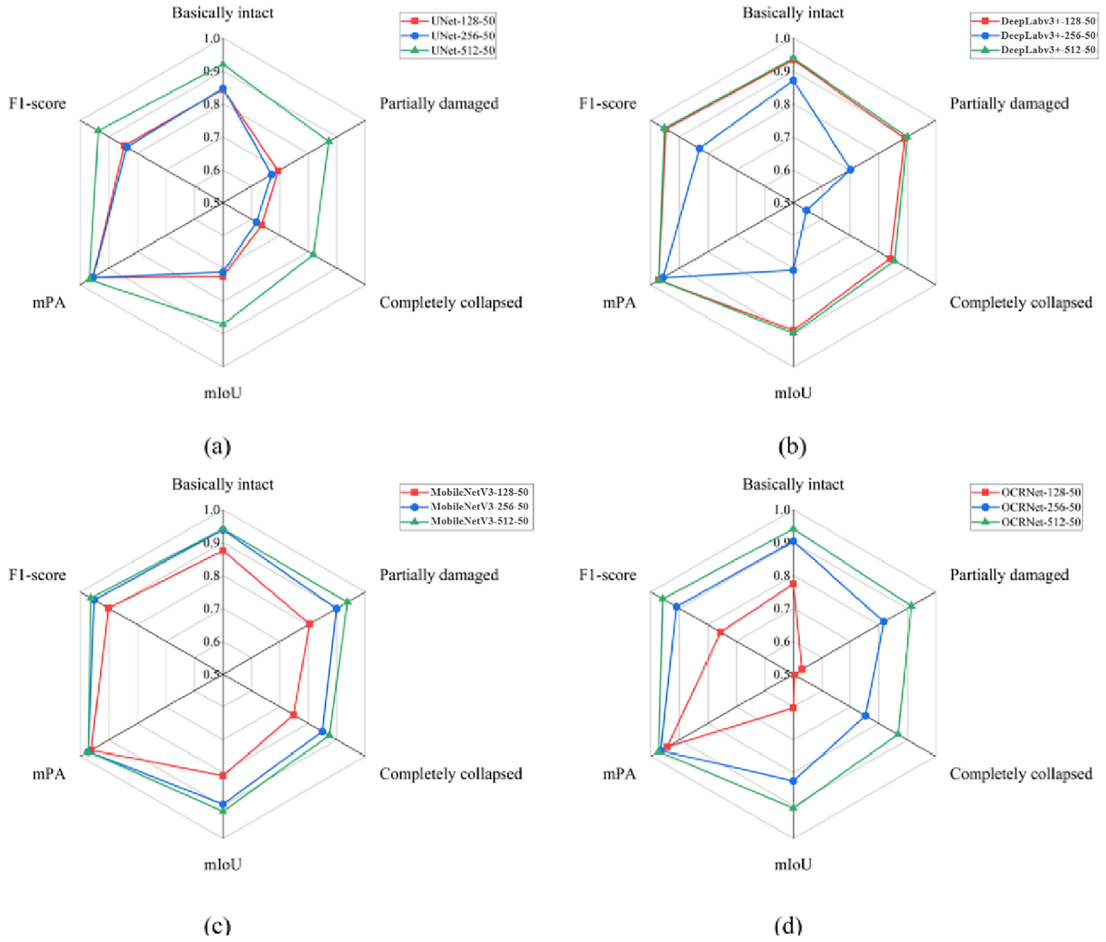


Figure 13: Comparison of input image sizes with overlapping sampling. (**a**) UNet; (**b**) DeepLabv3+; (**c**) MobileNetV3; (**d**) OCRNet.

DeepLabv3+ achieves a very high mIoU of 88.84% on the 128-50 dataset, but its performance drops abnormally on the 256-50 dataset, where the mIoU is only 70.64% and the IoU for "completely collapsed" is as low as 54.59%. This may be because at the size of 256, some key features fall on the edges of the overlapping regions, causing attention to be diluted. Nevertheless, on the 512-50 dataset, its performance recovers to 89.80%, ranking third after MobileNetV3 and OCRNet. This indicates that DeepLabv3+ is sensitive to input size and requires a sufficiently large receptive field to take full advantage of its atrous convolution.

MobileNetV3 achieves the highest mean mIoU, mPA, and F1-score among the four tested models. On the 512-50 dataset, its mIoU reaches 91.75%, mPA 97.26%, and F1-score 96.33%, ranking first in all three metrics. More notably, the IoU for the "partially damaged" category is as high as 93.70%, and that for the "completely collapsed" category reaches 87.38%, both higher than the corresponding values of the other models. This may reflect the ability of MobileNetV3, with its squeeze-and-excitation attention mechanism, to extract multi-scale features and recognize fragmented textures. Moreover, its lightweight architecture does not suffer from performance degradation when the amount of data increases due to overlapping sampling.

OCRNet also achieves strong performance under overlapping sampling. Its mIoU on the 512-50 dataset is 90.77%, close to that of DeepLabv3+ (89.80%) and slightly lower than that of MobileNetV3. The IoU for "completely collapsed" is 86.75% and that for "partially damaged" is 91.43%, showing a balanced performance. This may also benefit from the richer contextual information provided by the overlapping-sampling setting.

In summary, the overlapping-sampling setting is expected to reduce boundary truncation relative to fixed-grid cropping. At 512 × 512, mean IoU for the two damaged classes is about 45–50 percentage points higher than in Table 3. This comparison is not adjusted for the extra optimization steps. Among the four models, MobileNetV3 attains the highest mean mIoU, mPA, and F1-score on the 512-50 setting under the present protocol.

Post-earthquake inventories are typically left-tailed: collapsed buildings are far fewer than intact or moderately damaged ones. In Tables 3 and 4, mean IoU is consistently lower for the completely collapsed class than for the basically intact class. Identifying life-threatening collapse therefore remains the more difficult and more operationally important task; the present mean IoU on that class should not be read as predictive confidence, because no probability calibration was performed.

Mean IoU, mPA, and F1-score are higher under the overlapping-sampling configuration than under the non-overlapping configuration (Fig. 14); the comparison is not matched on the number of optimization steps. Fig. 15 compares the mean metrics of the four models under the two sampling configurations.

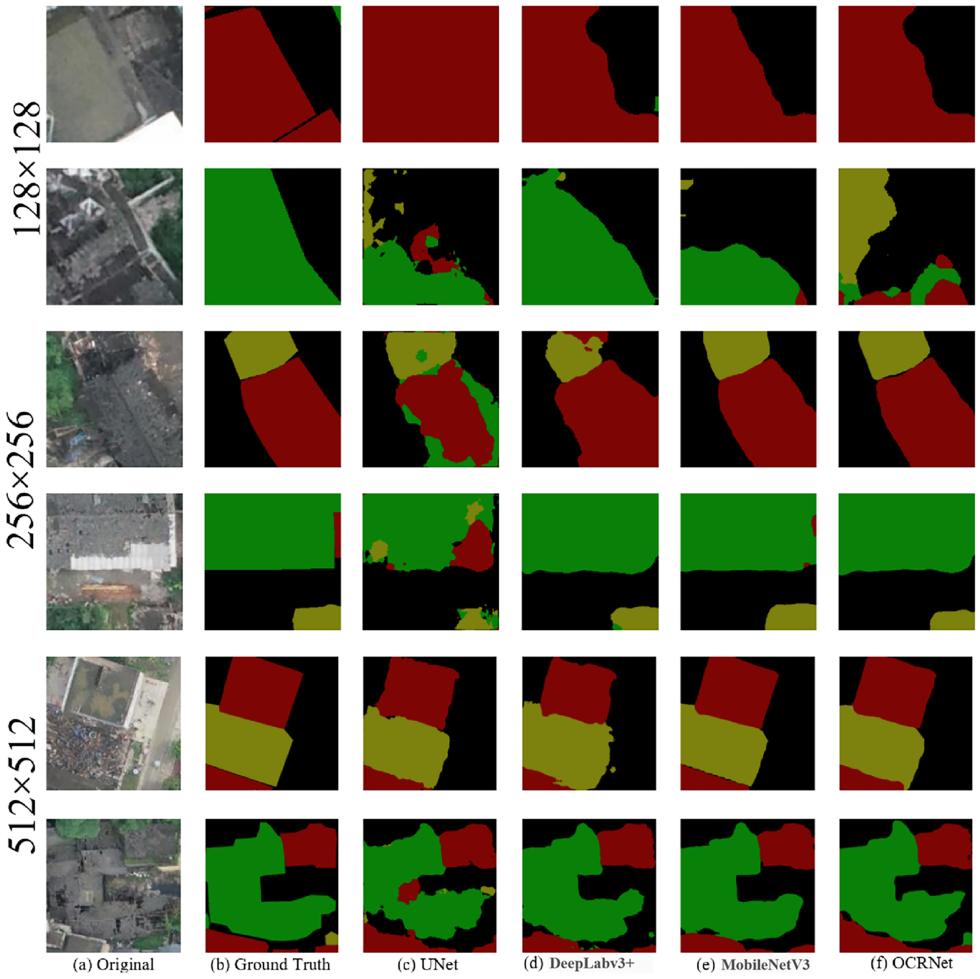


Figure 14: Prediction results for buildings under 50% overlapping sampling. Note: panels (**c**)–(**f**) show the predictions of UNet, DeepLabv3+, MobileNetV3, and OCRNet, respectively; the corresponding model and loss configurations are reported in Tables 4 and 5.

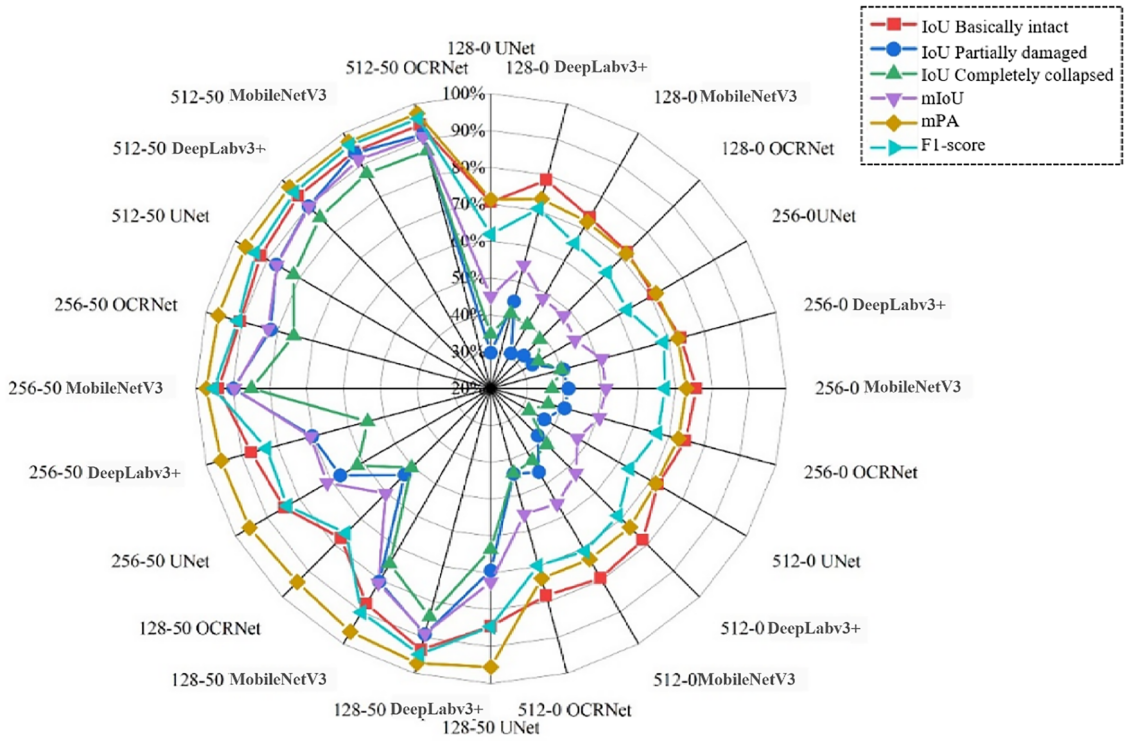


Figure 15: Comparison of models with different sampling methods. Note: values are arithmetic means of ten runs; run-level variability was unavailable.

The combination of overlapping sampling and the 512 × 512 image size produced higher mean metrics under the present protocol. The underlying reasons can be explained from three perspectives: data characteristics, sampling strategy, and model receptive field.

The dataset of this study originates from UAV orthophotos of mountainous towns after an earthquake. The number of building pixels is severely imbalanced compared with background pixels, with buildings themselves occupying only a small proportion. The damage caused by the earthquake is manifested as fractured and fragmented textures, where the spatial patterns of the "partially damaged" and "completely collapsed" categories are discontinuous and have complex edges. Under non-overlapping sampling, a fixed-size window slides with a stride equal to the window size, resulting in no overlap between adjacent samples. This approach causes many buildings, especially damaged ones, to be exactly cut by the boundaries of the cropping window, leading to truncated edge features and fracture textures, thus preventing the formation of complete semantic information. For a completely collapsed building, its debris area may be distributed across multiple discontinuous patches, with each patch showing only a part of the debris, making it difficult for the model to infer the overall collapse state from a single patch. This is the core reason why the IoU values for the "partially damaged" and "completely collapsed" categories are generally low under non-overlapping sampling.

Overlapping sampling sets the sliding stride to half the window size, creating a 50% overlap between adjacent patches. This improvement brings several benefits. Each building is covered by multiple patches from different offset positions, and each patch contains most or even the entire structure of the building, avoiding the truncation of features at boundaries. A building that would generate only one sample under non-overlapping sampling can generate up to nine samples under 50% overlapping sampling. In each of these samples, the position of the building relative to the patch centre is different, allowing the model

to learn the structural features of the same building from multiple perspectives. Overlapping sampling substantially increases the total number of training samples, which is an inherent part of its mechanism as a data augmentation strategy. Moreover, because the window shifts, the same building appears with different surrounding background information across different samples. We note that the increased sample count also leads to more optimization iterations under a fixed epoch setting, which may contribute to the performance improvement alongside the structural benefits of overlapping.

Regarding the choice of image size, comparative experiments with three sizes—128 × 128, 256 × 256, and 512 × 512—show that the 512 × 512 size achieves the best or near-best performance for all models. The damage features of post-earthquake buildings often require sufficient spatial context to be correctly identified. The 128 × 128 window is too small to accommodate an entire building, especially for larger buildings in mountainous towns; the model can only see local walls and struggles to distinguish "partially damaged" from "basically intact". The 256 × 256 size brings some improvement, but its receptive field is still insufficient for the completely collapsed scenes with scattered debris. The 512 × 512 window can cover the full extent of most individual buildings while retaining enough surrounding background information, enabling the model to make more accurate judgments by exploiting the relative relationships between buildings and adjacent rubble or debris. Furthermore, the 512 × 512 size synergises with 50% overlapping sampling. The larger window ensures that each crop contains rich structural and contextual information, while overlapping sampling guarantees that features at window boundaries are not lost. Together, they expose the model to more complete roof-view patterns of building damage within the training patches.

4.3 Optimization of Loss Function

A combination of Focal Loss and Dice Loss was compared with CE and Focal Loss alone on MobileNetV3. Loss function choice can influence training under class imbalance, but model architecture also shapes the representational capacity available for segmentation. In the field of deep learning-driven semantic segmentation, model architecture determines the upper bound of performance through three key aspects: (1) the capacity to construct feature representation spaces, (2) the mechanism for integrating multi-scale contextual information, and (3) the depth of spatial topological relationship modeling [46,47].

The role of the loss function is to guide the optimization process by defining the learning objective and shaping the feature space during training. While model architecture fundamentally defines the representational capacity of the model, the loss function plays a critical role in determining how effectively that capacity is utilized, particularly under class-imbalanced conditions. Table 5 reports that, on the 512-0 setting, replacing CE loss with Focal + Dice loss increased mean mIoU by 4.38 percentage points (from 56.06% to 60.44%) and mean mPA by 4.99 percentage points (from 73.51% to 78.50%). These are descriptive differences in the reported means and should not be interpreted as evidence of statistically established superiority.

Among the tested models, MobileNetV3 achieved the best segmentation performance on the 512 dataset. Therefore, this study selected MobileNetV3 to compare the loss functions. For tasks with scarce foreground categories or highly imbalanced classes, introducing Focal Loss and Dice Loss can further strengthen the focus on hard-to-classify samples. Focal Loss introduces a modulating factor to the cross-entropy loss, allocating more gradient contribution to pixels with lower confidence, thereby alleviating the problem that minority classes are overwhelmed by dominant classes in multi-category scenarios. Dice Loss directly optimizes the overlap between the predicted region and the ground truth region, showing stronger robustness to differences in class sample sizes. The experimental results are shown in Table 5. Fig. 16 compares the three loss-function configurations on the 512-0 and 512-50 settings.

On the 512-0 setting, mean mIoU increases from 56.06% with CE to 58.35% with Focal Loss and 60.44% with Focal + Dice (Table 5). On the 512-50 dataset, the Focal + Dice configuration produced mean mIoU,

mPA, and F1-score values that were 0.61, 0.06, and 0.44 percentage points higher than those obtained with CE loss, respectively. The mean IoU for the completely collapsed class was 1.77 percentage points higher. The Focal-only configuration produced a lower mean mIoU than CE (91.24% vs. 91.75%). Because run-level variability statistics are unavailable, it cannot be determined whether these differences exceed variation between repeated runs. These results indicate numerical differences in the reported means rather than statistically established superiority. Dice Loss optimizes region-level overlap. Focal Loss reweights hard pixels. Their combination is compared with CE and with Focal Loss alone in Table 5. The numerical differences are those reported above; they are not used here as evidence that class imbalance has been resolved.

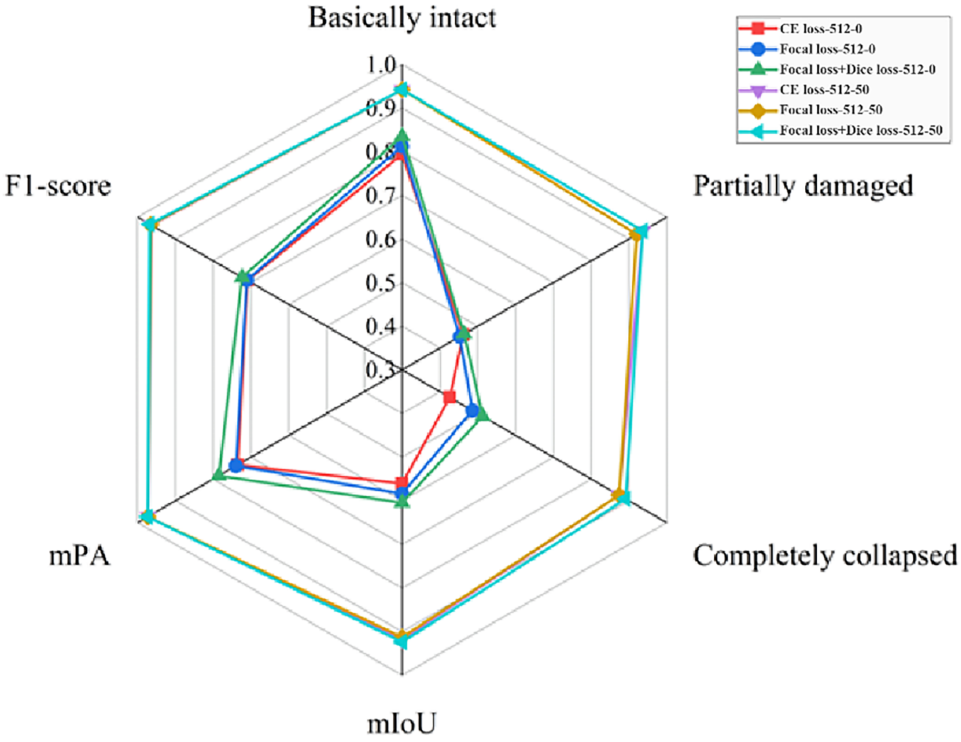


Figure 16: Comparison of different loss functions of MobileNetV3 model. Note: CE loss-512-0 and CE loss-512-50 denote the CE-loss experiments on the 512-0 (non-overlapping) and 512-50 (50% overlapping) settings, respectively; values are arithmetic means of ten runs, and run-level variability was unavailable.

4.4 Limitations

This study reports mean segmentation metrics for overlapping sampling and a hybrid loss on the Shuanghe Town UAV orthophoto. Nevertheless, several inherent limitations remain due to existing experimental constraints, and supplementary re-running of experiments cannot be performed.

First, class-wise pixel and patch statistics after overlapping cropping are unavailable. Because a uniform sliding window can replicate samples from all classes, overlapping cropping does not necessarily alter their relative distribution. The present study therefore does not establish that overlapping sampling improves class balance. The observed mean-performance differences may reflect the combined effects of increased sample redundancy, additional optimization updates, reduced boundary truncation, and altered spatial context. Moreover, under the same epoch setting, overlapping sampling increases both the number of training samples and the number of optimization iterations; thus, the current experiments cannot fully isolate the

effect of overlapping sampling itself from the effect of more training iterations. Future work should archive class-wise pixel and patch counts after cropping to permit this evaluation.

Second, the Focal-Dice hybrid loss in this work adopts a fixed weighting coefficient of $\lambda = 0.5$ for simple weighted combination. Hyper-parameter searching over alternative λ values was not conducted in the current experiments, and better segmentation performance could potentially be achieved by tuning this weighting hyper-parameter.

Third, this study lacks geographically-isolated or building-level independent validation data. Although block-wise spatial splitting is applied to reduce associated risks, potential spatial autocorrelation between training and test samples cannot be completely ruled out. In addition, architecture, input size, sampling scheme, and loss function were compared on this test block rather than on a nested validation split. Run-level outputs from the ten repeated experiments were not retained. Consequently, standard deviations, confidence intervals, and formal significance tests cannot be reconstructed. The reported comparisons should therefore be interpreted as descriptive differences between mean values rather than evidence of statistically reliable superiority.

Fourth, the dataset originates exclusively from Shuanghe Town for a single earthquake event. Damage identification is limited to the spatial extent of this mosaic and to roof-visible damage. Mean IoU on the completely collapsed class remains lower than on the basically intact class in Tables 3 and 4; this is a class-wise overlap score, not a calibrated probability of life-threatening damage. Transfer to other towns is not shown.

5 Conclusion

This study constructs a three-class building-damage dataset from the Shuanghe Town UAV orthophoto and compares four semantic segmentation models under non-overlapping and 50% overlapping cropping. Mean mIoU on 512 × 512 patches is about 35–40 percentage points higher under the overlapping setting; that comparison is not matched on the number of optimization steps. On the 512-50 setting with CE loss, MobileNetV3 produced the highest mean mIoU, mPA, and F1-score among the four architectures (91.75%, 97.26%, and 96.33%). Replacing CE with Focal + Dice changed these means to 92.36%, 97.32%, and 96.77%, and changed the mean IoU of the completely collapsed class from 87.38% to 89.15%. The largest mean-metric gains coincide with 50% overlapping cropping at 512 × 512. The present design confounds the overlap operation with a larger number of training patches and gradient updates, so the isolated contribution of overlap cannot be determined. On the 512-50 setting, replacing CE with Focal + Dice changes mean mIoU, mPA, and F1-score by 0.61, 0.06, and 0.44 percentage points, respectively; these are numerical differences in the reported means, not an attribution of the overall gain to the loss function. Key experimental and dataset-driven limitations of this work are discussed in Section 4.4 Limitations. Future work will extend the proposed approach to rapid building-damage assessment under other natural-disaster scenarios, carry out external cross-region validation to verify its generalisability, reduce the model's dependence on disaster-specific annotated datasets, and further evaluate its cross-scene performance under external validation.

Acknowledgement: The authors would like to thank all the pre-participants involved in this study, as well as Chongqing University Major Engineering Quality Inspection Co., Ltd., for their efforts in conducting the original post-earthquake building damage assessment and collecting the anonymized data used in this study. Their contributions made the secondary analysis in the present research possible.

Funding Statement: The funding for this research was provided by the National Key Research and Development Program of China (Grant No. 2025YFF0520001).

Author Contributions: Haijia Wen: methodology; supervision; project administration; funding acquisition; writing—review & editing. Wanying Zhang: methodology; data curation; validation; formal analysis; writing—original draft. Yingqi Zeng: data curation; writing—original draft. Fengguang Xiong: methodology; validation. Yujie Li: formal analysis; visualization. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The UAV orthophoto and the cropped training patches used in this study are not deposited as supplementary files. Access to the source imagery, if permitted, may be requested from the corresponding author.

Ethics Approval: Not applicable. This study exclusively involves the engineering assessment of post-earthquake structural damage to buildings and visual interpretation of UAV remote sensing data. No human participants, human subjects, or personal/identifiable data were involved in the research.

Conflicts of Interest: Given his role as First and Corresponding author of this journal, Haijia Wen had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Chen X, Li X, Liu Z. Evaluation of earthquake disaster recovery patterns and influencing factors: a case study of the 2008 Wenchuan earthquake. All Earth. 2023;35(1):132–48. doi:10.1080/27669645.2023.2207942.
2. Liang B, Hou J, He Z. Rapid assessment method to assess vulnerability of structures using vulnerability index and disaster matrix. Bull Earthq Eng. 2023;21(5):2691–722. doi:10.1007/s10518-023-01614-0.
3. Wen H, Zhou X, Zhang C, Liao M, Xiao J. Different-classification-scheme-based machine learning model of building seismic resilience assessment in a mountainous region. Remote Sens. 2023;15(9):2226. doi:10.3390/rs15092226.
4. Ahmad R. Smart remote sensing network for disaster management: an overview. Telecommun Syst. 2024;87(1):213–37. doi:10.1007/s11235-024-01148-z.
5. Kari SS, Raj AAB, Balasubramanian K. Evolutionary developments of today's remote sensing radar technology—right from the telemobiloscope: a review. IEEE Geosci Remote Sens Mag. 2024;12(1):67–107. doi:10.1109/mgrs.2023.3329928.
6. Ma X. High-resolution image compression algorithms in remote sensing imaging. Displays. 2023;79(4):102462. doi:10.1016/j.displa.2023.102462.
7. Saito K, Spence R, de C Foley TA. Visual damage assessment using high-resolution satellite images following the 2003 bam, Iran, earthquake. Earthq Spectra. 2005;21(1 Suppl):309–18. doi:10.1193/1.2101107.
8. Huang L, You J, Zhu X, Shuai T, Liao Y. Object-oriented polarimetric SAR image classification via the combination of a pixel-based classifier and a region growing technique. Eur J Remote Sens. 2023;56(1):2244149. doi:10.1080/22797254.2023.2244149.
9. Ali Rafiee A, Farhang M. A deep convolutional neural network for salt-and-pepper noise removal using selective convolutional blocks. Appl Soft Comput. 2023;145(1):110535. doi:10.1016/j.asoc.2023.110535.
10. Guo S, Feng Z, Wang P, Chang J, Han H, Li H, et al. Mapping and classification of the Liaohe Estuary wetland based on the combination of object-oriented and temporal features. IEEE Access. 2024;12(9):60496–512. doi:10.1109/ACCESS.2024.3389935.
11. Wen H, Hu J, Xiong F, Zhang C, Song C, Zhou X. A random forest model for seismic-damage buildings identification based on UAV images coupled with RFE and object-oriented methods. Nat Hazards. 2023;119(3):1751–69. doi:10.1007/s11069-023-06186-5.
12. Wen H, Wu J, Zhang C, Zhou X, Liao M, Xu J. Hybrid optimized RF model of seismic resilience of buildings in mountainous region based on hyperparameter tuning and SMOTE. J Build Eng. 2023;71(11):106488. doi:10.1016/j.jobe.2023.106488.

13. Zhang C, Wen H, Liao M, Lin Y, Wu Y, Zhang H. Study on machine learning models for building resilience evaluation in mountainous area: a case study of Banan district, Chongqing, China. Sensors. 2022;22(3):1163. doi:10.3390/s22031163.
14. Günen MA. Fast building detection using new feature sets derived from a very high-resolution image, digital elevation and surface model. Int J Remote Sens. 2024;45(5):1477–97. doi:10.1080/01431161.2024.2313991.
15. Gautam D, Bhattarai A, Rupakhety R. Machine learning and soft voting ensemble classification for earthquake induced damage to bridges. Eng Struct. 2024;303(4):117534. doi:10.1016/j.engstruct.2024.117534.
16. Shoaib M, Shah B, El-Sappagh S, Ali A, Ullah A, Alenezi F, et al. An advanced deep learning models-based plant disease detection: a review of recent research. Front Plant Sci. 2023;14:1158933. doi:10.3389/fpls.2023.1158933.
17. Wang Y, Gu L, Jiang T, Gao F. MDE-UNet: a multitask deformable UNet combined enhancement network for farmland boundary segmentation. IEEE Geosci Remote Sens Lett. 2023;20:3001305. doi:10.1109/LGRS.2023.3252048.
18. Chib PS, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: a survey. IEEE Trans Intell Veh. 2024;9(1):103–18. doi:10.1109/tiv.2023.3318070.
19. Feng S, Sun H, Yan X, Zhu H, Zou Z, Shen S, et al. Dense reinforcement learning for safety validation of autonomous vehicles. Nature. 2023;615(7953):620–7. doi:10.1038/s41586-023-05732-2.
20. Reda M, Onsy A, Haikal AY, Ghanbari A. Path planning algorithms in the autonomous driving system: a comprehensive review. Robot Auton Syst. 2024;174(10):104630. doi:10.1016/j.robot.2024.104630.
21. Xiong F, Wen H, Zhang C, Song C, Zhou X. Semantic segmentation recognition model for tornado-induced building damage based on satellite images. J Build Eng. 2022;61(2):105321. doi:10.1016/j.jobe.2022.105321.
22. Hong Z, Zhong H, Pan H, Liu J, Zhou R, Zhang Y, et al. Classification of building damage using a novel convolutional neural network based on post-disaster aerial images. Sensors. 2022;22(15):5920. doi:10.3390/s22155920.
23. Wang Y, Jing X, Xu Y, Cui L, Zhang Q, Li H. Geometry-guided semantic segmentation for post-earthquake buildings using optical remote sensing images. Earthq Engng Struct Dyn. 2023;52(11):3392–413. doi:10.1002/eqe.3966.
24. Xu Y, Li Y, Zheng X, Zheng X, Zhang Q. Computer-vision and machine-learning-based seismic damage assessment of reinforced concrete structures. Buildings. 2023;13(5):1258. doi:10.3390/buildings13051258.
25. Li T, Wang C, Zhang H, Wu F, Zheng X. DDFormer: a dual-domain transformer for building damage detection using high-resolution SAR imagery. IEEE Geosci Remote Sens Lett. 2023;20:3002205. doi:10.1109/LGRS.2023.3288007.
26. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 2018;106(7):249–59. doi:10.1016/j.neunet.2018.07.011.
27. Li Z, Wang H, Liu Y. Semantic segmentation of remote sensing image based on bilateral branch network. Visual Comput. 2024;40(5):3069–90. doi:10.1007/s00371-023-03011-9.
28. Li SQ, Zhang C, Zheng LL, Chen PC, Qin PF. Comparison of vulnerability models for masonry building portfolios considering different macroseismic intensity scales. J Build Eng. 2025;103(1):112066. doi:10.1016/j.jobe.2025.112066.
29. Gautam D, Fabbrocino G, Santucci de Magistris F. Derive empirical fragility functions for Nepali residential buildings. Eng Struct. 2018;171(3):617–28. doi:10.1016/j.engstruct.2018.06.018.
30. Xiong C, Li Q, Lu X. Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network. Autom Constr. 2020;109(4):102994. doi:10.1016/j.autcon.2019.102994.
31. Takhtkeshha N, Mohammadzadeh A, Salehi B. A rapid self-supervised deep-learning-based method for post-earthquake damage detection using UAV data (case study: sarpol-e zahab, Iran). Remote Sens. 2023;15(1):123. doi:10.3390/rs15010123.
32. Nex F, Duarte D, Steenbeek A, Kerle N. Towards real-time building damage mapping with low-cost UAV solutions. Remote Sens. 2019;11(3):287. doi:10.3390/rs11030287.
33. Liu X, Deng Z, Yang Y. Recent progress in semantic image segmentation. Artif Intell Rev. 2019;52(2):1089–106. doi:10.1007/s10462-018-9641-3.
34. Ren W, Tang Y, Sun Q, Zhao C, Han QL. Visual semantic segmentation based on few/zero-shot learning: an overview. IEEE. 2024;11(5):1106–26. doi:10.1109/jas.2023.123207.

35. Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst Appl. 2021;169(6214):114417. doi:10.1016/j.eswa.2020.114417.
36. Li R, Zheng S, Duan C, Su J, Zhang C. Multistage attention ResU-net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci Remote Sensing Lett. 2022;19:1–5. doi:10.1109/lgrs.2021.3063381.
37. Liu R, Tao F, Liu X, Na J, Leng H, Wu J, et al. RAANet: a residual ASPP with attention framework for semantic segmentation of high-resolution remote sensing images. Remote Sens. 2022;14(13):3109. doi:10.3390/rs14133109.
38. Akhtarmanesh A, Abbasi-Moghadam D, Sharifi A, Yadkouri MH, Tariq A, Lu L. Road extraction from satellite images using attention-assisted UNet. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:1126–36. doi:10.1109/jstars.2023.3336924.
39. Wang Y, Yang L, Liu X, Yan P. An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+. Sci Rep. 2024;14(1):9716. doi:10.1038/s41598-024-60375-1.
40. Du S, Du S, Liu B, Zhang X. Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images. Int J Digit Earth. 2021;14(3):357–78. doi:10.1080/17538947.2020.1831087.
41. Zhao S, Wang Y, Wu X, Zhang F. MAFNet: dual-branch fusion network with multiscale atrous pyramid pooling aggregate contextual features for real-time semantic segmentation. Complex Intell Syst. 2024;10(4):5107–26. doi:10.1007/s40747-024-01428-w.
42. Safavi F, Rahnemoonfar M. Comparative study of real-time semantic segmentation networks in aerial images during flooding events. IEEE J Sel Top Appl Earth Obs Remote Sens. 2023;16:4–20. doi:10.1109/JSTARS.2022.3219724.
43. Broni-Bediako C, Xia J, Yokoya N. Real-time semantic segmentation: a brief survey and comparative study in remote sensing. IEEE Geosci Remote Sens Mag. 2023;11(4):94–124. doi:10.1109/MGRS.2023.3321258.
44. Yuan Y, Chen X, Wang J. Object-contextual representations for semantic segmentation. In: Computer vision—ECCV 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 173–90. doi:10.1007/978-3-030-58539-6_11.
45. Yuan Y, Huang L, Guo J, Zhang C, Chen X, Wang J. OCNet: object context for semantic segmentation. Int J Comput Vis. 2021;129(8):2375–98. doi:10.1007/s11263-021-01465-9.
46. Zhang J, Jiang H, Shao H, Song Q, Wang X, Zong D. Semantic segmentation of in-vehicle point cloud with improved RangeNet++ loss function. IEEE Access. 2023;11:8569–80. doi:10.1109/access.2023.3238415.
47. Ganerød AJ, Franch G, Lindsay E, Calovi M. Automating global landslide detection with heterogeneous ensemble deep-learning classification. Remote Sens Appl Soc Environ. 2024;36(1):101384. doi:10.1016/j.rsase.2024.101384.