remote sensing

MDPI

Article

High-Quality Damaged Building Instance Segmentation Based on Improved Mask Transfiner Using Post-Earthquake UAS Imagery: A Case Study of the Luding Ms 6.8 Earthquake in China

Kangsan Yu, Shumin Wang *, Yitong Wang and Ziying Gu

Institute of Earthquake Forecasting, China Earthquake Administration, Beijing 100036, China; yukangsan@ief.ac.cn (K.Y.); wangyt@ief.ac.cn (Y.W.); guziying@ief.ac.cn (Z.G.)
* Correspondence: wangsm@ief.ac.cn

Abstract: Unmanned aerial systems (UASs) are increasingly playing a crucial role in earthquake emergency response and disaster assessment due to their ease of operation, mobility, and low cost. However, post-earthquake scenes are complex, with many forms of damaged buildings. UAS imagery has a high spatial resolution, but the resolution is inconsistent between different flight missions. These factors make it challenging for existing methods to accurately identify individual damaged buildings in UAS images from different scenes, resulting in coarse segmentation masks that are insufficient for practical application needs. To address these issues, this paper proposed DB-Transfiner, a building damage instance segmentation method for post-earthquake UAS imagery based on the Mask Transfiner network. This method primarily employed deformable convolution in the backbone network to enhance adaptability to collapsed buildings of arbitrary shapes. Additionally, it used an enhanced bidirectional feature pyramid network (BiFPN) to integrate multi-scale features, improving the representation of targets of various sizes. Furthermore, a lightweight Transformer encoder has been used to process edge pixels, enhancing the efficiency of global feature extraction and the refinement of target edges. We conducted experiments on post-disaster UAS images collected from the 2022 Luding earthquake with a surface wave magnitude (Ms) of 6.8 in the Sichuan Province of China. The results demonstrated that the average precisions (AP) of DB-Transfiner, AP_{box} and AP_{seg}, are 56.42% and 54.85%, respectively, outperforming all other comparative methods. Our model improved the original model by 5.00% and 4.07% in AP_{box} and AP_{seg}, respectively. Importantly, the AP_{seg} of our model was significantly higher than the state-of-the-art instance segmentation model Mask R-CNN, with an increase of 9.07%. In addition, we conducted applicability testing, and the model achieved an average correctness rate of 84.28% for identifying images from different scenes of the same earthquake. We also applied the model to the Yangbi earthquake scene and found that the model maintained good performance, demonstrating a certain level of generalization capability. This method has high accuracy in identifying and assessing damaged buildings after earthquakes and can provide critical data support for disaster loss assessment.

Keywords: damaged buildings; UAS; instance segmentation; Mask Transfiner; Luding Ms6.8 earthquake

check for updates

Citation: Yu, K.; Wang, S.; Wang, Y.; Gu, Z. High-Quality Damaged Building Instance Segmentation Based on Improved Mask Transfiner Using Post-Earthquake UAS Imagery: A Case Study of the Luding Ms 6.8 Earthquake in China. *Remote Sens.* **2024**, *16*, 4222. https://doi.org/10.3390/rs16224222

Academic Editor: Haigang Sui

Received: 7 October 2024
Revised: 4 November 2024
Accepted: 11 November 2024
Published: 13 November 2024

1. Introduction

China is one of the most earthquake-prone countries in the world, with earthquake-induced building damage being a major source of economic losses and casualties [1]. Identifying building damage information is critical for disaster assessment, providing essential decision support for emergency rescue and post-disaster reconstruction [2]. Traditional methods of obtaining building damage information primarily rely on manual surveys, which are costly, time-consuming, and lack timeliness. Due to their advantages of low cost, flexibility, rapid response, and the ability to capture images below cloud cover, UAS remote sensing has garnered increasing attention from scholars for disaster information collection and seismic damage assessment of buildings [3–9].

In recent years, researchers have carried out extensive studies on identifying damaged buildings using UAS remote sensing technology and deep learning methods [10–17]. Current approaches for damaged building identification can be broadly classified into object detection methods and semantic segmentation methods. Object detection uses disaster imagery to enable computers to locate damaged buildings by outputting horizontal bounding boxes and category labels for the identified regions [18]. Tilon et al. proposed a deep learning approach to post-disaster building damage detection using anomaly detection and generative adversarial networks. Through verification of the post-earthquake drone dataset, it was found that the model trained by this method can detect building damage caused by earthquakes [10]. Jing et al. proposed a network based on the YOLOv5s network that uses BiFPN for multi-scale feature fusion. The model has good accuracy and real-time performance [11]. Pi et al. applied a series of convolutional neural network (CNN) models to detect objects in drone orthophotos after disasters and found that these models can identify damaged and undamaged building roofs [12]. Focusing on the real-time performance of model recognition, Wang et al. proposed a real-time detection method for building damage areas suitable for embedded systems, which is suitable for practical applications in post-earthquake scenarios [13]. Semantic segmentation replaces the fully connected layers in CNN with convolutional layers, employing upsampling to restore image size and classifying each pixel to locate the information on damaged buildings [19]. Hong et al. proposed a convolutional neural network called EBDC-Net for building damage assessment. The network uses a feature extraction encoder module to extract semantic information, and a classification module to combine global features and contextual features to improve accuracy [14]. Zhang et al. utilized multi-scale segmentation and object-oriented classification methods to extract roof damage features from UAS oblique photography imagery. They introduced a normalized digital surface model to identify height-related damage features and used a lightweight CNN model for the preliminary evaluation of building facades [15]. Wang et al. proposed the QCNet-M-N model, which has a flexible configuration of M encoding stages and N embedded convolution operations. The model can identify earthquake-damaged buildings at the pixel level and achieve robust and stable segmentation accuracy under various weather conditions, such as abnormal lighting, rain, and fog [16]. Khankeshizadeh et al. proposed the WETUM, which predicts building damage maps by integrating three independently trained U-NET networks through a proposed grid search technique [17].

Although the above-mentioned UAV-related studies on post-earthquake damaged building assessment have achieved promising results, they still face some noteworthy challenges and gaps that should be well addressed. First, after an earthquake, quickly locating damaged buildings and accurately obtaining the contour information of damaged buildings are equally important for post-disaster rescue and reconstruction, but the above-mentioned methods for identifying damaged buildings usually only focus on one of the tasks. Object detection methods only provide bounding boxes for damaged buildings, while semantic segmentation methods provide segmentation masks. These approaches cannot simultaneously provide both the target range and accurate polygon information for the damaged buildings. In addition, due to the characteristics of UAS images and complex post-earthquake scenes, the contours of damaged buildings obtained by existing segmentation methods are coarse. The recognition results of the target edge are not fine enough; especially, there is a phenomenon of misidentifying multiple adjacent damaged buildings as one.

Instance segmentation can classify all pixels in an image, providing accurate polygons for targets as well as bounding boxes for detected objects. This technique can distinguish different instances of the same category [20] and better meet the actual task requirements of post-earthquake damaged building identification. A representative algorithm is Mask R-CNN [21], which builds on the ideas of Faster R-CNN [22]. It uses a region proposal network (RPN) to generate candidate regions with high recall at low cost and then detects the location of bounding boxes. Additionally, it adds a fully convolutional network (FCN) branch to segment each RoI region, resulting in the segmentation mask. Other algorithms such as PolarMask [23], YOLOACT [24], and SOLO [25] eliminate the proposal generation and feature re-pooling steps, integrating detection and segmentation into a single network to achieve more efficient results. Inspired by DETR [26], algorithms such as SOLQ [27], QueryInst [28], and FastInst [29] treat segmentation as a set prediction problem and use queries to represent the interested objects and jointly perform classification, detection, and mask regression on them. However, there is still a significant gap between the detection and segmentation performance for the algorithms, and the quality of the masks still needs improvement. In 2022, Ke et al. proposed a high-quality and efficient instance segmentation method called Mask Transfiner [30]. This network follows a coarse-to-fine feature extraction approach. Initially, the coarse mask of the target is obtained using a base detector, and then a Transformer [31] structure is employed to refine the coarse mask. Unlike methods that operate directly on dense tensors, this algorithm decomposes the image into a quadtree. Within the Transformer structure, only the error-prone nodes of the quadtree are corrected in parallel, rather than processing continuous images. Consequently, the model can predict highly accurate instance masks at a lower computational cost. The Mask Transfiner network was first applied in the field of marine pollution monitoring. Zou et al. introduced the Mask Transfiner into a floating-algae detection network and proposed CA-ResNet by integrating coordinate attention into the ResNet structure to model both the channel and position dependencies [32]. Subsequently, Yang et al. modified the classification branch of the Mask Transfiner by increasing the resolution of the feature map of each Region of Interest (RoI) region, incorporating a dual attention mechanism, and leveraging a center loss function, named RefinePod [33]. This model was used for high-throughput soybean pods, high-quality segmentation, and accurate seed-per-pod estimation. Panboonyuen et al. first applied the quadtrees in Mask Transfiner to anomaly detection and proposed a method for automotive damage recognition based on the Mask Attention Refinement with Sequential quadtree nodes (MARS) structure [34]. In this algorithm, MARS represented self-attention mechanisms to draw global dependencies between the sequential quadtree nodes layer and quadtree Transformer to recalibrate channel weights and predict highly accurate instance masks.

As a high-quality instance segmentation method, Mask Transfiner has a great advantage in obtaining accurate segmentation masks. In our study, we collected UAS remote sensing orthophotos of the Luding Ms6.8 earthquake in Sichuan Province, China. We conducted an in-depth analysis of the characteristics of damaged buildings in UAS images and attempted to transfer Mask Transfiner to the identification of damaged buildings. To address the challenges mentioned above, we proposed a high-quality instance segmentation network for extracting damaged buildings from post-earthquake UAS images. The primary contributions and innovations of this paper are as follows:

1. Different from the existing damaged building identification methods, this paper proposes a high-quality instance segmentation method to extract damaged buildings, which can accurately obtain the location and fine contour of damaged buildings. Each polygon predicted by the proposed method is almost consistent with the contours of damaged buildings.
2. To enhance the accuracy of collapsed building recognition, we use deformable convolution to replace standard convolution in the backbone part. This allows the network to capture more detailed features of irregularly shaped objects, thereby adapting to the arbitrariness of the shape of collapsed buildings.
3. An enhanced bidirectional feature pyramid network is proposed to fuse multi-scale features. It can enhance the feature expression ability of targets of different sizes, thereby improving the model's ability to recognize damaged buildings of different sizes.
4. We propose a more lightweight Transformer sequence encoder. This improves the efficiency of global feature extraction and the refinement of target edges when processing pixels in incoherent areas.

2. Study Area and Data

2.1. Study Area

At 12:52 PM on 5 September 2022, the Ms6.8 earthquake struck Luding County, Garze Tibetan Autonomous Prefecture, Sichuan Province, in China. The epicenter was located at 29.59°N latitude and 102.09°E longitude with a focal depth of 16 km. The maximum intensity of the earthquake reached IX, with areas experiencing intensity VII and above covering approximately 3608 km^2. Due to the sudden occurrence of the earthquake and the inadequate seismic resistance of rural buildings [35,36], substantial damage and even collapse of structures were observed in regions with intensity VI and above. According to local government statistics, the earthquake resulted in 93 deaths, 25 people missing, and over 420 injuries in Luding County. Additionally, more than 50,000 buildings were damaged, and the number of affected individuals exceeded 110,000 [37]. Luding County, being the most severely impacted area, experienced the highest intensity (IX) in most affected regions, including Moxi Town (a and c), Detuo Town (b, d, g), Fawang village (e), and Wandong village (f), as illustrated in Figure 1. The predominant building structures in the study area include frame, brick–concrete, beam–column wood, brick (earth)–wood, and flagstone types. Notably, 80% of the buildings have relatively low seismic resistance, comprising brick–concrete, beam–column wood, and brick (earth)–wood structures, and some self-built rural buildings fail to meet seismic fortification standards [38].

2.2. Damaged Buildings Dataset

The disaster imagery was acquired using the Dajiang Mavic 2 drone equipped with a Sony 35 mm lens. The imaging period was on 6–13 September 2022, with flying heights of 120 to 150 m. After mosaicking with Pix4Dmapper software (version 4.3.31), the generated orthophoto images have a resolution ranging from 0.03 to 0.15 m. Figure 2 illustrates a sample of the orthophoto images, with a total of seven regions being captured. Among these, images a–d were used to produce the damaged buildings dataset for model training, validating, and testing, while images e–g were taken to test the applicability of the proposed model to verify the identification performance and to test its robustness. The seven orthophoto images were cropped to a uniform size of 640 × 640 pixels for both model training and inference.

The damaged buildings within the study area primarily consist of brick–concrete, beam–column wood, and brick (earth)–wood structures. By comparing aerial imagery with field survey data, the extent of damage and key characteristics are illustrated in examples shown in Figure 2a,b. Unlike semantic segmentation tasks, instance segmentation tasks employ two annotation formats: bounding box annotations and mask segmentation annotations [39]. This study annotated each image in both formats by Labelme software (version 5.2.1), resulting in a dataset for model training. For each image, both types of annotation information are stored in a single JSON file. The visualized annotations are displayed in Figure 2c,d. After cropping images, a total of 10,876 images were obtained. After data cleaning, each annotated image was ensured to contain at least one damaged building, resulting in a final dataset of 704 images, including 1372 damaged buildings. The details of the damaged buildings dataset are shown in Table 1.

Table 1. Specifications of the subdatasets employed for training, validating, and testing. These samples originated from the four UAS images shown in Figure 1B(a–d).

Dataset-Labeled	Total Images	Total Sample
Training	480	935
Validation	120	231
Testing	104	206
Sum	704	1372

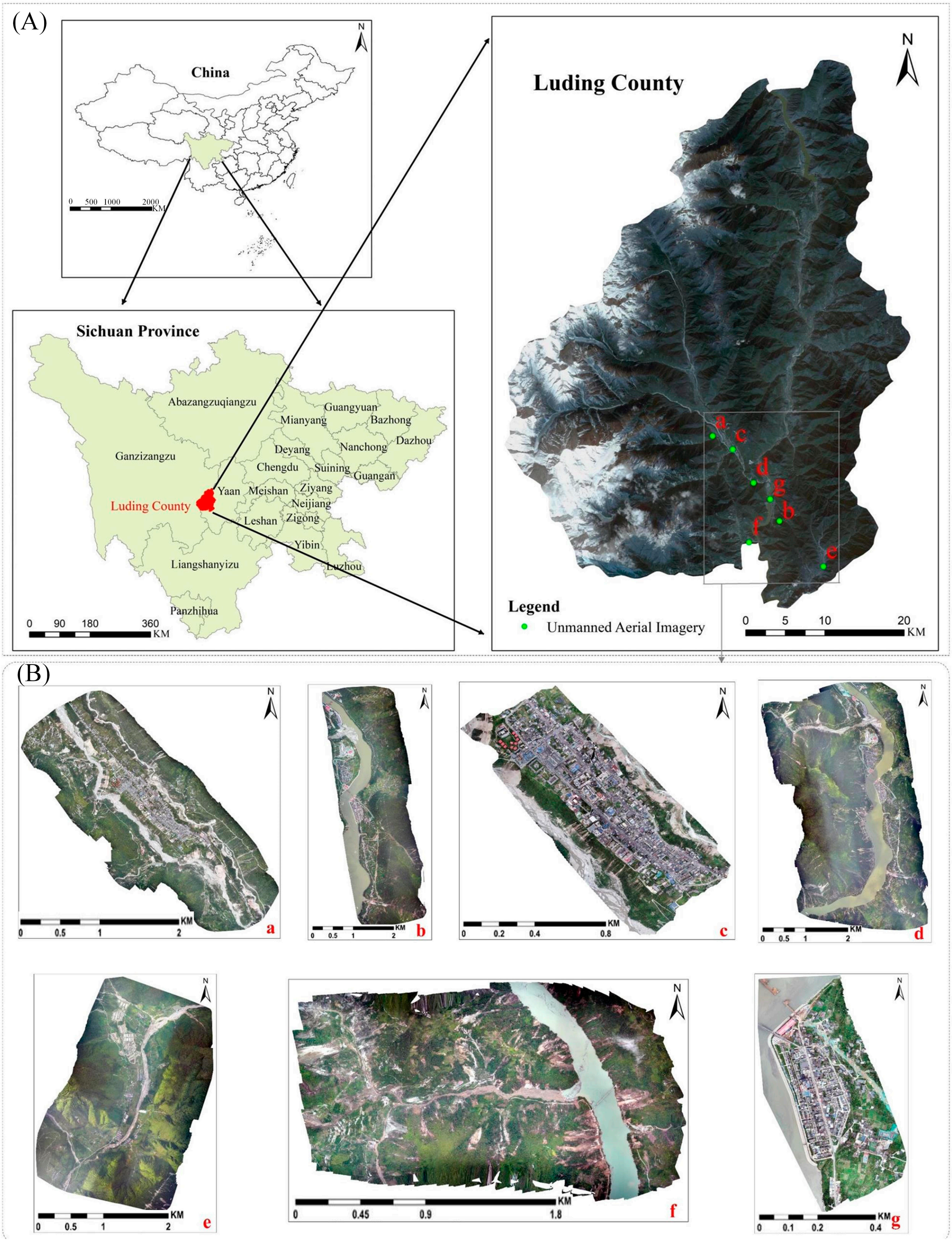

Figure 1. The study area and UAS orthophotos after the earthquake in Luding County, Sichuan Province. (**A**) study area; (**B**) UAS orthophotos: (**a**,**c**) Moxi town; (**b**,**d**,**g**) Detuo town; (**e**) Fawang village; (**f**) Wandong village.

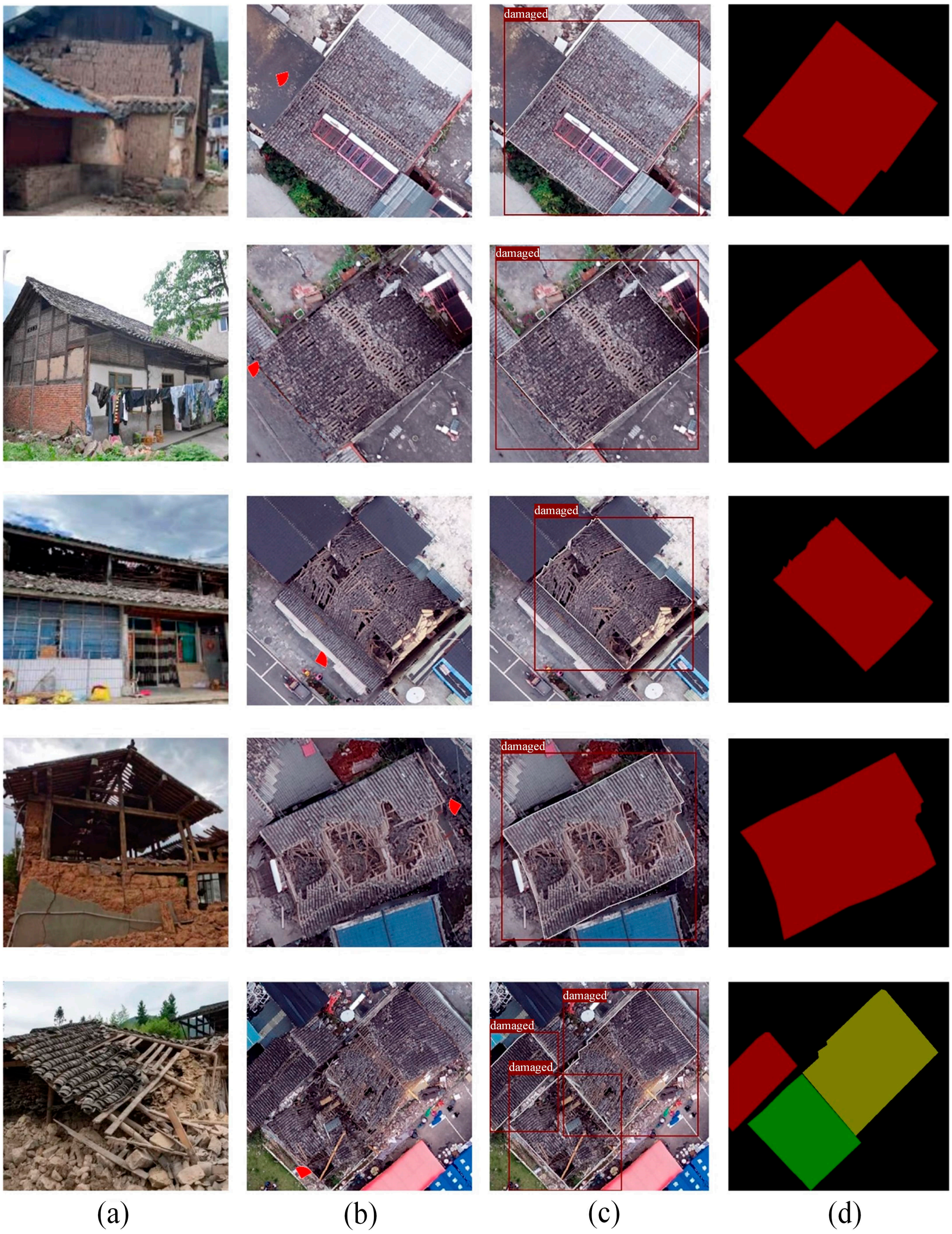

Figure 2. The samples of damaged buildings and labels: (**a**) Field investigation photos; (**b**) UAS images, the red fan-shaped marker representing the viewing angle of the observation location; (**c**) Labeled bounding boxes; (**d**) Labeled instance masks, the color of the polygon masks represents different instance objects.

3. Methodology

3.1. Overview of the Mask Transfiner Model

Unlike conventional convolutional networks, Mask Transfiner is a high-precision instance segmentation network based on Transformers [30]. In this network, the Transformer structure is not used as a feature extraction module within the backbone but rather as a mechanism to refine the predicted coarse masks. Existing instance segmentation models often exhibit roughness around object boundaries, resulting in a significant number of misclassified pixels. Mask Transfiner, utilizing Transformers, focuses solely on correcting these error-prone pixel nodes detected during the segmentation process, effectively addressing inaccuracies.

Specifically, as shown in Figure 3, the network begins with a base detector [21], which utilizes feature extraction via a ResNet-FPN method [40,41] and employs a two-stage instance segmentation strategy. Initially, it detects bounding boxes and subsequently predicts coarse masks for the objects. Based on these coarse masks and the hierarchical features from the base detector, an incoherence detector is employed to identify incoherent regions (error-prone pixels) and generate a quadtree [30]. The quadtree sequence is then input into a Transformer encoder and a decoder with small two-layer MLPs, which classify each pixel to produce a refined mask. Although these error-prone pixels represent only a small fraction of the total, they are crucial for the final mask quality, enabling Mask Transfiner to predict highly accurate instance masks with relatively low computational cost.

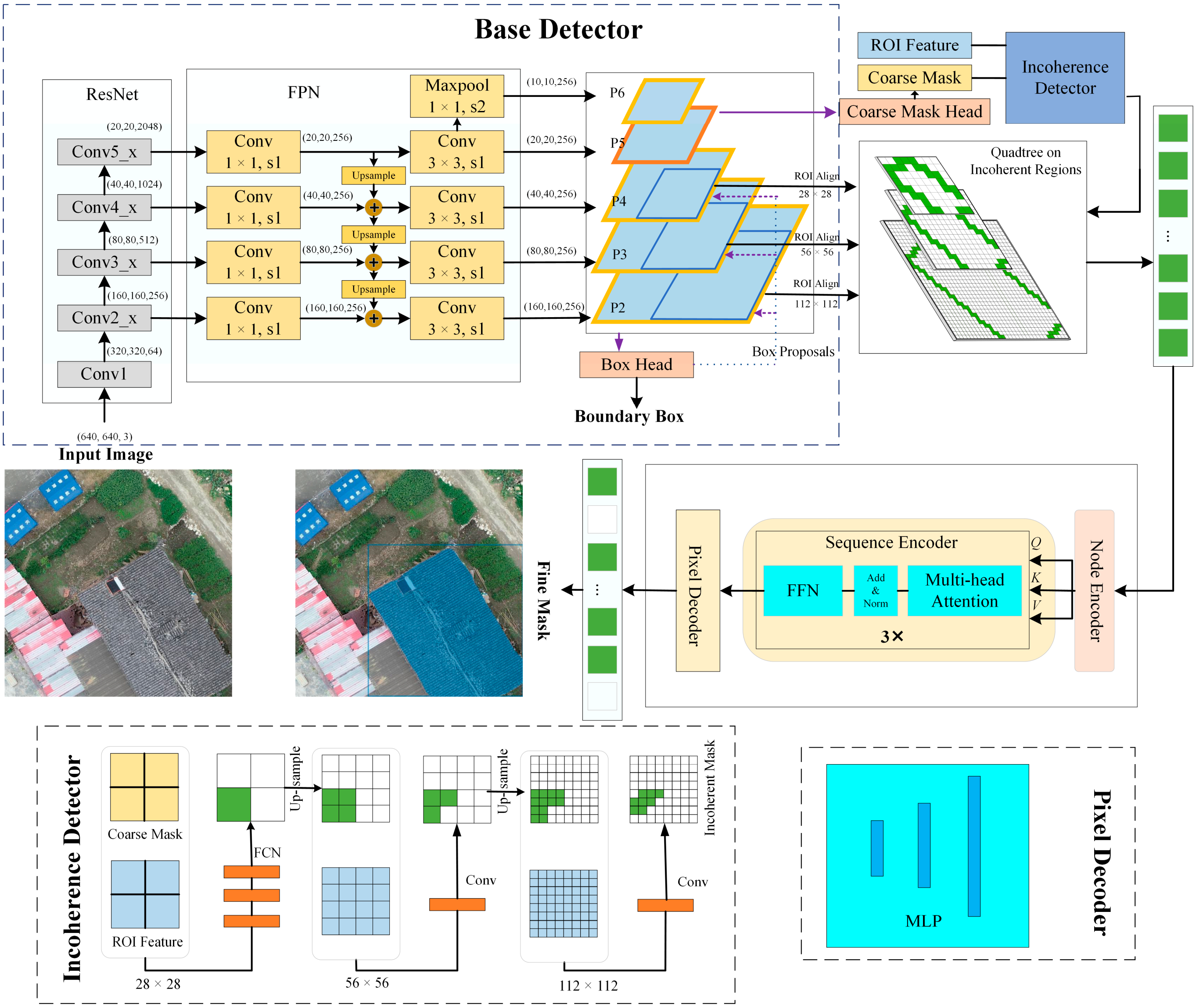

Figure 3. The network architecture of Mask Transfiner.

3.2. Improvement of Mask Transfiner

Mask Transfiner achieves high-quality segmentation masks and demonstrates superior segmentation and detection performance compared to contemporaneous methods. To adapt to the application of building damage identification in post-earthquake UAS imagery, this study has made enhancements to the Mask Transfiner model in three aspects: the CNN of the base detector, the feature pyramid network (FPN) of the base detector, and the sequence encoder Transformer of mask head. The modified model is named DB-Transfiner, and its network architecture is illustrated in Figure 4. The DB-Transfiner model primarily includes the following three modules:

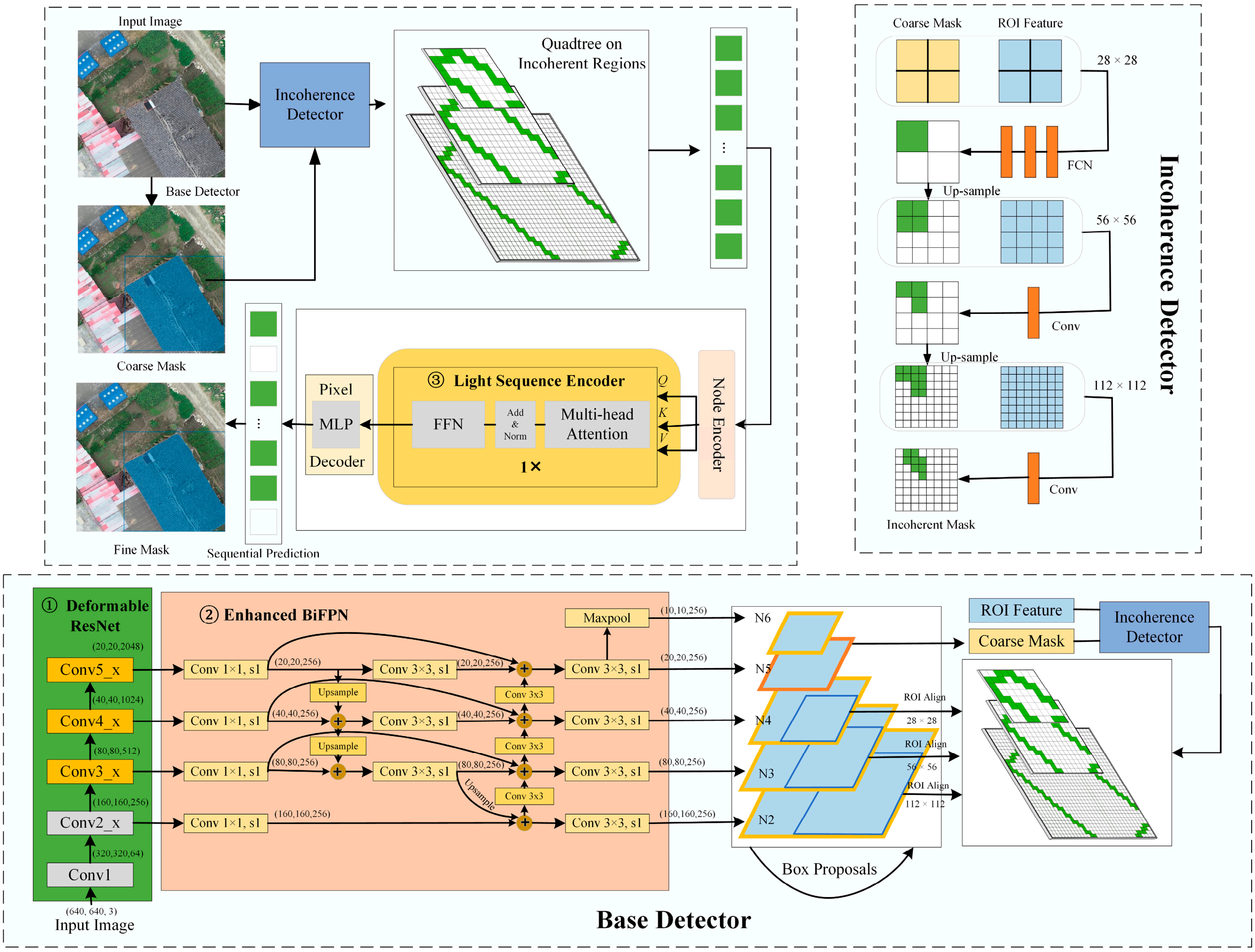

Figure 4. The improved network architecture for DB-Transfiner. Deformable convolution is employed in the backbone. The FPN is replaced by enhanced BiFPN to fuse the multi-scale features, and, in this study, a lightweight sequence encoder is adopted for efficiency.

1. Deformable Convolution Feature Extraction Module (as shown in Figure 4(①)): Improvements were made to the CNN component of the base detector in Mask Transfiner by replacing standard convolutions with deformable convolutions [42,43]. This enhancement allows the network to capture more detailed features of irregularly shaped targets, making it better suited for detecting collapsed building shapes with arbitrary forms.
2. Multi-Scale Feature Extraction and Fusion Module (as shown in Figure 4(②)): The FPN component of the base detector in Mask Transfiner was improved by proposing an enhanced bidirectional feature pyramid network (BiFPN) based on Path Aggregation Network (PANet) [44]. This modification facilitates multi-scale feature fusion, improving the model's ability to represent features of objects with various scales and enhancing its capability to recognize damaged buildings of different sizes.

3. Lightweight Transformer Global Feature Refinement Module (as shown in Figure 4(③)): The Transformer sequence encoder of the encoder was upgraded to use a lightweight Transformer sequence encoder. This improvement enhances the efficiency of global feature extraction and the refinement of object boundaries by processing incoherent region pixels.

3.2.1. Deformable Convolution Feature Extraction Module

In the Mask Transfiner model, standard convolutions are used in ResNet for feature extraction. However, the receptive field of standard convolutions cannot adapt to complex variations and fails to capture fine-grained features effectively [42]. Therefore, we introduce the concept of deformable convolutions within the ResNet architecture, designing a deformable convolution feature extraction module, as shown in Figure 5. Specifically, all 3 × 3 standard convolutions in stages 3 to 5 of ResNet (Conv3_x, Conv4_x, and Conv5_x) [40] are replaced with deformable convolutions, while stages 1 and 2 (Conv1 and Conv2_x) continue to utilize standard convolution operations.

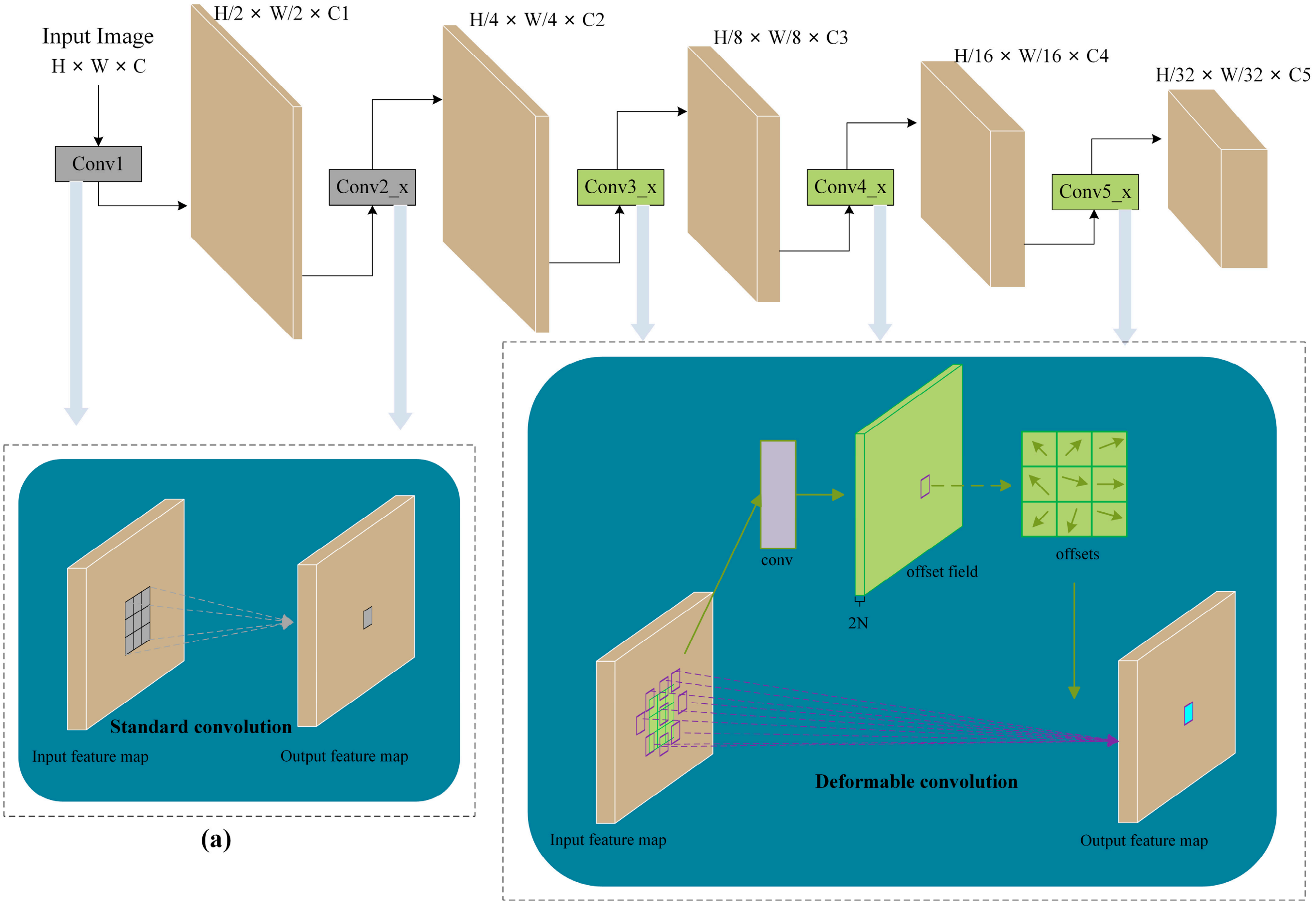

Figure 5. Deformable convolution feature extraction module. Arrows indicate the type of convolution used at each stage. The first two stages use standard convolution, and the last three stages use deformable convolution. (**a**) Standard convolution; (**b**) Deformable convolution.

As shown in Figure 5a, the standard convolution slides a fixed-size filter (convolution kernel) over the input image to compute features [45]. Taking a 3 × 3 convolution as an example, for each output $y(p_0)$, 9 positions are sampled from x. These 9 positions form a grid shape centered around $x(p_0)$, defined as follows:

$$R = \{(-1,-1),(-1,0),(-1,1),(0,-1),(0,0),(0,1),(1,-1),(1,0),(1,1)\} \tag{1}$$

For the feature output at point P_0, the standard convolution operates through the following formula:

$$y(p_0) = \sum_{p_n \epsilon R} \omega(p_n)\cdot x(p_0 + p_n) \tag{2}$$

However, standard convolution is limited by fixed convolution kernels when dealing with targets with significant geometric deformation, making it unable to adapt to deformations and irregular structures in the input image and effectively capture the features of deformed targets. Deformable convolution introduces learnable offsets Δp_n [42] on the basis of standard convolution, allowing the positions of the convolution kernels to be dynamically adjusted to accommodate geometric deformations and irregular structures in the input image, as shown in Figure 5b. For each output $y(p_0)$, the equation is as follows:

$$y(p_0) = \sum_{p_n \epsilon R} \omega(p_n)\cdot x(p_0 + p_n + \Delta p_n) \tag{3}$$

This computation method (DCN v1) [42] may introduce irrelevant contextual areas that interfere with feature extraction, potentially reducing the algorithm's performance. Therefore, in DCN v2 [43], a weight coefficient was added to differentiate whether the introduced areas are regions of interest, as shown in Equation (4):

$$y(p_0) = \sum_{p_k \epsilon R} \omega(p_k)\cdot x(p_0 + p_k + \Delta p_k)\cdot \Delta m_k \tag{4}$$

where Δp_k and Δm_k are the learnable offset and modulation parameter at the k-th position, with $\Delta m_k \in [0,1]$ and Δp_k being arbitrary values.

In this paper, the feature extraction module using deformable convolution adopts DCNv2, which adaptively adjusts the positions of the convolution kernels. This enables a more accurate focus on the features of damaged buildings, better capturing the geometric deformations and structural changes in the targets, thus improving the accuracy of target recognition.

3.2.2. Multi-Scale Feature Extraction and Fusion Module

Mask Transfiner employs FPN to fuse low-level features and high-level semantic features, thereby improving the detection of multi-scale objects. As shown in Figure 6, it utilizes a top-down pathway to extract multi-scale feature maps from different layers of the backbone network. Initially, a 1 × 1 convolution is used to reduce the channel dimensions of feature maps. Through successive upsampling, high-level feature maps are gradually upsampled to the same size as low-level feature maps. Lateral connections are then employed to fuse the upsampled high-level feature maps with the corresponding low-level feature maps, followed by a 3 × 3 convolution to produce the final features at that level. In neural networks, higher layers contain richer semantic information while lower layers capture more detailed information. However, for the identification of damaged buildings, precise information on object boundaries is crucial. Low-level networks contain abundant edge and detail features [44], which are highly valuable. In FPN, the long path for low-level information propagating to high-level levels can lead to significant detail loss. Moreover, FPN treats features of different scales equally during fusion, assuming that features at different levels contribute equally to the fused features [46], which may not balance the information from various scales effectively.

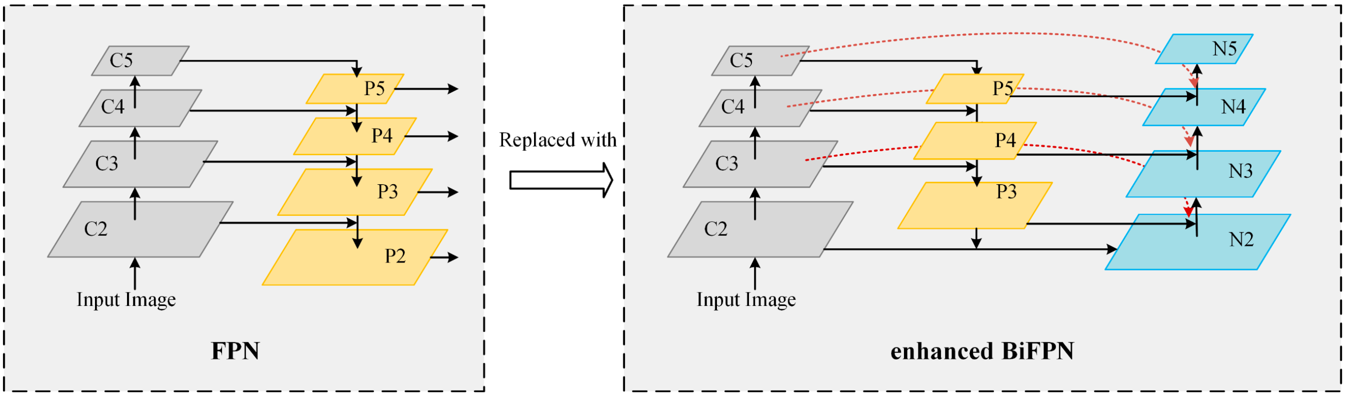

Figure 6. Replacing FPN with enhanced BiFPN to improve feature fusion network.

To address these issues, we introduce an enhanced BiFPN based on the concept of Path Aggregation Network (PANet) [46]. This enhancement includes a bottom-up pathway to strengthen low-level features and additional weights to allow the network to better understand the importance of different level features. The design of enhanced BiFPN is shown in Figure 6. Specifically, after high-level features are upsampled and fused with low-level features, we downsample low-level features and fuse the high-level features. To address the issue of feature sharing across levels, a weight-sharing strategy is employed to reduce network parameters. Additionally, three lateral connection paths are introduced to integrate more feature information without significantly increasing computational burden, enhancing the expressive capability of the feature maps and improving detector performance. Here, we take the C_i layer features and fuse them directly instead of using convolution and fusion. The module integrates more detailed information, allowing the network to better focus on small buildings while accurately identifying large objects, thus improving the model's detection accuracy.

Additionally, an adaptive feature pooling mechanism is employed. We leverage feature levels from N_2 to N_5 for mapping the RoI feature maps onto feature maps of different scales, thereby enhancing the flexibility and accuracy of feature aggregation. The construction method of the RoI feature pyramid is as follows:

$$k = \lfloor k_0 + log_2\left(\sqrt{WH}/size\right) \rfloor \tag{5}$$

where W and H represent the width and height of the RoI, respectively; *size* is set to 320, indicating half of the input image size; and k_0 is set to 4, representing the starting level of the feature pyramid.

According to Equation (5), when the size of an RoI exceeds 320^2, the features of the object will be mapped from N_4. When the size of a RoI is within the range of $[160^2,\ 320^2]$, the object features will be mapped from N_3. For RoIs smaller than 160^2, the object features will be mapped from N_2.

3.2.3. Lightweight Transformer Global Feature Refinement Module

Mask Transfiner employs Vision Transformer to refine the coarse masks predicted by the model. Its sequence encoder is composed of three Transformer structures, each utilizing four-headed self-attention to process different parts of the features in parallel [30]. Although the sequence encoder only calculates sparse feature points in incoherent regions, the large number of parameters in the Transformer structure results in high computational costs and extended training times. Additionally, the four-headed attention mechanism has limited capability in feature representation for damaged buildings. To quickly and accurately identify damaged buildings, we have designed a lightweight Transformer global feature refinement module to improve the efficiency of global feature extraction and edge refinement.

As shown in Figure 7, we employ a Transformer module with an eight-headed self-attention mechanism. We replace the three stacked Transformer modules with a single Transformer module to reduce computational load and improve the efficiency of global feature extraction. Given the uneven destruction of post-disaster buildings, we utilize an eight-headed self-attention mechanism to process different parts of the features in parallel [31], capturing richer features and thus enhancing the model's ability to refine the edges of the targets. This not only improves the efficiency of identifying damaged buildings but also increases the model's recognition accuracy.

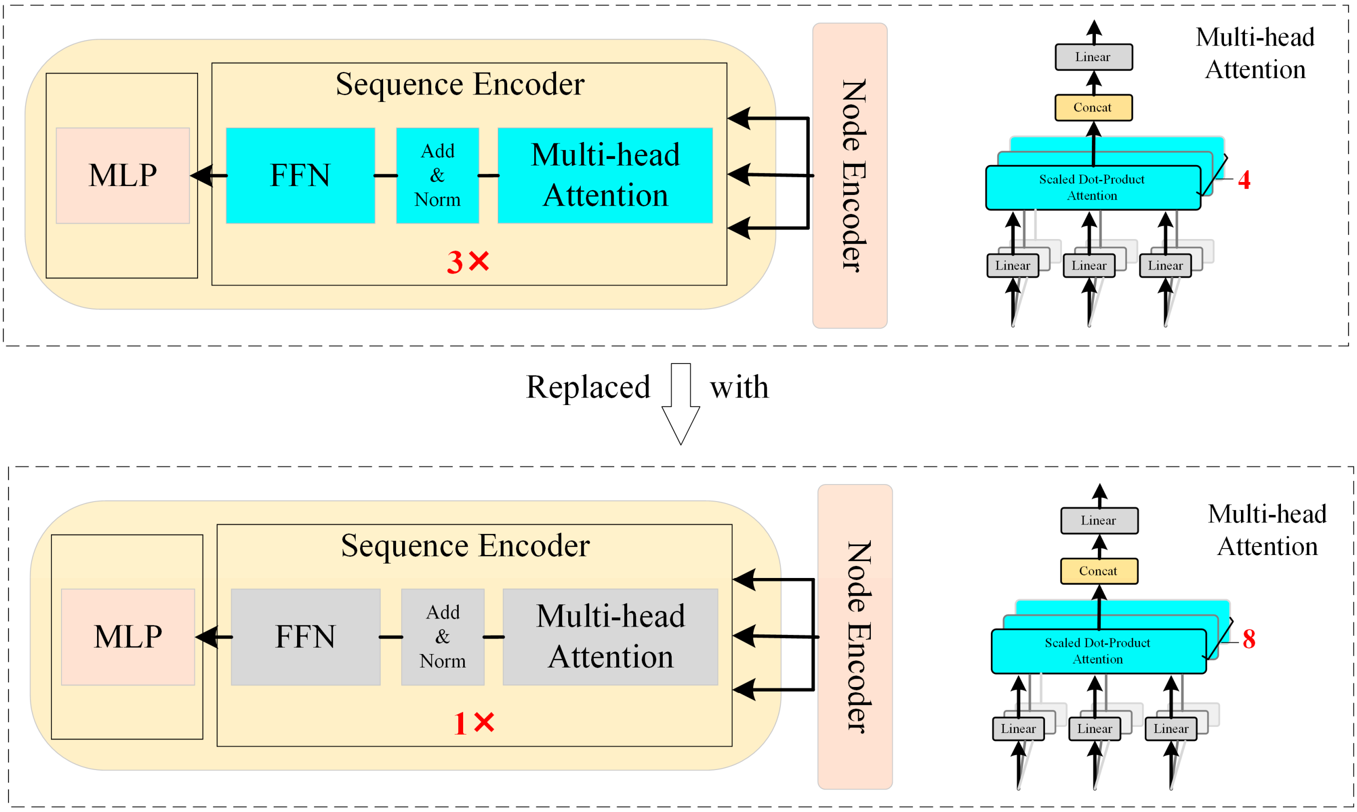

Figure 7. Lightweight sequence encoder to improve the efficiency of the network, using a Transformer structure with an eight-headed self-attention mechanism instead of three Transformer structures with four-headed self-attention mechanisms.

3.3. Implementation Details

The hardware configuration for this experiment includes an Intel Core i7-8700 @ 3.7 GHz six-core processor, 32 GB of RAM, and an NVIDIA A30 GPU. The software environment is a Linux operating system with GCC 5.5 as the compiler.

Our model is implemented using the Pytorch framework, with Python 3.7.16 as the programming language. The GPU computing platform is CUDA 11.3, and the CUDNN 8.0.5 deep learning library is used for GPU acceleration. Models in our experiment uniformly use ResNet50 as the backbone, which uses pretrained weights from ImageNet by Microsoft Research Asia. The model was trained using an initial learning rate of 0.0005, Batch Size 2. Adam's optimization method was used, with a maximum number of epochs of 100.

3.4. Evaluation Metrics

To measure the performance of the model, we need some quantitative evaluation metrics. According to the evaluation metrics of MS COCO, we comprehensively evaluate the model in this paper [41], mainly including the evaluation of the predicted bounding boxes and mask segmentation results. In this study, accuracy (Acc), kappa coefficient $(Kappa)$, and average precision (AP) were used as indicators to evaluate the model performance. The formulas are as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$IoU = \frac{TP}{(TP + FP + FN)} \tag{9}$$

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{10}$$

where true positive (TP) is the area that is actually a damaged building and is predicted as a damaged building, true negative (TN) is the area that is actually a non-damaged building and is predicted as a non-damaged building, false negative (FN) is the area that is actually a damaged building but is predicted as a non-damaged building, and false positive (FP) is the area that is actually a non-damaged building but is predicted as a damaged building. P_o is the observed agreement, which is the proportion of times the model's predictions match the true labels; P_e is the expected agreement, which is the proportion of agreement expected by chance based on the distribution of the classes.

Average precision (AP) is a mainstream evaluation metric for instance segmentation models, describing the model's prediction results based on *Precision* and *Recall*. The average precision for a given IoU (AP_{IoU}) can be expressed as follows:

$$AP_{IoU} = \int_0^1 p(r)dr \tag{11}$$

where r denotes *Recall*, while $p(r)$ represents the *Precision–Recall* (*PR*) curve, and the value of AP_{IoU} is the area under the *PR* curve. AP is the average of AP_{IoU} at 10 IoU thresholds ranging from 0.50 to 0.95 with a step size of 0.05, which can be obtained by the following:

$$AP = \frac{1}{10}\sum\nolimits_{IoU=0.5}^{0.95} AP_{IoU} \tag{12}$$

where $AP_{0.5}$ and $AP_{0.75}$ signify the computed AP at IoU thresholds of 0.5 and 0.75, respectively. Additionally, AP_L, AP_M, and AP_S, respectively, evaluate the models' ability to recognize large, medium, and small targets. The equation is as follows:

$$AP = \begin{cases} AP_S & if\ pixel\ area < 32^2 \\ AP_M & if\ 32^2 < pixel\ area < 96^2 \\ AP_L & if\ pixel\ area > 96^2 \end{cases} \tag{13}$$

We also use Frames Per Second (PFS) to evaluate the recognition efficiency of different methods. The formula is as follows, where T represents the time taken by the model to infer a single image.

$$PFS = \frac{1}{T} \tag{14}$$

4. Results

4.1. Comparison of Model Performance

To validate the accuracy of the models, the same damaged building dataset is used for training and validating. DB-Transfiner is a damaged building detection model built upon a deformable convolution feature extraction module (DCNM), multi-scale feature extraction and fusion module (MEFM), and lightweight Transformer global feature refinement module

(LTGM). We compared the performance of our model with other existing models based on several accuracy assessment metrics.

Figure 8 depicts the training set loss and validation set loss of the model during the training process. With the increase in training epochs, the training loss gradually decreases from 3.5 to approximately 0.25. The trends of training loss and validating loss are generally consistent, with the loss function curve gradually stabilizing. During the first 10 epochs, the loss quickly drops to around 1.25, and the loss function curve converges quickly, with significant fluctuations. Between 10 and 60 epochs, the model's convergence speed slows down, and there is fluctuation in the loss curve. From 60 to 100 epochs, the loss gradually becomes stable.

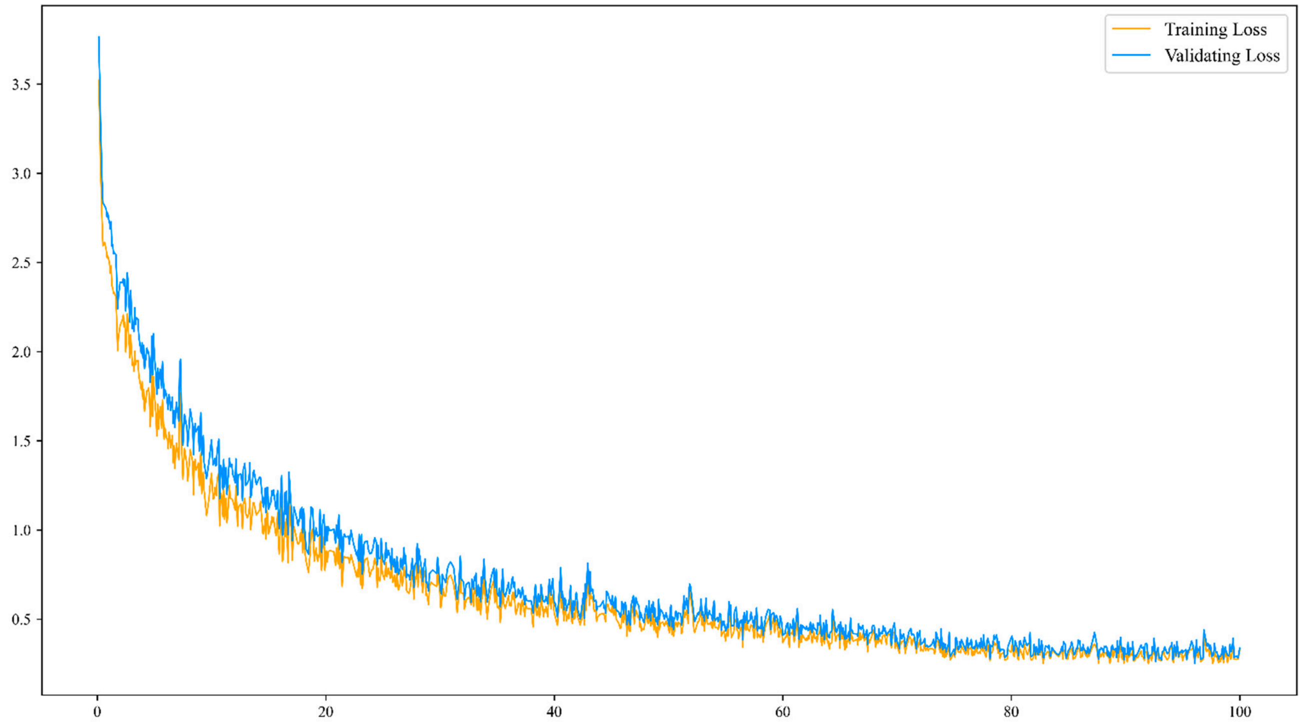

Figure 8. Loss curve during DB-Transfiner training.

The quantitative results are as follows in Table 2. From the results, it can be observed that our model demonstrated a significant advantage over other algorithms in terms of AP, Acc, and $Kappa$ evaluation metrics. Specifically, for AP, our model outperforms other methods in both segmentation (AP_{seg}) and detection (AP_{box}). The proposed DB-Transfiner with an AP_{seg} value of 54.85% managed to improve the AP_{seg} value of the Mask R-CNN, PolarMask, YOLOACT, SOLO, SOLQ, QueryInst, and FastInst networks by 9.07%, 12.20%, 13.62%, 7.40%, 5.35%, 5.07%, and 5.86%, respectively (see Figure 9). Moreover, compared to the widely adopted Mask R-CNN method, the DB-Transfiner model exhibited an enhancement of 13.58% and 0.21 in accuracy and Kappa values, respectively, indicating that the proposed approach substantially improves the accuracy of post-earthquake damaged building identification.

Notably, our model achieved an FPS (Frames Per Second) value of 14.1, which is an increase of 0.9 compared to the Mask Transfiner. This demonstrates that the DB-Transfiner model reduces computational and inference time to some extent, thereby enhancing model efficiency. This advantage is attributed to the use of a lightweight Transformer-based global module (LTGM) in DB-Transfiner, which employs fewer Transformer structures, consequently reducing the number of model parameters.

Table 2. Quantitative results for all models of instance masks. AP_S is null because the damaged buildings are more than 32 × 32 pixels in the UAS imagery. The best results are indicated in bold.

Model	AP_{seg} (%)	$AP_{seg}^{0.5}$ (%)	$AP_{seg}^{0.75}$ (%)	AP_{box} (%)	$AP_{box}^{0.5}$ (%)	$AP_{box}^{0.75}$ (%)	Acc_{seg} (%)	*Kappa* (%)	T (ms/img)	FPS (img/s)
Mask R-CNN	45.78	67.10	50.56	48.88	67.17	51.14	68.41	0.49	48.3	20.7
PolarMask	42.65	65.70	46.52	45.14	65.81	47.32	64.02	0.42	39.8	25.1
YOLOACT	41.16	65.24	45.35	44.32	65.30	46.10-	62.45	0.41	**12.1**	**82.8**
SOLO	47.45	67.88	53.74	49.74	68.15	53.94	70.68	0.53	32.1	31.2
SOLQ	49.50	69.23	56.10	53.12	69.40	56.12	73.92	0.58	73.5	13.6
QueryInst	49.78	69.79	56.45	52.09	69.89	56.55	74.17	0.58	51.3	19.5
FastInst	48.99	68.91	54.62	51.47	69.02	54.80	73.51	0.57	15.4	65.1
Mask Transfiner	50.78	70.72	57.62	51.42	70.49	57.60	75.88	0.60	75.8	13.2
DB-Transfiner (ours)	**54.85**	**70.75**	**62.20**	**56.42**	**71.97**	**60.50**	**81.99**	**0.70**	70.9	14.1

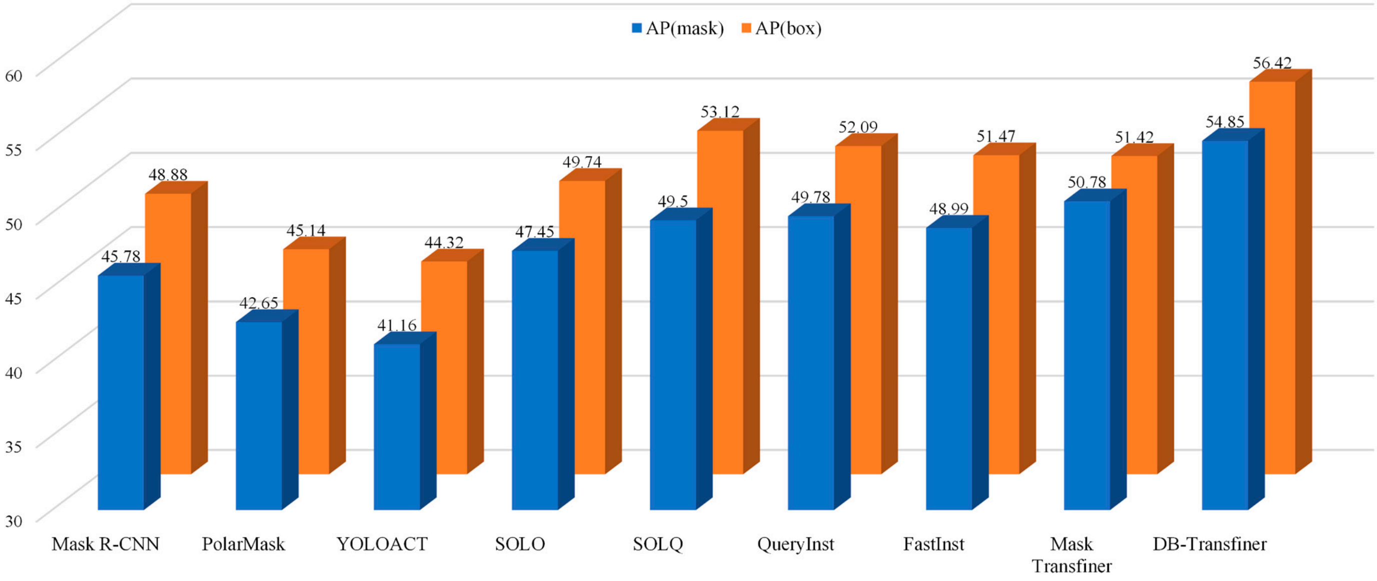

Figure 9. Comparison of the performance of all models based on the metrics *AP* (%).

We also qualitatively compared the inference results of the Mask R-CNN, Mask Transfiner, and DB-Transfiner models on the test set to visualize the effectiveness of damaged building recognition. By overlaying the predicted bounding boxes and binary masks onto the original images and similarly visualizing the labels, we can better assess the performance. This visualization is shown in Figure 10. We found that, compared to the Mask R-CNN method, both Mask Transfiner and DB-Transfiner accurately detect and segment multiple damaged buildings in post-earthquake scenes. Specifically, the results from Mask R-CNN are relatively coarse, with instances of missed or false detections, particularly in images containing multiple damaged buildings. In contrast, Mask Transfiner and DB-Transfiner show almost no missed detections or false alarms, indicating that the Transfiner mask branch performs better in instance segmentation.

Further, to validate the accuracy of the segmentation performance of DB-Transfiner, we visualized only the binary masks predicted by the model. The results demonstrate that the instance segmentation results of DB-Transfiner are closer to the contours of the original target and can identify damaged buildings with high quality. Mask R-CNN produces relatively blurry contours, struggles to distinguish the boundaries between damaged buildings and the background, and misses small targets (e.g., the target in the lower left corner of the third row of Figure 11). Although Mask Transfiner adapts to different-sized targets and generates relatively accurate results for single damaged buildings, its masks appear to stick together when identifying multiple damaged building targets, like Mask R-CNN. DB-Transfiner delivers superior recognition results for damaged buildings, with clearer boundaries that accurately distinguish different instances within the same damaged structure (e.g., targets①and②in the upper left corner of Figure 11). It also shows excellent adaptability for multi-target recognition and targets of various sizes. Our method not only avoids missing targets and false alarms but also produces better high-quality segmentation results.

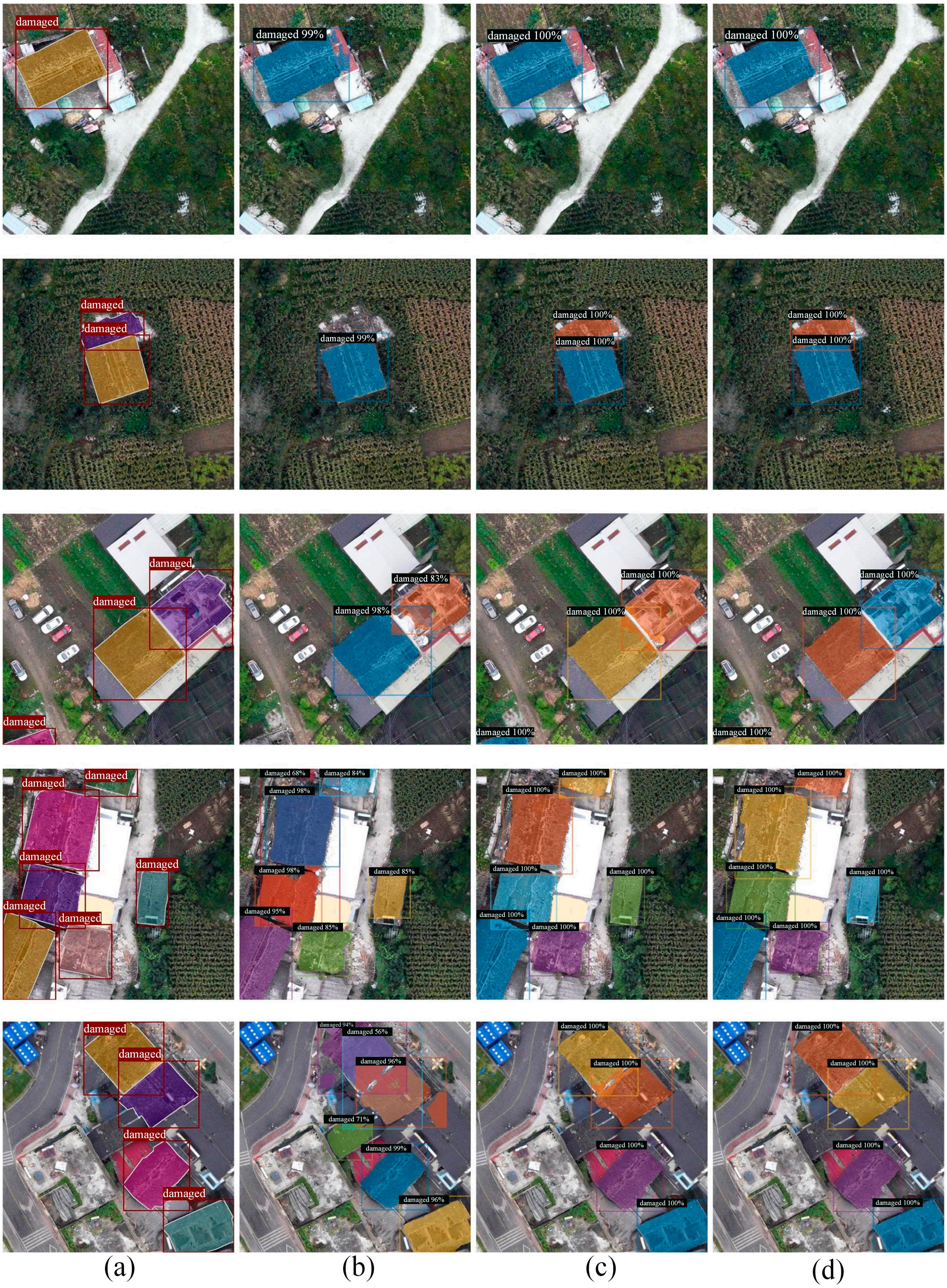

Figure 10. Visualization of the prediction results of different network models. The colored bounding boxes and polygons represent the detection and segmentation results, respectively. (**a**) Annotated images; (**b**) Mask R-CNN; (**c**) Mask Transfiner; (**d**) DB-Transfiner.

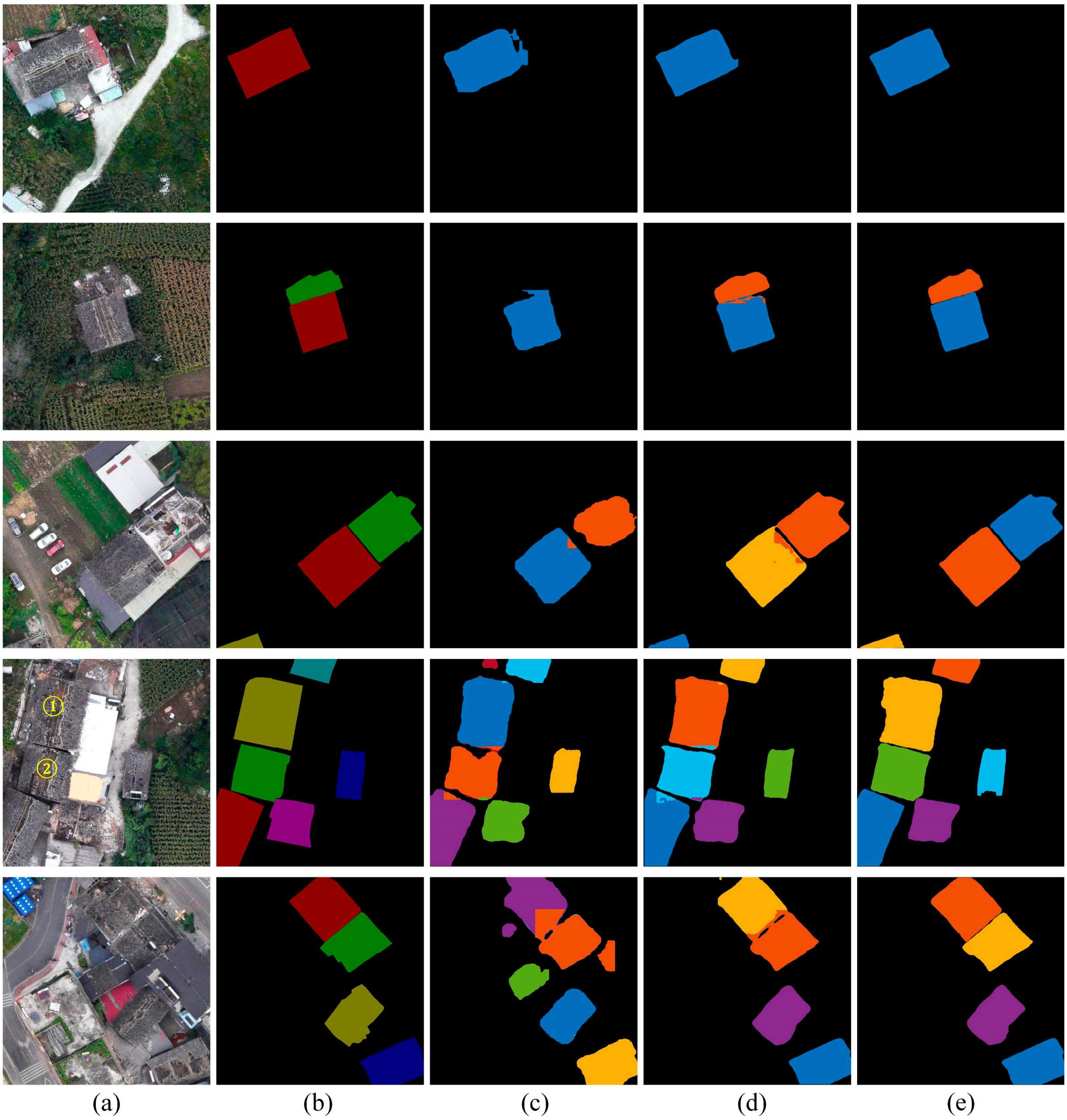

Figure 11. Visualization of instance mask results of different network models. The colored polygons represent the recognized instance objects. ① and ② represent two typical damaged buildings with the same level of destruction. (**a**) Original images; (**b**) Annotated results; (**c**) Mask R-CNN; (**d**) Mask Transfiner; (**e**) DB-Transfiner.

4.2. Ablation Study

To validate the effectiveness of the proposed DCNM, MEFM, and LTGM modules, we conducted ablation experiments on the damaged building dataset. The table below shows the results of progressively adding the three improvements. Here, the baseline represents the Mask Transfiner method, and "+" indicates the inclusion of the corresponding improvement.

The detecting and segmenting results are presented in Tables 3 and 4. Compared to the baseline method, the Mask Transfiner + DCNM model, which includes a deformable convolution feature extraction module, achieved significant improvements with AP for detection bounding boxes and segmentation masks increasing by 2.30% and 2.39%, respectively. After adding MEFM, the model showed notable improvements in the large object accuracy metric (AP_L) and medium object accuracy metric (AP_M) over only using baseline DCNM, with increases of 5.17% and 2.16% in detection bounding boxes, respectively. This demonstrates that the MEFM can extract more features of different sizes, enhancing the ability to perceive objects of varying scales in UAS remote sensing images. When using DCNM, MEFM, and LTGM on the model, the proposed model further improved the AP, $AP_{0.75}$, AP_L, and AP_M for segmentation masks by 4.07%, 4.58%, 4.13%, and 6.10%, respectively, compared to the baseline method.

Table 3. Ablation experiments of bounding boxes. The improvement is indicated in bold.

Model	AP **(%)**	$AP_{0.5}$ **(%)**	$AP_{0.75}$ **(%)**	AP_L **(%)**	AP_M **(%)**	***Accuracy*(%)**	***Kappa* (%)**
Baseline	51.42	70.49	57.60	59.01	31.42	76.14	0.61
Baseline + DCNM	53.72 **(+2.30)**	72.76 **(+2.27)**	60.06 **(+2.46)**	60.36 **(+1.35)**	35.60 **(+4.18)**	79.03 **(+2.89)**	0.66 **(+0.05)**
Baseline + DCNM + MEFM	55.20 **(+3.78)**	69.29 **(−1.20)**	60.47 **(+2.87)**	65.53 **(+6.52)**	37.76 **(+6.34)**	81.75 **(+5.61)**	0.70 **(+0.09)**
Baseline + DCNM + MEFM + LTGM	56.42 **(+5.00)**	71.97 **(+1.48)**	60.50 **(+2.90)**	66.72 **(+7.71)**	37.89 **(+6.47)**	82.93 **(+6.79)**	0.72 **(+0.11)**

Table 4. Ablation experiments of segmentation masks. The improvement is indicated in bold.

Model	AP **(%)**	$AP_{0.5}$ **(%)**	$AP_{0.75}$ **(%)**	AP_L **(%)**	AP_M **(%)**	***Accuracy*(%)**	***Kappa* (%)**
Baseline	50.78	70.72	57.62	59.82	27.42	75.88	0.60
Baseline + DCNM	53.17 **(+2.39)**	72.90 **(+2.18)**	61.30 **(+3.68)**	60.80 **(+0.98)**	30.87 **(+3.45)**	78.52 **(+2.64)**	0.67 **(+0.07)**
Baseline + DCNM + MEFM	54.50 **(+3.72)**	69.20 **(−1.50)**	61.84 **(+4.22)**	63.86 **(+4.04)**	32.90 **(+5.48)**	80.80 **(+4.92)**	0.69 **(+0.09)**
Baseline + DCNM + MEFM + LTGM	54.85 **(+4.07)**	70.75 **(+0.03)**	62.20 **(+4.58)**	63.95 **(+4.13)**	33.52 **(+6.10)**	81.99 **(+6.11)**	0.70 **(+0.10)**

We observed that with the addition of DCNM, MEFM, and LTGM one by one, the model's AP and $AP_{0.75}$ metrics increase in both segmentation and detection tasks, while $AP_{0.5}$ fluctuates and sometimes drops significantly. This is because the model has more stringent evaluation conditions at higher IoU thresholds (e.g., 0.75). This suggests that the model has improved detection and segmentation capabilities under stricter matching conditions (high IoU) but performs less effectively under more lenient conditions (low IoU). This reflects that the model generates more precise and compact prediction boxes and masks results, thus improving the overall accuracy.

4.3. Feature Maps Visualization

We observed that the accuracy of Mask Transfiner improved significantly after incorporating DCNM. Furthermore, the model's performance in recognizing damaged buildings of different sizes was notably enhanced by adopting the MEFM. To explain to our DB-Transfiner model how to make decisions, we visualized the heatmaps before and after the operations of these two modules using gradient-weighted class activation mapping (Grad-CAM) [47]. Class activation maps (CAMs) [48] display the regions of an image that contribute most to the prediction of a particular category, helping to understand the model's decision-making process. Grad-CAM is an improved method that generates more detailed class activation maps using gradient information. We presented key Grad-CAM heatmaps before and after employing DCNM and MEFM. The results are shown in Figure 12.

From these heatmaps, it is visually apparent which regions of the image the DB-Transfiner network focuses on when making classification decisions regarding damaged buildings. In the heatmap of the Conv2_x layer of DCNM, the high-intensity values are mainly distributed over the buildings, and some damaged buildings show high intensity values. This indicates a strong correlation between the network and building features at this layer. In the heatmap of the Conv5_x layer after DCNM using a deformable convolution operation, the high intensity values are concentrated on the damaged buildings, closely aligning with the targets. Collapsed buildings also exhibit high intensity values, suggesting that the model, after using deformable convolutions, can capture more features of collapsed buildings. However, there is still a slight deviation from the actual targets at this stage. When the MEFM is employed, the high intensity regions in the N_5 layer heatmap fall precisely on the damaged buildings. The heatmap shows areas that perfectly match the boundaries of the damaged buildings, with every damaged building of different sizes being highlighted.

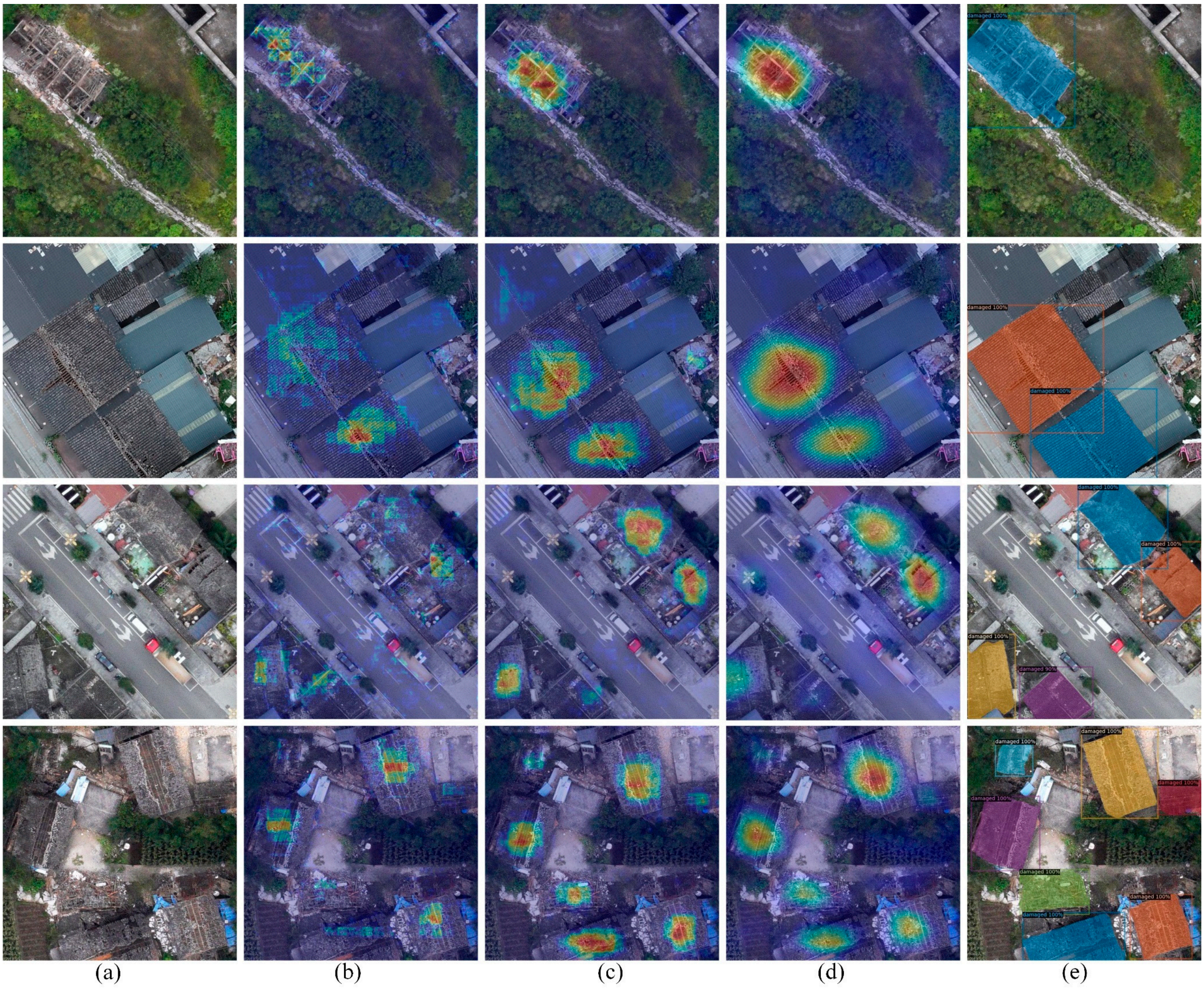

Figure 12. Visualization of heatmaps: (**a**) The original images; (**b**) The heatmaps of Conv2_x layer of the DCNM; (**c**) The heatmaps of Conv5_x layer of the DCNM; (**d**) The heatmaps of N_5 layer of the MEFM; (**e**) The final results. The colored borders represent the model's predicted different instance objects.

The base detector in DB-Transfiner extracts features and then refines the generated coarse segmentation masks. To explain how the LTGM, our model, was adopted, we visualized the feature maps of the coarse masks before and after the LTGM, as shown in Figure 13. We observed that, before using the LTGM, the model can focus on the polygonal shapes of the targets but also tends to highlight local similar features that are unrelated to the targets, leading to inaccurate segmentation masks. However, after refinement with the LTGM, the model pays more attention to the global information of the targets, resulting in contours that more closely resemble the actual targets.

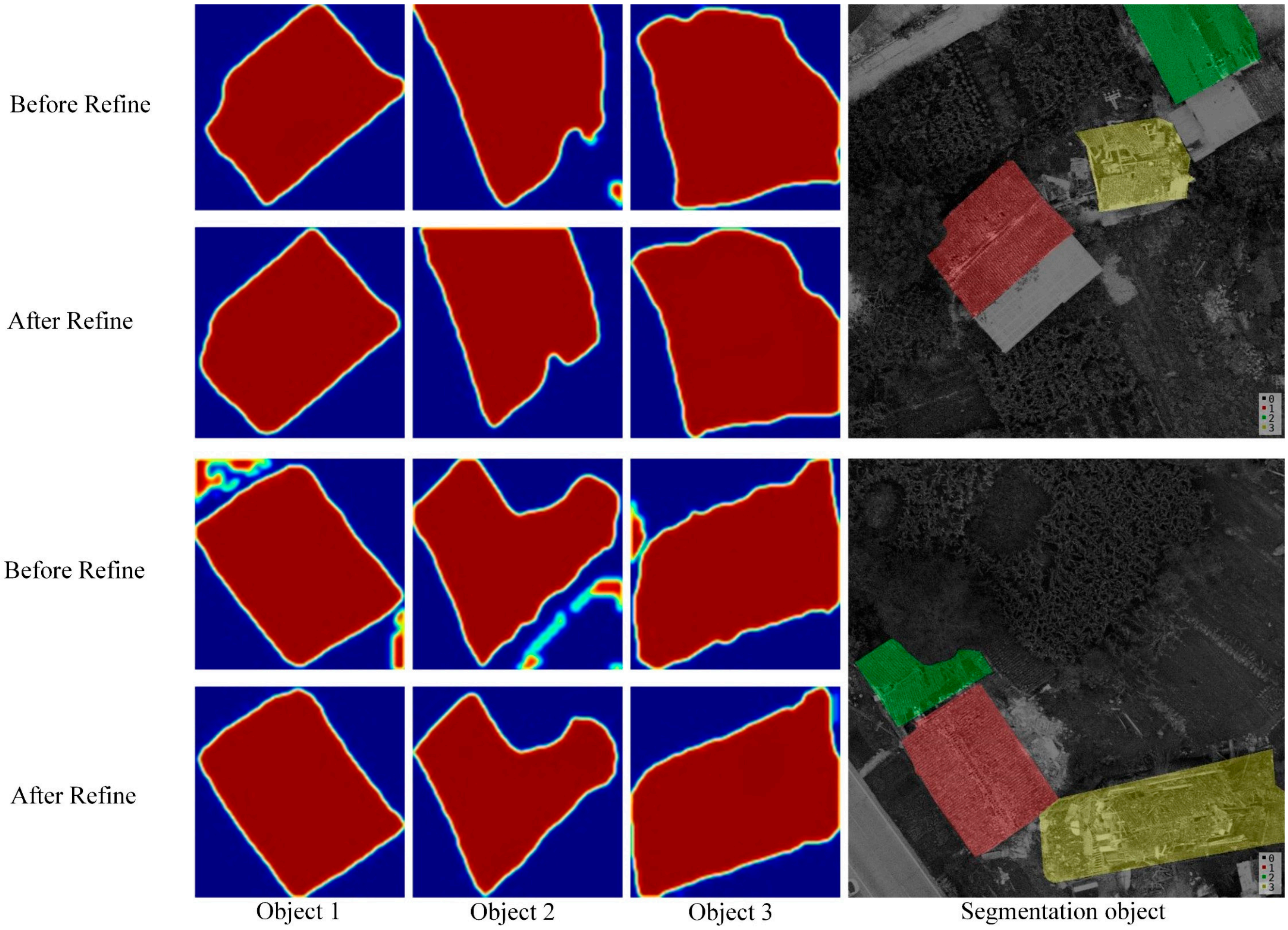

Figure 13. The visualization of feature maps before and after the LTGM. The colored borders represent the different instance objects.

4.4. Applicability of DB-Transfiner Model

To verify the identification performance of our model in recognizing damaged buildings in practical application, we used UAS images from the three areas mentioned in Section 2.2, as shown in Figures 14 and 15. We conducted a visual interpretation of the damaged buildings in the images of these three areas, identifying a total of 514 damaged buildings: 197 buildings in area e, 131 buildings in area f, and 186 buildings in area g. The corresponding statistical results for the three areas are presented in Table 5.

Table 5. Visual interpretation and automatic identification for damaged buildings in the 3 test areas (Figure 1B(e–g)). The time here is how long the DB-Transfiner model takes to identify damaged buildings in a certain area.

Test Area	**Ground Truth**	**Detection Number**	**Wrong Number**	**Omission Number**	**Time(s)**	**Correctness (%)**
e	197	162	6	35	249	82.23
f	131	108	3	23	192	82.44
g	186	164	5	22	86	88.17
Average Correctness	-	-	-	-	-	84.28

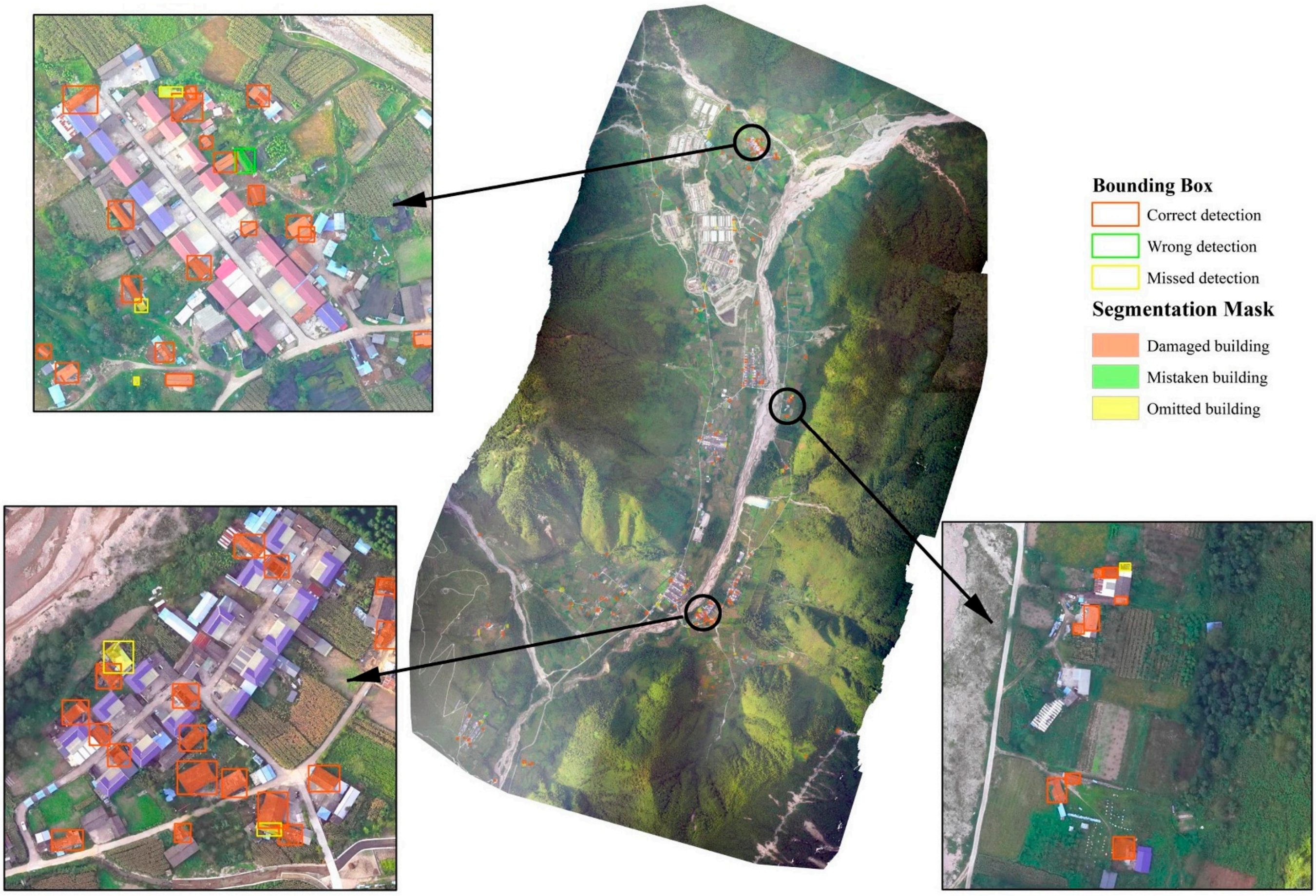

Figure 14. Results of damaged building classification in Fawang village (Figure 1B(e)). Red indicates correct detections, green indicates incorrect detections, and yellow indicates missed.

We found that most of the targets identified by the model are correct, with only a few misidentifications. The model accurately identified 162, 108, and 164 damaged buildings in areas e, f, and g, with correctness rates of 82.23%, 82.44%, and 88.17%, respectively, resulting in an average correctness rate of 84.28%. This demonstrates the better performance of DB-Transfiner in detecting and segmenting post-earthquake damaged buildings in UAS images. However, there were also some missed damaged buildings. The omissions were mainly due to two reasons: first, the poor imaging quality of certain dark areas in the images and second the obstruction by surrounding objects such as fallen walls, piled debris, and nearby trees.

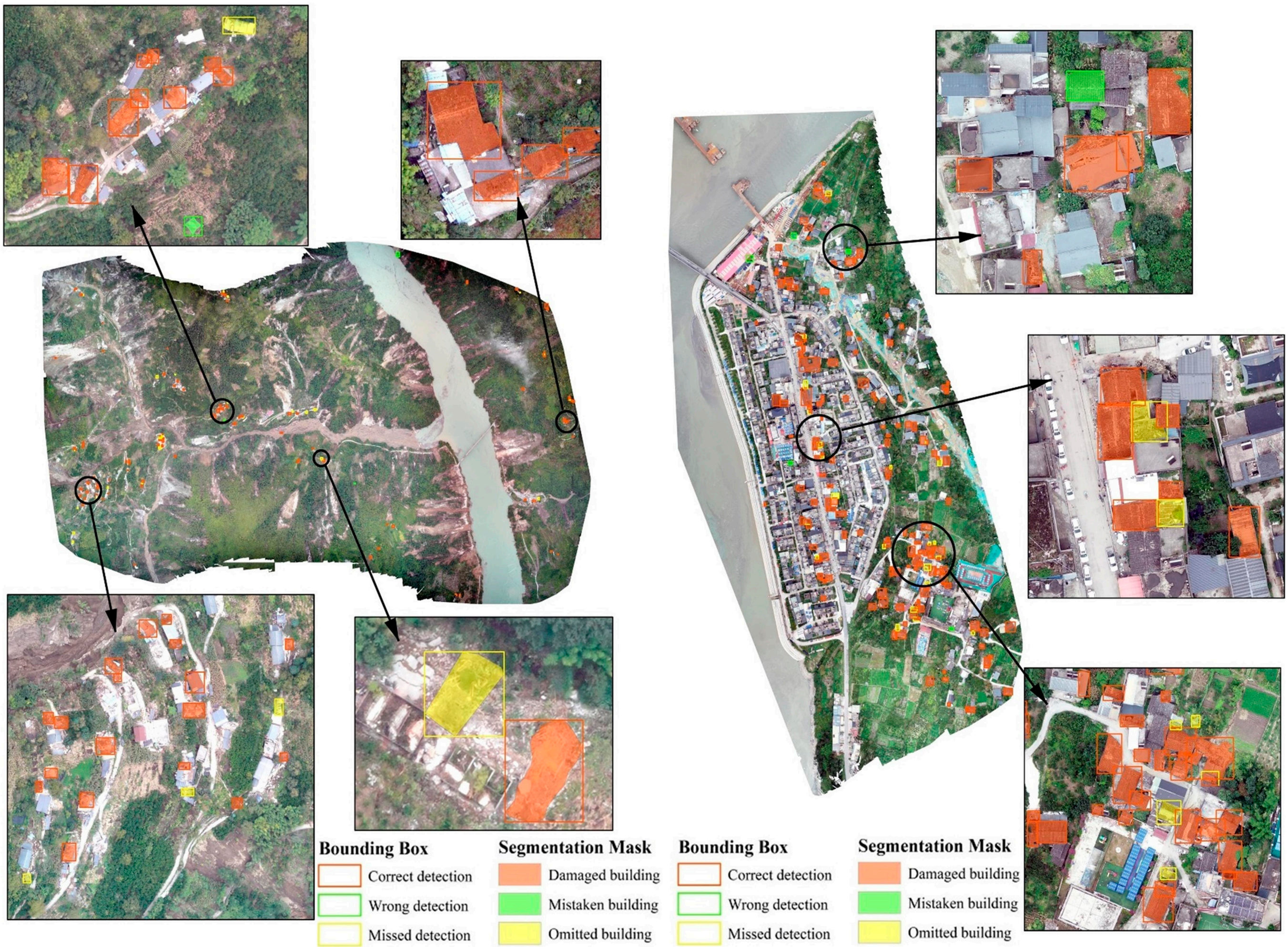

Figure 15. Results of damaged building classification in Wandong village and Detuo town (Figure 1B(f,g)). Red indicates correct detections, green indicates incorrect detections, and yellow indicates missed.

4.5. Generalization Capability of the Model in Yangbi Earthquake

In this study, we also tested the generalization of the model. We applied the DB-Transfiner model to the 21 May 2021, Yangbi M6.4 earthquake in China. As shown in Figure 16, we obtained two UAS remote sensing images and constructed a dataset of damaged buildings from the Yangbi earthquake to test the model's generalization capability. Image (a) is located in Huaian village, while image (b) is situated in the urban area of Yangbi town. The resolution of both images is 0.05 m.

We evaluated the model's performance on the test set, as presented in Table 6 and Figure 17. In the Yangbi earthquake, AP_{seg} scored 53.08%, slightly lower than the score of 54.85% for the Luding earthquake. *Accuracy* exceeded 80%, reaching 80.12%, demonstrating good recognition capability for targets. All of these results indicated that damaged buildings can be identified by our model with good accuracy, especially finer detection at the building edges, demonstrating that the model has a good generalization capability to new data.

Table 6. The generalization experiment of DB-Transfiner for segmentation masks.

Dataset	AP_{seg} (%)	$AP_{seg}^{0.5}$ (%)	$AP_{seg}^{0.75}$ (%)	AP_{seg}^{L} (%)	AP_{seg}^{M} (%)	*Accuracy* (%)	*Kappa* (%)
Luding earthquake	54.85	70.75	62.20	63.95	33.52	81.99	0.70
Yangbi earthquake	53.08	68.97	60.86	61.02	31.15	80.12	0.68

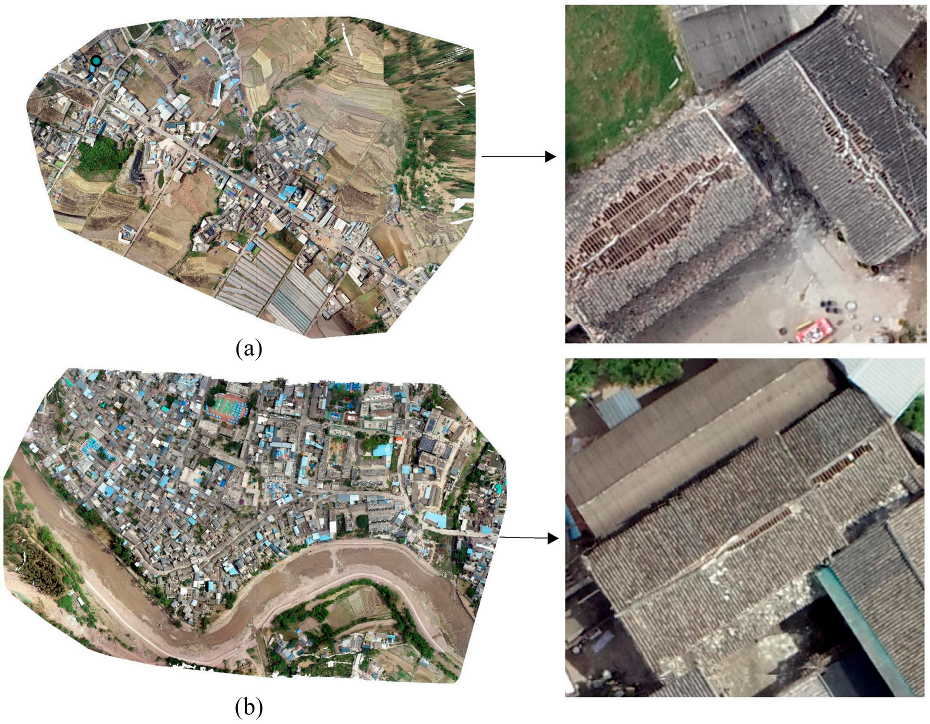

Figure 16. Example of UAV imagery from the Yangbi earthquake in Yunnan, China: (**a**) Huaian village; (**b**) Yangbi town.

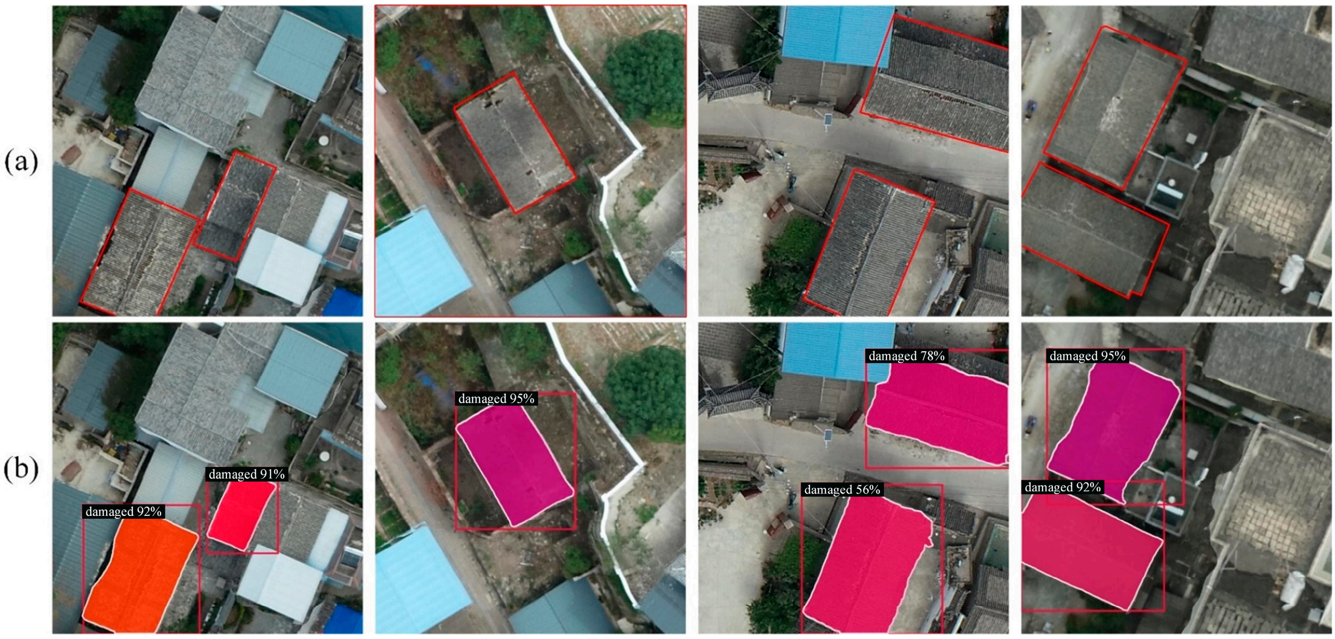

Figure 17. UAS imagery samples of damaged buildings from the Yangbi earthquake. (**a**) The red irregular polygons denote the damaged buildings. (**b**) The bounding boxes and polygon masks are the visualized results of our model. The colors represent different instance objects.

However, some limitations may be encountered when applying our method to different places, including similar spatial resolution, analogous backgrounds, and identical types of building structures. For example, in both the Luding earthquake and the Yangbi earthquake, the damaged buildings are mainly beam–column wooden structures.

5. Discussion

Accurately obtaining fine contour information of damaged buildings after earthquakes is a challenging problem. UAS can obtain high-resolution remote sensing images immediately after a disaster, which provides data support for earthquake damage assessment [3–5]. Existing deep learning-based methods [10–17] including object detection methods and semantic segmentation methods cannot simultaneously provide the accurate location of each target and the fine contours of damaged buildings, but this information is particularly important for accurate post-disaster loss assessment. We built a dataset for damaged buildings in UAS images. Based on this, we proposed DB-Transfiner, a high-quality instance segmentation network designed for post-earthquake damaged building detection in UAS-based optical remote sensing imagery.

We first addressed the challenge of irregular shapes in collapsed buildings, which standard convolutions struggle to capture effectively. We incorporated deformable convolutions in DCNM to enhance the ability to capture these deformation features, which increases the model's AP_{seg} and AP_{box} by 4.07% and 5.00%, respectively. In addition, we employed an enhanced BiFPN in MEFM for multi-scale feature fusion, which strengthens the representation of features for objects at different scales, resulting in a 4.13% and 6.10% improvement in the model's segmentation performance for AP_L and AP_M, respectively. Finally, in the segmentation mask head, we implemented LTGM with a lightweight Transformer encoder to handle pixels in incoherent regions defined by Mask Transfiner. This improves the efficiency of global feature extraction and object edge refinement, reducing the inference time by 4.9 ms per image.

In summary, as an instance segmentation method, the proposed network can not only accurately locate the location of damaged buildings but also obtain fine contours of damaged buildings. It is mentioned in the study by Jing et al. [11] that the proposed model cannot provide accurate footprints of damaged buildings, while our method does not have this limitation. Recently, Zou et al. [49] proposed an instance segmentation method for damaged building assessment. Although this method can better detect buildings with different damage levels, the outline of damaged buildings it identifies needs to be further improved. Our model can accurately predict the instance mask, and the segmented polygons are very consistent with the contours of the target building.

Despite the progress made, there remain several limitations and challenges that require further research. As shown in several cases in Figure 18, due to the varying perspectives of UAS images, occlusions and shadow effects caused by differences in object heights pose challenges for identifying the contours of damaged buildings in densely populated areas. Moreover, seismic damage assessment demands a balance between model accuracy and inference speed. Although the proposed model can process 0.9 more images per second compared to Mask Transfiner, it still incurs considerable computational cost when identifying damaged buildings across large areas, indicating a need to further optimize the processing speed. Notably, our method may encounter certain limitations when applied to different earthquakes, including similar damage levels, analogous backgrounds, and the same types of building structures. Our method performs well for obvious damage, but the model is less sensitive to some minor damage. This may be due to the roof being affected by accumulated debris. When the damaged building is other types of structures, such as brick–concrete or frame structures, the identification results of our method are not satisfactory. This is because the primary type of damaged buildings in the Luding earthquake were beam–column wooden structures [40], and we used these samples to train the model.

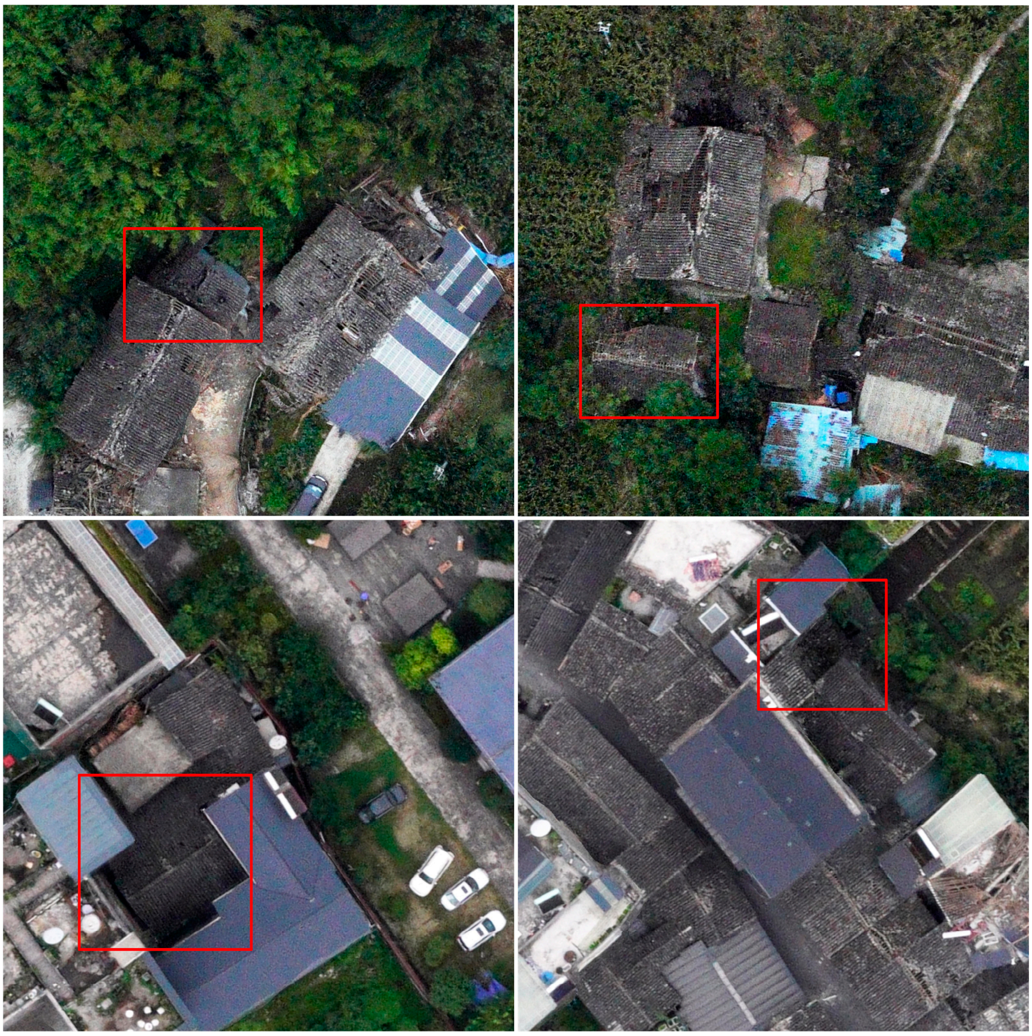

Figure 18. Examples of densely built-up areas. The red boxes indicate buildings with blurred contour information caused by shadows and occlusions.

In future work, we will explore several potential directions. First, we plan to employ rotated bounding boxes with angle information instead of horizontal bounding boxes [50] to reduce background interference and improve the model's performance in densely built areas. Additionally, we aim to simplify or replace some redundant components in the model to further advance research toward lightweight and efficient processing.

6. Conclusions

In this paper, we proposed a high-quality damaged building instance segmentation method, DB-Transfiner, to accommodate the problems of high but inconsistent spatial resolution, variations in different target scales, and complex post-earthquake scenes in aerial remote sensing imagery. To obtain better bounding boxes and fine segmentation masks of damaged buildings, we conducted DCNM, MEFM, and LTGM modules to study the Mask Transfiner network. DCNM can effectively capture the deformation characteristics of damaged buildings and enhance the network's ability to recognize the irregular shapes of completely collapsed buildings. The MEFM significantly improves the representation of features for objects of different sizes, thereby boosting the model's recognition performance. LTGM not only improves the model's global feature extraction and object edge refinement capabilities but also reduces the model's size and thus improves computational efficiency and inference speed. We implemented comprehensive experiments using UAS images from multiple regions affected by the Luding Ms6.8 earthquake in Sichuan Province.

The results demonstrated that this method can accurately obtain fine masks of damaged buildings and can identify damaged buildings with high quality. It shows high applicability and generalization capability in the identification and assessment of post-earthquake building damage.

Author Contributions: Conceptualization, S.W., K.Y. and Y.W.; methodology, K.Y. and S.W.; software, K.Y. and Y.W.; validation, K.Y. and S.W.; formal analysis, K.Y.; investigation, K.Y.; resources, S.W.; data curation, K.Y.; writing—original draft preparation, K.Y.; writing—review and editing, S.W., Z.G. and Y.W.; visualization, K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42271090; and Fundamental Research Funds of the Institute of Earthquake Forecasting, China Earthquake Administration, grant number CEAIEF20230202, CEAIEF20240508, CEAIEF2024030106 and CEAIEF20240405.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

References

1. Li, Q.; Mou, L.; Sun, Y.; Hua, Y.; Shi, Y.; Zhu, X.X. A Review of Building Extraction from Remote Sensing Imagery: Geometrical Structures and Semantic Attributes. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4702315. [CrossRef]
2. Valentijn, T.; Margutti, J.; van den Homberg, M.; Laaksonen, J. Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *Remote Sens.* **2020**, *12*, 2839. [CrossRef]
3. Nedjati, A.; Vizvari, B.; Izbirak, G. Post-earthquake response by small UAV helicopters. *Nat. Hazards* **2016**, *80*, 1669–1688. [CrossRef]
4. Xiong, C.; Li, Q.S.; Lu, X.Z. Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network. *Autom. Constr.* **2020**, *109*, 102994. [CrossRef]
5. Zhang, R.; Li, H.; Duan, K.F.; You, S.C.; Liu, K.; Wang, F.T.; Hu, Y. Automatic Detection of Earthquake-Damaged Buildings by Integrating UAV Oblique Photography and Infrared Thermal Imaging. *Remote Sens.* **2020**, *12*, 2621. [CrossRef]
6. Jhan, J.P.; Kerle, N.; Rau, J.Y. Integrating UAV and Ground Panoramic Images for Point Cloud Analysis of Damaged Building. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6500805. [CrossRef]
7. Xie, Y.; Feng, D.; Chen, H.; Liu, Z.; Mao, W.; Zhu, J.; Hu, Y.; Baik, S.W. Damaged Building Detection from Post-Earthquake Remote Sensing Imagery Considering Heterogeneity Characteristics. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4708417. [CrossRef]
8. Ge, J.; Tang, H.; Yang, N.; Hu, Y. Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 105–128. [CrossRef]
9. Wang, J.; Guo, H.; Su, X.; Zheng, L.; Yuan, Q. PCDASNet: Position-Constrained Differential Attention Siamese Network for Building Damage Assessment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622318. [CrossRef]
10. Tilon, S.; Nex, F.; Kerle, N.; Vosselman, G. Post-Disaster Building Damage Detection from Earth Observation Imagery Using Unsupervised and Transferable Anomaly Detecting Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 4193. [CrossRef]
11. Jing, Y.; Ren, Y.; Liu, Y.; Wang, D.; Yu, L. Automatic Extraction of Damaged Houses by Earthquake Based on Improved YOLOv5: A Case Study in Yangbi. *Remote Sens.* **2022**, *14*, 382. [CrossRef]
12. Pi, Y.; Nath, N.D.; Behzadan, A.H. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Adv. Eng. Inf.* **2020**, *43*, 101009. [CrossRef]
13. Wang, Y.; Feng, W.; Jiang, K.; Li, Q.; Lv, R.; Tu, J. Real-Time Damaged Building Region Detection Based on Improved YOLOv5s and Embedded System from UAV Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4205–4217. [CrossRef]
14. Hong, Z.; Zhong, H.; Pan, H.; Liu, J.; Zhou, R.; Zhang, Y.; Han, Y.; Wang, J.; Yang, S.; Zhong, C. Classification of Building Damage Using a Novel Convolutional Neural Network Based on Post-Disaster Aerial Images. *Sensors* **2022**, *22*, 5920. [CrossRef]
15. Zhang, T.; Zhang, X.; Zhu, P.; Tang, X.; Li, C.; Jiao, L.; Zhou, H. Semantic Attention and Scale Complementary Network for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Cybern.* **2022**, *52*, 10999–11013. [CrossRef]
16. Wang, Y.; Jing, X.; Cui, L.; Zhang, C.; Xu, Y.; Yuan, J.; Zhang, Q. Geometric consistency enhanced deep convolutional encoder-decoder for urban seismic damage assessment by UAV images. *Eng. Struct.* **2023**, *286*, 116132. [CrossRef]
17. Khankeshizadeh, E.; Mohammadzadeh, A.; Arefi, H.; Mohsenifar, A.; Pirasteh, S.; Fan, E.; Li, H.; Li, J. A Novel Weighted Ensemble Transferred U-Net Based Model (WETUM) for Postearthquake Building Damage Assessment from UAV Data: A Comparison of Deep Learning- and Machine Learning-Based Approaches. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701317. [CrossRef]

18. Li, X.; Yang, J.; Li, Z.; Yang, F.; Chen, Y.; Ren, J.; Duan, Y. Building Damage Detection for Extreme Earthquake Disaster Area Location from Post-Event UAV Images Using Improved SSD. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2674–2677.
19. Hussein, B.R.; Malik, O.A.; Ong, W.H.; Slik, J.W.F. Automated Extraction of Phenotypic Leaf Traits of Individual Intact Herbarium Leaves from Herbarium Specimen Images Using Deep Learning Based Semantic Segmentation. *Sensors* **2021**, *21*, 4549. [CrossRef]
20. Gu, W.; Bai, S.; Kong, L. A review on 2D instance segmentation based on deep neural networks. *Image Vision Comput.* **2022**, *120*, 104401. [CrossRef]
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
22. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
23. Xie, E.; Sun, P.; Song, X.; Wang, W.; Liu, X.; Liang, D.; Shen, C.; Luo, P. PolarMask: Single Shot Instance Segmentation with Polar Representation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12190–12199.
24. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-Time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October 2019; pp. 9156–9165.
25. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. SOLO: Segmenting Objects by Locations. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 649–665.
26. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229.
27. Dong, B.; Zeng, F.; Wang, T.; Zhang, X.; Wei, Y. SOLQ: Segmenting Objects by Learning Queries. In Proceedings of the Thirty-Fifth Annual Conference on Neural Information Processing Systems, New Orleans, LA, USA, 6–14 December 2021; pp. 4206–4217.
28. Fang, Y.; Yang, S.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; Liu, W. Instances as Queries. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6890–6899.
29. He, J.; Li, P.; Geng, Y.; Xie, X. FastInst: A Simple Query-Based Model for Real-Time Instance Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23663–23672.
30. Ke, L.; Danelljan, M.; Li, X.; Tai, Y.W.; Tang, C.K.; Yu, F. Mask Transfiner for High-Quality Instance Segmentation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4402–4411.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
32. Zou, Y.; Wang, X.; Wang, L.; Chen, K.; Ge, Y.; Zhao, L. A High-Quality Instance-Segmentation Network for Floating-Algae Detection Using RGB Images. *Remote Sens.* **2022**, *14*, 6247. [CrossRef]
33. Yang, S.; Zheng, L.; Wu, T.; Sun, S.; Zhang, M.; Li, M.; Wang, M. High-throughput soybean pods high-quality segmentation and seed-per-pod estimation for soybean plant breeding. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107580. [CrossRef]
34. Panboonyuen, T.; Nithisopa, N.; Pienroj, P.; Jirachuphun, L.; Watthanasirikrit, C.; Pornwiriyakul, N. MARS: Mask Attention Refinement with Sequential Quadtree Nodes for Car Damage Instance Segmentation. *arXiv* **2023**, arXiv:2305.04743.
35. Topics on Lu County "9•16" Rescue Attack. Available online: https://www.luxian.gov.cn/zwgk/fdzdgknr/zdmsxx/ylws/content_303681 (accessed on 17 May 2024). (In Chinese)
36. Gao, X.L.; Ji, J. Analysis of the seismic vulnerability and the structural characteristics of houses in Chinese rural areas. *Nat. Hazard* **2014**, *70*, 1099–1114. [CrossRef]
37. People First, Life First—The Seventh Diary of Sichuan Province's Response to the "Ninth Five-Year" Luding Earthquake. Available online: https://www.sc.gov.cn/10462/10464/10797/2022/9/12/5973fd88141145ea9f49477bb4f92c9d.shtml (accessed on 12 May 2024). (In Chinese)
38. Earthquake Experts: "9•5" Luding Earthquake Damage Has Five Characteristics. Available online: https://www.sc.gov.cn/10462/10778/10876/2022/9/14/1f2655ddc5394b1a989f22d1393560e8.shtml (accessed on 14 May 2024). (In Chinese)
39. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision-ECCV 2014, Zurich, Switzerland, 5–12 September 2014; pp. 740–755.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
42. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
43. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets V2: More Deformable, Better Results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9300–9308.

44. Liu, S.; Qi, L.; Qin, H.F.; Shi, J.P.; Jia, J.Y. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
45. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE.* **1998**, *86*, 2278–2324. [CrossRef]
46. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787.
47. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
48. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
49. Zou, R.; Liu, J.; Pan, H.; Tang, D.; Zhou, R. An Improved Instance Segmentation Method for Fast Assessment of Damaged Buildings Based on Post-Earthquake UAV Images. *Sensors* **2024**, *24*, 4371. [CrossRef]
50. Shi, P.; Zhao, Z.; Fan, X.; Yan, X.; Yan, W.; Xin, Y. Remote Sensing Image Object Detection Based on Angle Classification. *IEEE Access* **2021**, *9*, 118696–118707. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.