

Improved Swin Transformer-Based Semantic Segmentation of Postearthquake Dense Buildings in Urban Areas Using Remote Sensing Images

Liangyi Cui, Xin Jing, Yu Wang, Yixuan Huan, Yang Xu , and Qiangqiang Zhang 

Abstract—Timely acquiring the earthquake-induced damage of buildings is crucial for emergency assessment and post-disaster rescue. Optical remote sensing is a typical method for obtaining seismic data due to its wide coverage and fast response speed. Convolutional neural networks (CNNs) are widely applied for remote sensing image recognition. However, insufficient extraction and expression ability of global correlations between local image patches limit the performance of dense building segmentation. This paper proposes an improved Swin Transformer to segment dense urban buildings from remote sensing images with complex backgrounds. The original Swin Transformer is used as a backbone of the encoder, and a convolutional block attention module is employed in the linear embedding and patch merging stages to focus on significant features. Hierarchical feature maps are then fused to strengthen the feature extraction process and fed into the UPerNet (as the decoder) to obtain the final segmentation map. Collapsed and non-collapsed buildings are labeled from remote sensing images of the Yushu and Beichuan earthquakes. Data augmentations of horizontal and vertical flipping, brightness adjustment, uniform fogging, and non-uniform fogging are performed to simulate actual situations. The effectiveness and superiority of the proposed method over the original Swin Transformer and several mature CNN-based segmentation models are validated by ablation experiments and comparative studies. The results show that the mean intersection-over-union of the improved Swin Transformer reaches 88.53%, achieving an improvement of 1.3% compared to the original model. The stability, robustness, and generalization ability of dense building recognition under complex weather disturbances are also validated.

Index Terms—Attention mechanism, complex weather disturbances, dense seismic building segmentation, feature fusion, improved Swin Transformer, remote sensing images.

I. INTRODUCTION

EARTHQUAKES are one of the most severe natural disasters, and due to the recent acceleration of urbanization development, earthquake-induced building damage has become one of the most severe threats to human beings [1]. Therefore, after an earthquake occurs, it is crucial to recognize the number, location, and damage level of urban buildings rapidly to ensure postearthquake rescue and reconstruction [2]. The seismic damage-related data have been mainly collected via field investigation, which is labor-time-intensive and inefficient. In addition, particular circumstances, such as power facility destruction and communication system interruption caused by earthquakes, can bring additional challenges to conducting immediate field investigation. Therefore, an efficient and effective method that can meet the practical requirements of postearthquake rapid assessment and emergency rescue is urgently needed.

In recent years, with the development of satellite systems, remote sensing techniques have become increasingly popular in the field of natural disaster assessment [3]. The commonly-used remote sensing data [4], [5], can be roughly divided into three categories: synthetic aperture radar images [6], [7]; optical images [8]; and light detection and ranging data [9]. Among them, high-resolution optical images—which are easy to obtain and can provide rich information on postearthquake building attributes, such as color, texture, and shape—have been the most widely used [10]. Remote sensing images are wide-ranging, all-weather, unaffected by earthquakes, and accessible without onsite human inspection. In early-stage research, remote sensing image interpretation primarily relied on preset thresholds and handcrafted parameters and thus was highly affected by a subjective judgment in various application scenarios. In addition, the recognition speed and reliability highly depended on engineering experience and prior knowledge of image analysts. However, automatic extraction and autonomous recognition of seismic damage from remote sensing images have rapidly developed with advanced computer vision techniques, including image processing, machine learning, and deep learning.

Manuscript received 26 October 2022; revised 20 November 2022; accepted 24 November 2022. Date of publication 28 November 2022; date of current version 15 December 2022. This work was supported in part by the National Key Research and Development Program under Grant 2019YFC1511005, in part by the China Postdoctoral Science Foundation under Grant BX20190102 and Grant 2019M661286, and in part by the Heilongjiang Province Postdoctoral Funding under Grant LBH-TZ2016 and Grant LBH-Z19064 and in part by Heilongjiang Province Natural Science Funding under Grant LH2022E070. (Liangyi Cui and Xin Jing contributed equally to this work.) (Corresponding authors: Qiangqiang Zhang; Yang Xu.)

Liangyi Cui, Xin Jing, Yu Wang, Yixuan Huan, and Qiangqiang Zhang are with the School of Civil Engineering and Mechanics, Lanzhou University, Lanzhou 730000, China, and also with the Key Laboratory of Mechanics on Disaster and Environment in Western China, The Ministry of Education of China, Beijing 100816, China (e-mail: cuiy20@lzu.edu.cn; jingx21@lzu.edu.cn; wangyu16@lzu.edu.cn; huanyx21@lzu.edu.cn; zhangqq@lzu.edu.cn).

Yang Xu is with the School of Civil Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: xyce@hit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3225150

Compared to image processing and machine learning, deep learning has relatively better learning ability and stronger robustness against interference and variations in object size, position, shape, and geometry and thus can provide more accurate localization and damage information on dense seismic buildings [11]. Convolutional neural networks (CNNs) have been the most widely-used deep learning-based model for seismic damage data extraction from high-resolution remote sensing optical images. Currently, CNNs are widely applied to seismic damage identification from postearthquake remote sensing images. Cooner et al. [12] adopted CNNs to classify high-resolution seismic remote sensing imagery and quickly detect damaged buildings, achieving an accuracy of 55% for the 2010 Haiti earthquake with the 7.0 magnitude. Ma et al. [13] combined remote sensing images with block vector data and improved the Inception V3 architecture; a test accuracy of 90.07% on postearthquake aerial imagery of Yushu was achieved. Furthermore, Ji et al. [14] used the pretrained VGG model to recognize collapsed buildings in remote sensing images before and after the 2010 Haiti earthquake, concluding that the fine-tuned VGGNet model outperformed the original VGGNet model trained from scratch with an overall accuracy increasing from 83.38% to 85.19%. Xiao et al. [15] proposed a dynamic cross-fusion network to enable each task to share features from different CNN layers adaptively and achieved state-of-the-art performance. Zhan et al. [16] used the Mask R-CNN to extract information on damaged buildings from postearthquake remote sensing images and identify the damage level. An improved feature pyramid network (FPN) was designed, and a detection accuracy of 92% was achieved for the most severely damaged buildings (the overall classification accuracy for four damage classes was 88%).

However, conventional CNNs can focus only on a small range of pixel-level features, thereby providing insufficient information on global correlations between local pixels and lack the capacity to model global relationships between objects within an image and nonlocal relationships between pixels. In addition, the limited receptive field could not provide sufficient contextual features, which might have a significant impact on the damage assessment accuracy of dense seismic buildings [17]. Transformer-based models using global self-attentive mechanisms can compensate for the abovementioned shortcomings of conventional CNNs that focus only on local receptive fields without considering global features [18], [19], [20], [21], allowing each pixel to contain global correlations and thus improving generalization ability and interference robustness [22], [23], [24], [25], [26].

Dosovitskiy et al. [27] first present the vision transformer (ViT) models and utilized the transformer as the backbone network for image classification tasks. The ViT models tokenized the input image into fixed-size patches, which were then flattened as vectors and fed to the transformer backbone. Experimental results demonstrated that the ViT models pretrained on large-scale datasets could achieve better performance than the CNNs when migrated to the classification tasks on small-size and medium-size datasets. In recent years, several transformer-based vision models have been proposed for different computer vision tasks, such as target classification [28],

object detection [29], and semantic segmentation [30], [31], [32]. Despite the successful application of the transformer in the natural language processing field, there are two main challenges in its application to the visual domain from the original language domain. These challenges are introduced by significant differences in visual entity size among images and much higher resolutions of images compared to texts, which leads to an intensive computational cost.

To solve the above problem, Swin Transformer [33] is proposed with two principle improvements over conventional ViTs.

- 1) A hierarchical structure similar to the CNN structure is designed. This structure is very flexible in multiscale modeling and reduces the increase in computational complexity with the image size from square to linear.
- 2) The shifted window multihead self-attention (SW-MSA) block is proposed to reduce the computational cost while considering the information transferred between different windows.

Although the transformer-based models made a splash in computer vision, they have still been in the infancy phase for large-scale seismic disaster evaluation in urban areas. Da et al. [34] developed a two-stage damage assessment framework named the SDAFormer, which feeds pre-disaster and postdisaster images to the network separately for damage assessment. The SDAFormer won first place on the xBD (a large-scale building damage assessment dataset) and achieved a mean intersection-over-union (mIoU) improvement of 1.5% compared to the second-place method. Chen et al. [35] proposed a transformer-based damage assessment architecture consisting of a Siamese transformer encoder and a lightweight dual-tasks decoder, which outperformed traditional CNN models such as the Mask R-CNN and Siamese-UNet.

Although the CNN models have been extensively investigated for computer vision tasks, the feature extraction process of conventional CNN is always performed at a local region, and modelling the global correlation is challenging. Considering the characteristics of the investigated remote sensing images for postearthquake buildings in a city area, the buildings are densely distributed, and the structure style and damage type are often similar, which suggests that the small-region features are closely related and the global correlations should be significant for the recognition accuracy. Therefore, this article designs an integrated model using the improved Swin Transformer for global correlation modeling and CNN for local feature extraction to further enhance the recognition capacity of building damage states and location semantics, respectively.

Meanwhile, statistical analyses of previous studies have demonstrated that clouds approximately cover 70% of the Earth, which suggests that weather interferences of cloud or fog obscuration and illumination variances inevitably exist in remote sensing optical images [36]. In addition, postearthquake remote sensing images can suffer from light overexposure and darkness due to various illumination conditions. Therefore, accurate recognition of dense seismic buildings in images collected under strong weather disturbances represents a great challenge in semantic segmentation. However, research on semantic segmentation of postearthquake remote sensing images of dense urban

buildings with complex backgrounds and strong interferences is rather limited.

To address the abovementioned limitations, this article proposes a semantic segmentation method for seismic damage of large-scale dense buildings in large-scale urban areas with complex backgrounds and strong weather interferences. In addition, the opportunity of incorporating the transformer and CNN for seismic damage recognition from remote sensing images is analyzed.

The main contributions of this article can be summarized as follows.

- 1) An effective semantic segmentation method is proposed for high-resolution remote sensing optical images of dense buildings with complex backgrounds and strong weather interferences; this method can accurately and simultaneously extract the building damage state and location semantics.
- 2) An improved Swin Transformer with the encoder-decoder structure is proposed to simultaneously exploit multilevel local features and global correlations, which performs the multilevel feature fusion at each stage of the encoder, inserts convolutional block attention module (CBAM) in the linear embedding and patch merging modules, and uses the UPerNet as a decoder.
- 3) Two actual seismic scenarios of Yushu city and Beichuan city with different weather disturbances are used to simulate possible light overexposure, darkness, and fog occlusions and validate the effectiveness of the proposed method.
- 4) Ablation experiments are performed to demonstrate the efficacy and necessity of the proposed modules in the improved Swin Transformer. In addition, comparative studies are conducted to verify the superiority of the improved Swin Transformer over the original Swin Transformer and various mature CNN-based segmentation models.

The rest of the article is organized as follows. Section II describes the architecture of the improved Swin Transformer. Section III introduces the dataset and implementation details. Section IV presents the test results under two real-world seismic scenarios, ablation experiments, and comparative studies. Section V concludes the article.

II. PROPOSED METHOD

A. Overall Architecture

An improved Swin Transformer based on the encoder-decoder framework is proposed to realize accurate semantic segmentation of postearthquake dense buildings from remote sensing images with complex backgrounds and strong weather interferences. The overall architecture that uses the original Swin Transformer as a backbone of the encoder is presented in Fig. 1. As shown in Fig. 1, a feature fusion module is added to the end of the encoder to fully exploit the extracted features at various levels. In the proposed structure, hierarchical feature maps are concatenated using convolutions to enrich the transferable local features of different stages by multilevel feature fusion. In addition, the CBAM is inserted into the linear embedding and patch merging modules to alleviate feature leakage during the patch

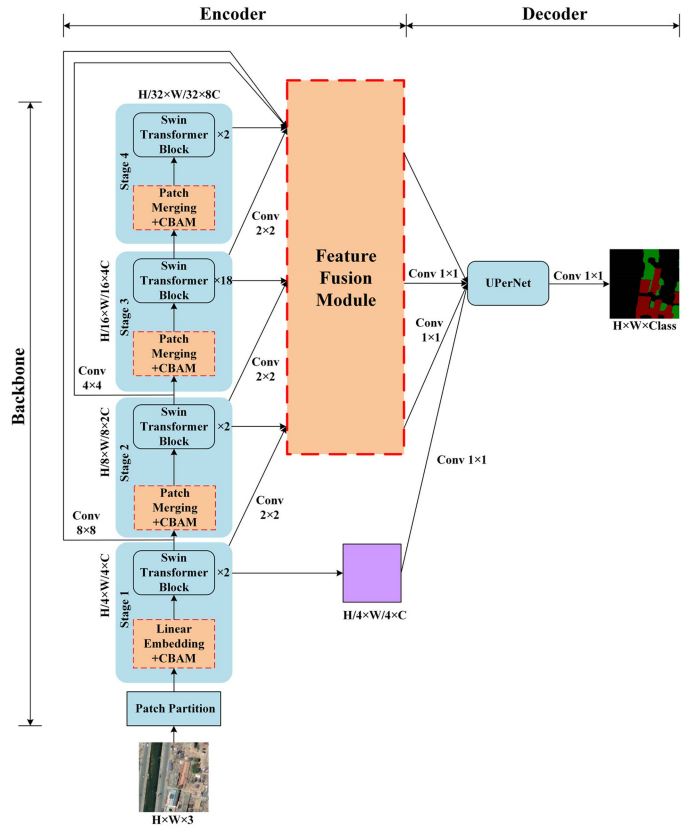


Fig. 1. Overall architecture of improved Swin Transformer for dense building segmentation.

downsampling process in the encoding stage. This enables the proposed model to distinguish different building damage states and location semantics, thus improving multiclass segmentation accuracy. Finally, the UperNet incorporating multilevel features is used as a decoder. Details on the feature fusion and CBAM modules are described in the following sections.

B. Swin Transformer Backbone

The Swin Transformer backbone includes an initial patch partition module and four different stages denoted by stages 1–4. Stage 1 consists of a linear embedding layer and two consecutive Swin Transformer blocks. Stage 2 consists of a patch merging module and two Swin Transformer blocks. Stage 3 consists of a patch merging module and 18 Swin Transformer blocks. Finally, Stage 4 consists of a patch merging module and two Swin Transformer blocks.

For the patch partition module, the input image with a size of $H \times W \times 3$ is split four times in the spatial directions and flattened in the channel direction, generating a patch size of $H/4 \times W/4 \times 48$. Then, the linear embedding layer projects the channel dimension to an arbitrary number denoted by C (in this article, $C = 128$) through the 1×1 convolution, generating a feature map with a size of $H/4 \times W/4 \times C$. The feature map of each stage is input into the patch merging module, and a half-patch-size downsampling process is performed by neighborhood sampling every two points, and thus the channel number quadruples. Then, a 1×1 convolution is utilized to adjust the

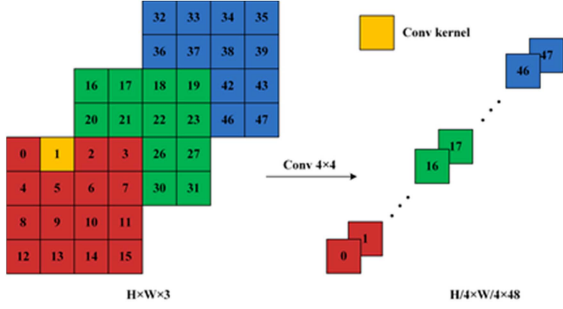


Fig. 2. Schematic of the patch partition module.

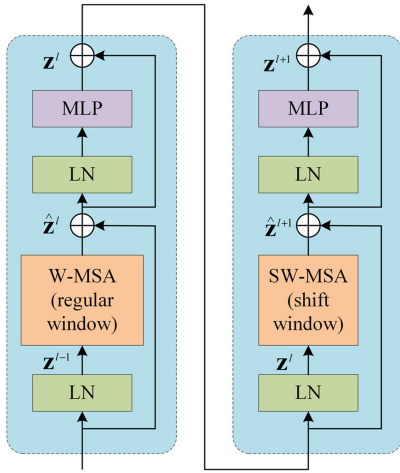


Fig. 3. Schematic of the fundamental component for Swin Transformer block.

channel number to double. The overall schematic of the patch partition module is presented in Fig. 2.

The schematic diagram of the Swin Transformer block, which is the fundamental component of the Swin Transformer, is presented in Fig. 3. Each Swin Transformer block includes a regular window and a shift window. The regular window consists of a layer-normalization (LN) layer, a window multihead self-attention (W-MSA) module, a residual connection, an LN layer, a multilayer perceptron (MLP), and a residual connection. The shift window has a similar structure as the regular window; the only difference is that an SW-MSA module is used instead of the W-MSA. The mathematical formula of the Swin Transformer block is expressed as follows:

$$\begin{aligned}
 \hat{Z}^l &= W - MSA[LN(Z^{l-1})] + Z^{l-1} \\
 Z^l &= MLP[LN(\hat{Z}^l)] + \hat{Z}^l \\
 \hat{Z}^{l+1} &= SW - MSA[E(Z^l)] + Z^l \\
 Z^{l+1} &= MLP[LN(\hat{Z}^{l+1})] + \hat{Z}^{l+1}
 \end{aligned} \quad (1)$$

where Z^{l-1} and Z^{l+1} denote the input and output of the Swin Transformer block, respectively. A detailed description of W-MSA, MLP, and SW-MSA can be found in the study of Han et al. [32].

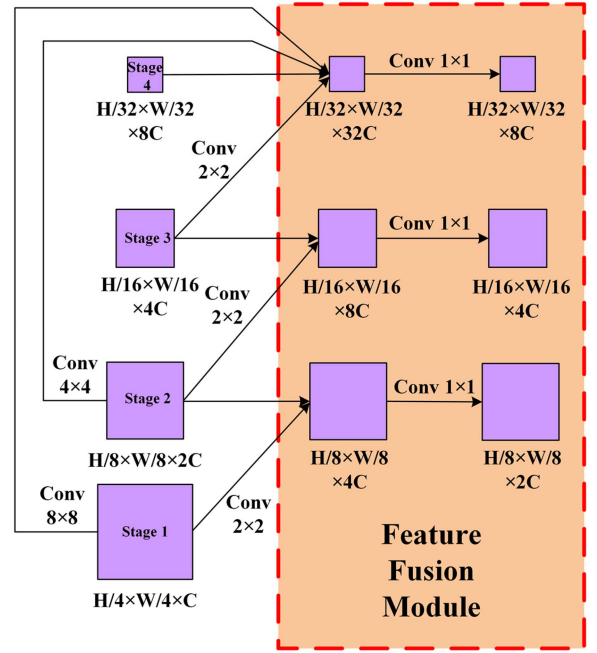


Fig. 4. Flowchart of the feature fusion module.

C. Feature Fusion Module

Compared with the traditional semantic segmentation task, the dataset investigated in this article consists of remote sensing images with complex backgrounds, and its unique characteristics are reflected in two aspects: images contain complex backgrounds, including several types of strong distractions, such as illumination variations and fog obscurations; and buildings in remote sensing images are in different geometries; particularly, shapes and sizes of collapsed and not-collapsed buildings are different.

A previous study has shown that using different convolution operators in the transformer architecture can provide information on both local and global features of the input image, significantly improving the semantic segmentation performance [37]. Inspired by this idea, a multilevel feature fusion module is designed after each Swin Transformer block to convolute the feature maps output by the previous levels to further enhance the extraction capability of local features and global correlations. Although the Swin Transformer has a hierarchical structure, there are no interactions between feature maps at any stages. Therefore, enriching the extracted features is essential considering that remote sensing images of postearthquake dense buildings contain various types of background distractions, including illumination overexposure, darkness, uniform fog, and non-uniform fog, and have a high diversity of geometric shapes and sizes.

The schematic diagram of the feature fusion module is shown in Fig. 4, where four feature maps from the corresponding stage of the Swin Transformer backbone are illustrated. The flat dimension of each stage is halved, and the channel dimension is doubled. The feature map of each stage is downsampled by a 2×2 convolutional kernel with a sliding stride of two and concatenated with that of the next stage in the channel direction.

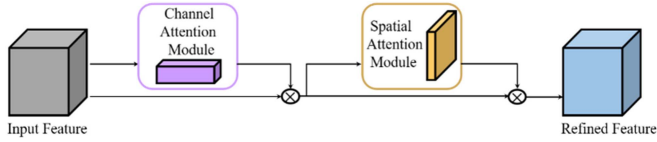


Fig. 5. Schematic of CBAM attention module.

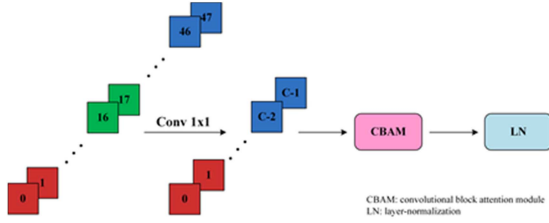


Fig. 6. Flowchart of inserting CBAM in linear embedding module.

Then, the channel number of the concatenated feature map is half reduced by the 1×1 convolution. Finally, feature maps from all stages are fused in the channel direction, and the channel size is quartered using a 1×1 convolutional kernel.

D. Convolutional Block Attention Module

The attention mechanism is a typical way to achieve adaptive attention inside a neural network, and the commonly-used attention mechanisms include channel attention and spatial attention. The channel attention aims to enable the network to focus on the category information inside an image by keeping the channel dimension unchanged and compressing the spatial dimension into a scalar. Furthermore, the spatial attention assists the network in paying more attention to the location information of targets inside an image by keeping the spatial dimension unchanged and compressing the multiple-channel dimension into one single channel. This article utilizes the CBAM by simultaneously combining channel attention and spatial attention and can distinguish significant feature maps of building damage states and location semantics. The schematic diagram of the CBAM, which is a lightweight attention mechanism module consisting of a channel attention part and a spatial attention part by Woo et al. [38], is presented in Fig. 5. Details of CBAM have been presented in [38] and omitted here.

The process of inserting the CBAM into the linear embedding module is illustrated in Fig. 6. The dimension of the feature map generated by the patch partition module is transformed to C by a 1×1 convolution block, and the CBAM module is inserted before the LN layer.

Conventional downsampling operations often use convolution, average pooling, and maximum pooling in a local region, which will inevitably cause feature leakage. Patch merging selects the neighborhood of every two pixels, reassembles them into a series of patches (the spatial size of patches is halved), and concatenates the patches in the channel dimension (the channel dimension is quadrupled), which is finally followed by a 1×1 convolution to adjust the channel dimension. Therefore, all the input information can be reserved, and no feature leakage occurs in patch merging. The process of inserting the CBAM into the

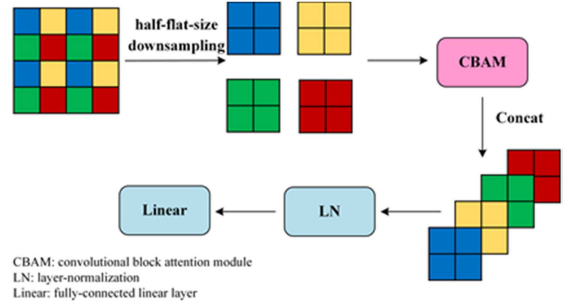


Fig. 7. Flowchart of inserting CBAM in patch merging module.

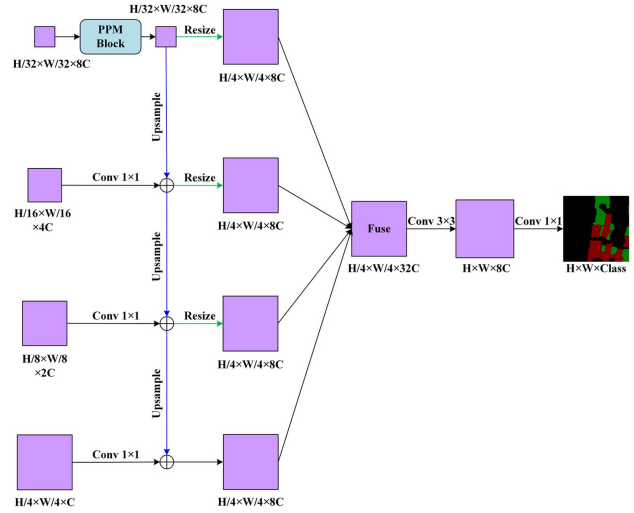


Fig. 8. UPerNet decoder architecture.

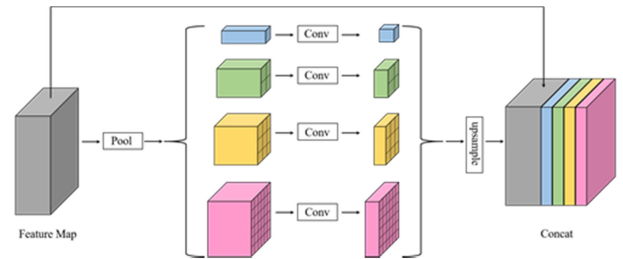


Fig. 9. Schematic of the PPM module in the decoder.

patch merging module is presented in Fig. 7. In each channel, neighborhood areas of every two points are reassembled into a patch (i.e., the flat size is halved). The reconstructed patches are fed into the CBAM module individually, and the output feature maps of the CBAM module are fused in the channel direction. The CBAM module is followed by an LN layer and a fully-connected linear layer.

E. UPerNet Decoder

For remote sensing images with complex backgrounds and small dense buildings, a multilevel segmentation predictor, the UPerNet [39], is employed to achieve full-scale coverage from low-level concrete features to high-level abstract features. The design of the UPerNet is based on the pyramid pooling module

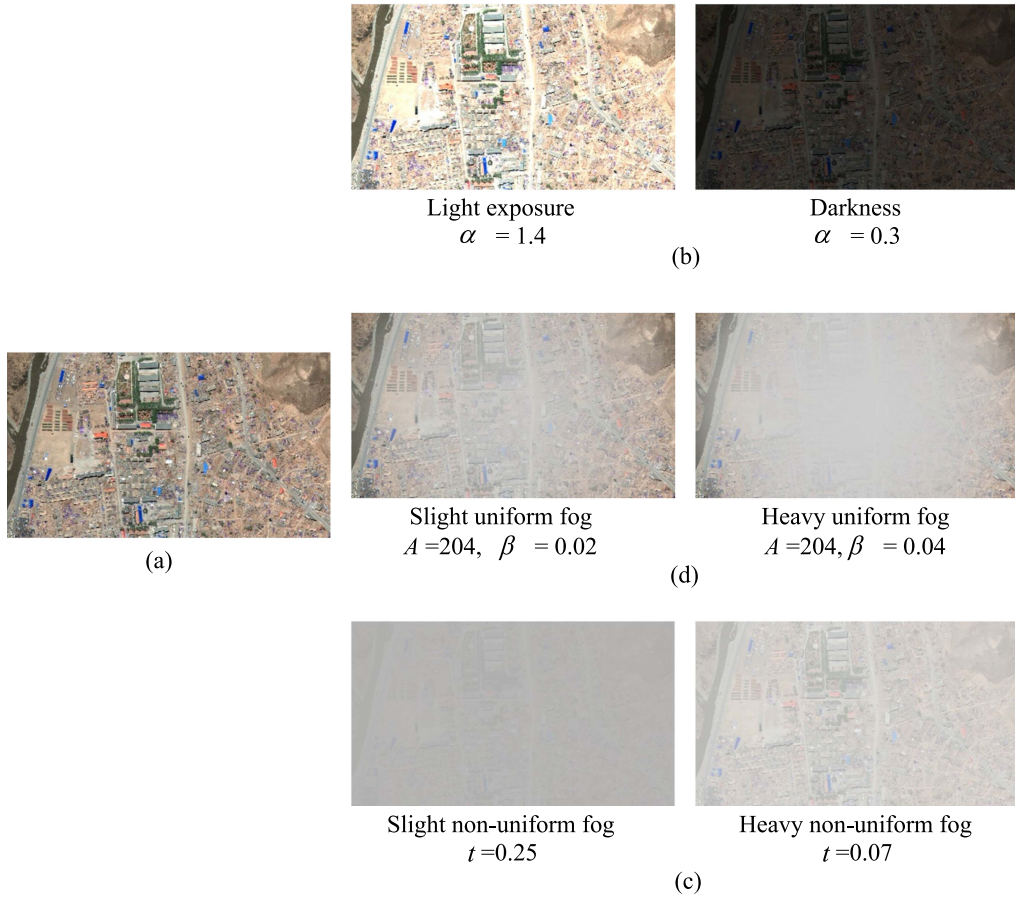


Fig. 10. Representative postearthquake remote sensing images with data augmentation. (a) Original image of Yushu city. (b) Brightness transformations of light overexposure and darkness. (c) Uniform fogging in light and heavy degrees. (d) Nonuniform fogging in light and heavy degrees.

(PPM) [40] and FPN, which fully integrates extracted features from different stages of the encoder. The architecture of the UPerNet decoder is shown in Fig. 8.

The PPM block utilizes pooling kernels covering different portions of the input feature map to generate multiscale correlations among different subregions. In this article, a four-level pyramid pooling is designed to individually perform the pooling operation for the whole, half of, a third of, and a sixth of the input feature map. Then, the channel dimensions are adjusted using 1×1 convolution, and the spatial dimensions are unified by bilinear interpolation upsampling. Finally, they are fused as the global prior and concatenated with the original feature map at the channel dimension, as shown in Fig. 9.

III. DATASET AND IMPLEMENTATION DETAILS

A. Dataset

In this article, 24 remote sensing city-scale images of the Yushu city and Beichuan city after Yushu and Wenchuan earthquakes with a resolution of 4608×2560 were used. The original images were downloaded from the Internet and manually pixel-wise labeled using “labelme” [41] to classify buildings into collapsed and non-collapsed buildings. Buildings with destructive shapes, severely-damaged roofs, columns, and beams were classified as collapsed, and other buildings were labeled as non-collapsed.

Data augmentation operations, including random flipping in the horizontal and vertical directions, brightness transformation, uniform fogging, and nonuniform fogging, were performed to expand the dataset and simulate possible light overexposure and darkness and fog occlusions in remote sensing images.

The brightness transformation is realized by rescaling the pixel intensity as follows:

$$\hat{I}(h, w) = \text{median} [0, \alpha \times I(h, w), 255] \quad (2)$$

where $I(h, w)$ and $\hat{I}(h, w)$ denote the image intensity at the pixel location (h, w) before and after brightness transformation, respectively; α denotes the rescaling coefficient controlling the light exposure and darkness; median operator ensures the transformed pixel intensity within the range of 0-255.

Based on the dark channel prior theory [42], dark pixels have very low intensity in at least one color channel of the RGB for most local regions that do not cover the sky; therefore, the non-uniform fogging operation is expressed by

$$\begin{aligned} \hat{J}(h, w) &= t(h, w)J(h, w) + [1 - t(h, w)] \times A \\ t(h, w) &= \exp[-\beta \times d(h, w)] \\ d(h, w) &= -\gamma \times \sqrt{\left(h - \frac{H}{2}\right)^2 + \left(w - \frac{W}{2}\right)^2} + W \end{aligned} \quad (3)$$

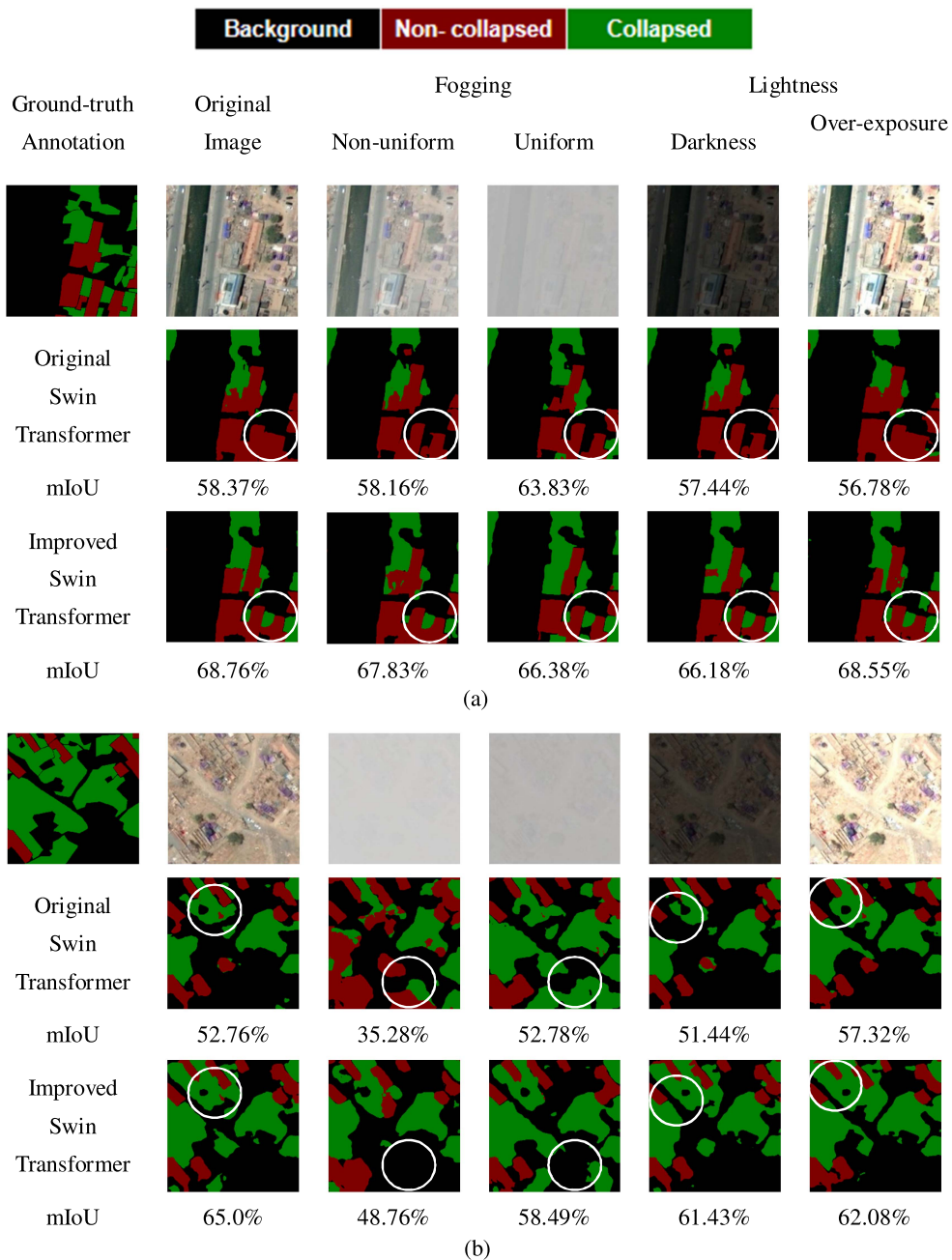


Fig. 11. Test results of 512×512 patches for Yushu city. (a) Patch 1. (b) Patch 2.

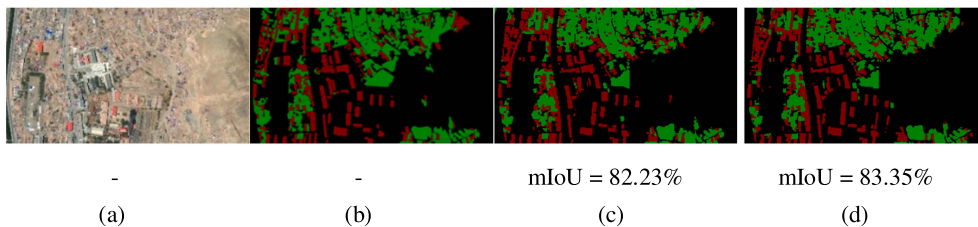


Fig. 12. Test results of large-scale remote sensing image for Yushu city. (a) Input Image. (b) Ground-truth Annotation. (c) Original Swin Transformer. (d) Improved Swin Transformer.

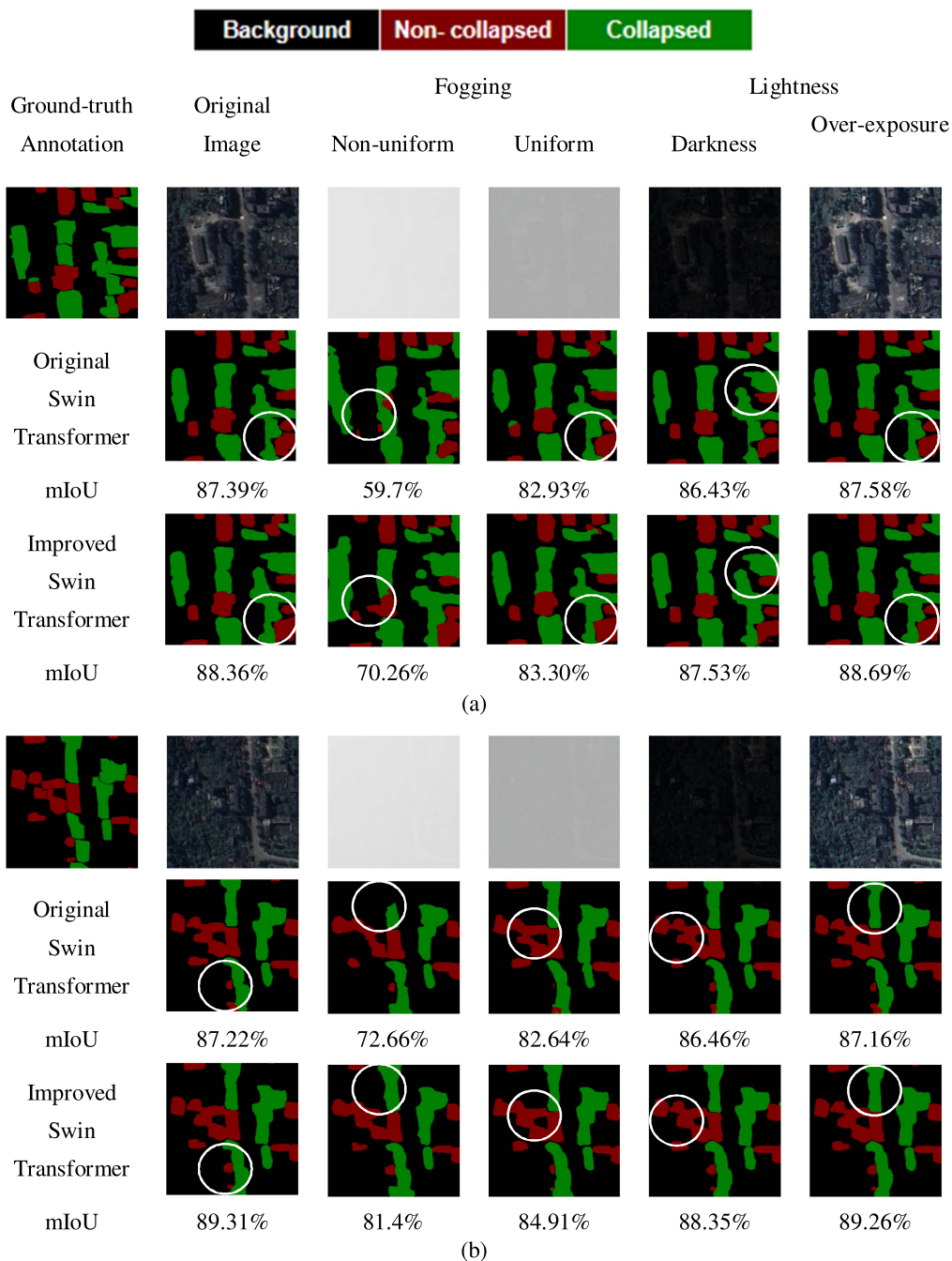


Fig. 13. Test results of 512×512 patches for Beichuan city. (a) Patch 1. (b) Patch 2.

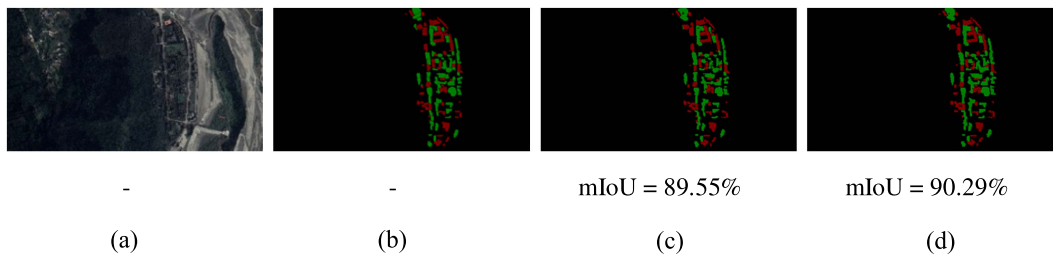


Fig. 14. Test results of large-scale remote sensing image for Beichuan city. (a) Input Image. (b) Ground-truth Annotation. (c) Original Swin Transformer. (d) Improved Swin Transformer.

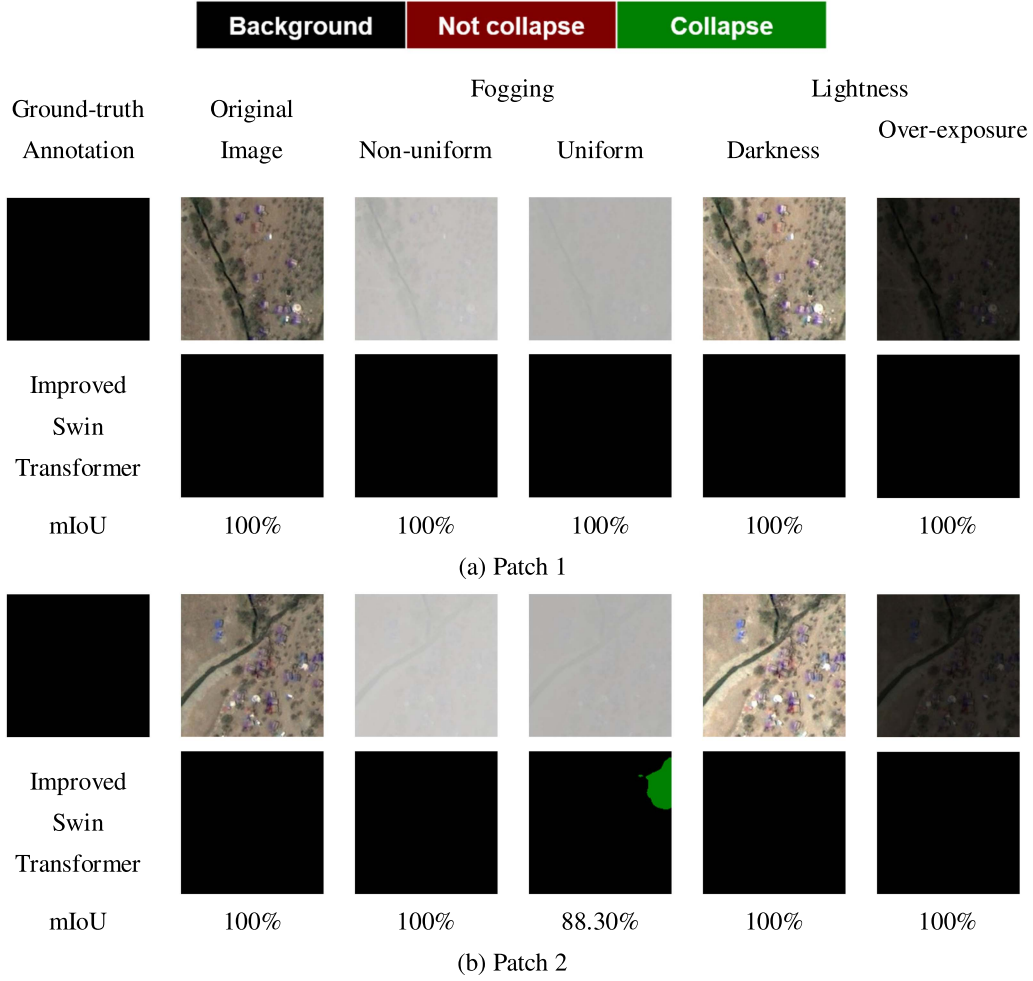


Fig. 15. Test results of negative objects for wild regions with trees, tents, and rivers. (a) Patch 1. (b) Patch 2.

where $J(h, w)$ and $\hat{J}(h, w)$ denote the image intensity before and after fogging transformation; h and w are the pixel indexes in the height and width directions, respectively; A denotes the fog brightness parameter, and its value is in the range of 0-255 corresponding to the grayscale intensity of fog changes from black to white; $t(h, w)$ represents the light transmittance; β denotes the fogging concentration factor; γ denotes the constant influence factor, and in this article $\gamma = 0.04$; $d(h, w)$ denotes the scene depth. H and W denote the height and width of the input image.

Considering that remote sensing images could be completely covered by a large area of clouds or fog, the uniform fogging operation is used to simulate possible scenarios and enhance the dataset as

$$\hat{J}(h, w) = tJ(h, w) + (1 - t) \times A. \quad (4)$$

Equation (4) is a particular case of (3) with a constant light transmittance at all pixel locations, where $\hat{J}(h, w)$ denotes the image intensity after uniform fogging transformation, $J(h, w)$ denotes the original image and $L(h, w)$ denotes a new image with the identical pixel value of 170 on three channels of RGB.

Fig. 10 shows some representative postearthquake remote sensing images with dense buildings after brightness, uniform,

and non-uniform fogging transformations with different configurations. After data augmentation, the original images were cropped to 512×512 patches with an overlap ratio of 50%. Finally, 8262 patches were obtained, 80% of which were used for training by random assignment, and the rest was used for validation.

B. Implementation Settings

The proposed method was implemented in PyTorch 1.7.0 on a workstation equipped with an i9-10900k CPU and a GeForce RTX 3090 GPU. The AdamW optimization algorithm was employed to update the model parameters under a learning rate of 0.0001, a batch size of 8, and a training epoch of 50. The mIoU between the predicted and ground-truth buildings was used as an evaluation metric of the proposed method and used the weights obtained from pre-trained on the ADE20K [43] dataset as pre-training weights for the model.

IV. RESULTS AND DISCUSSION

A. Test Results of Yushu City

Remote sensing seismic images of Yushu city, including various weather disturbances, were used to demonstrate the

recognition accuracy of the proposed method for postearthquake dense buildings. The test results obtained by the original and improved Swin Transformers on 512×512 patches of Yushu city collected under different weather disturbances are presented in Fig. 11. The results show that the proposed improved Swin Transformer achieved higher accuracy and better robustness against light overexposure, darkness, and fog occlusions than the original Swin Transformer with an average mIoU improvement of 0.83% for 512×512 patches. In Fig. 11, white circles in sub-figures present local details of predicted building corners and edges, indicating that the improved Swin Transformer could maintain better recognition ability under various weather disturbances than the original Swin Transformer. In addition, the improved Swin Transformer achieved better recognition on the fogging test images where the buildings were already difficult to distinguish, and the mIoU value improved by 1.18% compared to the original Swin Transformer. The test results on the large-scale image with a resolution of 4608×2560 are presented in Fig. 12, which shows that the improved Swin Transformer performed better than the original Swin Transformer.

B. Test Results of Beichuan City

Remote sensing seismic images of Beichuan city, which included various weather disturbances, were used to demonstrate the recognition accuracy of postearthquake dense buildings further. The test results obtained by the original Swin Transformer and improved Swin Transformer on the 512×512 patches of Beichuan city are presented in Fig. 13. The results in Fig. 13 show that the improved Swin Transformer achieved higher accuracy and better robustness against light overexposure, darkness, and fog occlusions than the original Swin Transformer with an average mIoU improvement of 1.05% for 512×512 patches. In addition, the improved Swin Transformer still achieved better recognition on the fogging test images, and the mIoU value improved by 1.77% compared to the original Swin Transformer. Additional test results on the 512×512 patches are given in Fig. 20. The test results of the two transformers on the large-scale image with a resolution of 4608×2560 are presented in Fig. 14, which shows that the improved Swin Transformer performed better than the original Swin Transformer.

C. Discussion of Test Results

For all test images of Yushu city and Beichuan city, the original Swin Transformer had more local misrecognition and larger prediction errors for building edges than the improved Swin Transformer, which resulted in the distinct shape variance of dense building regions. The original Swin Transformer tended to ignore unconnected pixels inside the building region and classified them into the same class. Moreover, the improved Swin Transformer achieved higher recognition accuracy than the original Swin Transformer for collapsed buildings with more irregular geometrical shapes. The recognition results of negative objects for wild regions with trees, tents, and rivers are shown in Fig. 15. The results show that negative objects are successfully classified into the background, and misrecognition rarely occurs.

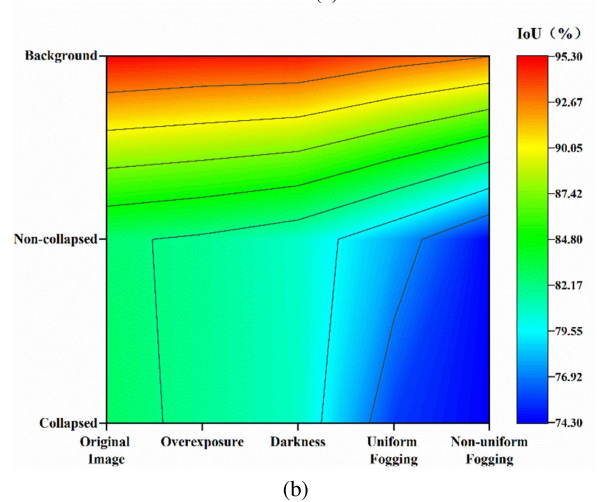
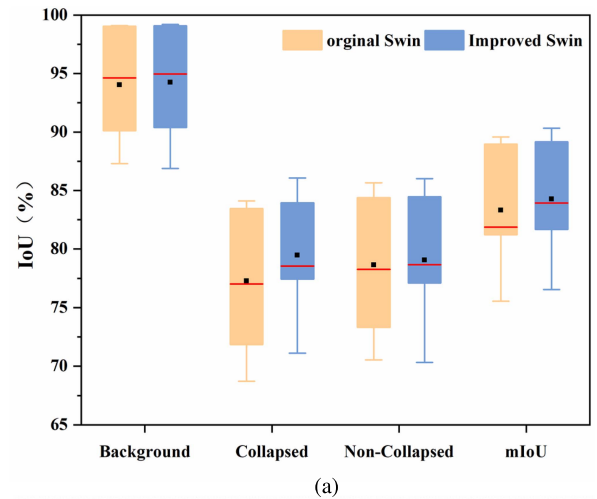


Fig. 16. Comparisons of category-wise IoU with different models and weather disturbances. (a) IoU comparisons for each category using Original and the improved Swin Transformer. (b) IoU contour plot for each category under different weather disturbances use the improved Swin Transformer.

It further indicates that the proposed model possesses good stability against complex environmental disturbances.

Under weather disturbances of fogging and brightness transformation, the misrecognition of the background of collapsed buildings and incomplete recognition of non-collapsed buildings often occurred. A possible reason may be that the fogging and brightness transformation introduced severe occlusion in certain areas, thus increasing the difficulty of accurate segmentation.

The comparison results of category-wise intersection-over-union (IoU) of the two models for different weather disturbances are presented in Fig. 16. As shown in Fig. 16(a), the proposed Swin Transformer improved the average segmentation IoU for each category with a lower volatility than the original Swin Transformer, suggesting the robustness and stability of the proposed method. Fig. 16(b) shows that the model performance decreased for each category when weather disturbances existed. Among the considered types of weather disturbances, the non-uniform fogging affected the model performance of the

TABLE I
COMPARISONS OF MODEL PERFORMANCE IN ABLATION EXPERIMENTS

Method	Back-ground IoU	Non-collapsed Building IoU	Collapsed Building IoU	mIoU
Original Swin Transformer	94.97%	84.5%	82.21%	87.23%
Transformer Swin +CBAM	95.10%	85.8%	83.56%	87.95%
Transformer Swin + Feature Fusion	95.29%	85.64%	84.15%	88.36%
Improved Swin Transformer (Full Model)	95.35%	86.07%	84.16%	88.53%

improved Swin Transformer the most, and the proposed model was less sensitive to brightness transformation than fogging occlusion.

It should be noted that the remote sensing images of Yushu city and Beichuan city had unique characteristics. In Yushu city, buildings were more densely distributed; intensities in the color space were similar to the background and plenty of tents and vehicles existed in the images, which increased difficulty in recognition. Although these factors could cause a slight decrease in average IoU, the improved Swin Transformer still achieved good recognition accuracy for each category. The results also indicated that the proposed method efficiently addressed the deficiencies of the original Swin Transformer and enhanced the edge smoothness and completeness of the results of geometrical shapes for postearthquake dense buildings. Therefore, the improved Swin Transformer had stronger robustness and resistance to different types of severe interferences under real-world scenarios than the original Swin Transformer.

D. Ablation Experiments and Comparative Studies

Ablation experiments were performed to demonstrate the effectiveness and necessity of the feature fusion and CBAM modules in the improved Swin Transformer. Besides the proposed model (including both the feature fusion module and the CBAM module), three additional models, namely the original Swin Transformer, the Swin Transformer + feature fusion module, and the original Swin Transformer + CBAM module, were trained using the same dataset, optimization algorithm, and training hyperparameters. Table I gives the comparison results of model performances in the ablation experiments. The results showed that both the feature fusion and the CBAM had certain contributions to the model performance improvement, but the effect of the feature fusion module was more significant. Accordingly, the feature fusion and CBAM modules improved

TABLE II
COMPARISONS OF MODEL PERFORMANCE WITH CNN-BASED METHODS

Method	Back-ground IoU	Non-collapsed Building IoU	Collapsed Building IoU	mIoU
PSPNet	91.73%	70.73%	73.65%	78.7%
DeepLab-V3+	93.16%	78.17%	77.8%	83.04%
UNet	93.86%	80.85%	79.05%	84.59%
Improved Swin Transformer (Proposed)	95.35%	86.07%	84.16%	88.53%

TABLE III
COMPARISONS OF DIFFERENT FEATURE FUSION MODULES

Module	Background IoU	Non-collapsed Building IoU	Collapsed Building IoU	mIoU
Feature Fusion-1	94.69%	84.62%	82.10%	87.13%
Feature Fusion-2	95.16%	85.11%	83.28%	87.85%
Proposed Feature Fusion	95.29%	85.64%	84.15%	88.36%

the segmentation accuracy of background, collapsed buildings, and noncollapsed buildings.

The full model achieved the highest improvements in segmentation IoU of background, collapsed buildings, and non-collapsed buildings by 0.38%, 1.57%, and 1.95%, respectively. The overall mIoU improvement was 1.3%, demonstrating that the improved Swin Transformer successfully integrated the advantages of feature fusion and CBAM modules. It further indicated the effectiveness of multilevel feature fusion in alleviating feature leakage and CBAM in focusing on small dense objects.

To verify the effectiveness of the improved Swin Transformer over conventional CNNs, several mature CNN-based semantic segmentation models, including the PSPNet [43], DeepLabV3+ [44], and UNet [45], were used for comparison. The dataset, optimization algorithm, and training hyperparameters were the same as those of the improved Swin Transformer. Table II gives a comparison of the performances of the improved Swin Transformer and several CNN-based models. The results showed that the UNet performed the best among the three CNN-based segmentation models but worse than the proposed Swin Transformer. Although the background IoU, noncollapsed IoU, collapsed IoU, and mIoU of the UNet reached 93.86%, 80.85%, 79.05%, and 84.59%, the improved Swin Transformer performed better in terms of all metrics by 2.49%, 5.22%, 5.11%, and 3.94%,

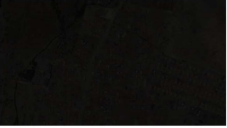
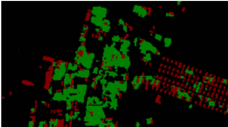
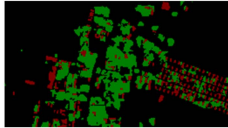

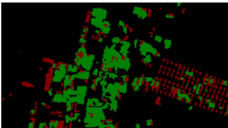
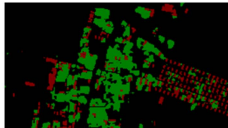

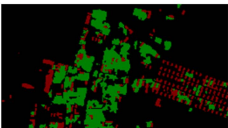
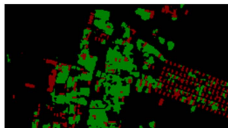

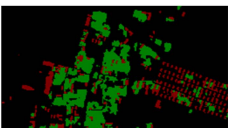
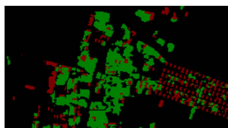

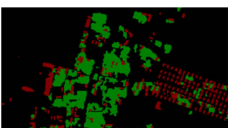
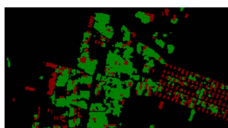

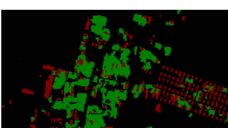
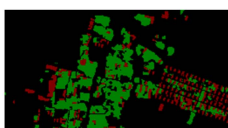

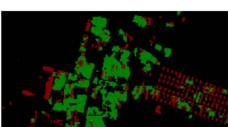
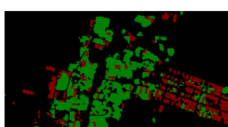
Test images under various lightness severities	lightness controlling parameter α	Background	Not collapse	Collapse	mIoU
		Ground-truth label	Prediction		
	0.1				78.40%
	0.4				82.41%
	0.7				83.36%
	1.0				83.95%
	1.3				83.07%
	1.6				78.52%
	1.9				67.29%

Fig. 17. Test results under different fogging conditions for a representative image.

respectively. This indicated that the proposed method integrating the Swin Transformer and CNN together enhanced the semantic segmentation accuracy of dense buildings in postearthquake remote sensing images compared to conventional CNN-based models.

The feature fusion module is designed to alleviate the possible feature leakage and enhance the multistage feature extraction. Even if some features at a particular stage are ignored, the feature fusion module can ensure that the information on missed features is retained and can be fed into the subsequent decoder. The authors admit that it is indeed challenging to determine which

feature stage is essential and should be enhanced in the feature fusion module. Therefore, the feature fusion model is designed in a two-step manner: the adjacent stages are fused to alleviate the feature leakage at the previous stage; and all the stages are fused at the final stage to take full advantage of the multistage features.

In addition, two comparative studies are performed to demonstrate the effectiveness of the proposed feature fusion module. First, the feature fusion module is only adopted at the final stage and ignored for the adjacent stages in the encoder, noted as feature fusion-1 in Table III.

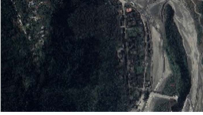
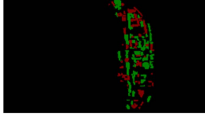
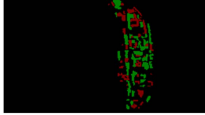

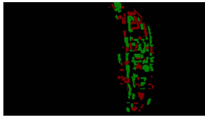
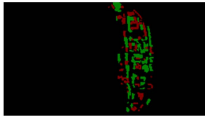

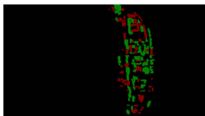
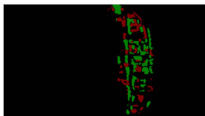

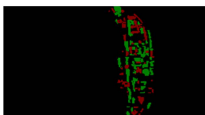
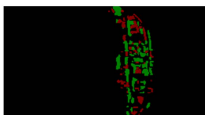

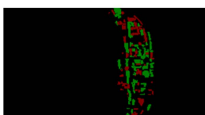
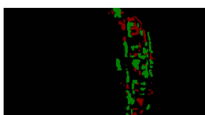
	Background	Not collapse	Collapse	
Test images under various Fogging severities	Fogging controlling parameter β	Ground-truth label	Prediction	mIoU
	0			90.29%
	0.02			89.58%
	0.03			87.48%
	0.04			84.50%
	0.05			72.17%

Fig. 18. Test results under different lightness conditions for a representative image.

Second, the feature fusion module is adopted both in the encoder and decoder, noted as feature fusion-2 in Table III. The encoder part is the same as Fig. 4; for the decoder part, feature maps of the first and second stages are downsampled by 2×2 convolution and concatenated in the channel dimension with those of the next stage. Afterward, the number of channels is halved by 1×1 convolution, and the residuals are finally added together. Table III gives the comparison results of these three different feature fusion modules, indicating that both insufficient (feature fusion-1) and excessive (feature fusion-2) feature fusion modules have negative impacts on recognition accuracy.

To explore the applicable range of controlling parameters under each weather condition, more experiments are performed, as shown in Figs. 17 and 18. Fig. 17 shows representative test results under various lightness conditions. It suggests that the controlling parameter α could be recommended in the range of 0.4–1.3 with a high mIoU over 0.8. When α is set as 1.9, a significant drop of about 19.85% in the prediction accuracy occurs. Fig. 18 shows representative test results under various fogging conditions. It suggests that the controlling parameter β could be recommended in the range of 0–0.04 with a high mIoU over 0.8. When β is set as 0.05, a significant drop of about 20% in the prediction accuracy occurs.

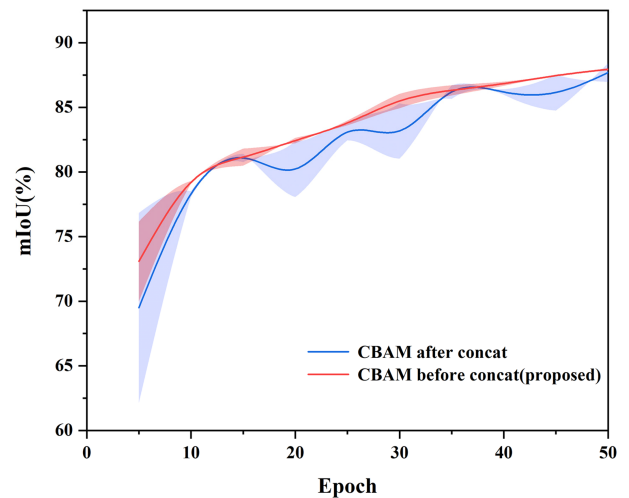


Fig. 19. Comparison of segmentation mIoU for different insertion positions of CBAM in original Swin Transformer backbone (using the average results of three independent experiments).

The mIoU increasing curves of the background, noncollapsed, and collapsed buildings for different CBAM insertion locations in the original Swin Transformer backbone are presented in

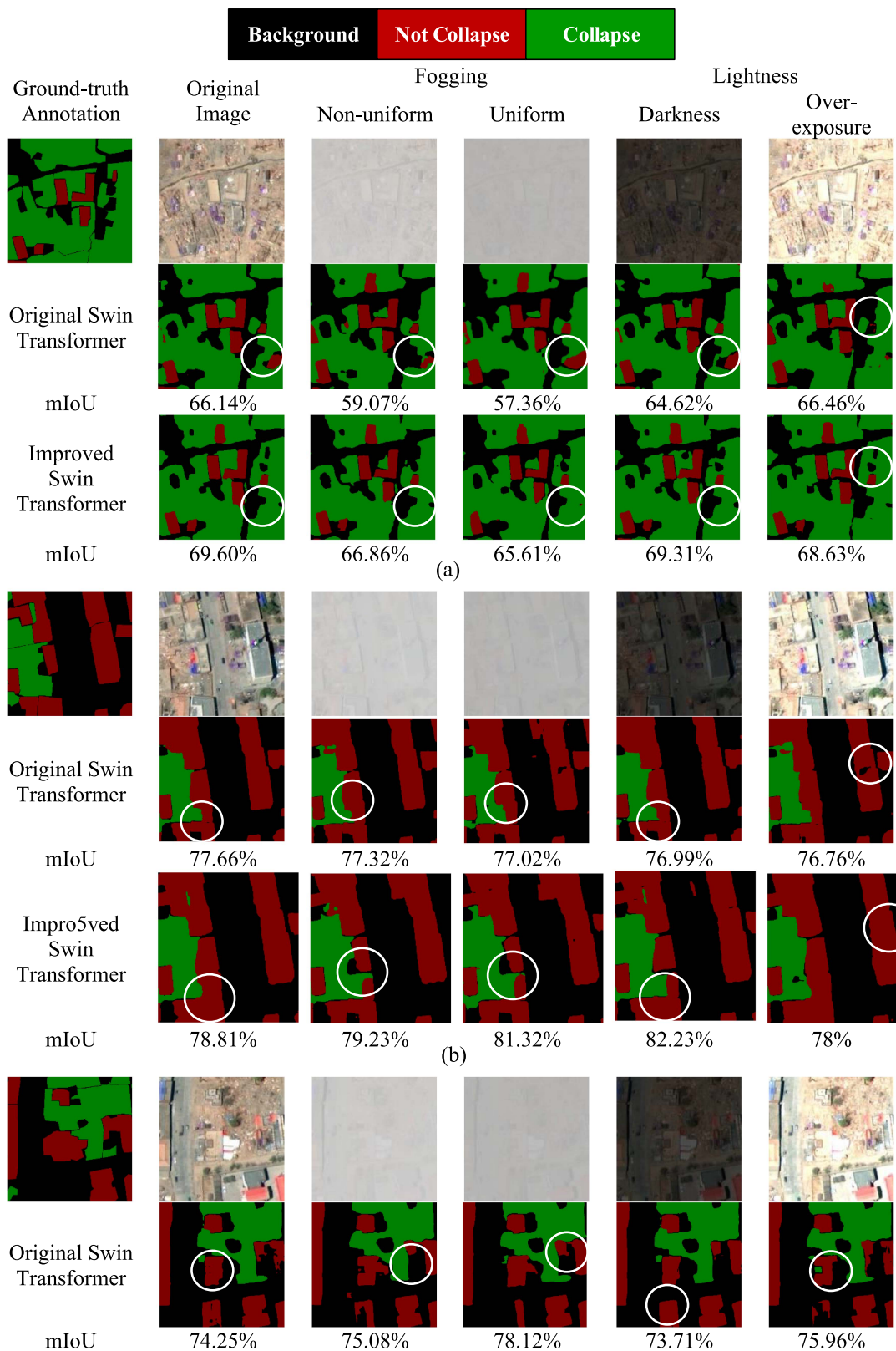


Fig. 20. Additional test results for 512×512 patches in Beichuan city and Yushu city (white circles compare local details of predicted building corners and edges improved by the proposed method). (a) Patch 1. (b) Patch 2. (c) Patch 3. (d) Patch 4. (e) Patch 5.

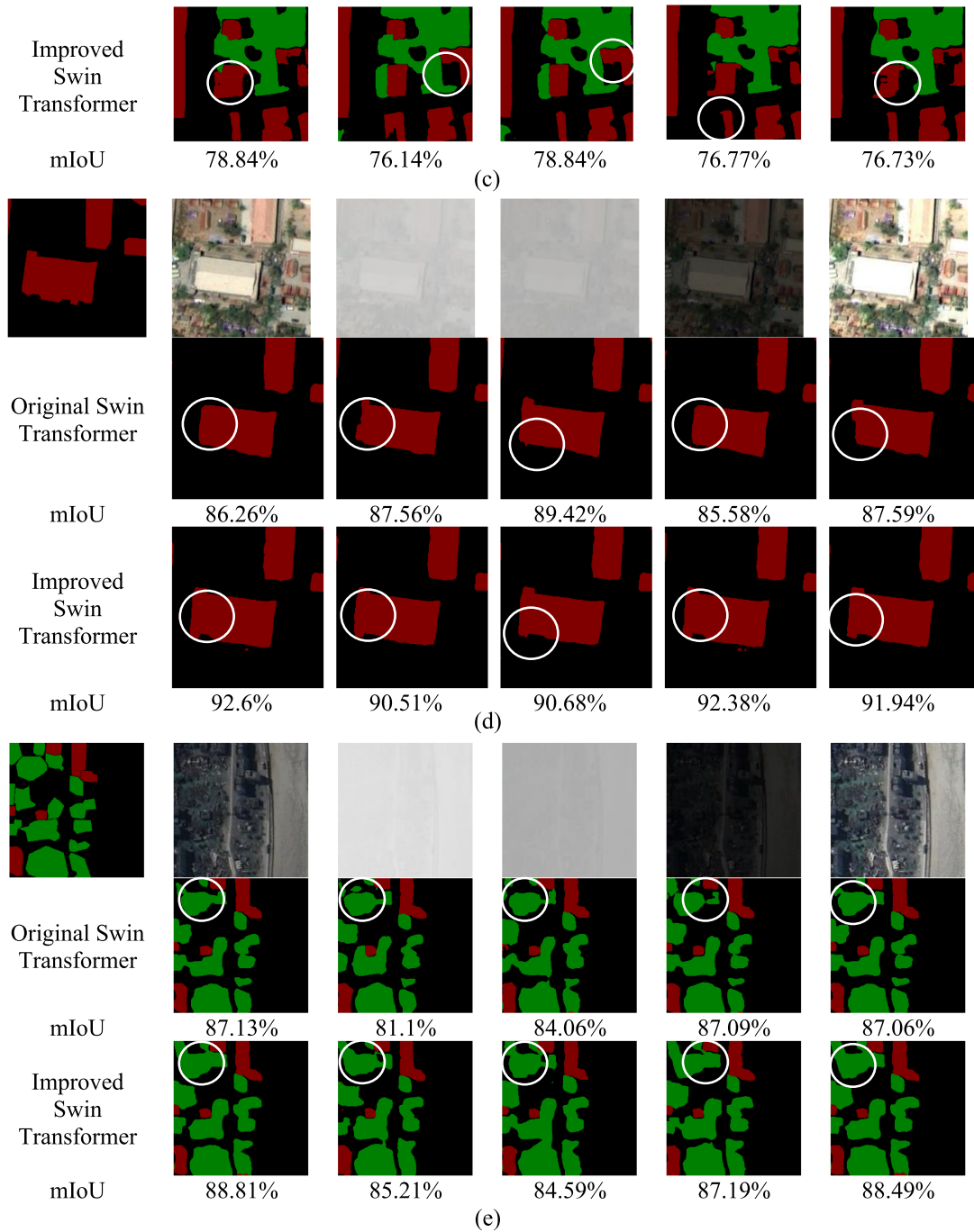


Fig. 20. (Continued).

Fig. 19, where “CBAM before concat” represents that all feature maps were first input into the CBAM module and then concatenated in the proposed patch merging block; “CBAM after concat” represents that all the related feature maps were concatenated before being input into the CBAM module in the Patch Merging block. The results indicated that the insertion strategy of CBAM before concatenation gained the higher training accuracy and lower diversity than inserting CBAM after concatenation.

V. CONCLUSION

This article proposed an improved Swin Transformer for remote sensing segmentation of postearthquake dense buildings in urban areas. The main contributions of this article are obtained as follows.

- 1) An improved Swin Transformer following the encoder-decoder framework was proposed to achieve accurate semantic segmentation of postearthquake dense buildings from remote sensing images under complex backgrounds

and strong weather interferences. The proposed structure performed multilevel feature fusion at each stage of the encoder, inserted the CBAM into the linear embedding and patch merging modules based on the original Swin Transformer backbone, and used the UPerNet as a decoder.

- 2) A total of 24 high-resolution remote sensing city-scale images were used to train and validate the proposed model. Different weather disturbances were considered by performing brightness transformation, uniform fogging, and nonuniform fogging to expand the dataset and simulate possible light overexposure, darkness, and fog occlusions under actual situations. The results showed that the improved Swin Transformer achieved higher recognition accuracy than the original Swin Transformer, especially for collapsed buildings with highly irregular geometrical shapes.
- 3) Ablation experiments were performed to demonstrate the effectiveness and necessity of the proposed modules in the improved Swin Transformer. The comparison results showed that the full model (i.e., the proposed model with feature fusion and CBAM) obtained the best segmentation IoU result of background of collapsed and noncollapsed buildings among all models, which further indicated the advantages of the multilevel feature fusion in alleviating feature leakage and the CBAM in focusing on small dense objects.
- 4) The comparison results showed that the improved Swin Transformer had distinct superiority over the original Swin Transformer and some mature CNN-based segmentation models, including the PSPNet, DeepLabV3+, and UNet. It indicated that the proposed method could enhance the semantic segmentation accuracy of dense buildings in postearthquake remote sensing images owing to the comprehensive extraction capability of local features and global correlations by organically integrating transformer and CNN structures.

In future work, the multiscale recognition of seismic disasters is supposed to be investigated using multisource data based on ViTs.

ACKNOWLEDGMENT

The authors highly appreciate Prof. Hui Li from Harbin Institute of Technology for the insightful comments and suggestions.

REFERENCES

- [1] Y. Wang, L. Y. Cui, C. Z. Zhang, W. L. Chen, Y. Xu, and Q. Q. Zhang, "A two-stage seismic damage assessment method for small, dense, and imbalanced buildings in remote sensing images," *Remote Sens.*, vol. 14, no. 4, Feb. 2022, Art. no. 1012.
- [2] C. Schweier and M. Markus, "Classification of collapsed buildings for fast damage and loss assessment," *Bull. Earthq. Eng.*, vol. 4, no. 2, pp. 177–192, Apr. 2006.
- [3] S. Voigt, T. Kemper, T. Riedlinger, R. Kiefl, K. Scholte, and H. Mehl, "Satellite image analysis for disaster and crisis-management support," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1520–1528, Jun. 2007.
- [4] L. G. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS J. Photogramm.*, vol. 84, pp. 85–99, Oct. 2013.
- [5] B. Adriano, J. Xia, G. Baier, N. Yokoya, and S. Koshimura, "Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 Sulawesi earthquake and tsunami in Palu, Indonesia," *Remote Sens.*, vol. 11, no. 7, pp. 886, Apr. 2019.
- [6] Y. S. Zhou, S. Zhang, X. K. Sun, F. Ma, and F. Zhang, "SAR target incremental recognition based on hybrid loss function and class-Bias correction," *Appl. Sci.*, vol. 12, no. 3, Jan. 2022, Art. no. 1279.
- [7] L. L. Li, X. G. Liu, Q. H. Chen, and S. Yang, "Building damage assessment from PolSAR data using texture parameters of statistical model," *Comput. Geosci.*, vol. 113, pp. 115–126, Apr. 2018.
- [8] X. Wang and P. J. Li, "Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 322–336, Jan. 2020.
- [9] K. Saito, R. J. S. Spence, C. Going, and M. Markus, "Using high-resolution satellite images for post-earthquake building damage assessment: A study following the 26 January 2001 Gujarat earthquake," *Earthq. Spectra*, vol. 20, no. 1, pp. 145–169, Feb. 2004.
- [10] P. Gamba and F. Casciati, "GIS and image understanding for near-real-time earthquake damage assessment," *Photogramm. Eng. Remote Sens.*, vol. 64, pp. 987–994, Oct. 1998.
- [11] W. J. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder–decoder network for automatic building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.
- [12] A. J. Cooner, Y. Shao, and J. B. Campbell, "Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake," *Remote Sens.*, vol. 8, no. 10, Oct. 2016.
- [13] H. J. Ma, Y. L. Liu, Y. H. Ren, D. C. Wang, L. J. Yu, and J. X. Yu, "Improved CNN classification method for groups of buildings damaged by earthquake, based on high resolution remote sensing images," *Remote Sens.*, vol. 12, no. 2, Jan. 2020.
- [14] M. Ji, L. F. Liu, R. C. Zhang, and M. F. Buchroithner, "Discrimination of earthquake-induced building destruction from space using a pretrained CNN model," *Appl. Sci.*, vol. 10, pp. 85–99, Jan. 2020.
- [15] H. Xiao, Y. Peng, H. Tan, and P. Li, "Dynamic cross fusion network for building-based damage assessment," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.
- [16] Y. H. Zhan, W. Liu, and Y. Maruyama, "Damaged building extraction using modified Mask R-CNN model using post-event aerial images of the 2016 Kumamoto earthquake," *Remote Sens.*, vol. 14, no. 4, Feb. 2022, Art. no. 1002.
- [17] W. J. Luo, Y. J. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *NIPS*, vol. 29, Dec. 2016, pp. 4905–4913.
- [18] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.
- [19] K. Zhang et al., "Practical blind denoising via swin-conv-UNET and data synthesis," 2022, *arXiv:2203.13278v2*.
- [20] C. Y. Si, W. H. Yu, P. Zhou, Y. C. Zhou, X. C. Wang, and S. C. Yan, "Inception transformer," 2022, *arXiv:2205.12956v2*.
- [21] R. L. Shao, Z. X. Shi, J. F. Yi, P. Y. Chen, and C. Hsieh, "On the adversarial robustness of vision transformers," 2021, *arXiv:2103.15670v3*.
- [22] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10231–10241.
- [23] S. Paul and P. Y. Chen, "Vision transformers are robust learners," *Proc. AAAI*, vol. 36, no. 2, pp. 2071–2081, 2022.
- [24] J. Y. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2021, pp. 12175–12185.
- [25] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," Apr. 2022, *arXiv:2009.06732*.
- [26] M. Raghu, T. Unterthiner, S. Kornblith, C. Y. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Mach. Learn.*, vol. 34, pp. 12116–12128, Aug. 2021.
- [27] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929v2*.
- [28] S. X. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2021, pp. 6881–6890.

- [29] T. Y. Lin, Y. X. Wang, X. Y. Liu, and X. P. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, Jun. 2022.
- [30] Y. Liu et al., "A survey of visual transformers," 2021, *arXiv:2111.06091v3*.
- [31] M. H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 1–38, Mar. 2022.
- [32] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: [10.1109/TPAMI.2022.3152247](https://doi.org/10.1109/TPAMI.2022.3152247).
- [33] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
- [34] Y. F. Da, Z. Y. Ji, and Y. S. Zhou, "Building damage assessment based on siamese hierarchical transformer framework," *Mathematics*, vol. 10, no. 11, Jun. 2022, Art. no. 1898.
- [35] H. Chen, E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley, "Dual-tasks siamese transformer framework for building damage assessment," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1600–1603.
- [36] I.-H. Lee and M. T. Mahmood, "Robust registration of cloudy satellite images using two-step segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1121–1125, May 2015.
- [37] M. Heidari et al., "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," 2022, *arXiv:2207.08518*.
- [38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [39] T. T. Xiao, Y. Liu, B. L. Zhou, Y. N. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434.
- [40] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, and J. Y. Jia, "Pyramid scene parsing network," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2017, pp. 2881–2890.
- [41] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 157–173, May 2008.
- [42] K. He, J. Sun, and X. O. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.
- [43] B. L. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2017, pp. 5122–5130.
- [44] L.-C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
- [45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *LNIP*, vol. 9351, pp. 234–241, Jan. 2015.



Liangyi Cui received the B.Sc. degree in civil engineering from Jilin University, Changchun, China, in 2020. He is currently working toward the M.Sc. degree with Lanzhou University of Civil Engineering, Lanzhou, China.

His research interests include computer vision in remote sensing.



Xin Jing received the B.Sc. degree in civil engineering from Lanzhou University, Lanzhou, China, in 2021. He is currently working toward the M.Sc. degree with Lanzhou University of Civil Engineering, Lanzhou, China.

His research interests include machine learning in seismic damage assessment.



Yu Wang received the B.Sc. degree in civil engineering from Jilin University, Changchun, China, in 2016 and the M.Sc. degree in civil engineering from Lanzhou University, Lanzhou, China, in 2019. He is currently working toward the Ph.D. degree with Lanzhou University of Civil Engineering, Lanzhou, China.

His research interests include deep learning and computer vision in disaster prevention.



Yixuan Huan received the B.Sc. degree in civil engineering from North China University of Technology, Beijing, China, in 2021. He is currently working toward the M.Sc. degree with Lanzhou University of Civil Engineering, Lanzhou, China.

His research interests include multimodal learning.



Yang Xu received the B.Sc., M.Sc., and Ph.D. degrees in civil engineering and engineering mechanics from Harbin Institute of Technology, Harbin, China, in 2012, 2014, and 2019, respectively.

He is currently an Assistant Professor with the School of Civil Engineering, Harbin Institute of Technology, China. His research topic is structural health diagnosis with computer vision and deep learning. As PI, he takes one Youth Project and one subproject of Major Program from National Natural Science Foundation of China, two subprojects of National Key R&D Program, one China National Postdoctoral Program for Innovative Talents, one general funding of Postdoctoral Science Foundation of China, one open funding of National Key Laboratory of China, one program from Natural Science Foundation of Heilongjiang Province, one special funding and one general funding of Heilongjiang Postdoctoral Program, respectively. He has authored or coauthored more than 40 papers in international journals and conferences (more than 20 SCI indexed, 4 selected as ESI, and Wiley Top Cited Papers) and authorized over 20 national patents and software copyrights.

Dr. Xu is currently a Member of the Early Career Researchers Committee of ISHMII, a Member of organizing committee in the international series competitions for structural health monitoring, the special session chair in the 8th World Conference on Structural Control and Monitoring, and the special issue guest editor of SCI journals.



Qiangqiang Zhang received the B.Sc. and M.Sc. degrees in civil engineering from Harbin Institute of Technology, China, in 2010 and 2012, respectively, and the Ph.D. degree in engineering mechanics from Harbin Institute of Technology, China, in 2016.

He is a "CuiYing Scholar" Professor with The college of Civil Engineering and Mechanics, Lanzhou University, China, the Dean of the Department of Civil Engineering, and a Doctoral Supervisor. He was a Visiting Researcher with Purdue University of Materials Science, USA. He won the support of

National Special Support Plan for High-Level Talents in 2022. He has conducted long-term research on the key scientific issues of smart structure and intelligent sensing techniques. His research work has been published in high-level journals (more than 40 SCI index) including Science and advanced materials and reported by nature, nature review materials, science, and other journal highlights.