Cascaded framework for earthquake building damage detection combining spatial and frequency domain feature integration

Dongping Ming, Shizhe Xie, Dehui Dong, Jing Zhang

School of Information Engineering, China University of Geosciences Beijing, People's Republic of China
Corresponding author: mingdp@cugb.edu.cn

Keywords: Earthquake, Building Collapse, Feature Knowledge Interaction, Frequency Domain, Semantic Segmentation, Classification of Building Damage.

Abstract

Building collapse is a major cause of casualties after an earthquake, so accurately extracting building damage information is critical for post-earthquake assessment and rescue. Currently, most deep learning methods focus on the end-to-end detection of building collapse. However, in real-world earthquake scenarios, the end-to-end computational process often lacks flexibility and struggles to meet the requirements of rapid emergency response. To address this issue, this paper proposes a cascaded framework that combines pre-earthquake building extraction and post-earthquake building damage classification. The proposed framework includes two sections: (1) Progressive building semantic segmentation model in the joint frequency domain. This model is designed to accurately extract buildings prior to an earthquake, with the goal of minimizing error propagation throughout the cascading process. The model addresses the spatial similarity of buildings under complicated backgrounds, as well as the high internal heterogeneity of buildings, by utilizing frequency domain techniques. It compensates for the shortcomings of traditional models in terms of incomplete information extraction through the effective integration of global and local information. Finally, the model employs edge priors for edge regularization. (2) Rapid building damage classification process. Based on the accurate building extraction results, a fast and efficient classification process is developed. This process uses a simple and lightweight classification network to effectively extract building damage information caused by the earthquake. The superiority of the proposed framework is validated through comparison with traditional cascading architectures and end-to-end models. The results show that the cascading framework not only provides accurate pre-earthquake building extraction, but also enables efficient and accurate post-earthquake damage classification, which meets the requirements of rapid post-earthquake emergency response. This balance of accuracy and speed is essential for effective disaster management and recovery.

1. Introduction

Earthquakes are among the world's most dangerous natural disasters, and building collapse has been identified as one of the most emblematic forms of seismic damage, leading directly to human casualties and significant property loss(Qu et al., 2023). Rapid assessment of earthquake-induced building damage is critical for effective emergency response and pre-rescue operations. The post-earthquake geological environment often presents significant hazards, making on-site investigations impractical. Therefore, the use of remote sensing data technology facilitates the rapid, efficient, and safe acquisition of information about post-earthquake building collapse(Xie et al., 2023). The use of automated and intelligent data mining and analysis increases the speed of disaster response and the efficiency of post-earthquake damage assessment, thereby reducing economic losses(Zhang et al., 2023).

In recent years, many researchers have explored the use of centimeter-level drone data for building damage detection. The ultra-high resolution of these data allows for a more detailed representation of building damage and provides high extraction accuracy. However, drone operators are often unable to reach hard-hit areas immediately after an earthquake, and some locations may be completely inaccessible. In addition, drones are limited in their ability to rapidly cover large areas, making them less effective for extensive data collection in disaster zones. As a result, sub-meter satellite imagery remains critical for rapid assessment of building collapse after an earthquake(Burke et al., 2019).

With the advancement of deep learning in computer vision, extensive applications in remote sensing building damage detection have emerged. Architecturally, building damage detection is mainly divided into end-to-end and cascaded frameworks. The end-to-end architecture typically employs Siamese-network structures that merge localization and classification tasks while sharing knowledge. The researchers used siamese networks to detect building damage(Sun et al., 2022; Chen et al., 2022; Seyed et al., 2024). However, current siamese networks are difficult to train due to the large amount of data, require precise image registration, and lack the flexibility of cascaded networks that allow for pre-earthquake building localization and rapid post-earthquake classification.

Cascaded architectures predominantly use object-based image analysis (OBIA) for segmentation. Patch-based CNNs integrated with OBIA primarily use superpixel segmentation to generate objects that are non-semantic with irregular geometric shapes(Zhang et al., 2018). However, semantic inconsistencies in building damage assessment occur in semantic and regularly shaped building objects, rendering traditional OBIA methods inapplicable. The crux lies in the fact that current OBIA only integrates process level with deep learning, lacking feature level interaction. Therefore, some researchers use used a fully convolutional network (FCN) for building localization (Gupta et al., 2019)and a patch-based CNN for damage classification (Qing et al., 2022a), but the limited parameterization of these methods fails to accurately represent building features, resulting in suboptimal accuracy in subsequent damage classification. All above, the paper presents the following innovations:

(1) To address the issues of lacking feature-level knowledge interaction and multi-target misclassification in OBIA cascade networks, a framework for building collapse detection is proposed, which utilizes the fusion of spatial and frequency domain features. This architecture generates objects with practical significance and refines the minimum unit of collapse detection.

(2) To improve the pre-earthquake building extraction and boundary accuracy, an advanced building semantic segmentation model combining spatial and frequency domain features is introduced. It includes the organic integration of global and local features and a building edge regularization module to better align segmentation results with actual building boundaries, thereby reducing error propagation in cascaded structures.

(3) To accurately classify building collapses after earthquakes, a simple and fast extraction method is proposed. The use of buffering strategies effectively reduces classification errors caused by registration issues. Simultaneously, to rapidly and effectively extract inter-channel deformation features, a lightweight, spatial-domain feature-enhanced deformable convolutional neural network is designed.

2. Method

A cascaded architecture for building collapse detection has been proposed, which decouples the task into pre-earthquake building localization and post-earthquake building collapse detection. This method utilizes domain feature enhancement to facilitate knowledge interaction between the two tasks, enabling more precise detection of building collapse information, and making the framework process more flexible The main workflow is illustrated in **Figure 1**.

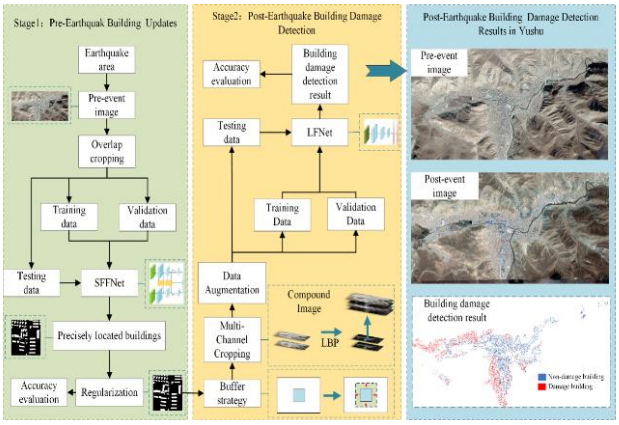

Figure 1. Overview of building collapse detection framework

2.1. Building collapse detection framework

The paper proposes a cascaded building collapse detection framework designed to flexibly extract building damage information after an earthquake. The process includes several key steps:

(1) Collection of pre-earthquake remote sensing images. The framework starts by collecting pre-earthquake remote sensing images of the affected area. To ensure accurate detection of buildings, these images typically require sub-meter spatial resolution. (2) Effective preprocessing methods. Research suggests that overlapping cropping and sample augmentation are effective preprocessing methods. These techniques prepare images for further processing and analysis. (3) Building Segmentation Using spatial and frequency domain feature-integrated building extraction Network (SFFNet). The pre-processed samples are then fed into SFFNet, a neural network, to obtain accurate building segmentation results. SFFNet is designed to effectively segment buildings from the remote sensing images. (4) Post-processing through connectivity analysis and regularization (Wei et al., 2020). After segmentation, the framework applies connectivity analysis and regularization to post-process the building detection results. This step refines the segmentation and isolates individual buildings. (5) Establishment of individual building buffer zones. Using the identified building vector positions, the framework creates buffer zones around individual buildings. These buffer zones are critical for isolating each building and its immediate environment for detailed analysis. (6) Creation of Multi-Channel Damage Detection Matrix Blocks. The buffer zones are then used to overlay pre-earthquake and post-earthquake images and pre- and post-earthquake Local Binary Pattern (LBP) (Ojala et al., 1994) texture features. This overlay creates multi-channel matrix blocks for damage detection, integrating different types of information for each building. (7) Classification with Lightweight fast network (LFnet). Finally, these multi-channel matrix blocks are fed into LFnet for classification. LFnet classifies the blocks and determines the extent of damage to each building.

2.2. Spatial and frequency domain feature-integrated building extraction network

Building vectors are critical for post-earthquake building damage detection. Due to variations in building materials, scale, and illumination, buildings exhibit significant differences in remote sensing imagery. Non-building structures such as parking lots and roads often appear similar to buildings, resulting in low inter-class variance and high intra-class variance in optical remote sensing imagery. Accurately and efficiently extracting building footprints from complex scenes remains a challenge, mainly due to insufficient feature extraction and inaccurate, irregular building boundary localization in semantic segmentation results. Therefore, in this paper, we propose an edge-prior-based progressive feature fusion network, as shown in **Figure 2**. In the network, we use U^{2}net (Qin et al., 2020) as the backbone and design a progressive space and frequency domain feature fusion block (PSFF Block). Specifically, based on the local detail information provided by spatial features, the Frequency Domain Global Information Extraction Module (FGIE Module) utilizes transformer in the frequency domain to obtain its global semantic information, and ultimately, the Adaptive Feature Fusion Module (AFF Module) performs feature fusion. In particular, by incorporating edge priors in the second layer, we enhance the extraction of regularized edge features of buildings, thereby improving segmentation accuracy and regularizing building boundaries.

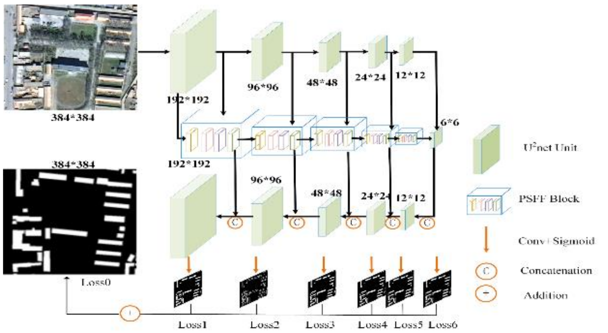

Figure 2. Spatial and frequency domain feature-integrated building extraction network

2.2.1 Progressive spatial and frequency domain feature fusion block

In order to solve the first problem, we adopt a progressive approach to semantic fusion as shown in **Figure 3**. As shown in **Figure 3(a)**, the U^{2}net unit is used to extract local feature representations of buildings in the spatial domain of the image. GFIE module is then used to capture global features in the frequency domain. Finally, an adaptive feature fusion module is used for layer-by-layer feature fusion to enrich the information content of the features. Based on the property that shallow features of CNNs are sensitive to high-frequency information, while deep features are sensitive to low-frequency features, we choose to supplement shallow layers of CNNs with high-frequency global information and deep layers with low-frequency global information, similar to the layer-by-layer feature extraction of CNNs. We perform truncation processing on frequency domain information at different levels to enable full integration of information. The GFIE module is shown in **Figure 3(b)**. The use of DCT for frequency domain transformation is mainly based on considerations of computational efficiency and suitability for real signals, as shown in Formula (1)-(3).

$$F(x,y)=\frac{2}{\sqrt{MN}}\sum_{V=0}^{M-1}\sum_{U=0}^{N-1} f(U,V)\cos\left[\frac{(2x+1)U\pi}{2M}\right]\cos\left[\frac{(2y+1)v\pi}{2M}\right] \tag{1}$$

$$U=0,1\ldots,M-1;V=0,1\ldots,N-1$$

$$f(x,y)=\frac{2}{\sqrt{MN}}\sum_{V=0}^{M-1}\sum_{U=0}^{N-1} C(U,V)\,F(U,V)\cos\left[\frac{(2x+1)U\pi}{2M}\right]\cos\left[\frac{(2y+1)v\pi}{2M}\right] \tag{2}$$

$$x=0,1\ldots,M-1;y=0,1\ldots,N-1$$

$$C(U)=\begin{cases}\frac{1}{M} & , \quad U=0\\ \sqrt{\frac{2}{M}} & ,U=1\ldots,M-1\end{cases} \qquad C(V)=\begin{cases}\frac{1}{M} & , \quad V=0\\ \sqrt{\frac{2}{M}} & ,V=1\ldots,M-1\end{cases} \tag{3}$$

where M and N represents the width and height of the image, *F*(*U,V*) is the frequency coefficient in the two-dimensional frequency domain, *f*(*x,y*)is the pixel value of the original image in the spatial domain, *U* and *V* is the coordinate in the frequency domain, *C*(*U*) and *C*(*V*) is the DCT transformation coefficient. After the DCT transformation, the energy percentage method is constructed to gradually remove high-frequency information, as shown in **Figure 3(c)**.

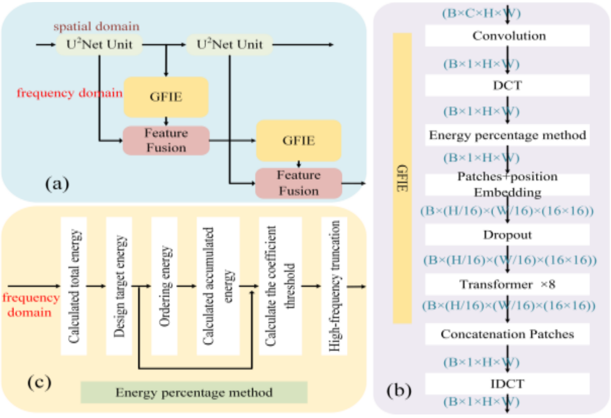

Figure 3. Progressive spatial and frequency domain feature fusion module

The energy percentage method calculates the total energy by constructing the energy of each coefficient in the frequency domain, and designs a truncation percentage to obtain the target energy for truncation. Then, all energy values are sorted in descending order and their cumulative sum is calculated. When the cumulative sum reaches the coefficient energy required for the target energy, the threshold for frequency domain truncation is obtained. This method has a certain adaptability compared to the traditional filter design, and uses this method to truncate high-frequency information by 2%, 4%, 6%, and 8%, as shown in in **Figure 4**.

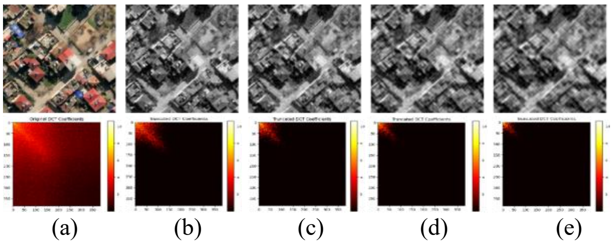

Figure 4. Images and visual frequency domain graphs under different high frequency truncation. (a) original image. (b) 2% truncation. (c) 4% truncation. (d) 6% truncation. (e) 8% truncation.

After feature extraction using the U^{2}net unit and the global information unit, the feature information is stored in multiple channels, which is not conducive to facilitating the distinction between buildings and background areas. Therefore, we use convolution operations to combine all the channel features into a single channel. For effective fusion of the extracted local and global information, a sigmoid function-guided feature fusion method is proposed. This method can effectively distinguish between buildings and non-buildings, which helps to guide the feature selection. The formula can be expressed as Formula (4):

$$F_C=F_P\times S_P+(1-S_P)\times F_G \tag{4}$$

where F$_C$ represents fusion features, F$_P$ represents local information extracted by u^{2}net, F$_G$ represents global features extracted in frequency domain, and S$_P$ represents the use of sigmoid to predict probability.

2.2.2 Edge control strategy

Considering that buildings, as man-made structures, have distinct geometric features, an edge-prior-based adaptive regularization method for building edges is proposed to address inaccuracies and irregularities often found in building semantic segmentation results. Deep networks, which focus on abstract semantic features, tend to miss finer details, while initial shallow layers retain excessive details, leading to noise. To address this, the planar semantic information of the second layer is converted to edge semantic information, guided by edges extracted from building labels. This improves the network's ability to extract building edge features. When extracting edge information from labels, a broadened label edge strategy is used because of the difficulty in training networks with too fine edges. In addition, a weighted loss function is used to balance the samples for edge learning. The edge loss function for the second layer is as follows:

$$Weight_N=\frac{Num(Pixel_N)}{Num(Pixel_N)+Num(Pixel_p)} \tag{5}$$

$$Weight_P=\frac{Num(Pixel_P)}{Num(Pixel_N)+Num(Pixel_p)} \tag{6}$$

$$\begin{aligned}loss_2 &= L_{bce}(X2_p,X_{edge})\times Weight_N\\ &+L_{bce}(X2_N,X_{edge})\times Weight_p\end{aligned} \tag{7}$$

where *loss*$_2$ represents the second layer loss function, *L$_{bce}$* represents the use of the binary cross-entropy loss, *Weight$_N$* represents the background weight, *Weight$_P$* represents the target weight, and *Num* represents the total number of pixels.
The overall network loss function is as follows:

$$\begin{aligned}loss_{fused} &= L_{bce}(X_1,X_{gt})+L_{SSIM}(X_1,X_{gt})\\ &+loss2+\textstyle\sum_{i=3}^{6}L_{bce}(X_i,X_{gt})\end{aligned} \tag{8}$$

where *X$_i$* represents the result of the i-th layer, *X$_{gt}$* represents the ground truth, *L$_{bce}$* represents the use of the binary cross-entropy loss, and L_{SSIM} is structural similarity loss.

2.3. Post-earthquake building collapse detection

Traditional change detection typically involves three main tasks: (1) detecting the transition of buildings from intact to damaged, (2) assessing whether buildings that were intact before the earthquake remained mostly intact after the earthquake, and (3) detecting irrelevant background. However, the collapse of buildings after an earthquake is often irregular in extent, which may lead to background changes unrelated to building damage, thereby reducing the accuracy of building damage detection.

In response to this problem, this paper takes a novel approach by focusing on individual buildings rather than background changes. By using the pre-earthquake detection results of individual buildings, this paper simplifies the aforementioned tasks into two more specific objectives: (1) identifying buildings that have transitioned from an intact state to a damaged state, and (2) determining which buildings that were intact before the earthquake have remained largely intact. The advantage of this approach is that by focusing on the state changes of individual buildings, it effectively avoids misjudgments caused by changes in the background, thus improving the accuracy of building damage detection. In addition, this method makes the detection tasks more precise and focused, helping to improve the overall performance of change detection.

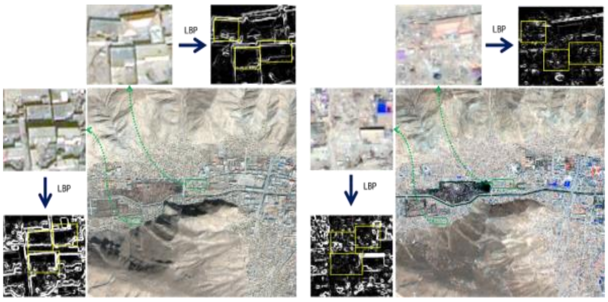

Figure 5. Schematic diagram of LBP changes

As shown in **Figure 5**, given the difficulty of perfectly aligning pre-earthquake and post-earthquake images and the distinct contextual features of collapsed buildings, a buffering strategy is used to ensure the integrity of buildings in the samples and to capture more features of collapsed structures. Since the most obvious post-collapse features are building boundary and texture, Local Binary Patterns (LBP) images of buildings are used to enhance bands in pre-earthquake and post-earthquake images and guide the classification network to learn texture features. Notably, the proposed LFNet is simple, fast, and has high classification accuracy. Among them, Resnet is mainly used as the basic network, and deformable convolution (Dai et al.,2017) is added to each layer of Resnet to adapt to the changes between channels. Therefore, a building earthquake damage detection process is constructed, as shown in **Figure 6**.

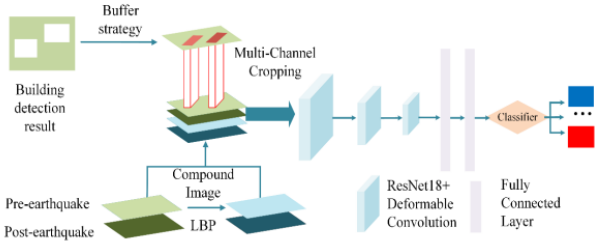

Figure 6. Building fall damage detection process

3. Experiments and Results

3.1. Study area data

The study area is located in Guangjie Town, Yushu City, Qinghai Province, China, as shown in **Figure 7**. A magnitude 7.1 earthquake occurred here on April 14, 2010, resulting in extensive structural damage, 2,220 deaths, and thousands of injuries. The experimental data are from Google Earth imagery with a spatial resolution of 0.6 meters (panchromatic and multispectral fusion imagery) and an image size of 20,000 pixels × 12,000 pixels. Details of the image data are shown in **Table 1**.

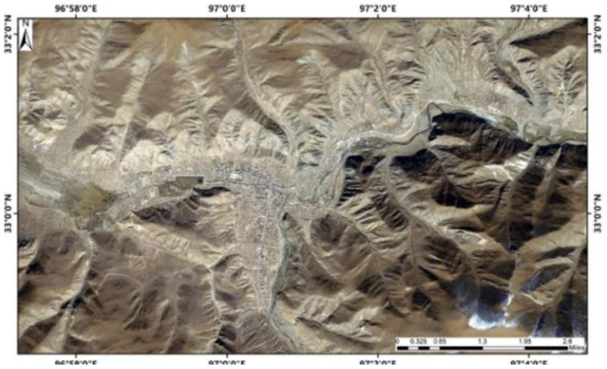

Figure 7. Study area, Yushu City

Test area	Type	Acquisition time	Source	Spatial resolution
Yushu	Pre-earthquake Image	2008.12.08	Google earth	0.6m
	Post-earthquake Image	2010.4.17	Google earth	0.6m

Table 1. Description of the data used in the study

3.2. Evaluation metrics

In terms of evaluation metrics, the system commonly used in semantic segmentation - Precision, Recall, F1 and IoU - has been adopted.

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - Score = 2 \times \frac{precision \times Recall}{precision + Recall} \tag{11}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{12}$$

where TP is the number of pixels correctly extracted as buildings, FP is the number of other object pixels extracted as buildings, and FN is the number of building pixels extracted as other objects.

3.3. Pre-earthquake building extraction result analysis

This section reports experiments conducted on two datasets. In order to ensure fairness, the experimental data set was cropped to the same size and the same method of data enhancement was used. Our model was thoroughly compared with state-of-the-art

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-503-2024 | © Author(s) 2024. CC BY 4.0 License.

methods(Chen et al., 2021; Li et al., 2022; Wang et al., 2022b; Zhou et al., 2022) to demonstrate the segmentation quality and to assess the capabilities of our model.

3.3.1 Experimental detail

In the experiment, the sample was cropped to 384*384. During training, the Adam optimizer was used with default parameters (initial learning rate = 1e-4, betas = (0.9, 0.999), eps = 1e-8, weight decay = 0). The network was trained with a batch size of 4 and a termination iteration of about 300 epochs. The training process was performed on a platform with an I7-10700 CPU and 3090 GPU, with 24G of memory.

In the Yushu dataset, building detection is challenging due to the complex background and dense urban areas with shadows from buildings and trees, as well as significant size variations among buildings, making small structures difficult to extract. As shown in **Figure 8**, where TP (white) means the number of pixels correctly extracted as buildings, FP (blue) means the number of other object pixels extracted as buildings, and FN (red) means the number of building pixels extracted as other objects. The first two rows depict images of urban areas in Yushu City, where buildings are prominent against the city background with orderly arrangements. From the five contrast results, it can be observed that our method achieves higher precision in extracting building edges. The latter three rows show images of rural areas, where building layouts are less organized. This leads to issues of missed and false detections due to minimal differences between buildings and background, as well as challenges in accurately segmenting tightly spaced buildings. Based on these experimental results, compared to using spatial-domain deep learning methods alone, our approach utilizes the Discrete Cosine Transform (DCT) to transform spatial-domain signals into the frequency domain. This enables the exploration of building features in complex backgrounds, reducing missed and false detections. Furthermore, by selecting high-frequency and low-frequency features, we address the segmentation challenges posed by adjacent buildings.

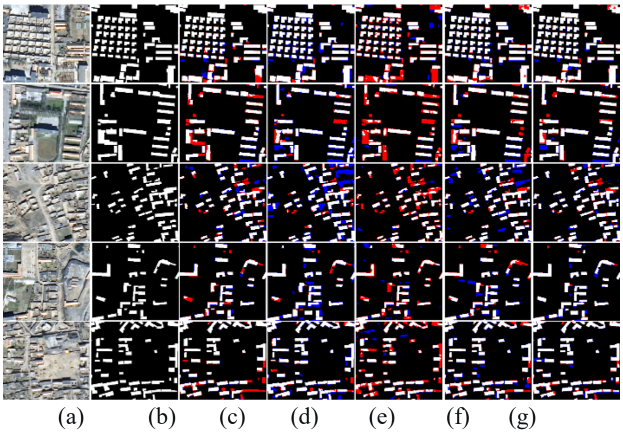

Figure 8. Building extraction results. (a) original image. (b) Ground Truth. (c) MANet. (d) SGCN. (e) TransUnet. (f) UNetFormer. (g) Ours.

The accuracy results of the five sets of experiments are shown in **Table 2**. Although our proposed method did not perform as well as other methods in terms of accuracy, it achieved a more balanced recall rate, indicating a better restriction of false extractions, resulting in better F1 and IoU scores. Specifically, compared to MANet, our method improved IoU and F1 by 1.58% and 2.28%, respectively, indicating the effectiveness of our network. Compared to SGCN, there was an increase of 3.04% and 4.32%, respectively, showing better overall performance. Compared to TransUnet, the increases were 4.01% and 5.66%, and compared to UNetFormer, there were improvements of 2.02% and 2.90%, respectively. This suggests that using transformers to find features in the frequency domain is more accurate than direct extraction in the image domain.

Method	Precision (%)	recall (%)	F1 (%)	IoU (%)
MANet	80.27	82.21	81.23	68.39
SGCN	73.43	87.32	79.77	66.35
TransUnet	82.68	75.26	78.80	65.01
UNetFormr	79.96	81.63	80.79	67.77
Ours	82.34	83.30	82.81	70.67

Table 2. Quantitative evaluation of different methods.

3.3.2 Ablation study

In this section, separate ablation studies are performed on two datasets to assess the effectiveness of each critical component of the model. The U2net is used as a baseline, and additional modules are progressively integrated. We focus on the visualization of the penultimate layer in the decoder, as shown in **Figure 9**. It can be seen that as the number of components increases, there is a tendency for more buildings to be identified in the feature map. After integrating the GFIE, but opting for Concatenation Fusion instead of the AFF module, there is a noticeable improvement in the detection efficiency of small target buildings. However, this approach also results in a higher false positive rate for building detection. The use of the AFF module allows effective control over global and local feature fusion, which helps to reduce some false detections. Finally, the implementation of edge priors helps to refine building boundaries, thereby improving accuracy.

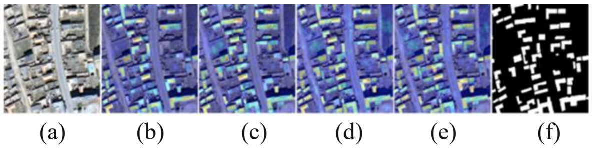

Figure 9. The visualization of ablation study. (a) Original image. (b) U^{2}net. (c) U^{2}net+ GFIE. (d) U^{2}net+ GFIE+AFF. (e) U^{2}net+ GFIE+AFF+ Edge Prior. (f) Ground Truth.

As shown as **Table 3**, compared to the U^{2}net, the inclusion of GFIE results in a slight increase in both F1 score and IoU. There is an increase of 0.76% in F1 score and 0.32% in IoU. These improvements demonstrate the effectiveness of using the frequency domain to extract global information. With the addition of AFFM, which allows for a more effective fusion of global and local information, there is a significant improvement in both F1 score and IoU. The F1 score and IoU increase by 1.22% and 1.29%. Finally, the implementation of edge priors further improves the overall building segmentation accuracy. These metrics show the cumulative benefits of each component in improving the model's performance for building segmentation tasks.

	U^{2}net	GFIE	AFFM	Edge Prior	F1(%)	IoU(%)
(a)	√	×	×	×	79.96	68.83
(b)	√	√	×	×	80.72	69.15
(c)	√	√	√	×	81.94	70.44
(d)	√	√	√	√	82.81	70.67

Table 3. Quantitative evaluation of ablation study.

3.4. Post-earthquake building extraction result analysis

To validate the effectiveness of the building damage extraction framework proposed in this paper, we selected two types of cascaded building damage information extraction frameworks(Chen and Liu, 2021; Qing et al., 2022b) and two end-to-end building damage information extraction methods for comparison(Caye Daudt et al., 2018; Yan et al., 2022).

During training, the Adam optimizer was used with default parameters (initial learning rate lr = 5e-4, weight decay = 5e-2). LFnet was trained with a batch size of 32 and a termination iteration of about 100 epochs. The training process was performed on a platform with an I7-10700 CPU and 3090 GPU, with 24G of memory.

Using LFnet to classify building collapses in the Yushu area after the earthquake, the precision for collapsed building detection is 91.24%, the recall is 90.52%, and the F1 score is 90.88%. **Figure 10** shows the assessment results of building damage after the earthquake, where red represents collapsed buildings, blue indicates intact buildings, and white represents the background.

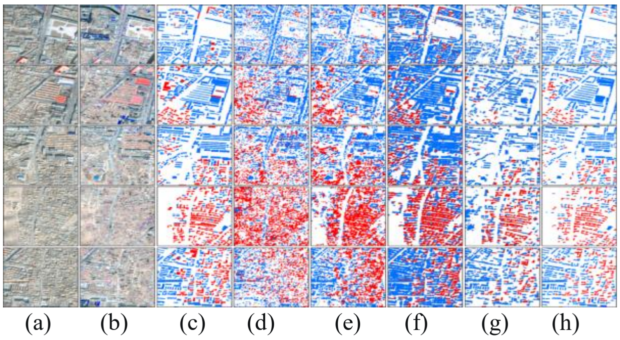

Figure 10. Building Collapse Detection results for Yushu Dataset. (a) Pre-earthquake image. (b) Post-earthquake image. (c) Ground Truth. (d) SLIC +SVM. (e) SLIC +CNN. (f) EF. (g) FTN. (h)Ours.

The first two groups of images are from the Yushu urban area, where relatively few buildings are damaged, while the last three groups are from the urban-rural interface and villages, where more buildings are damaged. It can be seen from the images that the object-based segmentation performance is poor in the two groups. This is mainly because the superpixel segmentation used in these experiments did not incorporate semantic information from the images during segmentation, resulting in fragmented buildings and inaccurate boundary positioning. As a result, buildings tend to stick together during the post-classification clustering process. Furthermore, in the first set of experiments, classification was based on representing an area with a single point, and the representativeness of this point is key to classification accuracy. However, a single point lacks semantic relationships with its surroundings and similarly categorized pixels, and it cannot effectively address issues such as small variance between background and damaged buildings. Therefore, in areas with more damaged buildings, the number of false detections increases significantly. In the second set of experiments, the classification involved cropping small patches within a spot and using a deep learning network to determine the category of these small patches, incorporating some semantic information between categories. However, these small patches also have errors, as each patch can contain more than one category (ground, damaged building, intact building). As a result, while the accuracy of this experiment is significantly better than that of the first group, it still does not solve the problems of buildings sticking together and false detections. The third and fourth groups use an end-to-end classification method. It is obvious that the main factor affecting the accuracy of end-to-end methods is the accuracy of the building extraction by the main network. If the accuracy of building extraction is low, the results tend to be poor. The accuracy rating, as shown in **Table 4**, includes C1 for background, C2 for intact buildings, and C3 for damaged buildings.

Method		SLIC +SVM	SLIC+CNN	EF	FTN	Ours
Precision (%)	C1	88.36	93.94	97.99	94.74	94.60
	C2	21.00	21.99	40.61	55.63	69.38
	C3	21.78	29.28	35.48	66.05	74.24
Recall (%)	C1	68.70	72.40	74.71	95.07	95.72
	C2	23.49	30.71	82.79	68.07	69.44
	C3	26.23	29.08	79.34	56.61	68.16
F1 (%)	C1	77.30	81.78	84.78	94.90	95.16
	C2	22.18	25.63	54.49	61.22	69.40
	C3	23.80	29.18	49.03	60.97	71.07
IoU (%)	C1	63.00	69.17	73.58	90.30	90.72
	C2	12.47	14.70	37.45	44.13	53.15
	C3	13.51	17.08	32.48	43.85	55.12

Table 4. Quantitative evaluation of Building Collapse Detection results.

4. Conclusions

In this study, a cascaded architecture for building collapse detection, which integrates spatial and frequency domain target feature knowledge interaction, has been proposed. Based on experimental results and analysis, the conclusions can be drawn as follows.

(1) Compared to end-to-end direct damage detection networks, the cascade framework clarifies the tasks and overcomes the training and transfer difficulties of direct detection networks, resulting in a more flexible overall process. Compared to traditional OBIA cascade networks, our method fully extracts semantic information at each stage, forming feature-level knowledge interaction, generating objects with practical significance, and improving the detection rate of building damage.

(2) Accurate building detection is the foundation of this framework because it can reduce loss propagation within the framework. The building detection method introduced in this paper validates that the effective combination of spatial and frequency domains in complex backgrounds can improve the accuracy of building extraction. This method addresses the problem of high heterogeneity leading to small inter-class variance and large intra-class variance and utilizes adaptive feature selection of global and local information to address the challenge of inferring image content from distant context. Additionally, the edge verification module can further improve the accuracy of boundary detection.

(3) Building on accurate building extraction, a fast and simple method for building collapse classification is proposed. By utilizing a buffering strategy to reduce errors caused by registration, and constructing a multi-channel classification network based on texture priors, rapid detection of building collapses is achieved.

In summary, the newly proposed cascaded building collapse detection framework is a workflow with clear tasks, flexible processes, and high detection accuracy, which can better serve the emergency management domain. Moreover, the building extraction network as the core method of the workflow can also be introduced to other areas of geoscience applications, such as the semantic segmentation of other land cover features (roads, farmlands, etc.).

However, due to the limited resolution and vertical field of view of remote sensing, it is not possible to observe the lateral damage of building walls, which leads to the omission of intermediate levels of damage in our damage detection. This is a key issue that requires further research.

Fundings

This research was funded by the National Natural Science Foundation of China (42371379), the National Key R&D Program of China (2022YFB3903604), the “Deep-time Digital Earth” Science and Technology Leading Talents Team Funds for the Central Universities for the Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing) (Fundamental Research Funds for the Central Universities; grant number: 2652023001).

References

Ahmadi S A, Mohammadzadeh A, Yokoya N, Ghorbanian A, 2024. BD-SKUNet: Selective-Kernel UNets for Building Damage Assessment in High-Resolution Satellite Images. Remote Sensing, 16(1): 182.

Burke, C., McWhirter, P.R., Veitch-Michaelis, J., McAree, O., Pointon, H.A.G., Wich, S., Longmore, S., 2019. Requirements and Limitations of Thermal Drones for Effective Search and Rescue in Marine and Coastal Areas. Drones 3, 78.

Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully Convolutional Siamese Networks for Change Detection, in: 2018 25th IEEE International Conference on Image Processing (ICIP). Presented at the 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063–4067.

Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L., 2022. Dual-Tasks Siamese Transformer Framework for Building Damage Assessment, in: IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 1600–1603.

Chen, J., Liu, D., 2021. Bottom-up image detection of water channel slope damages based on superpixel segmentation and support vector machine. Adv. Eng. Inform. 47, 101205.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.

Chen, Q., Wang, L., Waslander, S.L., Liu, X., 2020. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. ISPRS J. Photogramm. Remote Sens. 170, 114–126.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable Convolutional Networks. Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

Ding, L., Tang, H., Liu, Y., Shi, Y., Zhu, X.X., Bruzzone, L., 2022. Adversarial Shape Learning for Building Extraction in VHR Remote Sensing Images. IEEE Trans. Image Process. 31, 678–690.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 10–17.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D., 2023. Transformers in medical image analysis. Intell. Med. 3, 59–78.

Jiang, X., Zhang, X., Xin, Q., Xi, X., Zhang, P., 2021. Arbitrary-Shaped Building Boundary-Aware Detection With Pixel Aggregation Network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2699–2710.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2022. Multiattention Network for Semantic Segmentation of Fine-Resolution Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Li, W., Xue, L., Wang, X., Li, G., 2023. ConvTransNet: A CNN–Transformer Network for Change Detection With Multiscale Global–Local Representations. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Li, Y., Miao, N., Ma, L., Shuang, F., Huang, X., 2023. Transformer for object detection: Review and benchmark. Eng. Appl. Artif. Intell. 126, 107021.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Ojala, T., Pietikainen, M., Harwood, D., 1994. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Proceedings of 12th International Conference on Pattern Recognition. Presented at the Proceedings of 12th International Conference on Pattern Recognition, pp. 582–585 vol.1.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M., 2020. U2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognit. 106, 107404.

Qin, Z., Zhang, P., Wu, F., Li, X., 2021. FcaNet: Frequency Channel Attention Networks. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 783–792.

Qing, Y., Ming, D., Wen, Q., Weng, Q., Xu, L., Chen, Y., Zhang, Y., Zeng, B., 2022a. Operational earthquake-induced building damage assessment using CNN-based direct remote sensing change detection on superpixel level. Int. J. Appl. Earth Obs. Geoinformation 112, 102899.

Qing, Y., Ming, D., Wen, Q., Weng, Q., Xu, L., Chen, Y., Zhang, Y., Zeng, B., 2022b. Operational earthquake-induced building damage assessment using CNN-based direct remote sensing change detection on superpixel level. Int. J. Appl. Earth Obs. Geoinformation 112, 102899.

Qu, Z., Zhu, B., Cao, Y., Fu, H., 2023. Rapid report of seismic damage to buildings in the 2022 M 6.8 Luding earthquake, China. Earthq. Res. Adv. 3, 100180.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net:

Convolutional Networks for Biomedical Image Segmentation.

Sandryhaila, A., Moura, J.M.F., 2014. Discrete Signal Processing on Graphs: Frequency Analysis. IEEE Trans. Signal Process. 62, 3042–3054.

Jain P, Tyagi V, 2013. Spatial and frequency domain filters for restoration of noisy images. IETE Journal of Education. 54(2): 108-116.

Sun C, Du C, Wu J, Chen H, 2022. SUDANet: A Siamese UNet with Dense Attention Mechanism for Remote Sensing Image Change Detection. Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer Nature Switzerland. 78-88.

Varghese, J., 2016. Adaptive threshold based frequency domain filter for periodic noise reduction. AEU - Int. J. Electron. Commun. 70, 1692–1701.

Wang, L., Fang, S., Meng, X., Li, R., 2022a. Building Extraction With Vision Transformer. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022b. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.

Wang, Z., Zhou, Y., Wang, F., Wang, S., Qin, G., Zou, W., Zhu, J., 2023. A Multi-Scale Edge Constraint Network for the Fine Extraction of Buildings from Remote Sensing Images. Remote Sens. 15, 927.

Wei, S., Ji, S., Lu, M., 2020. Toward automatic building footprint delineation from aerial images using CNN and regularization. IEEE Trans. Geosci. Remote Sens. 58, 2178–2189.

Xie, S., Ming, D., Yan, J., Yang, H., Liu, R., Zhao, Z., 2023. Research on Fine Estimation of People Trapped after Earthquake on Single Building Level Based on Multi-Source Data. Appl. Sci. 13, 5430.

Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.-K., Ren, F., 2020. Learning in the Frequency Domain. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1740–1749.

Yan, T., Wan, Z., Zhang, P., 2022. Fully Transformer Network for Change Detection of Remote Sensing Images. Presented at the Proceedings of the Asian Conference on Computer Vision, pp. 1691–1708.

Yang, Y., Soatto, S., 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70.

Zhang, C., Wang, X., Zhang, Hongye, Zhang, Hanyu, Han, P., 2021. Log Sequence Anomaly Detection Based on Local Information Extraction and Globally Sparse Transformer Model. IEEE Trans. Netw. Serv. Manag. 18, 4119–4133.

Zhang, W., Wang, R., Chen, X., Jia, D., Zhou, Z., 2023. Systematic assessment method for post-earthquake damage of regional buildings using adaptive-network-based fuzzy inference system. J. Build. Eng. 78, 107682.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.

Zhou, G., Chen, W., Gui, Q., Li, X., Wang, L., 2022. Split Depth-Wise Separable Graph-Convolution Network for Road Extraction in Complex Environments From High-Resolution Remote-Sensing Images. IEEE Trans. Geosci. Remote Sens. 60, 1–15.