CSA-Net: Complex Scenarios Adaptive Network for Building Extraction for Remote Sensing Images

Dongjie Yang, Xianjun Gao, Yuanwei Yang, Minghan Jiang, Kangliang Guo, Bo Liu, Shaohua Li, and Shengyan Yu

Abstract—**Building extraction is significant for the intelligent interpretation of high-resolution remote sensing images (HRSIs). However, in some complex scenarios where the features of the building and its adjacent ground objects are similar, the current segmentation model cannot distinguish them effectively. Therefore, we propose a complex scenarios adaptive network (CSA-Net) for building extraction. CSA-Net is comprised of the hierarchical-context feature extraction (HFE) module, the global-local feature interaction (GFI) module, and the multiscale-adaptive feature fusion (MFF) structure. The HFE obtains high-level semantic information at different levels and fuses it with low-level detailed information by skipping connections to enhance the reasoning and perception ability of building structure in complex scenes. Then, the GFI acquires global-local features of buildings and their surrounding environment via dense multiscale dilated convolution. The information can be shared through efficient interaction among features, and irrelevant backgrounds can be suppressed. Then, in the up-sampling process, the MFF alleviates the feature loss and enhances the robustness of the network by using feature fusion after layer-by-layer adaptive weight allocation. Experiments show that CSA-Net outperforms other comparable methods, with intersection over union values of 79.99%, 89.75%, and 73.59%, respectively, on the Google Arlinton, WHU, and Massachusetts building datasets. The visual comparison results demonstrate that our method can enhance the accuracy of building extraction in complex scenes. Meanwhile, the efficiency results indicate our approach strikes a balance between calculation parameters and time and achieves high levels of efficiency.**

Manuscript received 27 March 2024; revised 17 May 2024; accepted 5 June 2024. Date of publication 13 June 2024; date of current version 8 July 2024. This work was supported in part by Open Fund of Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake, Ministry of Natural Resources under Grant MEMI-2021-2022-08, in part by Open Fund of Hunan Provincial Key Laboratory of Geo-Information Engineering in Surveying, Mapping and Remote Sensing, Hunan University of Science and Technology under Grant E22205, in part by Open Fund of National Engineering Laboratory for Digital Construction and Evaluation Technology of Urban Rail Transit under Grant 2023ZH01 and Grant 2021ZH02, in part by the National Natural Science Foundation of China under Grant 4217021074, in part by Tianjin Science and Technology Plan Project under Grant 23YFYSHZ00190 and Grant 23YFZCSN00280, and in part by Hunan Provincial Natural Science Foundation Project Department Union Fund under Grant 2024JJ8327. *(Corresponding author: Xianjun Gao.)*

Dongjie Yang, Xianjun Gao, Yuanwei Yang, Minghan Jiang, Kangliang Guo, and Shaohua Li are with the School of Geosciences, Yangtze University, Wuhan 430100, China (e-mail: 2023730054@yangtzeu.edu.cn; junxgao@yangtzeu.edu.cn; yyw_08@163.com; minghanjiang@qq.com; 596692956@qq.com; lish@yangtzeu.edu.cn).

Bo Liu is with the Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China (e-mail: liubo@ecut.edu.cn).

Shengyan Yu is with the Inner Mongolia Autonomous Region Surveying and Mapping Geographic Information Center, Hohhot 010050, China (e-mail: 1746685539@qq.com).

Data is available on-line at Github: https://github.com/329719386/CSA-Net.

Digital Object Identifier 10.1109/JSTARS.2024.3413987

Index Terms—**Attention, building extraction, deep learning, global-local features, remote sensing images.**

I. Introduction

The automatic extraction of buildings from high-resolution remote sensing images (HRSIs) is widely used in urban planning [1], [2], population estimation, and digital city implementation [3]. In fact, building scenarios are complex and diverse. For example, some buildings are surrounded by roads and squares of similar spectra and texture features, leading to false distinguishing of buildings. In addition, trees and shadows around the building also block the building information in HRSI, affecting the integrity of the building extraction. Therefore, the automatic extraction of buildings in complex scenes is still very challenging.

Traditional building extraction primarily uses pixel-based or object-based classification methods [4]. By manually designing and selecting features, the classifier is trained to classify buildings [5]. Postprocessing is designed to improve classification accuracy based on region growth, morphological processing, and other methods [5]. However, these methods cannot be adaptively adjusted as the environment changes. Therefore, the application of these methods is minimal, and it is challenging to achieve high-precision and efficient building extraction in large-scale, complex scenes [6].

Convolutional neural networks (CNNs) can automatically obtain multilevel features from low to high, thereby reducing the impact of artificial factors on the feature acquisition process [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17]. The end-to-end pixel-level semantic segmentation method, such as CNNs [18], [19], fully convolutional network [20], and U-Net [21], has become the prevailing framework for deep learning-based building extraction in HRSI [22]. To improve the accuracy of building semantic segmentation, researchers often created more dataset [23], [24], [25], enhance advanced network architecture [26], and develop building characterization modules [27], [28]. Although the above methods have obtained remarkable achievements in building segmentation, the following challenges remain in complex scenes

The first challenge is that it is difficult to extract effective discriminative features when the spectral and texture feature information of buildings and adjacent nonbuildings is continuous. As shown in Fig. 1(a), the parking lot around the building is easily recognized as a building. Meanwhile, white objects with different spectra and textures inside the building roof are easily recognized as nonbuildings due to the limitations of local features. Multiscale feature extraction and attention

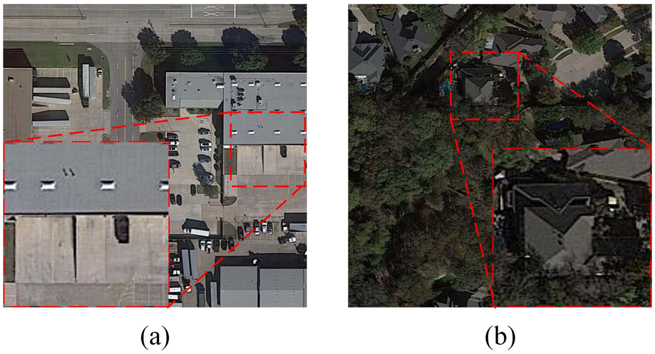


Fig. 1. Complex building scenes from the GA dataset. The red boxes mark areas that are challenging for building extraction in HRSI. (a) Spectrum and texture of the neighborhood and interior parts influence building identification. (b) Buildings are easily affected by trees and shadows.

mechanisms [22] are explored to capture more local and global feature information. However, these methods cannot effectively distinguish complex buildings from adjacent nonbuildings with similar features.

The second challenge is occlusion by trees or shadows. As shown in Fig. 1(b), the shadows of trees or buildings will block partial building information, resulting in incomplete building extraction. Existing methods generally enhance detailed information by focusing on boundaries [29] or generating building vectors from corner points [30]. They always rely on prior parameter settings. However, in complex scenes, the appeal methods are difficult to effectively extract the occluded part of the building.

Aiming to solve the building extraction challenges in complex scenes, we propose a complex scenarios adaptive network, CSA-Net. The main contributions are illustrated as follows.

1) We designed the global-local feature interaction (GFI) module for the first challenge. The GFI module mainly comprises the global-local feature extraction (GFE) module and the efficient channel interaction (ECI) structure. The GFE uses string and parallel dilated convolution architecture to obtain more abundant local and global features of buildings and the surrounding environment. The ECI shares global-local feature information through channel interaction. Finally, building features in complex scenes are enhanced by assigning higher weights to buildings in global-local features.
2) We designed the hierarchical-context feature extraction (HFE) module for the second challenge. The HFE uses two hierarchies features as input to obtain richer high-level features, such as shape, structure, and other abstract features. Meanwhile, we conduct an interactive combination of input information to reduce noise, such as inaccurate edge or occluding information, and suppress background noise. Thus, the HFE helps the network to extract the building more completely by enhancing the reasoning and perception of the overall structure of buildings.
3) We designed the multiscale-adaptive feature fusion (MFF) structure to adapt to complex scenes. Since the critical feature information received by the GFI and HFE modules is partly lost during the upsampling process, the information recovery of complex scene features is unstable. The MFF supplements features progressively from the bottom to the top to mitigate the loss of critical feature information. The transmission of essential semantic information is strengthened by assigning weights to different layers.

The rest of this article is organized as follows: Section II summarizes some previous research results and problems with building extraction according to the research direction. Section III presents the GFI, the HFE modules, and the MFF structure in CSA-Net. Section IV presents the experimental results. Section V concludes the article.

II. Related Work

In this section, we will briefly introduce the related work of building extraction by traditional and deep learning methods.

A. Traditional Building Extraction Methods

The traditional building extraction method can usually be divided into two steps: feature information extraction and classification to distinguish buildings from the background [5], [31], [32]. In feature extraction, many manually designed features were explored. Huang and Zhang [33] introduced morphological index MBI. Turker and Koc-San [34] utilized building features such as morphology and texture. Zhang et al. [35] used various features like spectrum and texture to obtain more building information. In addition, some scholars have also tried to utilize nonbuilding features, such as the normalized difference vegetation index, to distinguish buildings and backgrounds [36]. Further, by inputting the extracted features into a classifier, such as the support vector machine [37] or artificial neural networks [38], the classifier can be trained for building classification [39], [40]. However, traditional methods are affected by artificial prior knowledge, and when faced with building extraction in large-scale and complex scenes, the results are challenging to meet actual needs [41].

B. Deep Learning Building Extraction Methods

Numerous scholars have researched and explored building extraction based on deep learning [42], [43], [44], [45], [46], [47], [48], [49]. They mainly explored building extraction methods from three aspects: multilevel feature extraction, GFE, and upsampling optimized structure. This section will briefly review the relevant work in these three directions.

1) Multilevel Feature Extraction: Contextual feature supplementation benefits semantic segmentation [50]. U-Net [21] supplements contextual information by skipping connection. Seyedhosseini and Tasdizen [51] optimized the network to segment images of different resolutions by obtaining features at different hierarchies. However, features at different levels are frequently different. The irrelevant noise weakens the semantic segmentation ability of the network [52]. For example, high-level features contain information such as occlusion, edges, corners, contours, and shapes. The structure and shape features can enhance the reasoning ability of regular buildings. However, inaccurate edge and occlusion information will also affect the complete extraction of buildings. Therefore, some researchers suppress irrelevant background noise by introducing attention [53], [54], [55], [56]. Wang et al. [57] adopted a two-dimensional importance weight map to reweight the original features after learning a space-wise importance score using an encoder-decoder structure. Chen et al. [23] added an attention mechanism to obtain more refined features during image size restoration. Gong et al. [58] introduced a context–content aware module (C^2AM) to acquire contextual information and enhance building dependencies. The above method proves the importance of multilevel feature extraction for semantic segmentation. However, previous models still lack efficient extraction of high-level features, resulting in difficulty in describing the complete inference of complex scenes and increasing calculation parameters.

2) Global-Local Feature Extraction: The effective extraction of building information relies on the efficient mining and comprehensive utilization of both global and local data from HRSI [59], [60]. ParSeNet [61] provides a new direction for subsequent semantic segmentation by introducing global concepts. However, in real scenarios, extracting global features places higher requirements on the receptive field [62]. Some scholars improve the receptive field through dilated convolution to obtain more global–local features [22], [63], [64]. Nevertheless, the characteristics are frequently discontinuous. PSPNet [65] utilizes the pyramid pooling module to get more complete global features. However, the receptive field of the network is still limited. Transformer introduces a novel approach for global feature acquisition by establishing global feature correlation through self-attention [66], [67]. Researchers are attempting to address the challenge of effectively capturing various global features in CNN by including vision transformer (ViT) models in classification tasks [59], [68], [69]. Nevertheless, the substantial computational expense restricts the use of ViT in classification problems. Furthermore, ViT presents challenges in efficiently obtaining features of the local context. Liu et al. [70] obtained rich global–local building features through three parallel architectures. However, when the network acquires additional global features, it unintentionally brings more background noise. Therefore, GFE networks are frequently required for enhanced representation of building aspects via combining attention [44]. Liu et al. [46] combine attention and multiscale nested characteristics to more effectively represent multiscale features and spatial connections between buildings and surrounding areas. In summary, fusing global and local features is a practical building extraction method, but it is still ineffective for accurate feature extraction for complex scenes of buildings. ECA [71] efficiently acquires important features through feature interactions within a 1D convolution, and such interactions are crucial for enhancing the extraction of compelling discriminative features in complex building scenes. Therefore, inspired by ECA [71] and HDC [64], we constructed the GFI module. The GFI mainly uses the GFE module to extract GFE by successive parallel dilated convolutions. Besides, the GFI uses the original features to lessen the effects of feature discontinuity during global feature extraction. Finally, the GFI performs adaptive one-dimensional convolution via the ECI structure to interactively manipulate the global-local fusion features, thus enhancing the discrimination of building features in complex scenes.

3) Upsampling Optimized Structure: Networks with encoding and decoding structures will suffer the loss of features during upsampling [72]. SegNet [73] obtains the maximum value position through the MaxPooling operation and restores it by upsampling to increase the accuracy. The above methods fail to fully utilize the features generated by each layer during the upsampling process. FPN [74] enhances the feature utilization of each layer in the upsampling process by predicting and performing feature fusion layer by layer from the bottom to the top layer. Ji et al. [75] enhanced the upsampling process in a scale-invariant manner and applied it to building extraction. Inspired by distillation, Ma et al. [76] improved the inference speed by removing the first layer of upsampling. However, simple deletion may also lead to losing some critical information. Thus, in complex scenes, utilizing the features of each upsampling layer efficiently will help alleviate the loss of crucial information.

TABLE I
HFE ALGORITHM

```
Hierarchical-Context Feature Extraction
Input: Different hierarchical features $x_1x_2$, $x_1x_2 \in R^{C\times H\times W}$
Output: Hierarchical-context feature
1 Aggregate Hierarchical-context feature respectively:
  $y = f_{GAP}(x)$, $y_1y_2 \in R^{C\times 1\times 1}$
2 Obtain the weight of Hierarchical-context:
  $\omega = \sigma(C1D_k(y_1) + C1D_k(y_2))$
    $= \sigma(\sum_{j=1}^{k} W^j y_i^j + \sum_{m=1}^{k} W^m y_n^m), y_i^j \in \Omega_i^k, y_n^m \in \Omega_n^k$
3 Enhance Hierarchical-context Feature: $f_{HFE} = x_2\omega$
Return $f_{HFE}$
```

III. Methodology

A. Model Overview

The CSA-Net is proposed for better building extraction in complex scenes. Fig. 2 shows the overall structure of CSA-Net. It consists of the encoder, the HFE module, the GFI module, the MFF structure, the decoder, and the skip connections.

B. Hierarchical-Context Feature Extraction

Inspired by ECA-Net [71] and CHM [51], we design the HFE module to fully leverage crucial structural features and augment the discernibility amidst intricate backgrounds. Fig. 3 and Table I show the structure and algorithm of HFE, respectively. Specifically, we first carry out global average pooling of two hierarchies features $x_1, x_2 \in R^{C\times W\times H}$ in the same downsampling stage to obtain aggregate features y_1 and y_2. Then, 1D convolution $C1D$ is used to interact with k channels adjacent to y_1 and y_2, respectively, and weights W^j and W^m are assigned to the channels, and the mixed weight ω is obtained by merging the interaction results as follows:

$$\begin{aligned}\omega &= \sigma(C1D_k(y_1) + C1D_k(y_2))\\ &= \sigma\left(\sum_{j=1}^{k} W^j y_i^j + \sum_{m=1}^{k} W^m y_n^m\right), y_i^j \in \Omega_i^k, y_n^m \in \Omega_n^k \quad (1)\end{aligned}$$

where y_1 and y_2 represent the result of $f_{\text{GAP}}(\cdot)$ (global averaging pooling) for input features $x_1, x_2 \in R^{C\times W\times H}$. $C1D(\cdot)$ denotes 1D convolution. σ is the sigmoid activation process. Ω_i^k indicates the set of k adjacent channels of y_i [71], and k is calculated in (2). W^j and W^m represent the weight of the corresponding channel

$$k = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{\text{odd}} \quad (2)$$

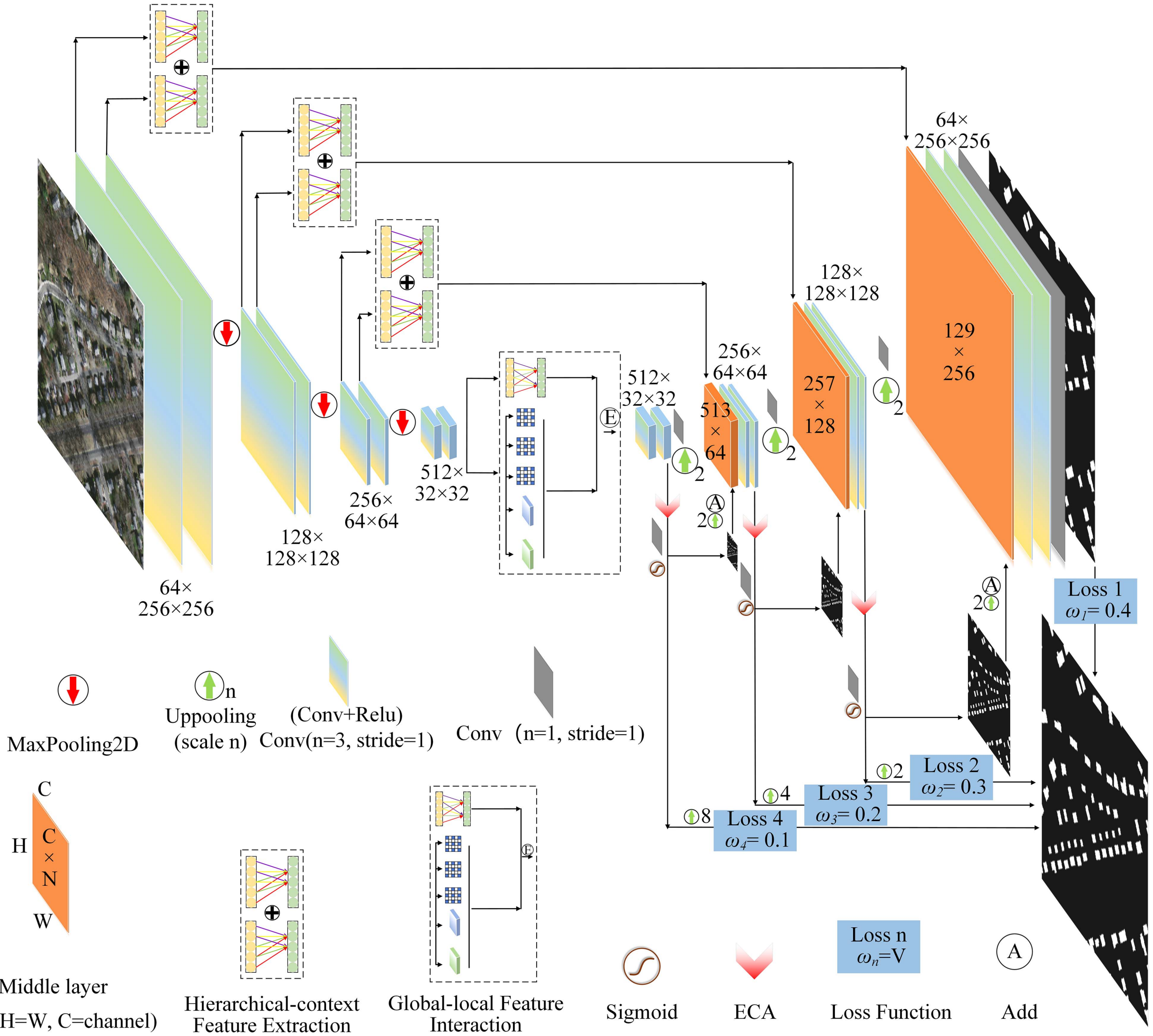


Fig. 2. Overall structure of CSA-Net. The HFE is at the skip connection, the GFI is mosaiced with the middle of the model, and the MFF is fused in the image restoration stage.

where$C$ is the channel dimension, $|t|_{\text{odd}}$ is the nearest odd number of t. In this study, we kept $\gamma$and b at 2 and 1, respectively, throughout all the experiments.

Finally, given that deeper layers contain richer high-level semantic information, the fusion weight ω is dot-multiply with the deeper feature input x_2 as follows:

$$f_{\text{HFE}} = x_2 \omega \tag{3}$$

where f_{HFE} is the output of the HFE module, $x_2 \in R^{C\times W\times H}$ represents deeper feature input, and ω is the fusion weight.

With the HFE, important structural features can be guided by jump connections to lower-level features and enhance the derivation of building integrity.

C. Global-Local Feature Interaction

To discern building from nonbuilding areas by leveraging the abundant high-level features of the underlying layer, we explored the GFI module inspired by HDC [64] and ECA-Net [71]. As shown in Fig. 4 and Table II, the GFI module consists of two main parts: the GFE module and the ECI structure.

Initially, the GFE module is employed to extract comprehensive global-local features. The GFE, defined in (4), comprises five parallel structures. To obtain rich and relatively continuous global-local feature information, the first three parallel structures incorporate continuous dilated convolution, with dilation rates set at (1, 2, 3), (1, 3, 5), and (1, 3, 9), respectively. The fourth structure utilizes global average pooling to complement channel-level features. To mitigate the feature discontinuity caused by dilated convolution, the fifth structure is employed to reintegrate original data and supplement the features

$$\begin{aligned} f_{\text{GFE}} = {} & ((\phi_3^3\phi_3^2\phi_3^1\theta_1^1(x)) + (\phi_3^5\phi_3^3\phi_3^1\theta_1^1(x)) \\ & + (\phi_3^9\phi_3^3\phi_3^1\theta_1^1(x)) + (\phi_1^1 f_{\text{GAP}}(\theta_1^1(x))) + \theta_1^1(x))\theta_1^1 \end{aligned} \tag{4}$$

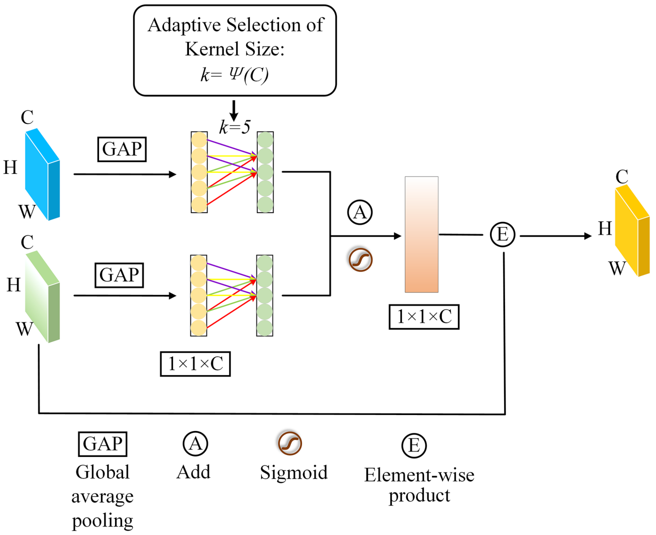


Fig. 3. Structure of the HFE module.

TABLE II
GFI ALGORITHM

```
Global-Local Feature Interaction
Input: Output feature of the decoder $x$ , $x \in R^{C\times H\times W}$
Output: Enhance Global-local feature
1 Global-local feature extraction:
  $f_{GFE} = ((\phi_3^3\phi_3^2\phi_3^1\theta_1^1(x)) + (\phi_3^5\phi_3^3\phi_3^1\theta_1^1(x)) +$
  $(\phi_3^9\phi_3^3\phi_3^1\theta_1^1(x)) + (\phi_1^1 f_{\mathrm{GAP}}(\theta_1^1(x))) + \theta_1^1(x))\theta_1^1$
2 Obtain the weight of global-local feature by the ECI:
  $\omega=\sigma(C1D_k(f_{GAP}(f_{GFE}(x))) + C1D_k(f_{GAP}(x)))$
  $\quad =\sigma(C1D_k(y_a) + C1D_k(y_b))=\sigma(\omega_a+\omega_b)$
3 Enhance Global-local feature: $f_{GFI} = \omega f_{GFE}(x)$
Return $f_{GFI}$
```

where f_{GFE} represents the result of GFE. ϕ_K^d denotes dilated convolution and batch normalization processing. K is the size of the convolution kernel, and d is the dilate rate. θ_1^1 represents a convolution operation with a convolution kernel size of 1 × 1. $f_{\mathrm{GAP}}(\cdot)$ represents a global average pooling operation and $x \in R^{C\times W\times H}$ is an input feature.

The weight ω through the ECI will then be used to reinforce the area of concern. Specifically, through $f_{\mathrm{GAP}}(\cdot)$, the global-local features obtained by the GFE are fused to get y_a. Adaptive 1D convolution is employed to integrate the features of k channels adjacent to y_a, with a heightened emphasis on the building area by assigning a higher weight. The output is denoted as ω_a. However, f_{GFE} is not completely continuous. Consequently, the original input feature x for global average pooling is used to derive y_b, followed by interactive operations to yield the weighted result ω_b. Finally, ω_b is employed to complement ω_a, resulting in the acquisition of the fused weight ω as follows

$$\begin{aligned}\omega &= \sigma(C1D_k(f_{\mathrm{GAP}}(f_{\mathrm{GFE}}(x))) + C1D_k(f_{\mathrm{GAP}}(x))) \\ &= \sigma(C1D_k(y_a) + C1D_k(y_b)) = \sigma\left(\omega_a + \omega_b\right)\end{aligned} \tag{5}$$

where ω is the fused weight, $C1D_k(\cdot)$ is 1D convolution.$f_{\mathrm{GAP}}(\cdot)$ represents global average pooling operation.$x \in R^{C\times W\times H}$ denotes input feature. $f_{\mathrm{GFE}}(\cdot)$ indicates the GFE process. k is the adaptive convolution kernel, calculated in (2) [71]. ω_a is the weight of y_a after the interaction. ω_b is the weight of y_b after the interaction. σ is the sigmoid activation function.

Finally, f_{GFE} obtained after the GFE module is dot-multiplied by the global-local interactive fusion weight ω, and the feature f_{GFI} extracted by the GFI module is output

$$f_{\mathrm{GFI}} = \omega f_{\mathrm{GFE}}(x) \tag{6}$$

where f_{GFI} is the result of the GFI module. ω is the weight. $f_{\mathrm{GFE}}(\cdot)$ is the GFE process.$x \in R^{C\times W\times H}$ is the input feature.

D. Multiscale-Adaptive Feature Fusion

In recovering feature maps through the U-shaped structure, encompassing layers from low to high, and adaptively amalgamating crucial features from each layer, we introduce the MFF structure. This design enhances the robustness of the network for extracting buildings within complex scenes through an extended fusion approach.

Fig. 5 illustrates the MFF structure. Inspired by FPN [74], the MFF aims to fully leverage the features of each layer during the upsampling to mitigate the loss of crucial features. Initially, the MFF employs effective attention mechanisms to eliminate interference. Moreover, a 1 × 1 convolutional layer and a sigmoid function, denoted by the purple arrows, are positioned at the end of each output path to generate predictions. The anticipated results, excluding the final level, undergo double upsampling. The blue arrow signifies the merging of the upsampled result with the output from the skip connection. As demonstrated by the orange arrows, the output of each level is subjected to a loss function with distinct weights and is then restored to the original image size through upsampling.

E. Multilevel Loss Function

The loss function plays a pivotal role in quantifying the disparity between the actual value and the predicted value, significantly influencing the training of neural networks. The loss functions for building extraction is binary cross entropy loss (BCE loss) [77] as follows

$$\mathcal{L}_B = -\frac{1}{N}\sum_{i=1}^{N}(G_i \times \log p_i + (1 - G_i) \times \log(1 - p_i)) \tag{7}$$

where N is the total number of pixels in the picture. $\mathcal{L}_B$ is the BCE loss. G_i represents whether the ith pixel belongs to a building or not.$G_i = 1$ if it is part of a building; otherwise, $G_i = 0$. p_i is the likelihood that a building will appear as the ith pixel in the anticipated outcome.

Given that CSA-Net employs the MFF structure and forecasts outcomes at each level of the expansive route, the ultimate loss function must be established by harmonizing the loss functions associated with each expected result. The $\mathcal{L}_{\mathrm{CSA-Net}}$ is defined as

$$\mathcal{L}_{\mathrm{CSA-Net}} = \sum_{n=1}^{4}\omega_n\mathcal{L}_n. \tag{8}$$

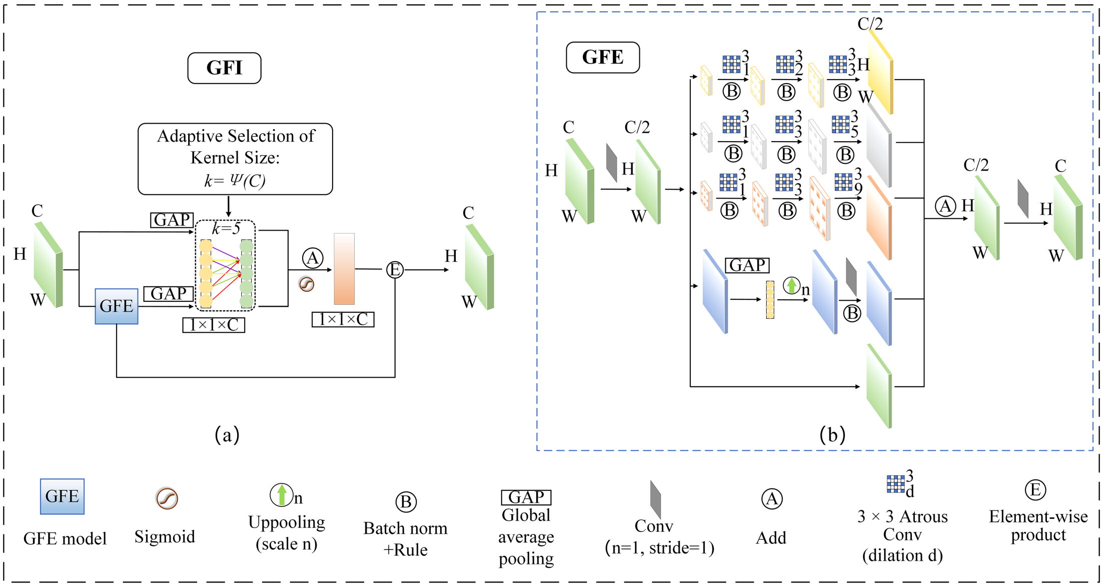


Fig. 4. GFI overall structure. (a) Structure of the GFI. (b) Structure of the GFE.

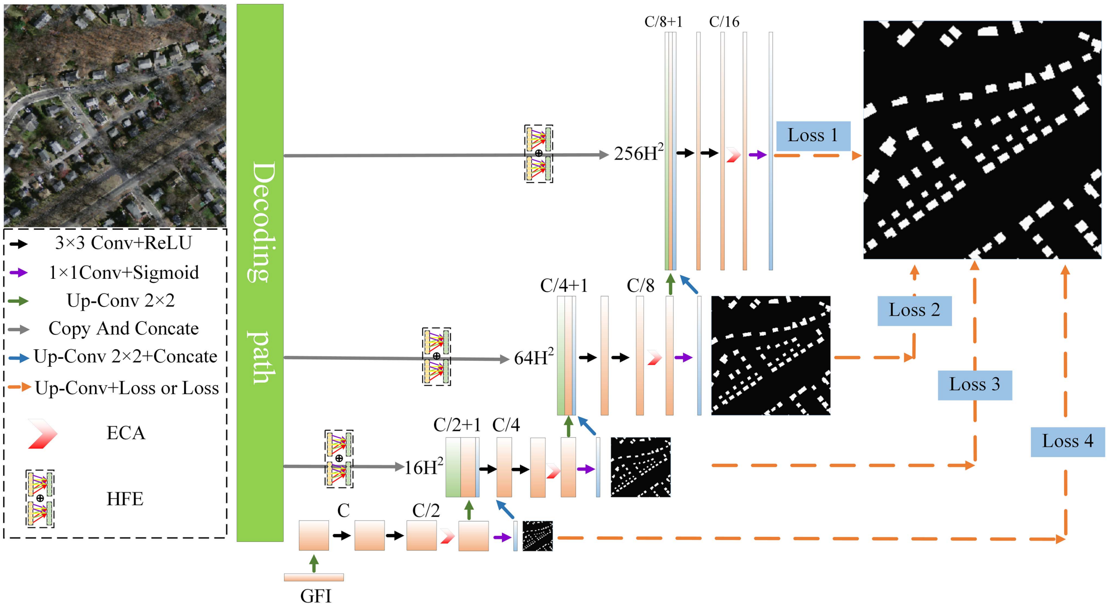


Fig. 5. Architecture of the MFF.

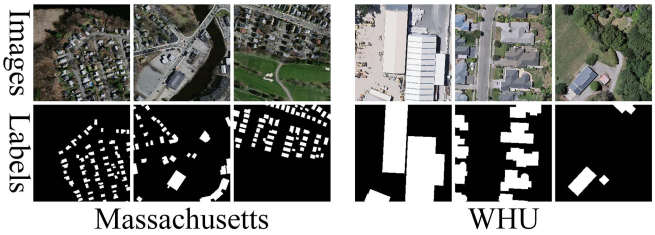


Fig. 6. Some samples of the WHU and Massachusetts datasets.

where $\mathcal{L}_n$ represents the loss function of CSA-Net from the bottom of the network to the output level n at the final output, and ω_n represents the weight of the output level n.

IV. Experiments and Results

A. Experimental Datasets Introduction

The GA, the WHU [78], and the Massachusetts [79] dataset were the three building datasets we employed in this experiment. Table III provides the detailed dataset configuration.

1) Public Dataset: Fig. 6 shows some samples of the WHU and Massachusetts datasets. The WHU dataset [78] has a resolution of 0.3 mers and an area of 450 km^2. A total of 8189 images with a size of 512 × 512 are included. We downsized all images from 512 × 512 to 256 × 256. Finally, it contains 18 944, 4144, and 9664 images in training, validation, and test sets.

The Massachusetts dataset constructed by Mnih [79] has a resolution of 1m, a size of 1500 × 1500, and a total of 151

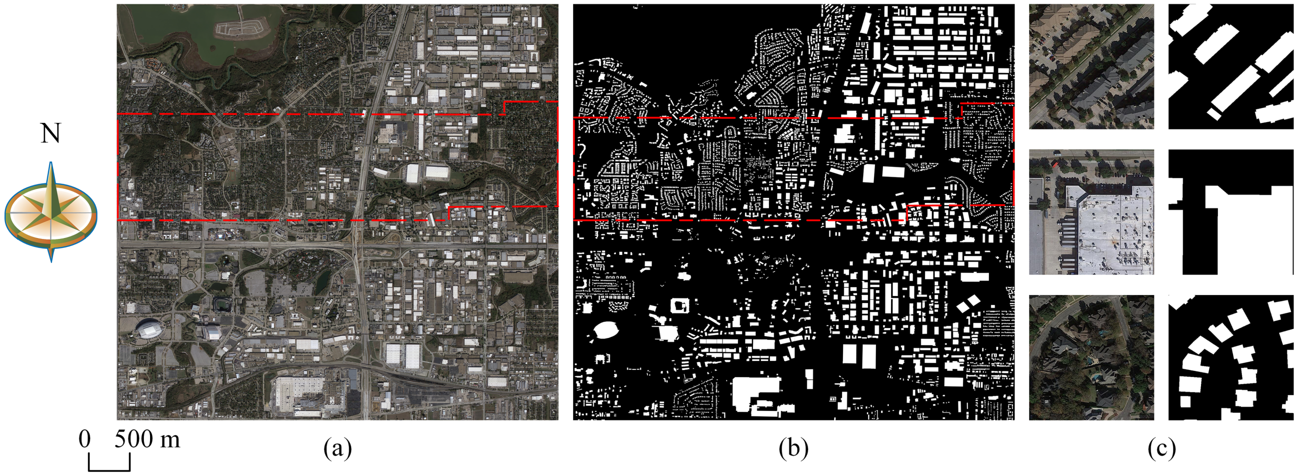


Fig. 7. Information of GA dataset. (a) and (b) Overviews of the GA dataset, with the light red portion being the test set. (c) Displays some samples from the GA dataset.

TABLE III
DATASET CONFIGURATION

Datasets	Resolution	Train	Validation	Test	Size(pixel)
WHU	0.3m	18944	4144	9664	256×256
Massachusetts	1m	4392	144	360	256×256
GA	0.46m	5112	532	2000	256×256

images. We cropped all of the images to a size of 256 × 256. We have 360 images for testing, 144 for validation, and 4392 for training.

2) Google Arlinton Building Dataset: To better test the generalization ability of our model in practical applications, we created a Google Arlinton (GA) dataset, as shown in Fig. 7. The GA is collected from Northeast Arlington and obtained via Google Earth with a resolution of 0.46 m. We divided the whole image into 7644 RGB images of the same size 256 × 256. Concurrently, they are separated into 2000 test sets, 532 validation sets, and 5112 training sets. Fig. 7 depicts the test set with a red dashed box. Because the buildings in the test area are relatively concentrated, and the building scenes contain more tree occlusion, we can better evaluate the generalization of the model in complex scenes.

B. Implementation Details

We apply the TensorFlow 1.14 and Keras 2.24 frameworks to implement the CSA-Net and nine additional contrast designs. Additionally, this computer has NVIDIA RTX 2080 Ti GPUs for training acceleration. The parameter settings are shown in Table IV.

Five metrics are defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{9}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{10}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

TABLE IV
PARAMETER SETTINGS

Parameter Settings	
Optimizer	Adam [80]
Initial Learning rate	0.0001
Batch size	6
Epoch for the Massachusetts dataset	200
Epoch for the WHU dataset	50
Epoch for the Owner-drawing dataset	200
Evaluation Indicators	OA, Precision, Recall, F_1-score, and IoU

$$\mathrm{F_1{-}score} = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{12}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{13}$$

where TP is true-positive, FP represents false-positive, TN denotes true-negative, and FN is false-negative.

C. Detail Settings of CSA-Net

We also provide detailed parameters for the CSA-Net in Table V, which include encode, skip, decode, and upsample-optimization.

D. Test Result and Analysis

In this part, we use nine comparative methods to prove the effectiveness of CSA-Net on three datasets and conduct a comparative analysis of the results of each method.

1) Test on the WHU Building Dataset: Table VI displays quantitative results on the WHU dataset. Regarding the intersection over union (IoU) and F_1-score, our proposed method outperforms the basic model (U-Net [21]) by 4.41% and 2.51%, respectively. In addition, the IoU is 6.14% higher than DeepLabV3+ [63] and 2.93% higher than ASF-Net [23], while the F_1-score by CSA-Net is 3.53% better than DeepLabV3+ and 1.65% better

TABLE V
CSA-Net Detailed Settings

Layer#		Layer Type	Output shape
	Input		3×H×W
ENCODE	E1	2×Conv2D(Input,relu,f=64,k=3,d=1,s=1)	64×H×W
	E2	Maxpooling2D(E1,pool_size=(2,2))	64×H/2×W/2
	E3	2×Conv2D(E2,relu,f=128,k=3,d=1,s=1)	128×H/2×W/2
	E4	Maxpooling2D(E3,pool_size=(2,2))	128×H/4×W/4
	E5	2×Conv2D(E4,relu,f=256,k=3,d=1,s=1)	256×H/4×W/4
	E6	Maxpooling2D(E5,pool_size=(2,2))	256×H/8×W/8
	E7	2×Conv2D(E6,relu,f=512,k=3,d=1,s=1)	512×H/8×W/8
	E8	GFI(E7,f=1024)	1024×H/8×W/8
SKIP	S1	HFE(Input, E1)	64×H×W
	S2	HFE(E2, E3)	128×H/2×W/2
	S3	HFE(E4, E5)	256×H/4×W/4
	D1	2×Conv2D(E8,relu,f=512,k=3,d=1,s=1)	512×H/8×W/8
	D2	Upsampling2D(D1,size=(2,2))	512×H/4×W/4
	D3	Conv2D(D2,relu,f=256,k=1,d=1,s=1)	256×H/4×W/4
	D4	Concatenate(S3, D3, U4)	513×H/4×W/4
DECODE	D5	2×Conv2D(D4,relu,f=256,k=3,d=1,s=1)	256×H/4×W/4
	D6	Upsampling2D(D5,size=(2,2))	256×H/2×W/2
	D7	Conv2D(D6,relu,f=256,k=1,d=1,s=1)	128×H/2×W/2
	D8	Concatenate(S2, D7, U8)	257×H/2×W/2
	D9	2×Conv2D(D8,relu,f=128,k=3,d=1,s=1)	128×H/2×W/2
	D10	Upsampling2D(D9size=(2,2))	128×H×W
	D11	Conv2D(D10,relu,f=64,k=1,d=1,s=1)	64×H×W
	D12	Concatenate(S1, D11, U12)	129×H×W
	D13	2×Conv2D(D12,relu,f=64,k=3,d=1,s=1)	64×H×W
	D-Out	Conv2D(D13,sigmoid,f=1,k=1,d=1,s=1)	1×H×W
UPSAMPLE|OPTIMIZATION	U1	ECA(D1)	512×H/8×W/8
	U2	Conv2D(U1,sigmoid,f=1,k=1,d=1,s=1)	1×H/8×W/8
	U3	Upsampling2D(U2,size=(8,8))	1×H×W
	U4	Upsampling2D(U2,size=(2,2))	1×H/4×W/4
	U5	ECA(D5)	256×H/4×W/4
	U6	Conv2D(U5,sigmoid,f=1,k=1,d=1,s=1)	1×H/4×W/4
	U7	Upsampling2D(U6, size=(4,4))	1×H×W
	U8	Upsampling2D(U6, size=(2,2))	1×H/2×W/2
	U9	ECA(D9)	128×H/2×W/2
	U10	Conv2D(U9,sigmoid,f=1,k=1,d=1,s=1)	1×H/2×W/2
	U11	Upsampling2D(U10,size=(2,2))	1×H×W
	U12	Upsampling2D(U10,size=(2,2))	1×H×W
	U-Out	[U3, U7, Us11]	

TABLE VI
Evaluation Results of Ten Networks on the WHU Dataset (%)

Methods	OA	Precision	Recall	IoU	F$_1$-score
U-Net[21]	98.20	90.25	**94.00**	85.34	92.09
SegNet [73]	97.95	89.12	92.87	83.42	90.96
DeepLabV3+[63]	97.97	89.25	92.97	83.61	91.07
MAP-Net [44]	98.18	91.88	91.74	84.86	91.81
BRRNet [45]	98.23	92.12	91.95	85.24	92.03
ASF-Net [23]	98.41	91.93	93.98	86.82	92.95
FSAU-Net [81]	97.86	94.36	85.90	81.71	89.93
DMU-Net [82]	97.46	85.14	93.45	80.35	89.10
CFENet [83]	97.86	89.75	91.15	82.55	90.44
CSA-Net	**98.81**	**95.41**	93.80	**89.75**	**94.60**

Bold and underlined metrics are superior*.

TABLE VII
Evaluation Results of Ten Networks on the GA Dataset (%)

Methods	OA	Precision	Recall	IoU	F$_1$-score
U-Net [21]	94.63	88.85	80.48	73.10	84.46
SegNet [73]	95.31	89.62	83.88	76.45	86.65
DeepLabV3+[63]	94.75	92.93	76.90	72.65	84.16
MAP-Net [44]	95.49	**93.07**	81.22	76.59	86.74
BRRNet [45]	95.45	90.74	83.43	76.88	86.93
ASF-Net [23]	95.35	91.28	82.22	76.23	86.51
FSAU-Net [81]	94.71	85.84	84.85	74.44	85.34
DMU-Net [82]	93.86	81.76	85.20	71.59	83.44
CFENet [83]	94.67	82.82	**89.15**	75.24	85.87
CSA-Net	**96.05**	90.77	87.07	**79.99**	**88.88**

Bold and underlined metrics are superior*.

TABLE VIII
Evaluation Results of Ten Networks on the Massachusetts Dataset (%)

Methods	OA	Precision	Recall	IoU	F$_1$-score
U-Net [21]	94.01	85.33	82.06	71.91	83.66
SegNet [73]	93.54	82.83	82.55	70.49	82.69
DeepLabV3+[63]	93.45	82.60	82.24	70.10	82.42
MAP-Net [44]	93.74	82.41	84.54	71.62	83.46
BRRNet [45]	94.12	85.46	82.57	72.40	83.99
ASF-Net [23]	94.20	83.87	**85.38**	73.34	84.62
FSAU-Net [81]	93.44	84.49	79.45	69.34	81.89
DMU-Net [82]	91.29	79.93	71.29	60.47	75.36
CFENet [83]	92.07	78.37	79.45	65.17	78.97
CSA-Net	**94.47**	**87.27**	82.44	**73.59**	**84.79**

Bold and underlined metrics are superior*.

than ASF-Net. Thus, the CSA-Net outperforms other approaches in both IoU and F$_1$-score.

Figs. 8 and 9 are the outcomes of the building extraction of CSA-Net and the comparison method. Fig. 8 is the extensive area building extraction result, and Fig. 9 is the typical building extraction example. As shown in Figs. 9(a), (b), and 8(a), the comparison method easily classifies some roofs as nonbuildings due to the distinct spectral textures of the building roofs. Figs. 9(c), (d), (e), and 8(b) illustrate that some places have a spectrum and texture similar to buildings due to the closeness of materials to buildings, leading to false identification. In response to the issues above, the GFI module enhances the ability of CSA-Net to identify buildings by enhancing the acquisition of global-local features and integrating attention. The visualization results outperform the comparison method.

As illustrated in Fig. 9(f), when the foliage of trees obscures the structure of a building, the covered portion becomes less readily identifiable by the network. By obtaining contextual features at different levels, the HFE module improves the reasoning capacity of CSA-Net and reduces the interference caused by trees and shadows. The MFF structure further enhances the adaptability of the network to complex building scenarios by mitigating the loss of essential features.

2) Test on the GA Dataset: To further verify the generalization ability of the network in complex scenes, we tested the GA datasets with poor lighting and more occlusion and shadows. Table VII presents the quantitative comparative outcomes of

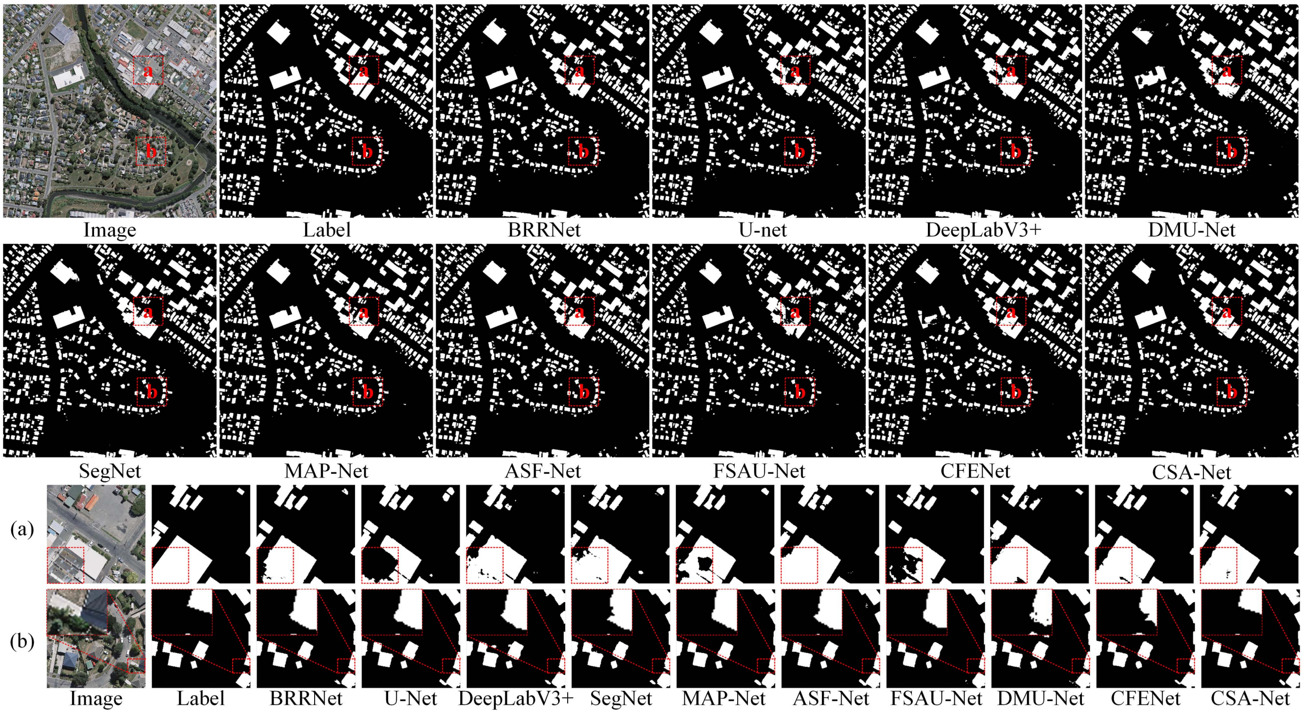


Fig. 8. Large-area building extraction findings from several approaches on the WHU dataset. (a) and (b) Enlarged images of the dotted red boxes.

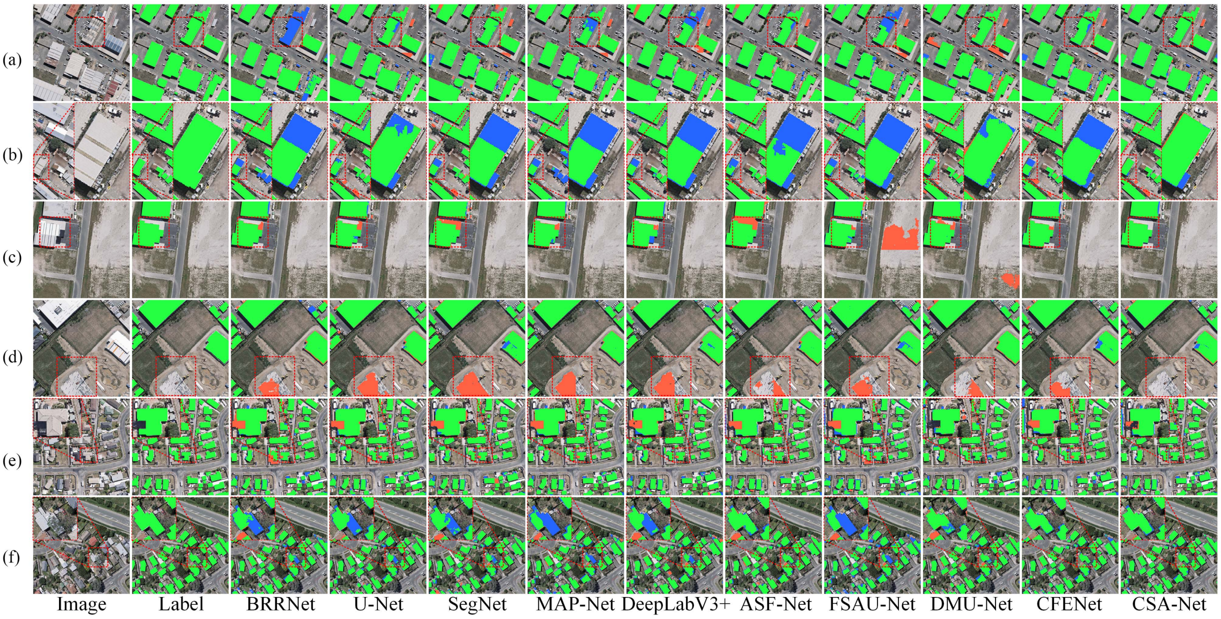


Fig. 9. Visual comparison of ten methods conducted on the WHU dataset. In the diagram, green is TP, red is FP, and blue is FN.

the ten approaches on the GA dataset, while Figs. 10 and 11 depict the visual results. As can be seen in Table VII, CSA-Net performs the best across all approaches, outperforming SegNet and DeepLabV3+ in the IoU by 3.54% and 7.34%, respectively, and outperforming the baseline model (U-Net) by 6.89%.

As can be seen from Figs. 10(a) and 11(a)–(c), (f), occlusion and shadow easily affect the integrity of building extraction. CSA-Net uses the HFE to help the network extract a more complete building, reducing the interference of occlusion and shadows. As shown in Figs. 10(a), (b) and 11(a), (b), (d)–(f), when the internal spectral texture of the building roof is different but similar to the surrounding area, it is easily identified as a nonbuilding. When the surrounding region has characteristics similar to those of buildings, it is easy to identify it as a building. The CSA-Net fusion of the GFI module improves building recognition by assessing global–local features and focusing on

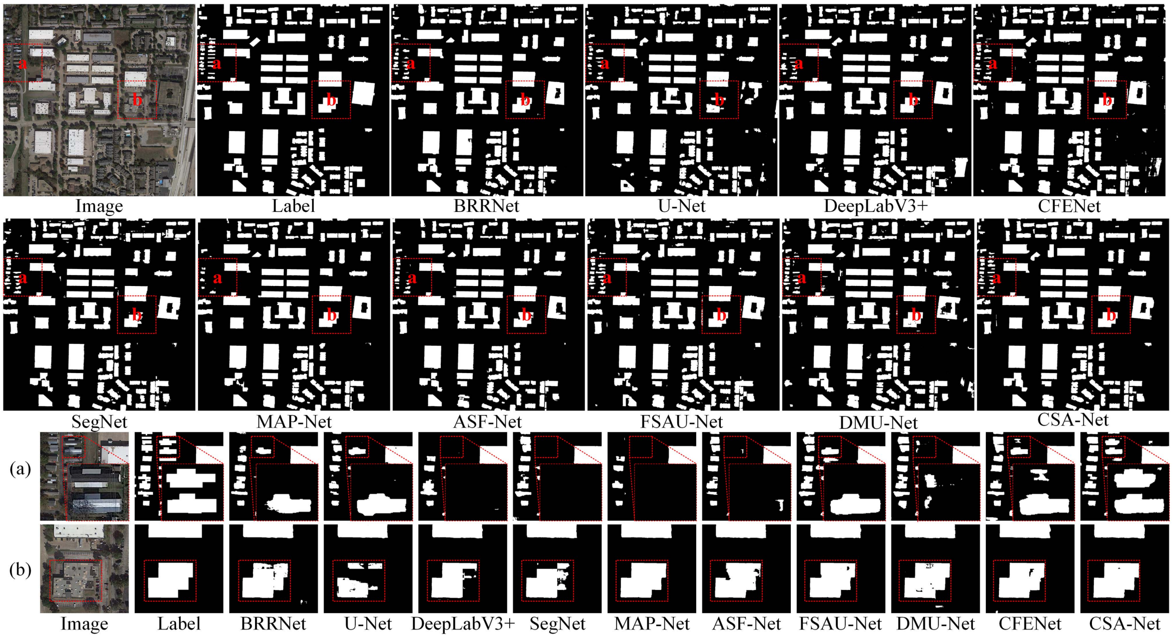


Fig. 10. Large-area building extraction findings from several approaches on the GA dataset. (a) and (b) Enlarged images of the dotted red boxes.

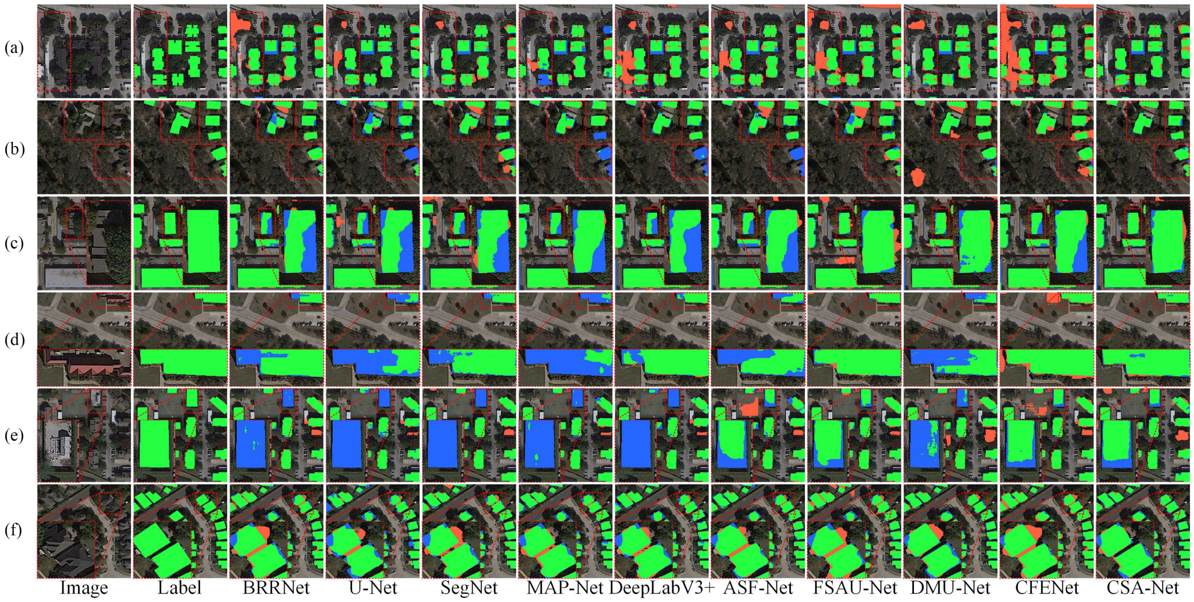


Fig. 11. Visual comparison of ten methods conducted on the GA dataset. In the diagram, green is TP, red is FP, and blue is FN.

the building component among them. Finally, combined with the MFF structure, our method further enhances its robustness in complex environments.

3) Test on the Massachusetts Dataset: Table VIII shows the evaluation results of ten networks on the Massachusetts dataset, and Figs. 12 and 13 are visualizations of the results. The evaluation results for the Massachusetts dataset are lower than for the WHU and GA datasets because they have a lower image resolution. Nevertheless, CSA-Net continues to perform best across all evaluation measures.

The visualization results show that CSA-Net has the least FP and FN. In red-framed regions in Figs. 12(a), (b), and 13(a)–(c), the buildings have similar textures and spectra to the surrounding areas. Fig. 13(d) and (e) depicts a foreign object on the roof of a building. The scenario above poses difficulty in accurately extracting buildings using the nine comparison approaches. CSA-Net employs the GFI to optimize the extraction of global–local features and integrates feature interactions to enhance building judgment. According to Fig. 13(f), CSA-Net incorporates the HFE modules to mitigate

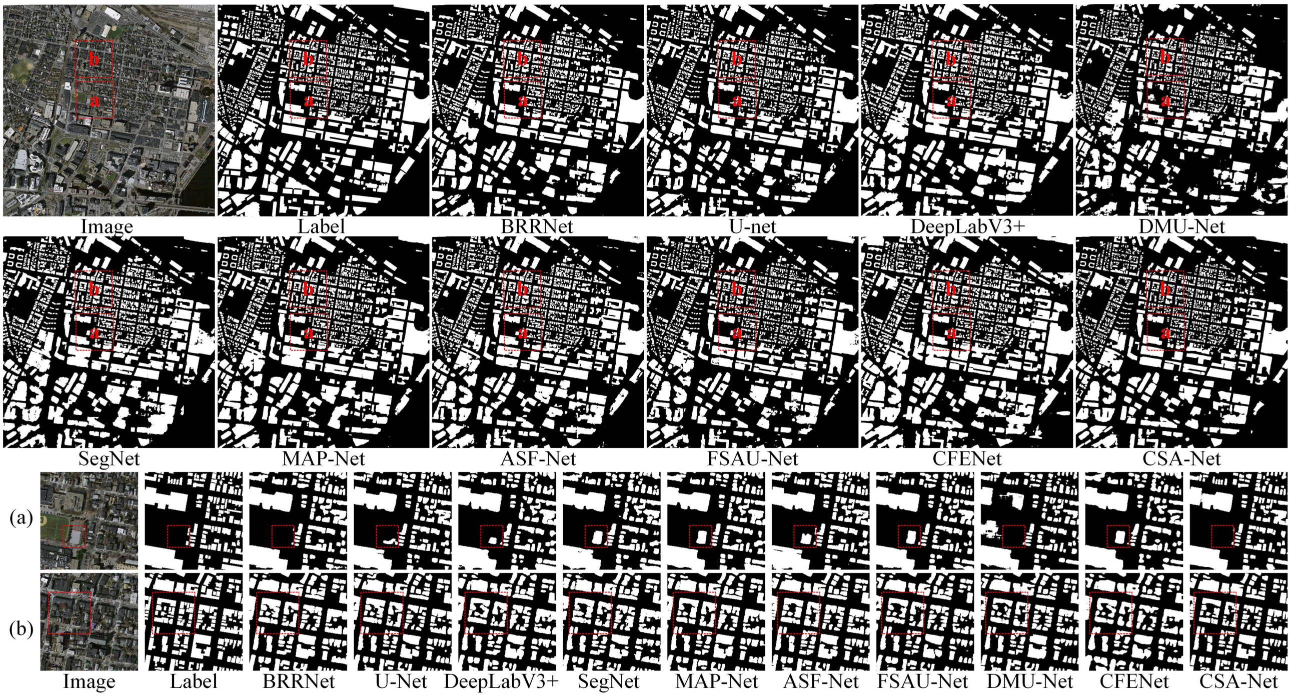


Fig. 12. Large-area building extraction findings from several approaches on the Massachusetts dataset. Lines (a) and (b) are enlarged images of the dotted red boxes.

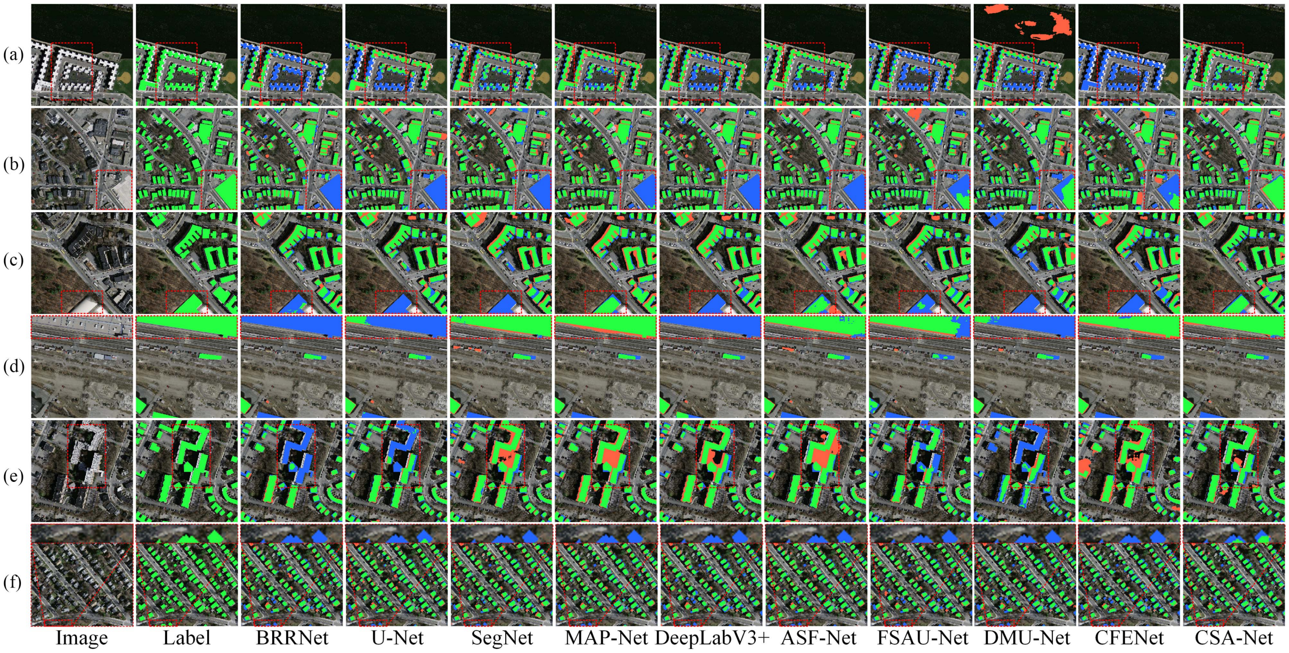


Fig. 13. Visual comparison of ten methods conducted on the Massachusetts dataset. In the diagram, green is TP, red is FP, and blue is FN.

interference from trees and shadow occlusion. In addition, by combining the MFF, our approach further improves segmentation accuracy.

E. Ablation Experiments of the CSA-Net

1) Ablation Experiments for Modules and Structures: To prove the effectiveness of the proposed HFE module, the GFI module, and the MFF structure for CSA-Net, we chose U-Net as the baseline model and conducted ablation experiments on the WHU dataset. Table IX shows the quantitative evaluation results of the continuous addition of modules. Fig. 14 shows the visualization results of the ablation experiment, and the areas of interest are marked in the red frame.

Network (c) demonstrates that when the GFI module was included in the U-Net, the baseline went from 92.09 and 85.34 in terms of F_1-score and IoU to 92.26 and 85.63. As evident from the red frame in the first and second lines of Fig. 14, some areas with similar characteristics to buildings are easily identified as buildings, while buildings with different internal textures

TABLE IX
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT MODULES ON THE WHU DATASET (%)

Networks	a	b	c	d	e	f	g	h
U-Net	✓	✓	✓	✓	✓	✓	✓	✓
HFE	×	✓	×	×	✓	×	✓	✓
GFI	×	×	✓	×	×	✓	✓	✓
MFF	×	×	×	✓	✓	✓	×	✓
OA	98.20	98.32	98.25	98.25	98.55	98.51	98.51	98.81
Precision	90.25	91.01	90.94	90.21	93.70	92.62	93.35	95.41
Recall	94.00	94.26	93.61	94.51	93.20	94.07	93.25	93.80
IoU	85.34	86.23	85.63	85.72	87.70	87.51	87.44	89.75
F_1-score	92.09	92.61	92.26	92.31	93.45	93.34	93.30	94.60

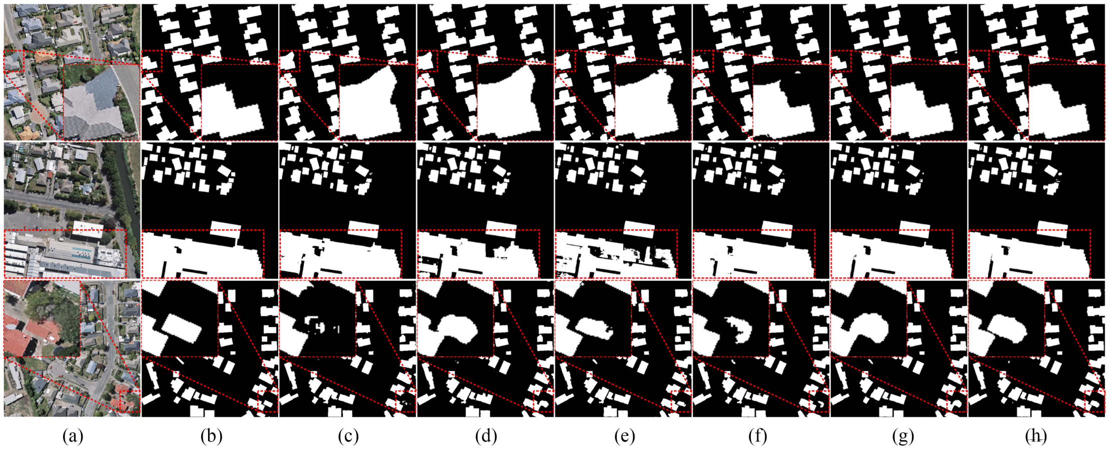


Fig. 14. Visualized results of ablation experiments on the WHU. (a) Image. (b) Label. (c) U-Net. (d) U-Net+ HFE. (e) U-Net+ HFE +MFF. (f) U-Net+GFI. (g) U-Net+GFI+MFF. (h) CSA-Net.

TABLE X
MODULE WEIGHT ABLATION EXPERIMENTS (%)

Weight	OA	Precision	Recall	IoU	F_1-score
GFI-Weight	98.55	94.68	92.13	87.59	93.39
CSA-Net	98.81	95.41	93.80	89.75	94.60
HFE-Weight	98.51	92.62	94.07	87.51	93.34
CSA-Net	98.81	95.41	93.80	89.75	94.60

and spectra are easily identified as nonbuildings. The baseline network acquires more global features after adding the GFI, thus enhancing the identification of building parts. However, there are still some areas where the identification is wrong. As shown in Fig. 14 and network (f), the recognition ability of buildings is further improved when the MFF is added to the network.

Network (b) shows that adding the HFE to U-Net increases the F_1-score and IoU by 0.52% and 0.89%, respectively. As seen from the red frame in the third row in Fig. 14, it is easy to miss extraction when trees are around the building. CSA-Net mitigates this phenomenon by combining contextual reasoning with the HFE modules to help the network extract a more complete building. Meanwhile, introducing the MFF can further strengthen the robustness of the network. As shown in Fig. 14 and Network (e), when the MFF is added to the network, the semantic segmentation accuracy is further improved, which helps CSA-Net to extract building information more thoroughly.

Ablation experiments show that the performance of CSA-Net can be effectively improved by adding the HFE, the GFI modules, and the MFF structures one by one.

2) Ablation Experiments for Maximum Dilated Rate and Module Weight: To verify the robustness and feasibility of the maximum expansion rate and weights for the accuracy of CSA-Net, we do the corresponding ablation experiments on the WHU dataset. It is known through Table XI that segmentation accuracy is highest when the maximum dilated rate is 9. It is known through the weight ablation experiment in Table X that the addition of weights in the modules GFI and HFE improves the IoU by 2.16% and 2.24%, respectively. It is demonstrated through Table XI that successive parallel dilated convolution with a maximum dilated rate of 9 is more effective in obtaining global–local features. It is proved through Table X that the weights are essential for the robustness and classification accuracy of the network.

3) Ablation Experiments for Different Epochs: To verify the effect of different epochs on accuracy, we conducted ablation experiments with different epochs on different datasets.

TABLE XI
ABLATION EXPERIMENTS FOR MAXIMUM DILATED (%)

Dilate rate	OA	Precision	Recall	IoU	F$_1$-score
(1, 2, 3) (1, 3, 5) and (1, 3, 7)	98.66	93.96	94.01	88.65	93.99
(1, 2, 3) (1, 3, 5) and (1, 3, 9)	98.81	95.41	93.80	89.75	94.60
(1, 2, 3) (1, 3, 5) and (1, 3, 11)	98.79	94.66	94.46	89.68	94.56

TABLE XII
EPOCH ABLATION EXPERIMENT ON THE WHU DATASET (%)

Epoch	OA	Precision	Recall	IoU	F$_1$-score
10	97.31	85.20	91.72	79.12	88.34
20	98.29	95.02	89.36	85.36	92.10
30	97.86	87.22	**94.66**	83.13	90.79
40	98.78	94.45	94.59	89.62	94.52
50	**98.81**	**95.41**	93.80	**89.75**	**94.60**
60	98.54	92.67	94.38	87.82	93.52

Bold and underlined metrics are superior*.

TABLE XIII
EPOCH ABLATION EXPERIMENT ON MASSACHUSETTS DATASET (%)

Epoch	OA	Precision	Recall	IoU	F$_1$-score
40	93.74	**87.85**	77.18	69.74	82.17
80	94.09	86.16	81.48	72.05	83.75
120	94.13	86.98	80.67	71.97	83.70
160	94.37	86.35	**82.98**	73.36	84.63
200	**94.47**	87.27	82.44	**73.59**	**84.79**
240	94.36	86.57	82.63	73.25	84.56

Bold and underlined metrics are superior*.

TABLE XIV
EPOCH ABLATION EXPERIMENT ON GA DATASET (%)

Epoch	OA	Precision	Recall	IoU	F$_1$-score
40	95.33	90.27	83.26	76.40	86.62
80	95.90	90.87	86.06	79.21	88.40
120	95.86	**93.22**	83.24	78.49	87.95
160	96.01	91.73	85.78	79.62	88.65
200	**96.05**	90.77	**87.07**	**79.99**	**88.88**
240	96.03	92.94	84.55	79.45	88.55

Bold and underlined metrics are superior*.

TABLE XV
EVALUATION OF THE COMPUTATIONAL EFFECTIVENESS OF VARIOUS APPROACHES

Methods	Params (M)	FLOPs(M)	Time cost (h)
U-Net [21]	28.9	57.9	7.6
SegNet[73]	15.3	58.9	17.8
DeepLabV3+[63]	28	56.1	8.4
MAP-Net [44]	24	47.9	11.7
BRRNet [45]	17.3	17.3	15.2
ASF-Net [23]	31.3	62.7	12.0
FSAU-Net [81]	11.7	23.5	6.1
DMU-Net [82]	12.5	25.0	8.1
CFENet [83]	94.8	189	5.1
CSA-Net	20	40.1	9.5

Table XII shows the CAS-Net segmentation accuracy of the WHU dataset at different epochs. Although the segmentation accuracy of CSA-Net occasionally fluctuates as the epoch increases, the overall trend is to improve and stabilize gradually. The overall accuracy is highest at epoch 50, while the recall value is not the highest. Therefore, the result of the experiment indicates that epoch 50 is the most appropriate. Tables XIII and XIV show the segmentation results on the Massachusetts and GA datasets with different epochs, which indicate that epoch 200 is more suitable.

F. Comparison of Training Times and Parameters for Different Methods

Table XV presents a comparative analysis of the computational efficiency of different methods, encompassing model parameters and FLOPs. Additionally, Table XV provides insights into the total time required for each approach to complete 50 iterations of training on the WHU buildings dataset.

These test results show that CSA-Net outperforms existing state-of-the-art building extraction methods while requiring less parameterization than the U-Net. This is mainly because we replaced the last layer of U-Net with GFI. Our approach exhibits faster execution compared to SegNet, MAP-Net, BRRNet, and ASF-Net, albeit with a slight decrease in speed compared to DeepLabV3+. Simultaneously, the parameter number and FLOPs of our method are lower than that of DeepLabV3+, MAP-Net, SegNet, and ASF-Net and only marginally higher than BRRNet. Although the computational efficiency of FSAU-Net and DMU-Net is higher than our method, there is a relatively large gap in accuracy. In summary, CSA-Net strikes a commendable balance between efficiency and accuracy.

V. CONCLUSION

This study recommends a novel CSA-Net based on the HFE module, the GFI module, and the MFF structure for building extraction. The HFE modules are embedded in skipping connections to help networks derive more complete buildings in complex scenes by acquiring richer and more effective high-level features. The GFI module is explored to obtain rich global and local features in the encoding process and suppress the background part to enhance the characterization of building features. The primary objective in devising the MFF structure is to mitigate feature loss during the upsampling process, aiming to augment the network robustness, particularly in complex scenarios. In the three datasets used in this article, the IoU of CSA-Net is 79.99%, 89.75%, and 73.59%, respectively, outperforming other advanced methods with nearly the fewest parameter indicators. Moreover, the results of the visualization experiments demonstrate that CSA-Net is capable of resolving some issues associated with building extraction in complex scenes.

Although our method can achieve good extraction results for complex building scenes, our method requires a large number of labels to train, so semisupervised methods that reduce manual labels will be the focus of future building extraction research.

REFERENCES

[1] X. Gao, M. Wang, Y. Yang, and G. Li, "Building extraction from RGB VHR images using shifted shadow algorithm," *IEEE Access*, vol. 6, pp. 22034–22045, 2018.

[2] Z. Lv, T. Yang, T. Lei, W. Zhou, Z. Zhang, and Z. You, "Spatial–spectral similarity based on adaptive region for landslide inventory mapping with remote-sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4405111.

[3] B. Zhang, C. Wang, Y. Shen, and Y. Liu, "Fully connected conditional random fields for high-resolution remote sensing land use/land cover classification with convolutional neural networks," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1889.

[4] N. Zerrouki and D. Bouchaffra, "Pixel-based or object-based: Which approach is more appropriate for remote sensing image classification?," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2014, pp. 864–869.

[5] J. Li et al., "A review of building detection from very high resolution optical remote sensing images," *GIScience Remote Sens.*, vol. 59, no. 1, pp. 1199–1225, 2022.

[6] Y. Wang et al., "B-FGC-Net: A building extraction network from high resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 2, 2022, Art. no. 269.

[7] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[9] X. Zhang, W. Yu, and M.-O. Pun, "Multilevel deformable attentionaggregated networks for change detection in bitemporal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621518.

[10] Z. Lv, Z. Lei, L. Xie, N. Falco, C. Shi, and Z. You, "Novel distribution distance based on inconsistent adaptive region for change detection using hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4404912.

[11] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411115.

[12] Z. Lv, H. Huang, W. Sun, T. Lei, J. A. Benediktsson, and J. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2505405.

[13] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5612911.

[14] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intraimage and interimage context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4403712.

[15] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412012.

[16] Y. Zhang, Y. Yang, X. Gao, L. Xu, B. Liu, and X. Liang, "Robust extraction of multiple-type support positioning devices in the catenary system of railway dataset based on MLS point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5702314.

[17] X. Gao et al., "Two-stage domain adaptation based on image and feature levels for cloud detection in cross-spatiotemporal domain," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610517.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[22] Y. Zhou et al., "BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618617.

[23] J. Chen, Y. Jiang, L. Luo, and W. Gong, "ASF-Net: Adaptive screening feature network for building footprint extraction from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4706413.

[24] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[25] Z. Lv, P. Zhang, W. Sun, T. Lei, J. A. Benediktsson, and P. Li, "Sample iterative enhancement approach for improving classification performance of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 2500605.

[26] Y. Yu et al., "Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 895–899, May 2021.

[27] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, and T. Lei, "Novel land-cover classification approach with nonparametric sample augmentation for hyperspectral remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4407613.

[28] Q. Hu et al., "Automated building extraction using satellite remote sensing imagery," *Automat. Construction*, vol. 123, Mar. 2021, Art. no. 103509.

[29] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5215512.

[30] Y. Zhu, B. Huang, J. Gao, E. Huang, and H. Chen, "Adaptive polygon generation algorithm for automatic building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4702114.

[31] J. Chang, Y. Jiang, J. Li, M. Tan, Y. Wang, and S. Wei, "Building height extraction based on joint optimal selection of regions and multi-index evaluation mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5603113.

[32] J. Chang et al., "Object-oriented building contour optimization methodology for image classification results via generalized gradient vector flow snake model," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2406.

[33] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," *Photogram. Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.

[34] M. Turker and D. Koc-San, "Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 58–69, 2015.

[35] T. Zhang, X. Huang, D. Wen, and J. Li, "Urban building density estimation from high-resolution imagery using multiple features and support vector regression," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3265–3280, Jul. 2017.

[36] Y. Li et al., "An integrated system on large scale building extraction from DSM," *Int. Arch. Photogram., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 35–39, 2010.

[37] F. Dornaika et al., "Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors," *Expert Syst. Appl.*, vol. 58, pp. 130–142, 2016.

[38] M. Teimouri, M. Mokhtarzade, and M. J. Valadan Zoej, "Optimal fusion of optical and SAR high-resolution images for semiautomatic building detection," *GIScience Remote Sens.*, vol. 53, no. 1, pp. 45–62, 2016.

[39] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogram. Remote Sens.*, vol. 117, pp. 11–28, 2016.

[40] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sens. Environ.*, vol. 118, pp. 259–272, 2012.

[41] L. Zhang et al., "An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN," *Sensors*, vol. 20, no. 5, 2020, Art. no. 1465.

[42] Y. Li et al., "SSDBN: A single-side dual-branch network with encoder–decoder for building extraction," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 768.

[43] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[44] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[45] Z. Shao et al., "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 12, no. 6, 2020, Art. no. 1050.

[46] T. Liu et al., "Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, 2022, Art. no. 102768.

[47] X. Liu et al., "A lightweight building instance extraction method based on adaptive optimization of mask contour," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103420.

[48] W. Li et al., "Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, 2022, Art. no. 102970.

[49] S. Ran et al., "Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2794.

[50] Q. Geng, H. Zhang, X. Qi, G. Huang, R. Yang, and Z. Zhou, "Gated path selection network for semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2436–2449, 2021.

[51] M. Seyedhosseini and T. Tasdizen, "Semantic image segmentation with contextual hierarchical models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 951–964, May 2016.

[52] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "Decoupling semantic and edge representations for building footprint extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5613116.

[53] Q. Zhu et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS J. Photogram. Remote Sens.*, vol. 175, pp. 353–365, 2021.

[54] T. Yu et al., "ConvBNet: A convolutional network for building footprint extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2501205.

[55] D. Feng et al., "GCCINet: Global feature capture and cross-layer information interaction network for building extraction from remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 114, 2022, Art. no. 103046.

[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[57] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458.

[58] M. Gong et al., "Context-content collaborative network for building extraction from high-resolution imagery," *Knowl.-Based Syst.*, vol. 263, 2023, Art. no. 110283.

[59] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[60] Z. Lv, M. Zhang, W. Sun, J. A. Benediktsson, T. Lei, and N. Falco, "Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4411911.

[61] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: https://arxiv.org/abs/1506.04579

[62] W. Luo et al., "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4905–4913.

[63] L.-C. Chen et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[64] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.

[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.

[66] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[67] D. He et al., "Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113884.

[68] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo, "BCTNet: Bi-Branch cross-fusion transformer for building footprint extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402014.

[69] S. Chan, Y. Wang, Y. Lei, X. Cheng, Z. Chen, and W. Wu, "Asymmetric cascade fusion network for building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004218.

[70] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6106–6120, Jul. 2021.

[71] Y. Qiu et al., "AFL-Net: Attentional feature learning network for buildingextraction from remote sensing images," *Remote Sens.*, vol. 15, no. 1, Jan. 2023, Art. no. 95.

[72] X. Yin, Y. Li, and B.-S. Shin, "TGV upsampling: A making-up operation for semantic segmentation," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–12, Aug. 2019.

[73] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[74] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 936–944.

[75] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3308–3322, 2019.

[76] J. Ma et al., "Building extraction of aerial images by a global and multi-scale encoder-decoder network," *Remote Sens.*, vol. 12, no. 15, 2020, Art. no. 2350.

[77] P.-T. De Boer et al., "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, 2005.

[78] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[79] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Univ. Toronto, Toronto, ON, Canada, 2013.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

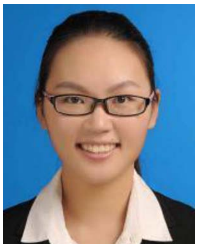
[81] M. Hu et al., "FSAU-Net: A network for extracting buildings from remote sensing imagery using feature self-attention," *Int. J. Remote Sens.*, vol. 44, no. 5, pp. 1643–1664, 2023.

[82] P. Li et al., "DMU-Net: A dual-stream multi-scale U-Net network using multi-dimensional spatial information for urban building extraction," *Sensors*, vol. 23, no. 4, 2023, Art. no. 1991.

[83] J. Chen et al., "A context feature enhancement network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2276.

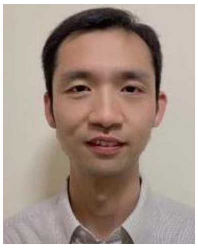
Dongjie Yang received the B.S. degree in mechanical manufacture and automation from the Changsha University, Changsha, China, in 2012. He is currently working toward the Master and Ph.D. degrees in geological engineering with the School of Geosciences, Yangtze University, Wuhan, China.

His research interests include intelligent interpretation of remote sensing Images and deep learning.



Xianjun Gao received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from the Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

She is currently an Associate Professor with the School of Geosciences, Yangtze University, Wuhan, China. She has more than 50 research papers. She is the holder of eight patents. Her major research interests include multisource remote sensing data automatic processing and high resolution images intelligent interpretation.



Yuanwei Yang received the Ph.D. degree in cartography and geographic information engineering from the Wuhan University, Wuhan, China, in 2016.

He is an Associate Professor and deputy Dean with the Department of Geographic Information Science, School of Geosciences, Yangtze University, Wuhan, China. He is the coauthor of more than 40 papers. His current research interests include mobile laser scanning data analysis, railway scene understanding, and semantic segmentation.

Minghan Jiang received the B.S. degrees in computer science and technology from the Hunan Institute of Science and Technology, Yueyang, China, in 2022. He is currently working toward the Master degree in geomatics engineering with the School of Geosciences, Yangtze University, Wuhan, China.

His major research interests include remote sensing content generation and multisource remote sensing data automatic processing.

Kangliang Guo received the Ph.D. degree in paleontology and stratigraphy from the University of Geology Beijing, Beijing, China, in 2004.

He is currently a Professor with the School of Geosciences, Yangtze University, Wuhan, China. He has more than 20 research papers. His current research interests include the geographic information system, geology of oil, and gas field development.

Bo Liu received the B.S. and M.S. degrees in cartography and geographic information engineering from the East China Institute of Technology, Jiangxi, China, in 2004 and 2007, and the Ph.D. degree in cartography and geographic information engineering from the Wuhan University, Wuhan, China, in 2016.

From 2007 to 2013, he was a Lecturer, and from 2013 to 2020, he was an Associate Professor, and since 2020, he has been a Professor with the East China University of Technology. He is the author of more than 30 peer reviewed journal articles. His research interests include theories of geographic information engineering, artificial intelligence, and LiDAR point clouds processing.

Shaohua Li received the Ph.D. degree in mineral survey and exploration from the China Petroleum Exploration and Development Research Institute, Beijing, China, in 2003.

He is a Professor with the School of Geosciences, Yangtze University, Wuhan, China. He has more than 160 research papers. His current research interests include geostatistics, reservoir modeling, and artificial intelligence for oil and gas.

Shengyan Yu received the B.S. degree in remote sensing science and technology and M.D. degree in geomatics engineering from the Wuhan University, Wuhan, China, in 2010 and 2014, respectively.

She is currently working with the Surveying and Mapping Geographic Information Center of Inner Mongolia Autonomous Region, Hohhot, China. She has five research papers and one patent. Her major research interests include natural resources remote sensing monitoring.