BEMRF-Net: Boundary Enhancement and Multiscale Refinement Fusion for Building Extraction From Remote Sensing Imagery

Shaohan Cao, Dejun Feng, Suning Liu, Wanqi Xu, Hongyu Chen, Yakun Xie, Heng Zhang, Saied Pirasteh, *Senior Member, IEEE*, and Jun Zhu

***Abstract*—Deep learning methods are widely used in building information extraction from remote sensing images (RSIs). However, this task still faces great challenges. First, it is difficult to perform accurate boundary localization due to the complex contextual relationship of the building boundary area. Second, the heterogeneity caused by different materials on building tops, as well as environmental factors such as climate and vegetation, complicates the extraction of top information. Finally, given the complexity of the RSIs, the accuracy and generalization of the existing method still need to be further enhanced. This study introduces the boundary enhancement and multiscale refinement fusion network to overcome these challenges. The boundary-aware self-attention module is initially proposed to refine the network's boundary detection capabilities. It applies the distance transform algorithm across both channel and spatial dimensions, reducing boundary fluctuation and enhancing the accuracy of building boundary extraction. Subsequently, we present the multiscale refined fusion module to resolve discontinuities within buildings caused by rooftop obstructions. This module effectively merges high and low-level features, overcoming information gaps through the strategic integration of multiscale data. To demonstrate the efficacy of our methodology, we conducted comprehensive experiments and analyses using the WHU Building Dataset and the Inria Aerial Image Labeling Dataset, both known for their complexity. Our method surpassed 16 contemporary state-of-the-art techniques, achieving IoU scores of 91.15% and 79.52% for each dataset, respectively, marking the highest accuracy levels reported. Furthermore, in-depth discussions on efficiency, generalizability, and ablation studies emphasized our method's robustness and adaptability.**

***Index Terms*—Boundary enhancement, building extraction, deep learning, multiscale refinement fusion.**

I. Introduction

BUILDING extraction from remote sensing images (RSIs) is of great significance, offering widespread applications, including urban planning analysis, population estimation, and the protection of resources and the environment [1], [2], Despite remote sensing technology's substantial potential for building extraction, challenges persist due to the inherent diversity of building structures and the variability in color, texture, and spectral features associated with different building materials [3]. These challenges necessitate ongoing technological innovation and enhancement.

Prior to the widespread adoption of machine learning technologies, extraction of buildings using traditional digital image processing techniques depended on manually predefined features or models. Researchers developed various methods, including image segmentation and edge detection, to separate buildings from their backgrounds by leveraging building edge information [4], [5]. Additionally, some approaches utilized the spectral and texture information of image pixels, analyzing each pixel's characteristics in remote sensing imagery to identify buildings [6], [7], [8]. However, these methods required extensive data volumes, hindering quick and convenient building information extraction. Generally, traditional techniques were subject to human subjectivity and data complexity, lacking in generalization capacity and computational efficiency to meet demands effectively.

Deep learning has made significant advances in computer vision in recent decades, garnering widespread interest among researchers [9], [10], [11], [12]. These models excel at autonomously learning and extracting complex features, thereby substantially enhancing performance in building identification and extraction tasks [13], [14]. Specifically, fully convolutional networks (FCNs) [15] have been widely used in processing RSIs. Presently, convolutional neural network (CNN) architectures such as ResNet [16], Faster R-CNN [17], and ViT [18]

Received 25 April 2024; revised 14 June 2024 and 7 July 2024; accepted 19 August 2024. Date of publication 22 August 2024; date of current version 20 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42301473, Grant 42201445, Grant 42171355, and Grant U22A20565, in part by the Postdoctoral Innovation Talents Support Program under Grant BX20230299, in part by the China Postdoctoral Science Foundation under Grant 2023M742884, in part by the Natural Science Foundation of Sichuan Province of China under Grant 24NSFSC2264, in part by the Natural Science Foundation of Tianjin under Grant 23JCQNJC01130, and in part by the Key Research and Development Project of Sichuan Province under Grant 24ZDYF0633. *(Corresponding author: Yakun Xie.)*

Shaohan Cao, Dejun Feng, Suning Liu, Wanqi Xu, Hongyu Chen, Yakun Xie, and Jun Zhu are with the Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China (e-mail: shcao@my.swjtu.edu.cn; djfeng@swjtu.edu.cn; sakura_ningning@163.com; xwq1207@my.swjtu.edu.cn; chy0519@my.swjtu.edu.cn; yakunxie@163.com; zhujun@swjtu.edu.cn).

Heng Zhang is with the China Railway Design Corporation, Tianjin 300251, China, also with the National Engineering Research Centerfor Digital Construction and Evaluation Technology of Urban Rail Transit, Tianjin 300251, China, and also with the School of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 610031, China (e-mail: zhangheng3s@163.com).

Saied Pirasteh is with the Institute of Artificial Intelligence, Shaoxing University, Shaoxing 312010, China, and also with the Department of Geotechnics and Geomatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai 602105, India (e-mail: sapirasteh1@usx.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3447788

are extensively utilized in remote sensing. These models have emerged as fundamental technologies for the classification and recognition of RSIs, effectively overcoming the constraints of traditional approaches by accurately identifying and classifying surface features [19], [20], [21], [22].

In the field of building extraction, many scholars have utilized the advantages of CNNs to propose new convolutional networks for achieving better extraction results [23], [24]. Specifically, the effect of building boundary extraction significantly impacts the overall extraction performance, leading numerous researchers to improve building segmentation accuracy through boundary information. Some research approaches attempt to simplify models' ability to capture boundary features by enhancing the expression of boundary information [25], [26], [27]. Xu et al. [28] introduced the MDBES-Net, which lowers the difficulty of extracting building boundary features, Although this method simplifies the capture of boundary information, it also results in a more complex network structure and a higher computational load. Other research approaches attempt to extract building boundary information by capturing boundary features at different resolutions [29], [30]. Lin et al. [31] addressed the insufficient feature extraction and poor model generalization of traditional methods by introducing the BEARNet, This approach enhances focus on boundary information; however, it does not resolve the issue of weakened boundary localization capabilities, which are diminished by the complex background information surrounding the boundaries, thereby reducing the effectiveness of boundary extraction.

Beyond boundary extraction, the integrity of building tops is another core focus in the field. Some studies address this by enhancing the overall description of building tops to capture more global information, though this often comes at the expense of local detail in RSIs [32], [33]. Subsequent researchers, such as Chen et al. [34], have attempted to prioritize local detail features in neural network models. Although this approach effectively utilizes different levels of feature layer, it overlooks the potential benefits of feature fusion across different scales. Later methods, including the GCCINet proposed by Feng et al. [35], employ continuous atrous convolutions and multiscale fusion to address issues such as missing small buildings and irregularities in building appearance extraction. This method enhances boundary extraction accuracy while mitigating the global relevance among distant pixels. Despite the improvements, relying solely on simple linear fusion of features with different resolutions can result in distortion or loss of features during the fusion process. This method struggles to address occlusions or heterogeneity at the top of buildings, often leading to incomplete extraction of rooftop details.

In complex scenes, building boundaries can cause confounding in the surrounding environment. As depicted in Fig. 1, the information on buildings is remarkably similar to that of roads, leading to unclear boundary information. Existing detection methods scatter attention around the building boundaries, causing fluctuations within a certain range. This results in weak capabilities for precise boundary localization and, consequently, an inability to accurately extract boundary information. Additionally, regarding integrity, current methods typically employ multiscale fusion techniques to address issues related to the completeness of building tops. However, challenges such as shadows obscuring the tops of buildings and the heterogeneity of those tops make it difficult for multiscale approaches to prevent issues such as internal holes or incomplete extractions. There is an insufficient focus on context-sensitive information, and the integration of feature information lacks refinement, as illustrated in Fig. 2. This often results in the presence of hole information at the tops of buildings.

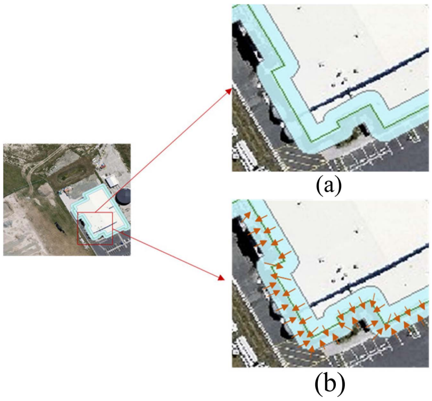


Fig. 1. (a) Part covered in blue is the perimeter of the building's boundary, within which the position of the boundary fluctuates (b) contraction of the range of boundary fluctuations of a building toward the boundary line.

To address the aforementioned issues, we propose the boundary enhancement and multiscale refinement fusion network (BEMRF-Net). This network employs the distance transform method to the network to concentrate more precisely on boundary localization. It utilizes multiscale information for refined fusion to enhance the network's capability to extract continuous information on building tops. The main contributions of this article are as follows.

1) BEMRF-Net is designed to address the challenge of how to efficiently extract buildings from RSIs. The method focuses on precise boundary localization of buildings and the extraction of internal contextual information. Extensive experiments and analyses were conducted on two datasets, comparing them with 16 state-of-the-art (SOTA) models and conducting a variety of challenging experiments to demonstrate the method's effectiveness and robustness.
2) The boundary-aware self-attention module (BSA) is proposed to solve the problem of weak positioning capabilities for building boundaries. This method uses the distance transform technique to direct the network model focus on the boundary messages, thereby narrowing the fluctuation range for locating building boundaries. Simultaneously, it combines boundary features with deep semantic features as auxiliary features of the model heightens the model's congnition of complex scenes and enhances the model's ability to accurately locate boundaries.
3) The multiscale refined fusion module (MSRF) is specifically designed to enhance the integrated extraction of buildings. Due to the complexity of scenes at the rooftop scenes, which often lead to holes in the extraction results,

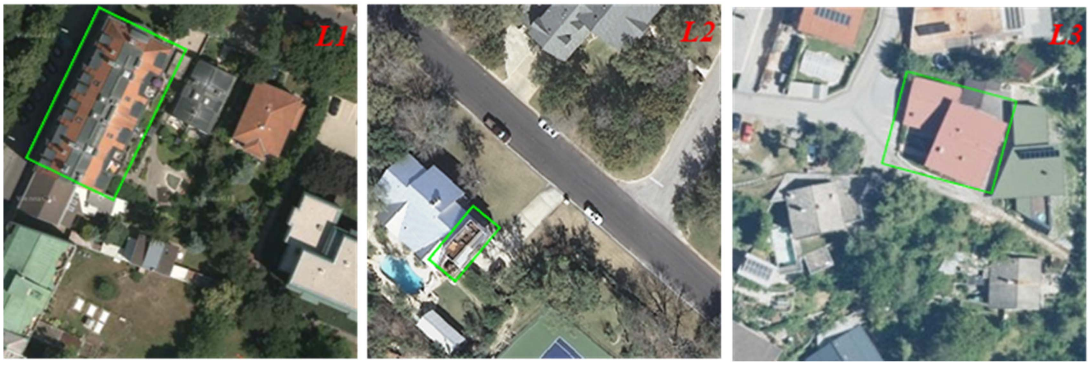


Fig. 2. (L1) Heterogeneity at the top of the building, (L2) clutter covering the top of the building, and (L3) shadow blocking at the top of the building.

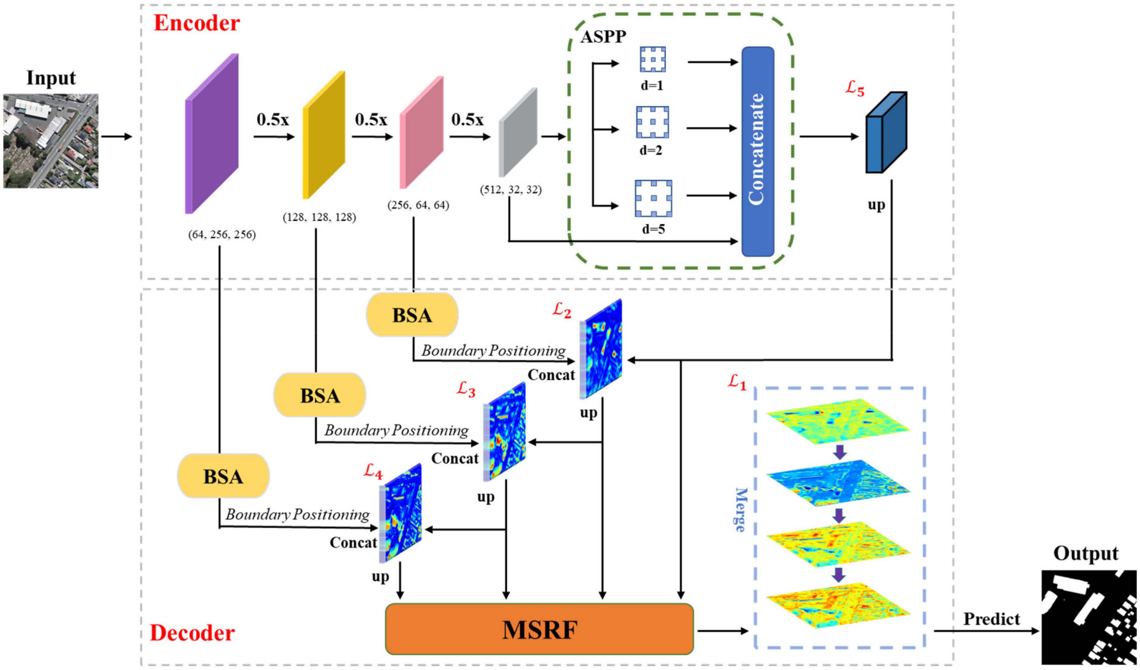


Fig. 3. Overview of the proposed BEMRF-Net.

this method achieves a refined fusion of different level features of the building target area. By obtaining local information while avoiding a decrease in the ability of extract global features, it effectively integrates contextual information at the tops of buildings, enhancing the completeness of information at the building tops.

II. Methodology

A. Network Overall

The entire framework of BEMRF-Net is illustrated in Fig. 3, primarily consisting of the BSA module and MSRF module. Initially, for feature extraction, the encoding phase utilizes the first four layers of VGG16, VGG16 as an encoder can provide a stable and powerful feature extraction base, and its simple and efficient structure can maintain good feature extraction efficiency, which is very suitable for dealing with building extraction tasks in remote sensing data. We employ the ASPP [36] after the encoder to facilitate global contextual inference between the encoder and decoder stages. Subsequently, to enhance the capability in distinguish buildings from their surroundings, the outputs from the first three layers of the encoder are input into the BSA module, which sharpens the delineation of building boundaries. Finally, our MSRF module is designed to preserve high-level semantic insights while retaining intricate low-level details, a balance that is crucial for accurate building extraction. The proposed modules' contributions to the building extraction task and their detailed functionalities are elaborated in the following.

B. Boundary-Aware Self-Attention Module

In order to accurately identify different ground buildings, boundary information is important for differentiating buildings from their surroundings and enhancing segmentation accuracy.

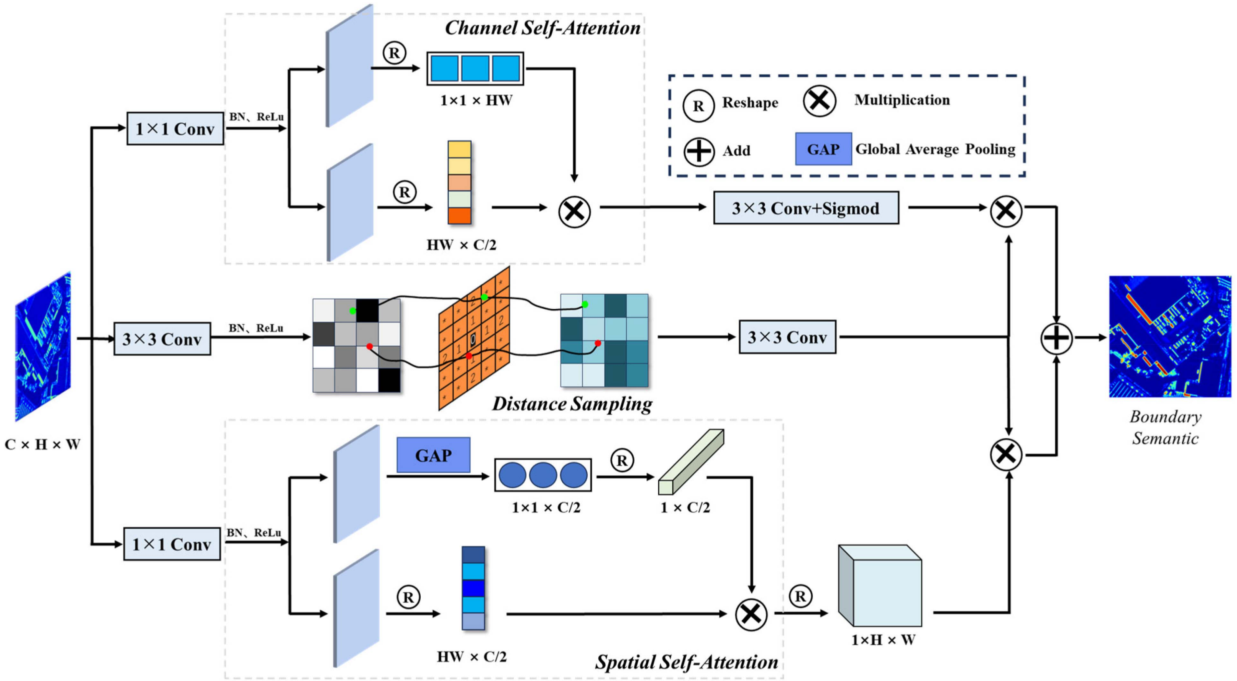


Fig. 4. Structure of the BSA module.

Background noise can distract the network model's focus away from the building itself. However, attention mechanisms can direct the network toward boundary information, effectively filtering out irrelevant data [37]. We introduce the BSA module, which utilizes the Euclidean distance transformation algorithm to emphasize boundary details, thereby increasing the model's sensitivity to boundary information. Unlike traditional attention mechanisms, our BSA module applies self-attention across both spatial and channel dimensions. This approach preserves a high internal resolution, enabling the capture of detailed low-level boundary features. By integrating the distance transform algorithm, the model's focus on building boundaries is sharpened, reducing the variability in boundary localization and achieving refined segmentation. As depicted in Fig. 4, tensors from the encoder (C, H, W) are directed into both channel and spatial self-attention branches following a 1 × 1 convolution and a boundary information branch that utilizes distance transformation after a 3 × 3 convolution.

Different weights are obtained after two parallel reshape operations in the channel self-attention branch, and the attention weights with channel information (C,11) are acquired after the Sigmoid function and multiplication. As shown in the following equations:

$$CA = R_1(f) \times R_2(f) \tag{1}$$

$$F_c = Sig\left[Conv_{3\times3} CA\left(Conv_{1\times1}(f)\right)\right] \tag{2}$$

where F_c represents the output feature map $F_c \in \mathbb{R}^{C\times1\times1}$ in the channel dimension, f represents the input feature map $f \in \mathbb{R}^{C\times H\times W}$, $CA(\cdot)$represents the channel self-attention, $R_i \cdot (i = 1, 2, 3, 4)$ represents the reshape operation and converts the feature maps into different formats, $Conv_{j\times j}(\cdot)(j = 1, j = 3)$ represents convolutional layers with different convolutional kernel sizes, and $Sig(\cdot)$ represents the Sigmoid activation function.

We first apply maximum downsampling and reshape operations to the tensor in the spatial self-attention branch, thereby generating attention weights with spatial information (1, H, W) through a subsequent reshape operation. Following this, we integrate the distance-transformed boundary information by multiplying it with the channel and spatial information separately, then aggregate these products. This results in a feature tensor where boundary information is intricately fused across both channel and spatial dimensions. The process is shown in the following equations:

$$SA = R_3(Gp(f)) \times R_4(f) \tag{3}$$

$$F_s = R_3\left[SA(Conv_{1\times1}(f))\right] \tag{4}$$

$$F_{final} = F_c \ \times \varphi \oplus F_s \times \varphi \tag{5}$$

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{6}$$

where F_s represents the feature $F_c \in \mathbb{R}^{1\times H\times W}$ output in the spatial dimension, $SA(\cdot)$ stands for Spatial Self-Attention, $Gp(\cdot)$ stands for Global Average Pooling, F_{final} stands for the final output of the module, φ stands for the distance employing the correction value obtained by the algorithm, and $\oplus$ stands for the element-by-element summation.

It is worth noting that the Euclidean distance transform algorithm is particularly well-suited for processing high-resolution RSIs. This is due to its effectiveness in feature extraction and its computational efficiency when dealing with irregularly shaped objects, compared with other distance measurement methods.

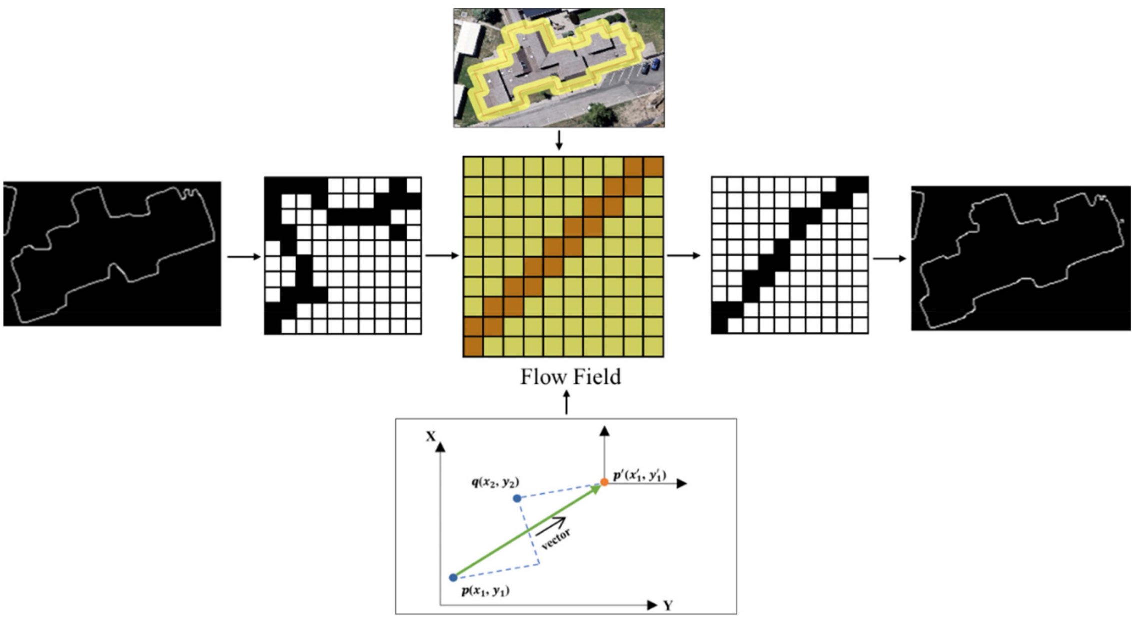


Fig. 5. Flow Field is generated by the distance transform algorithm and the actual boundary position information. The Flow Field guides the predicted boundary pixels to move toward the actual boundary pixels, narrowing the fluctuation range of boundary localization.

The Euclidean distance Transform algorithm enhances image analysis by updating each pixel's value to represent the nearest distance to a foreground pixel. Specifically, As illustrated in Fig. 5, to convert the images to binary format, we first reduce the feature map's channel count to one using a 1 × 1 convolution. Following this, we generate a Flow Field of identical dimensions to the input feature map. For every foreground pixel, we calculate the Euclidean distances (vectors) to the closest background pixel by (6), where p and q represent different pixels, and these distances are recorded within the Flow Field. This process assigns a semantic offset from each foreground pixel to its nearest background counterpart, effectively highlighting the semantic gap from the boundaries to the background. Subsequently, we iterate over each pixel in the feature map, adjusting it according to the semantic offsets derived from the Flow Field. This distance sampling algorithm amplifies the image's boundaries and narrows the model's variability in identifying building boundaries, thereby improving boundary localization in complex backgrounds. The pseudo-code for the Euclidean distance transform algorithm is provided in the following.

C. Multiscale Refined Fusion Module

The presence of irregularly shaped trees that obscure the tops of buildings can diminish the network's accuracy and robustness, resulting in discontinuities within the buildings' representations. To mitigate these disturbances, we enhance the model by integrating an ensemble of contextual information. Furthermore, we implement a multiscale fusion approach to prevent gradient vanishing or explosion, issues that stem from constraints associated with single-layer networks, and increased network depth [38].

Therefore, we developed the MRSF module. As illustrated in Fig. 6, this module takes the outputs from four different layers of the encoder as its inputs. Initially, feature fusion occurs between adjacent scale features in deeper layers. Subsequently, these fused features undergo further fusion and interaction with the features of the preceding layer, aiming to attenuate redundant features. This process facilitates the refined integration of multiscale features. The specific formulations are presented in the following equations:

$$\lambda_1 = SF\left(Up\left(V_1\right), V_2\right) \oplus Cbr_{1\times 1}\left(V_4\right) \tag{7}$$

$$\lambda_2 = SF\left(Up\left(\lambda_1\right), Up\left(V_3\right)\right) \oplus Cbr_{1\times 1}\left(Cat\left(Up\left(V_3\right), V_4\right)\right) \tag{8}$$

where $V_i(i = 1, 2, 3, 4)$ represents the high and low-level feature maps $V_i \in \mathbb{R}^{C\times H\times W}$ output from the encoder, $\lambda_i(i = 1, i = 2)$ represents the output of the ith time fusion, $SF(\cdot)$represents fusion submodule, $Cbr_{1\times 1}(\cdot)$represents the convolution, normalization, and activation, and $Up(\cdot)$ represents the upsampling operation.

To preserve the local details of the image, we combined low-level semantic features with the fusion results. Subsequently, we integrated the most profound semantic features with the outcomes of the two previous fusions, thereby reinforcing the model's emphasis on deep-level semantic information. Following the multiscale feature fusion, we obtained the final output feature map with 1 × 256 × 256 dimensions through a 1 × 1 convolution. The specifics of this process are detailed in (9), where Λ represents the final prediction map, and Cat represents the splicing operation.

$$\Lambda = Conv_{1\times 1}\left(Cat\left(Up\left(V_1\right), \lambda_2\right)\right). \tag{9}$$

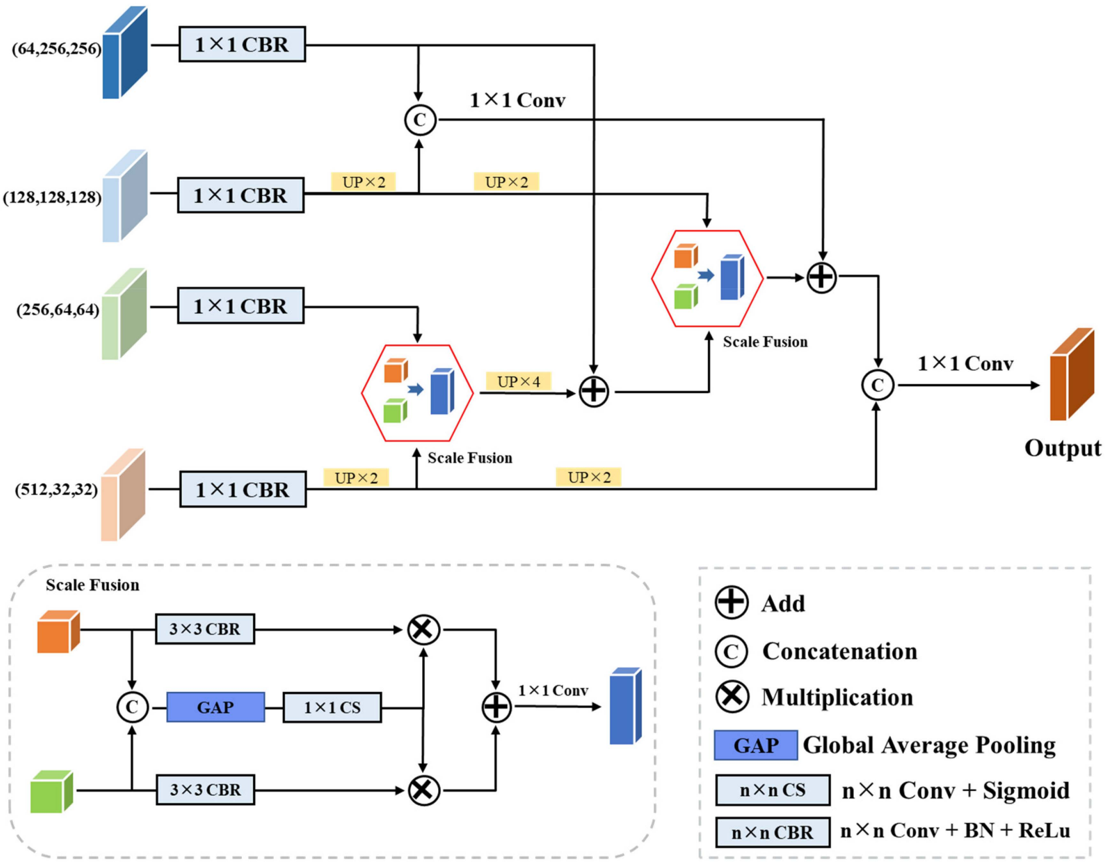


Fig. 6. Structure of the MSRF module.

As shown in the bottom half of Fig. 6, scale fusion is a fusion submodule. Initially, two input feature maps are linked in series along the channel dimension to merge the high and low features This is followed by global average pooling, 1 × 1 convolution, and application of the Sigmoid function, which are manipulated in such a way as to integrate semantic information from different feature layers. Subsequently, the feature information fused in the initial phase and the global context weights derived in the first phase are multiplied and then aggregated. The resultant feature maps, enriched with cross-layer interactions, embody information richness across multiple scales. The resultant representation of the scale fusion output is given in the following equations:

$$\rho = Sig\left[Conv_{1\times 1}\left(Gp\left(Cat\left(e_1, e_2\right)\right)\right)\right] \tag{10}$$

$$\psi = Cbr_{3\times 3}\left(e_1\right) \times \rho \oplus Cbr_{3\times 3}\left(e_2\right) \tag{11}$$

where $e_i (i = 1, i = 2)$ represents the two feature maps of the input submodule, ρ represents the result of the first step, and ψ represents the final output of the fusion submodule.

III. Data and Experimental Settings

A. Data

This article evaluates the effectiveness of BEMRF-Net using two datasets: the WHU Building Dataset [39] and the Inria Aerial Image Labeling Dataset [40], these datasets are used frequently in the reverse side of semantic segmentation. It should be noted that to enhance training outcomes, we have implemented data augmentation processes on both datasets. These include image-level flipping, center cropping, and grid distortion. These operations do not alter the datasets themselves but increase the training complexity and duration for the network. This approach enables the network model to more effectively discover the optimal fitting function.

WHU dataset: This dataset covers approximately 1 87 000 buildings. The data have a spatial resolution of 0.3 m and are clipped into 8189 samples of 512 × 512 pixels. We cropped the training and validation set images into 256 × 256 pixels to fully exploit the performance of our network.

Inria Aerial Image Labeling Dataset: This dataset covers several cities in Europe and the United States with a spatial resolution of 0.3 m. It consists of 360 images with a 1:1 ratio between the training and test sets, and the size of each image is 5000 × 5000 pixels. We crop each image to 256 × 256 pixels to enlarge the dataset.

B. Experiment Setting

The server configuration used for the experiments in this article is NVIDIA GeForce RTX 3090 24G GPU. Furthermore, a

Algorithm 1: Euclidean Distance Transform EDT.

```
Input: binary_image (2D array of 0s and 1s, where 0 = background, 1 = foreground)
Output: distance_map (2D array of distances to the nearest foreground pixel)
1. Initialize distance_map:
  for each pixel (y, x) in binary_image:
    if binary_image[y][x] = = 0:
      distance_map[y][x] = ∞
    else:
      distance_map[y][x] = 0
2. Forward pass:
  for y from 1 to height - 1:
    for x from 1 to width - 1:
      distance_map[y][x] = min(distance_map[y][x], distance_map[y-1][x] + 1, distance_map[y][x-1] + 1)
3. Backward pass:
  for y from height - 2 down to 0:
    for x from width - 2 down to 0:
      distance_map[y][x] = min(distance_map[y][x], distance_map[y+1][x] + 1, distance_map[y][x+1] + 1)
4. Final adjustment:
  for each pixel (y, x) in distance_map:
    distance_map[y][x] = sqrt(distance_map[y][x])
Return distance_map
```

deep supervision strategy was utilized to facilitate faster model convergence. Following iterative testing and adjustments, the batch size was set to 8, and Adam as the optimizer. The model was trained for 100 epochs.

We use the Binary Cross-entropy (BCE) loss function, commonly employed in binary classification tasks, as the loss function. Specifically, the loss function consists of five components: the loss of the final result prediction $Loss_1$, the losses of low-level fusion features at three stages $Loss_2$, $Loss_3$, $Loss_4$, and the loss of the ASPP output $Loss_5$ The loss function formula is as follows, where α and β are hyperparameters. After multiple tests, we found that setting them to 0.3 and 0.2, respectively, yields the best experimental results. We will specifically describe the selection process of these two hyperparameters in Section V-D.

$$Loss = 0.5 * Loss_1 + \alpha \sum_{i=2}^{4} Loss_i + \beta Loss_5. \tag{12}$$

C. Evaluation Metrics

We adopt the following five evaluation metrics: overall accuracy (OA), precision, recall, F1 score (F1), and intersection over union (IoU). Specifically, TP_{pixel} is True Positives of the pixel, FP_{pixel} is False Positives of the pixel, TN_{pixel} is True Negatives of the pixel, and FN_{pixel} is False Negatives of the pixel. Their mathematical expressions are as follows:

$$OA = \frac{TP_{pixel} + TN_{pixel}}{TP_{pixel} + TN_{pixel} + FP_{pixel} + FN_{pixel}} \tag{13}$$

$$Precision = \frac{TP_{pixel}}{TP_{pixel} + FP_{pixel}} \tag{14}$$

$$Recall = \frac{TP_{pixel}}{TP_{pixel} + FN_{pixel}} \tag{15}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precicion} \tag{16}$$

$$IoU = \frac{TP_{pixel}}{TP_{pixel} + FN_{pixel} + FP_{pixel}}. \tag{17}$$

Beyond the standard evaluation criteria previously mentioned, we adopt the Hausdorff distance (HD) [41] and the structural similarity index (SSIM) [42] to further assess our method's performance in boundary delineation and building integrity. These metrics specifically evaluate boundary consistency and visual similarity, offering more precise insights into the success of our building boundary segmentation approach.

HD is a measure used to quantify the difference between two sets of points and is commonly applied in image analysis, computer vision, and computational geometry. Specifically, given two sets of points M and N, HD can be shown by the following equations:

$$p = \sup_{m \in M} inf_{n \in N} d\left(m, n\right) \tag{18}$$

$$q = \sup_{n \in N} inf_{m \in M} d\left(m, n\right) \tag{19}$$

$$h\left(M, N\right) = \max\left\{p, q\right\}. \tag{20}$$

Here, sup represents the smallest upper limit in the set; inf represents the largest lower limit in the set; d represents the Euclidean distance. In order to reduce the bias caused by extreme values, HD is multiplied by 95% to serve as the final evaluation metric (95% HD).

SSIM is a number used to represent the degree of similarity between two images as viewed by the human eye, focusing primarily on the similarity of luminance, contrast, and structure. For two images, $\eta$and θ, the formula is as follows:

$$SSIM\ \left(\eta, b\right) = \left(\frac{2\mu_\eta \mu_\theta + H_1}{\mu_\eta^2 \mu_\theta^2 + H_1}\right) \times \left(\frac{2\sigma_{\eta\theta} + H_2}{\sigma_\eta^2 \sigma_\theta^2 + H_2}\right) \tag{21}$$

where μ_η, μ_θ are the average luminance of images η and θ, respectively; μ_η^2, μ_θ^2 are the variances of images η and θ, respectively; $\sigma_{\eta\theta}$ is the covariance of images a and b; and H_1, H_2 are small constants added to avoid division by zero.

IV. Results

A. Experimental Results

We compare six methods to verify the reliability of our method, The models selected for comparison were DeepLabv3+ [43], SegNet [44], BMFR-Net [45], GCCINet [35], C3Net [46], and MEC-Net [47]. The results of this comparative analysis are detailed in Tables I and II.

TABLE I
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE WHU DATASET

Method	Common metrics					Boundary metrics	
	IoU(%)↑	F1(%)↑	Precision(%)↑	Recall(%)↑	OA(%)↑	95%HD↓	SSIM（%）↑
DeepLabv3+	86.55	93.56	94.22	93.11	98.29	98.61	89.24
SegNet	86.75	92.54	93.88	92.35	98.44	97.73	90.92
BMFR-Net	89.22	94.25	93.78	94.91	98.73	91.22	91.98
GCCINet	90.83	95.20	95.16	**95.23**	98.93	80.41	93.57-
C^3Net	87.58	93.38	95.05	91.11	98.56	80.46	92.92
MEC-Net	90.55	95.04	94.93	95.16	98.90	79.89	93.37
Ours	**91.15**	**95.49**	**95.77**	94.83	**98.96**	**79.61**	**94.63**

The bold values mean that the value has the best performance in this column of data.

TABLE II
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE INRIA DATASET

Method	Common metrics					Boundary metrics	
	IoU(%)↑	F1(%)↑	Precision(%)↑	Recall(%)↑	OA(%)↑	95%HD↓	SSIM（%）↑
DeepLabv3+	74.55	87.55	83.33	78.41	96.14	319.28	78.94
SegNet	73.65	84.56	86.25	82.70	94.49	322.71	78.16
BMFR-Net	75.76	86.39	87.52	85.27	94.29	281.15	80.52
GCCINet	78.88	88.19	89.09	**87.31**	96.77	283.66	83.79
C^3Net	76.23	86.51	85.94	87.10	96.26	296.12	82.49
MEC-Net	79.20	87.39	89.66	87.17	96.64	282.29	84.16
Ours	**79.52**	**88.40**	**90.56**	86.35	**96.88**	**280.06**	**84.97**

The bold values mean that the value has the best performance in this column of data.

1) Quantitative Analysis: Table I clearly demonstrates BEMRF-Net's superior performance across a range of metrics, significantly outperforming other methods. For standard metrics, BEMRF-Net achieved the highest scores in IoU, F1, precision, and OA, recording 91.15%, 95.49%, 95.77%, and 98.96%, respectively. Notably, improvements ranged from 0.08% to 4.6% for IoU, 0.16% to 2.95% for F1, 0.59% to 2.73% for precision, and 0.03% to 0.67% for OA. Regarding boundary metrics, BEMRF-Net also excelled, securing scores of 79.71% in 95%HD and 94.63% in SSIM.

Table II shows the extraction results of our method on the Inria dataset. BEMRF-Net scored the highest in the standard metrics of IoU, F1, precision, and OA, with scores of 79.52%, 88.40%, 90.56%, and 96.88%, respectively, showing improvements of 0.21%–7.76%, 0.21%–4.84%, 0.9%–7.23%, and 0.11%–2.59%. In terms of boundary metrics, BEMRF-Net achieved 280.06 in 95%HD and 84.97% in SSIM, representing the best and most favorable outcomes in these categories, respectively.

2) Qualitative Analysis: We conducted a comprehensive qualitative analysis of six outstanding models on both datasets, focusing on boundary extraction, background interference, and continuity within buildings. As illustrated in Figs. 7 and 8, groups (a) through (f) serve as control groups for this experiment. The methodology and experimental setup were aligned with those described in the original papers.

The WHU dataset comprises a diverse array of buildings and building clusters set against complex scenes. We selected six representative images for testing, with the outcomes displayed and comparatively analyzed. As illustrated in Fig. 7, we visualized our prediction results, demonstrating the model's capability in various scenarios. Groups (a), (b), and (c) showcase the model's proficiency in extracting building boundaries amidst complex backgrounds, where environmental factors notably impact detection accuracy. Particularly, in groups (a) and (b), the similarity in color and texture between roads and buildings presents a significant challenge, yet BEMRF-Net distinctively outperforms six other methods by accurately distinguishing between roads and building boundaries. In group (c), our method's resilience to shadow interference is evident, maintaining accurate boundary delineation where other methods falter. This success is attributed to the BSA module's focus on semantic boundary information at multiple levels.

Furthermore, Groups (d), (e), and (f) illustrate the network model's capability in extracting the interiors of buildings. Specifically, group (d) reveals that the extraction results from other methods exhibit holes and boundary gaps due to color differences at the tops of buildings, issues that BEMRF-Net effectively mitigates. In groups (e) and (f), buildings obscured by trees and debris demonstrate significant gaps or complete omissions in the extraction results of other methods, contrary to

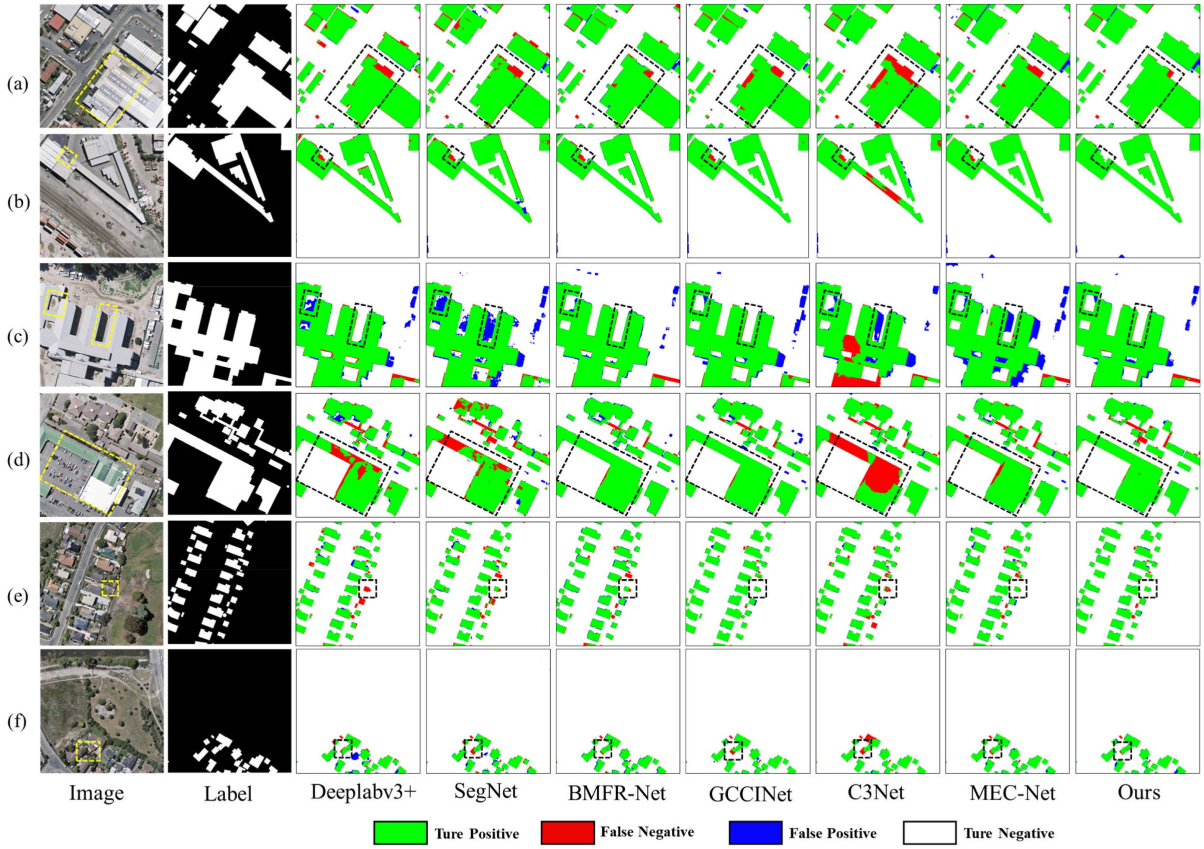


Fig. 7. Qualitative comparison with SOTA methods on the WHU dataset.

our approach, which ensures more thorough coverage and yields more comprehensive outcomes. This superiority is attributed to the MRSF module's efficient integration of deep and shallow semantic information, which prevents the loss of contextual data due to multiscale variances. Overall, our method demonstrates robust performance.

As illustrated in Fig. 8, we analyze the qualitative results of BEMRF-Net on the Inria dataset, which encompasses aerial images from five distinct cities, each with significant geographical and environmental variations. This diversity makes the Inria dataset a more challenging and illustrative testbed for assessing our model's generalization capabilities, particularly when compared with the WHU dataset. BEMRF-Net demonstrates a superior ability to extract complete building boundaries amidst irregular borders and dense vegetation, outperforming other models for both large and small structures, as evident in groups (a) and (b). The impact of shadows on building boundary segmentation, as demonstrated in group (c), shows that only our method's extraction results are minimally affected by shadows, leading to a more precise delineation of building boundaries. This accuracy is largely attributed to the BSA module, which focuses on semantic boundary information.

Vegetation obscuring building tops poses additional detection challenges, as evidenced in groups (d) and (e). While other models struggle with omissions and misjudgments, failing to distinguish between buildings and nonbuildings, BEMRF-Net effectively minimizes vegetation interference, improving foreground and background differentiation. Group (f) reveals that significant texture variations on building tops often lead to incomplete extractions by other models, characterized by missing boundaries and internal voids. Conversely, BEMRF-Net, supported by the MRSF module's integration of multiscale information, substantially reduces such inaccuracies, ensuring the internal continuity of buildings. Overall, our comprehensive qualitative analysis underscores BEMRF-Net's superiority in addressing the complexities of building extraction from aerial imagery, validating its enhanced performance across varied environmental conditions.

B. Comparison With Excellent Methods

To comprehensively demonstrate the advanced capabilities and superiority of our method, we compared it with recently developed building extraction models. We adhere to the configuration parameters specified in their original papers for methods where the source code is publicly available. In addition, a punctuation mark "-" in the table indicates that the metric was not reported or tested in the original paper.

Table III illustrates that compared with other methods, BEMRF-Net achieved the highest scores on all evaluation metrics except precision and recall. Although it does not lead in these two metrics, BEMRF-Net demonstrates the most effective balance among them. These results represent significant improvements over existing building extraction techniques Furthermore, the experimental outcomes corroborate the methodological advancement and superiority of BEMRF-Net in extracting buildings from RSIs.

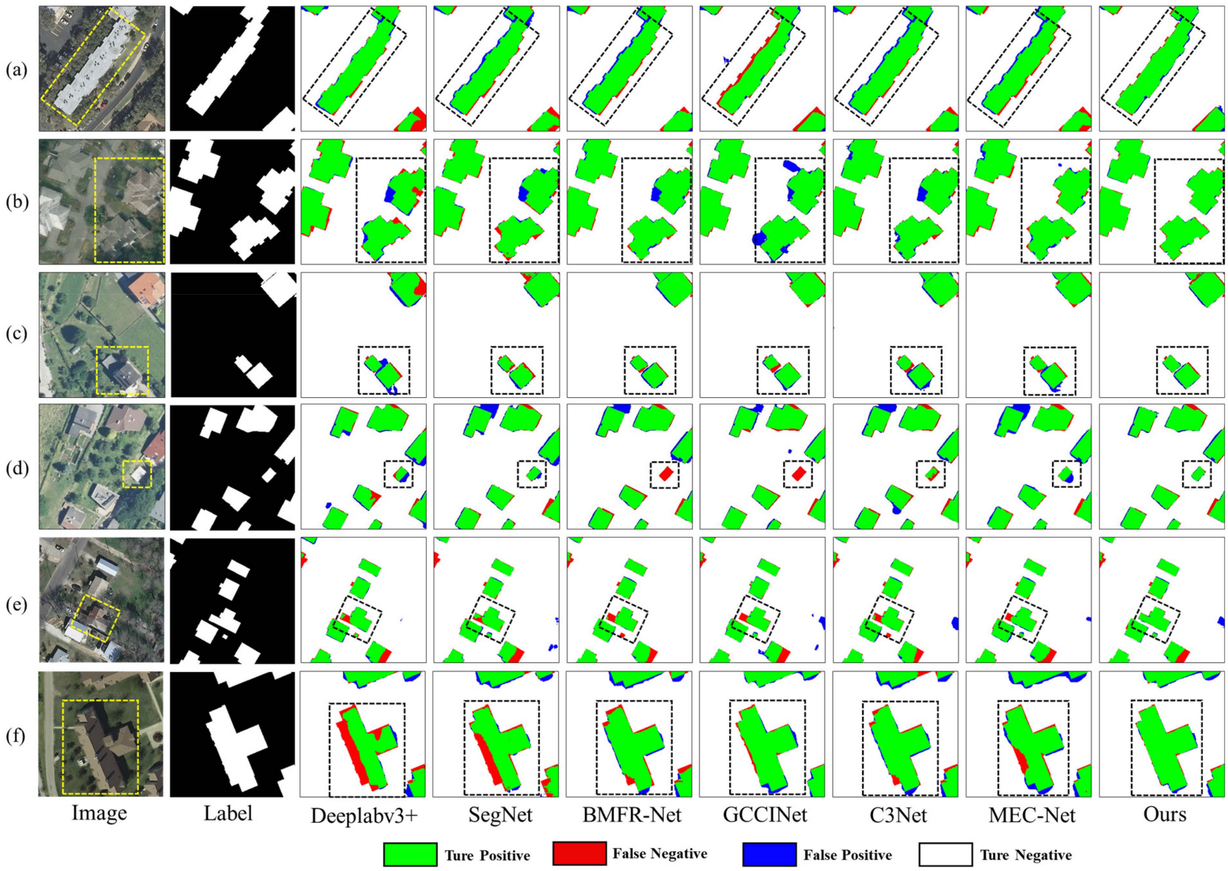


Fig. 8. Qualitative comparison with SOTA methods on the INRIA dataset.

TABLE III
Quantitative Comparison With Methods in Recent Studies on the WHU Dataset

Method	IoU(%)	F1(%)	Precision(%)	Recall(%)	OA(%)
STTNet [48]	90.10	94.79	94.09	95.50	98.83
MAPNet [49]	90.58	93.71	93.22	**95.82**	98.64
BOMSC-Net [50]	90.15	94.80	95.14	94.50	98.20
MA-FCN [51]	90.70	–	95.20	95.10	–
MFA-Net [52]	91.07	95.33	94.61	96.02	–
SCA-Net [53]	89.90	93.87	95.18	92.59	–
B-FGC-Net [54]	90.04	94.76	95.03	94.49	98.90
LiteST-Net [55]	90.13	94.22	93.75	95.06	98.19
RSR-Net [56]	88.32	–	94.92	92.63	–
DS-Net [57]	90.84	94.31	95.01	94.93	–
Ours	**91.15**	**95.49**	**95.77**	95.23	**98.93**

The bold values mean that the value has the best performance in this column of data.

V. Discussion

A. Attention Module Comparison Analysis

To comprehensively demonstrate the role and indispensable contribution of the BSA module, we conducted comparative experiments with the BSA module against the MSA [58] module, CBAM [59], and SE [60] module to illustrate the superiority of the BSA module in building extraction. We added different attention modules only to the baseline network. To ensure that the experiment is reliable, we experimented with the WHU dataset. As shown in Table IV, the experimental results indicate that in general metrics, the SE module had the highest F1 score, the MSA module scored highest in IoU, precision, and OA, and the BSA module only had the highest recall. However, the BSA module performed best across all indicators in boundary metrics.

TABLE IV
QUANTITATIVE ANALYSIS USING DIFFERENT ATTENTION MODULES

Module	Common metrics					Boundary metrics	
	IoU(%)↑	F1(%)↑	Precision(%)↑	Recall(%)↑	OA(%)↑	95%HD↓	SSIM(%)↑
SE	90.03	**95.06**	94.27	94.24	98.83	95.17	81.10
CBAM	90.32	94.51	95.66	94.02	97.49	92.34	86.87
MSA	**90.60**	94.75	**95.79**	94.35	**98.91**	96.69	79.53
BSA	90.41	94.96	95.40	**94.52**	97.88	**85.49**	**92.00**

The bold values mean that the value has the best performance in this column of data.

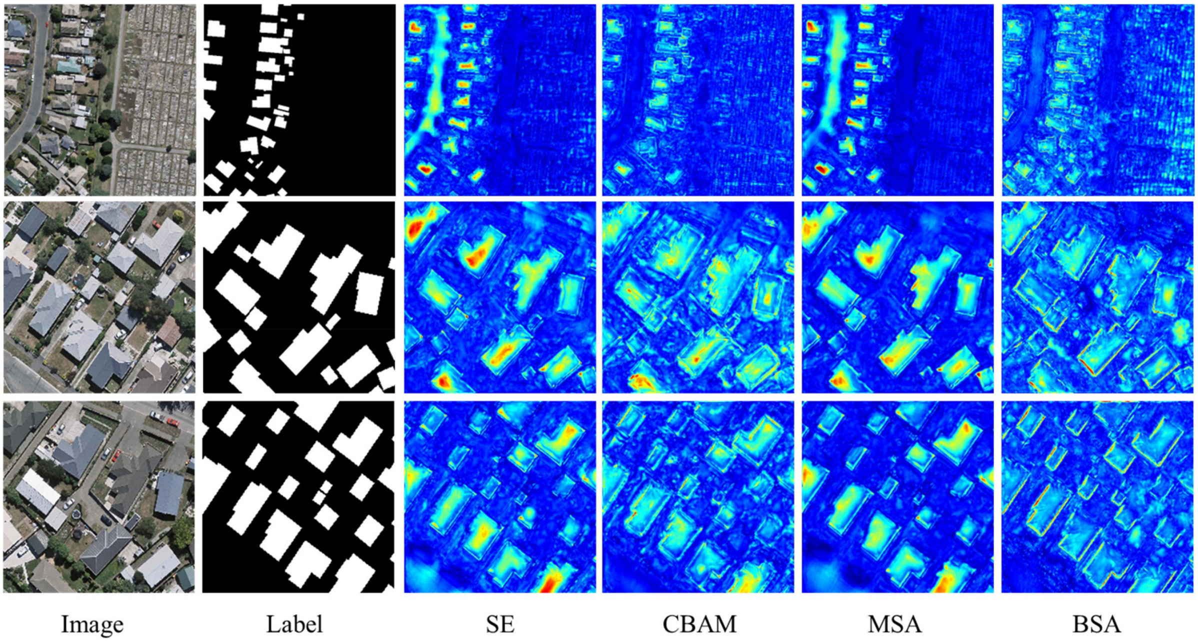


Fig. 9. Feature maps derived after sending the features from the encoder into different attention modules indicate that the blue areas represent areas of low model concern and red areas represent areas of high model concern.

We visualized the feature map outputs after each module to provide a more intuitive demonstration of different attention modules' impact on building boundaries. As shown in Fig. 9, compared with the other three attention modules, the BSA module shows a more pronounced focus on building boundaries, with the boundaries mostly highlighted in yellow or red. This is primarily because the BSA module is designed specifically for building boundaries in conjunction with the distance transform algorithm, reducing attention interference around the boundaries and focusing more on the boundaries. In contrast, the SE module, CBAM, and MSA module do not specifically target boundaries, resulting in their attention being dispersed across various positions of the building.

B. Complexity Analysis

To rigorously evaluate our approach's computational efficiency, we analyzed the number of parameters (Params) and floating-point operations (FLOPs) across our method and other methods. We benchmarked the performance of each model to assess their capabilities. Fig. 10 illustrates that GCCINet boasts the lowest parameter count, whereas MEC-Net demonstrates the minimal requirement for FLOPs. Despite our method achieving the highest score exclusively in IoU, it records moderate scores across other metrics, reflecting our focus on enhancing operational efficiency without compromising accuracy. Consequently, BEMRF-Net achieves an optimal balance between accuracy, parameter count, and computational demands. However, as shown in the figure, a notable disadvantage of BEMRF-Net is its high parameter count, which leads to slower operation. In future work, we will aim to design a more lightweight module to improve the network's efficiency.

C. Generalization Ability Analysis

To further evaluate our method's generalizability and applicability, we applied our model to an additional satellite imagery dataset with a 0.5 m resolution [61]. This dataset comprises satellite images from a specific region in Beijing, characterized by building styles markedly different from those found in the WHU and Inria datasets. It includes 344 images without a separate validation set. For a comprehensive assessment, the training and test sets were utilized solely for evaluation.

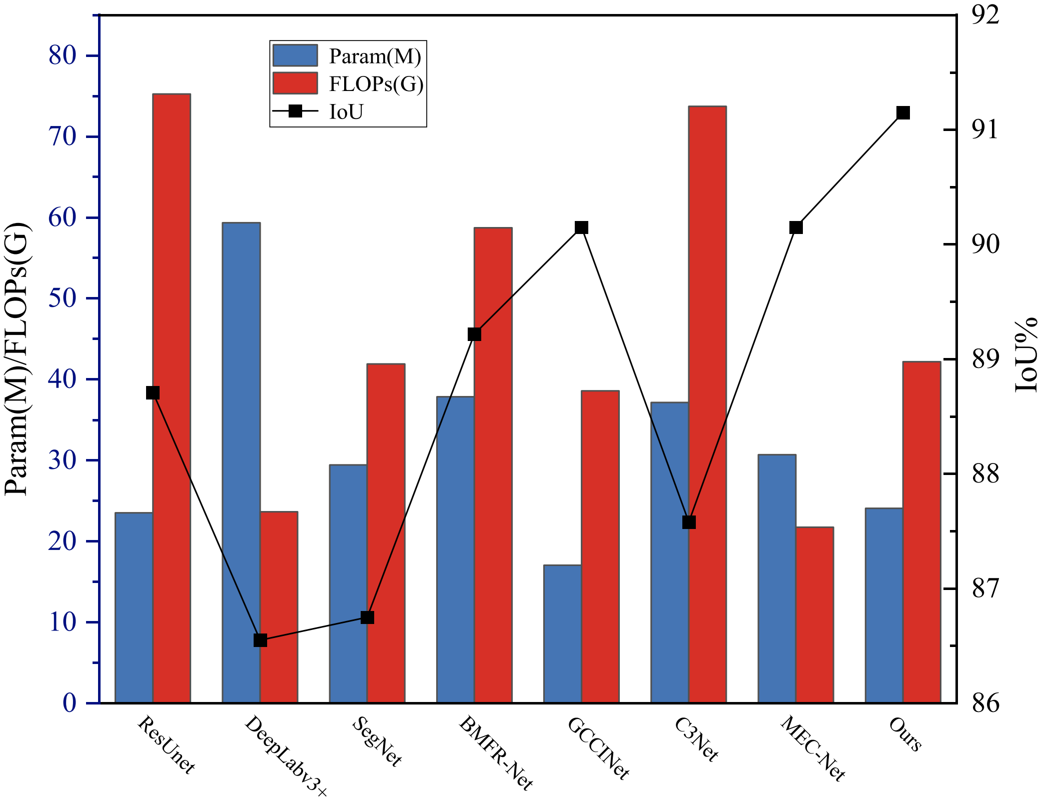


Fig. 10. Params and accuracy of BEMRF-Net compared with SOTA models.

TABLE V
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE NEW DATASET

Method	Metrics				
	IoU(%)	F1(%)	Precision(%)	Recall(%)	OA(%)
DeepLabv3+	67.77	80.29	85.08	76.91	93.23
SegNet	70.60	85.77	89.26	84.33	93.73
BMFR-Net	83.63	90.09	93.23	89.04	97.77
GCCINet	86.21	92.59	**93.97**	91.26	97.30
C^{3}Net	81.63	89.89	89.75	90.02	96.25
MEC-Net	85.33	91.50	93.00	**92.05**	97.91
Ours	**86.58**	**92.64**	93.68	91.67	**97.94**

The bold values mean that the value has the best performance in this column of data.

Table V presents the test results for the new dataset, revealing that BEMRF-Net maintains outstanding performance. Except for the precision and recall metrics, our method outperforms six other evaluated methods across all other metrics, evidencing BEMRF-Net's robust generalizability.

Fig. 11 presents a qualitative analysis of our experimental results. Observations from groups (a) and (b) reveal that, despite the challenges posed by roads and shadows, our method consistently ensures accurate boundary localization and extraction in the new dataset. Furthermore, in groups (c) and (d), BEMRF-Net demonstrates its effectiveness in identifying building tops, effectively reducing the impact of vegetation cover and heterogeneity. When compared with six other methods, our approach exhibits superior building extraction performance.

D. *Ablation Analyses*

Table VI shows the results of our tests on the weights of the loss function, illustrating the effect of different loss function settings on the performance of the method. Our weighted loss function proves to be effective and the weight settings are reasonable.

To better understand the function of each block in the BEMRF-Net, we conducted detailed ablation studies. These studies were carried out separately on the WHU dataset and the Inria dataset.

This experiment commenced with an evaluation of BEMRF-Net's basic network, stripped of all modules to retain only the core backbone. Subsequently, individual modules were incrementally integrated into the basic network to assess their

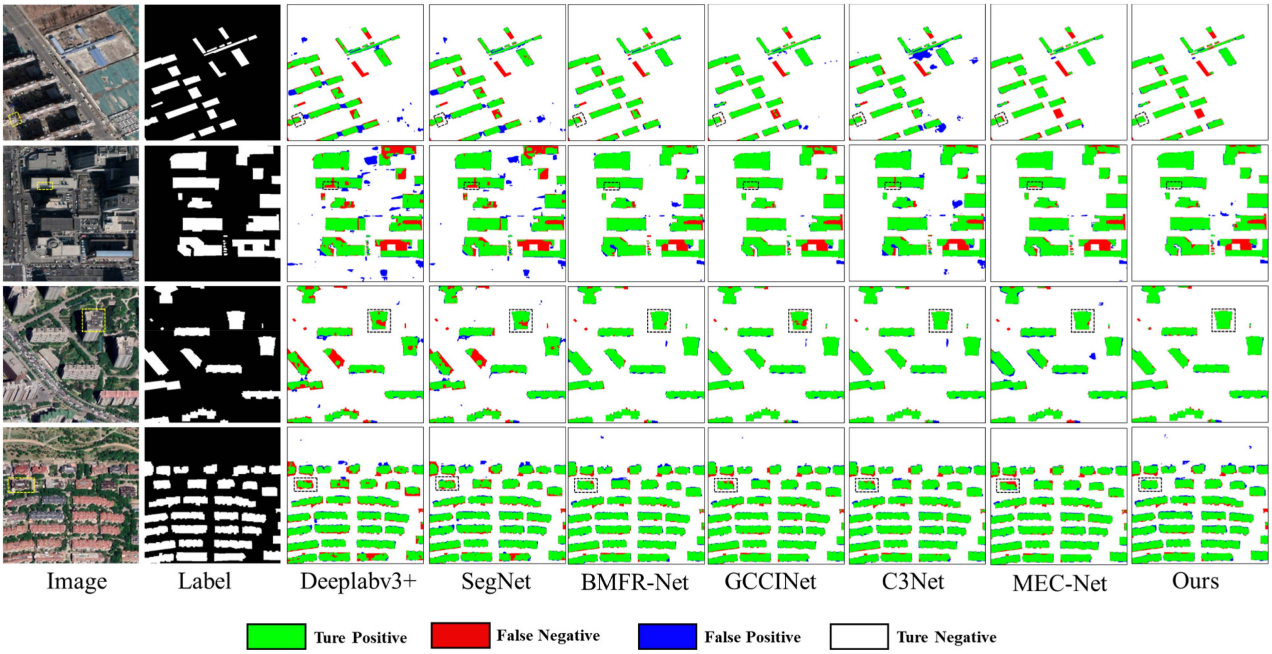


Fig. 11. Qualitative comparison with SOTA methods on the new dataset.

TABLE VI
Comparison of Different Weight Settings

Settings	IoU(%)	F1(%)	Precision(%)	Recall(%)	OA(%)
Common BCE	90.64	95.09	95.27	94.91	**98.95**
Weighted($\alpha = 0.1, \beta = 0.4$)	90.36	94.93	95.24	94.62	98.87
Weighted($\alpha = 0.2, \beta = 0.3$)	91.06	95.10	95.60	94.60	98.91
Weighted($\alpha = 0.4, \beta = 0.1$)	90.98	94.85	94.86	94.84	98.85
Weighted($\alpha = 0.3, \beta = 0.2$)	**91.15**	**95.49**	**95.77**	**95.23**	98.93

The bold values mean that the value has the best performance in this column of data.

contribution. As indicated in Table VII, the increase in IoU by 3.73%, F1 by 2.22%, and Recall by 2.17% with the addition of the BSA module; incorporating the MRSF module led to increases of 3.72% in IoU, 1.96% in F1, and 1.92% in Recall. Similarly, in the Inria dataset, integration of the BSA module improved IoU, F1, and Recall by 3.35%, 1.77%, and 1.14%, respectively, while the addition of the MRSF module saw improvements of 2.46% in IoU, 1.58% in F1, and 1.57% in Recall. These enhancements affirm the significant role of both modules in elevating standard and boundary metric performances. Detailed descriptions of these modules' functionalities are provided as follows:

1) Influence of BSA: To evaluate the effectiveness of the BSA module, we compared the heat maps of the final results of the basic network and the network containing only the BSA module, as shown in Fig. 12. The analysis reveals that before integrating the BSA module, the model's ability to delineate building boundaries was compromised by roads and shadows, resulting in a limited focus on the actual boundaries. Incorporating the BSA module shifts the model's focus away from nonboundary elements, directing attention toward the critical boundary regions. Specifically, in the first image row, the proximity of boundaries to shadows diminishes boundary detection, which the BSA module effectively counters by enhancing boundary emphasis. In the second row, the presence of roads distracts the model, a challenge the BSA module overcomes by reducing road-related attention. Similarly, in the third row, despite the high resemblance of building boundaries to the background, the BSA module succeeds in sharpening the model's focus on the boundaries, thereby improving the distinction between the foreground and background within boundary zones.

2) Influence of MSRF: Similarly, we conducted heat map comparison experiments on the WHU dataset between the basic network and a version with only the MSRF module, as shown in Fig. 13. Prior to integrating the MSRF module, the presence of building top covers and heterogeneity within the input images impaired the model's ability to accurately recognize buildings, leading to a dispersion of focus toward the background. Upon incorporating the MSRF module, the heatmaps clearly demonstrate a redirected focus toward the buildings themselves. Analysis of the first and second rows of figures reveals how vegetation and shadows obscure building tops, resulting in inaccuracies

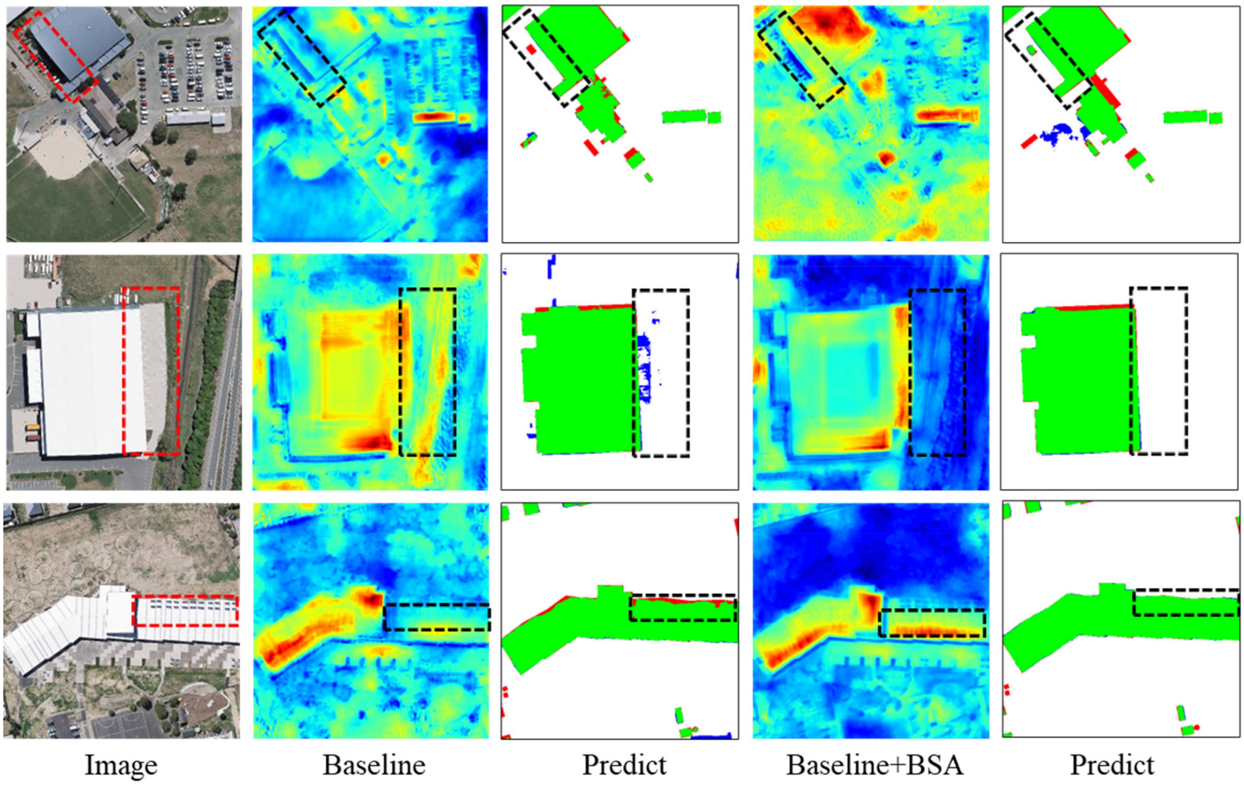


Fig. 12. Heat map visualization based on BSA feature maps.

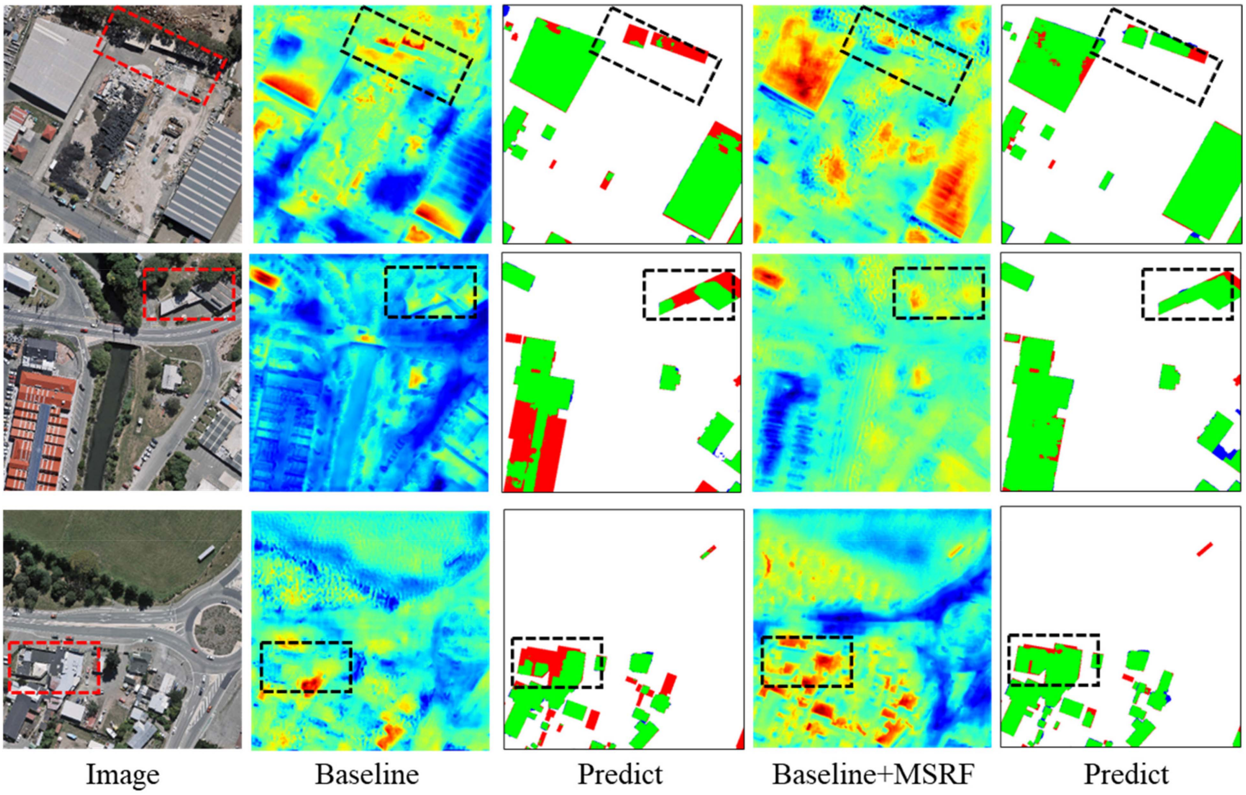


Fig. 13. Heat map visualization based on MRSF feature maps.

TABLE VII
Comparison of Ablation Experimental Results of Different Modules on WHU Dataset

	Method	Common metrics					Boundary metrics	
		IoU(%)↑	F1(%)↑	Precision (%)↑	Recall (%)↑	OA(%)↑	95%HD↓	SSIM(%)↑
WHU dataset	Baseline	86.68	92.74	91.20	92.35	94.58	95.62	87.54
	+ASPP	87.56	93.22	92.19	95.80	95.93	93.64	89.08
	+BSA	90.41	94.96	95.40	94.52	97.88	85.49	92.00
	+MRSF	89.95	94.70	95.14	94.27	97.82	84.07	91.12
	+ASPP+BSA	90.53	95.03	94.85	95.21	98.89	86.89	90.30
	+ASPP+MRSF	90.39	94.95	95.16	94.75	98.55	90.36	88.71
	BEMRF-Net	**91.15**	**95.49**	**95.77**	**94.83**	**98.96**	**79.71**	**94.63**
Inria dataset	Baseline	74.85	85.62	86.69	84.75	94.63	319.32	79.22
	+ASPP	75.80	86.24	85.83	85.65	95.18	313.01	81.56
	+BSA	78.20	87.39	88.80	86.16	96.08	291.24	84.44
	+MRSF	77.31	87.20	88.11	86.32	95.59	297.65	83.69
	+ASPP+BSA	78.09	87.69	89.27	86.18	96.66	295.33	83.43
	+ASPP+MRSF	77.60	87.39	89.81	85.09	96.61	307.94	80.16
	BEMRF-Net	**79.52**	**88.40**	**90.56**	**86.35**	**96.88**	**280.16**	**84.97**

The bold values mean that the value has the best performance in this column of data.

in the baseline network's extraction outcomes. The addition of the MSRF module mitigates these distractions by focusing the model's attention on the buildings, significantly improving extraction accuracy. Furthermore, the third row illustrates that the baseline network struggles with the heterogeneity of building tops, failing to fully capture their details. With the MSRF module, the model exhibits enhanced attention to both the exterior and interior aspects of buildings, effectively minimizing extraction errors and bolstering its identification capabilities.

VI. Conclusion

This article introduces the BEMRF-Net, a novel network designed to extract building information from RSIs. Initially, the BSA module enhances the accuracy of building boundary localization across both channel and spatial dimensions, utilizing the distance transformation method to focus the network's attention on boundary details. Additionally, the MSRF module, introduced at the decoding layer, addresses discontinuities within buildings caused by occlusions. This module integrates high-level and low-level semantic information, captures comprehensive global context, and preserves local positional details, thereby mitigating internal discrepancies and enhancing the integrity of rooftop information. A comparative analysis against 16 SOTA models on two public datasets demonstrates the method's significant advantages through both quantitative and qualitative assessments. Additionally, discussions on the method's efficiency, generalizability, and ablation studies further validate its effectiveness. Given the complexity of remote sensing imagery, future work will focus on enhancing network efficiency and method generalization while maintaining high accuracy.

References

[1] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 96–115, Feb. 2022.

[2] Y. Zhao, G. Sun, L. Zhang, A. Zhang, X. Jia, and Z. Han, "MSRF-Net: Multiscale receptive field network for building detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515714.

[3] H. Liu et al., "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 20, Oct. 2019, Art. no. 2380.

[4] R. Zhou, H. Yu, Y. Cheng, and F. Li, "Quantum image edge extraction based on improved Prewitt operator," *Quantum Inf. Process.*, vol. 18, pp. 1–14, Jul. 2019.

[5] H. Ying, G. Chen, and Q. Chen, "A novel approach of edge detection based on gray correlation degree and kirsch operator," in *Appl. Mechanics Materials*, 2015, pp. 169–172.

[6] 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. Geoinformation*, vol. 34, pp. 58–69, Feb. 2015.

[7] J. Zhu, J. Zhang, H. Chen, Y. Xie, H. Gu, and H. Lian, "A cross-view intelligent person search method based on multi-feature constraints," *Int. J. Digit. Earth*, vol. 17, no. 1, 2024, Art. no. 2346259.

[8] Y. Xie et al., "Landslide extraction from aerial imagery considering context association characteristics," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 131, pp. 1569–8432, 2024.

[9] Y. Xie et al., "Efficient video fire detection exploiting motion-flicker-based dynamic features and deep static features," *IEEE Access*, vol. 8, pp. 81904–81917, 2020.

[10] S. Pirasteh et al., "Developing an algorithm for buildings extraction and determining changes from airborne LiDAR, and comparing with R-CNN method from drone image," *Remote Sens.*, vol. 11, no. 11, May 2019, Art. no. 1271.

[11] D. Feng et al., "Boundary-semantic collaborative guidance network with dual-stream feedback mechanism for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4706317.

[12] 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.

[13] D. Feng, X. Shen, Y. Xie, Y. Liu, and J. Wang, "Efficient occluded road extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 13, no. 24, Dec. 2021, Art. no. 4974.

[14] M. Varshosaz, M. Sajadian, S. Pirasteh, and A. Moghimi, "Automated two-step seamline detection for generating large-scale orthophoto mosaics from drone images," *Remote Sens.*, vol. 16, no. 5, Mar. 2024, Art. no. 903.

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

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

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] Y. Xie et al., "An enhanced relation-aware global-local attention network for escaping human detection in indoor smoke scenarios," *ISPRS J. Photogrammetry Remote Sens.*, vol. 186, pp. 140–156, Apr. 2022.

[20] H. Chen et al., "Slice-to-slice context transfer and uncertain region calibration network for shadow detection in remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 203, pp. 166–182, Sep. 2023.

[21] D. Feng et al., "Regularized building boundary extraction from remote sensing imagery based on augment feature pyramid network and morphological constraint," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12212–12223, 2021.

[22] R. Hang, G. Li, M. Xue, C. Dong, and J. Wei, "Identifying oceanic eddy with an edge-enhanced multiscale convolutional network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 9198–9207, 2022.

[23] E. Khankeshizadeh et al., "A novel weighted ensemble transferred U-Net based model (WETUM) for post-earthquake building damage assessment from UAV data: A comparison of deep learning-and machine learning-based approaches," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701317.

[24] 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.

[25] Y. Xie et al., "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, 2020.

[26] Y. Hu, Z. Wang, Z. Huang, and Y. Liu, "PolyBuilding: Polygon transformer for building extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 199, pp. 15–27, May 2023.

[27] W. Xu et al., "Building height extraction from high-resolution single-view remote sensing images using shadow and side information," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6514–6528, 2024.

[28] S. Xu, M. Du, Y. Meng, G. Liu, J. Han, and B. Zha, "MDBES-Net: Building extraction from remote sensing images based on multiscale decoupled body and edge supervision network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 519–534, 2024.

[29] G. Yang, Q. Zhao, and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sens.*, vol. 12, no. 13, Jul. 2020, Art. no. 2161.

[30] W. Li, K. Sun, H. Zhao, W. Li, J. Wei, and S. Gao, "Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 113, Sep. 2022, Art. no. 102970.

[31] H. Lin, M. Hao, W. Luo, H. Yu, and N. Zheng, "BEARNet: A novel buildings edge-aware refined network for building extraction from high-resolution remote sensing images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 20, 2023, Art. no. 6005305.

[32] J. Wang, X. Yan, L. Shen, T. Lan, X. Gong, and Z. Li, "Scale-invariant multi-level context aggregation network for weakly supervised building extraction," *Remote Sens.*, vol. 15, no. 5, Mar. 2023, Art. no. 1432.

[33] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sens*, vol. 12, no. 6, Mar. 2020, Art. no. 1050.

[34] Z. Chen, Y. Luo, J. Wang, J. Li, C. Wang, and D. Li, "DPENet: Dual-path extraction network based on CNN and transformer for accurate building and road extraction," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 124, Nov. 2023, Art. no. 103510.

[35] D. Feng, H. Chen, Y. Xie, Z. Liu, Z. Liao, and J. Zhu, "GCCINet: Global feature capture and cross-layer information interaction network for building extraction from remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 114, Nov. 2022, Art. no. 103046.

[36] L.-C. Chen, G. Papandreou, L. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[37] Y. Xie et al., "Clustering feature constraint multiscale attention network for shadow extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705414.

[38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, vol. 9, pp. 249–256.

[39] 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.

[40] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[41] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[44] 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.

[45] S. Ran, X. Gao, Y. Yang, S. Li, G. Zhang, and P. Wang, "Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 14, Jul. 2021, Art. no. 2794.

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

[47] Z. Wang et al., "A multi-scale edge constraint network for the fine extraction of buildings from remote sensing images," *Remote Sens.*, vol. 15, no. 4, Feb. 2023, Art. no. 927.

[48] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sens.*, vol. 13, no. 21, Nov. 2021, Art. no. 4441.

[49] 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.

[50] 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.

[51] 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.

[52] S. Li, T. Bao, H. Liu, R. Deng, and H. Zhang, "Multilevel feature aggregated network with instance contrastive learning constraint for building extraction," *Remote Sens.*, vol. 15, no. 10, May 2023, Art. no. 2585.

[53] Y. Wang, Q. Zhao, Y. Wu, W. Tian, and G. Zhang, "SCA-Net: Multiscale contextual information network for building extraction based on high-resolution remote sensing images," *Remote Sens.*, vol. 15, no. 18, Sep. 2023, Art. no. 4466.

[54] Y. Wang, X. Zeng, X. Liao, and D. Zhuang, "B-FGC-Net: A building extraction network from high resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 2, Jan. 2022, Art. no. 269.

[55] W. Yuan, X. Zhang, J. Shi, and J. Wang, "LiteST-Net: A hybrid model of lite swin transformer and convolution for building extraction from remote sensing image," *Remote Sens.*, vol. 15, no. 8, Apr. 2023, Art. no. 1996.

[56] H. Huang, Y. Chen, and R. Wang, "A lightweight network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614812.

[57] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, “Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images,” *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589.

[58] A. Vaswani et al., “Attention is all you need,” *arXiv*, 2017, doi: 10.48550/arXiv.1706.03762.

[59] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

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

[61] L. Xia, X. Zhang, J. Zhang, H. Yang, and T. Chen, “Building extraction from very-high-resolution remote sensing images using semi-supervised semantic edge detection,” *Remote Sens.*, vol. 13, no. 11, Jun. 2021, Art. no. 2187.

Shaohan Cao received the B.S. degree in surveying and mapping engineering from the School of Surveying, Mapping and Land Information Engineering, Henan Polytechnic University, Jiaozuo, China, in 2022. He is currently working toward the master’s degree in surveying andmapping engineering with Southwest Jiaotong University, Chengdu, China.

His research interests include remote sensing image processing and GIS.

Dejun Feng received the M.S. and Ph.D. degrees in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2001 and 2004, respectively.

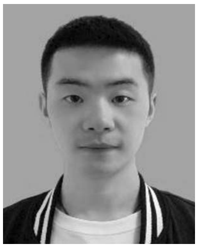
He is currently an Associate Professor with the Faculty of Geosciences and Engineering, Southwest Jiaotong University. His research interests include geographic information systems and remote sensing image processing.

Suning Liu received the B.S. degree in surveying and mapping engineering from the School of Resource and Environmental Engineering, Anhui University, Hefei, China, in 2021. She is currently working toward the master’s degree in surveying andmapping engineering with Southwest Jiaotong University, Chengdu, China.

Her research interests include remote sensing image processing and GIS.

Wanqi Xu received the B.S. degree in surveying and mapping engineering from the School of Civil Engineering and Transportation, Shandong Jiaotong University, Jinan, China, in 2022. She is currently working toward the master’s degree in surveying andmapping engineering with Southwest Jiaotong University, Chengdu, China.

Her research interests include remote sensing image processing and GIS.



Hongyu Chen received the B.S. degree in surveying and mapping engineering from the College of Hydraulic Engineering, Hebei University of Water Resources and Electric Engineering, Cangzhou, China, in 2021. He is currently working toward the master’s degree with Southwest Jiaotong University, Chengdu, China.

His research interests include computer vision and remote sensing image processing.

Yakun Xie received the B.S. degree in survey engineering from the School of Survey Engineering, Henan University of Urban Construction, Pingdingshan, China, in 2015, the M.S. degree in geodesy and survey engineering from the Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, China, in 2018, and the Ph.D. degree in surveying and mapping from Southwest Jiaotong University, in 2022.

He is currently an Assistant Professor with the Faculty of Geosciences and Engineering, Southwest Jiaotong University. His research interests include remote sensing image processing, deep learning, and computer vision.

Heng Zhang received the Ph.D. degree in surveying and mapping science and technology from Southwest Jiaotong University, Chengdu, China, in 2016.

He is currently with China Railway Design Corporation, where he specializes in 3D GIS and photogrammetry. In this project, his main contribution is to provide experimental data for this paper and guide the experimental process. His expertise ensures the technical accuracy and innovative approach of the project, significantly enhancing the research outcomes.

Saied Pirasteh (Senior Member, IEEE) received the first Ph.D. degree in remote sensing and GIS (geology) in 2004 from Aligarh Muslim University, India, and also the Ph.D. degree in geomatics and GeoAI (geography) from the University of Waterloo, Waterloo, ON, Canada, in 2018.

He is currently a Professor in Geomatics and Geospatial Artificial Intelligence and the Associate Dean with the Institute of Artificial Intelligence, Shaoxing University, Shaoxing, China. He is also a Visiting Professor and Scientist with the Geospatial Intelligence and Mapping Lab, University of Waterloo. He is also an Adjunct Professor with the Department of Geotechnics and Geomatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, India. He has coauthored more than 250 publications. His research interests include Geospatial Data Science & GeoAI, Remote Sensing Computer Vision including satellite, drone, LiDAR data processing, GIS and Geospatial Information analysis and modeling and developing algorithms for their novel applications in hazards & disasters beyond.

Dr. Pirasteh is currently an Advisory Board and Expert Member of the United Nations Global Geospatial Information Management (UN-GGIM) Academic Network and the Chair of the ISPRS ICWG III/IVa on Disaster Management for 2022–2026. He is an Associate Editor for IEEE TRANSACTIONS OF GEOSCIENCE AND REMOTE SENSING. He founded the Geospatial Infrastructure Management Ecosystem (GeoIME) (www.geoime.ca) and invented and commercialized this Technology System product and software tools. He has supervised and cosupervised many bachelor students’ projects, particularly more than 100 master/Ph.D. students, postdoc fellows, and visiting scholars from various countries. He has delivered many keynotes at international events and organized many events worldwide.

Jun Zhu received the M.S. degree in geodesy and survey engineering from Southwest Jiaotong University, Chengdu, China, in 2003, and the Ph.D. degree in cartography and geographic information systems from the Chinese Academy of Sciences, Beijing, China, in 2006.

From 2007 to 2008, he was a Postdoctoral Research Fellow with The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with the Faculty of Geosciences and Engineering, Southwest Jiaotong University. His research interests include computer vision, 3-D GIS technology, and virtual geographic environments.