LFEMAP-Net: Low-Level Feature Enhancement and Multiscale Attention Pyramid Aggregation Network for Building Extraction From High-Resolution Remote Sensing Images

Yu Liu, Erzhu Li, Wei Liu, Xing Li, and Yuxuan Zhu

***Abstract*—With the rapid development of Earth observation technology and deep learning, building extraction from remotely sensed imagery based on deep convolutional neural networks has attracted wide attention in recent years. However, due to the heterogeneity of building shapes and sizes and the complexity of the surrounding objects, current building extraction methods still have challenges in boundary accuracy and complete building extraction. For these purposes, we proposed a low-level feature enhancement and multiscale attention pyramid aggregation network (LFEMAP-Net) that considers building boundary information and multiscale feature expression to obtain higher accuracy building extraction. First, a low-level feature enhancement model is proposed based on the prior edge information to enhance the representation of spatial details, effectively addressing issues related to information loss and fuzzy boundaries. Additionally, a multiscale attention pyramid aggregation model is developed during the decoding stage to facilitate the fusion of features from different scales, thereby enhancing the extraction of building features. The experimental results on two publicly available datasets validate that LFEMAP-Net can overcome building extraction interruptions and boundary blur in complex scenes, and achieve boundary optimization and complete segmentation of buildings and achieve better performance than other advanced semantic segmentation models.**

***Index Terms*—Building extraction, deep learning, edge extraction, feature enhancement, multiscale attention.**

I. Introduction

BUILDINGS, as primary features in high-resolution remote sensing images (HRSI), are closely related to human activities, urban development, and societal functions. Their footprint information plays a fundamental role in comprehending the complex interactions between human endeavors and environmental influences and is an important component in applications, including urban planning [1], land use [2], and disaster response [3]. Over the past few decades, extracting the accurate building footprint information from HRSI has attracted wide attention and has obtained great progress in various applications. With the rapid advancement of Earth observation technology, it becomes possible to extract very detailed building footprint information, thanks to HRSI with rich spatial and structural information. However, in most urban areas, remote sensing ground objects are artificial surfaces, exhibiting complex composition in HRSI. Specifically, some buildings and other impervious surfaces share similar spectral and spatial features, such as some city roads and building roof, making it difficult to capture their unique features. Furthermore, due to the variability in building sizes, the diverse distribution of buildings, and the complex environmental surrounding, it presents a significant challenge for accurately building extraction task based on HRSI [4]. Thus, there exists an exigent need to develop precise and efficient building extraction methodologies that effectively leverage the features of HRSI to enhance the quality of building footprint information obtained from remotely sensed imagery in urban areas.

Since the emergence of high-resolution remote sensing technology, considerable endeavors have been committed to developing building extraction methods based on HRSI. Traditional approaches early applied for building extraction rely on manually designed features. Characterize buildings by establishing representative architectural features from characteristics, such as spectrum, texture, and geometry [5], [6]. However, because of occlusion by trees, shadows, and other factors, these methods cannot fully utilize the various information available in buildings, which limits their feature extraction capabilities. Additionally, some researchers have developed template libraries using prior knowledge of building shapes and, subsequently, incorporated them into active contour models to guide the evolution of segmentation curves [7]. Nevertheless, this approach has limitations in dealing with a wide range of complex and diverse building shapes. To this end, some works have integrated multiple sources of GIS and auxiliary data to enrich building features [8], [9], significantly improving the robustness of building recognition but often come with high data costs and complex algorithms. Therefore, these methods are still difficult to meet research requirements [10], [11].

Manuscript received 29 October 2023; revised 10 December 2023; accepted 20 December 2023. Date of publication 25 December 2023; date of current version 10 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42371465, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231353, and in part by the Natural Science Research of Jiangsu Higher Education Institutions of China under Grant 23KJB420002. *(Corresponding author: Erzhu Li.)*

The authors are with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China (e-mail: liuy101004@163.com; lierzhu2008@126.com; liuw@jsnu.edu.cn; lixing@jsnu.edu.cn; zhuyuxuannjust@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3346454

In recent years, due to the continuous progress in deep convolutional neural networks (DCNNs), a series of deep-learning-based approaches have gained significant traction in the remote sensing community [12], [13], [14]. Compared with traditional methods, deep learning makes full use of multilayer structures to extract high-level abstract features from spatial data, thus enhancing classification and detection accuracy [15], [16], [17]. This end-to-end deep network, which automatically adapts its parameters to capture features, proves more efficient than the manual design of features. Benefiting from the fully convolutional networks (FCNs) [18], data-driven DCNNs can automatically identify distinct objects within remotely sensed imagery through extensive training on labeled samples. This breakthrough enables dense predictions on large-scale remote sensing images and provides an efficient solution for extracting building features. For instance, Shrestha and Vanneschi [19] improved FCN with conditional random fields for boundary refinement, Deng et al. [20] used an encoder–decoder with attention gates and spatial pyramids for multiscale feature capture, and Chen et al. [21] combined deeplabv3+ with dense connections and ResNet for enhanced performance.

However, there are still challenges [22], [23], [24] in extracting buildings based on DCNNs. On the one hand, the network architectures tend to prioritize high-level semantic features, potentially sacrificing the finer edge and shape details, resulting in the loss of local detail features and edge information, leading to blurred boundaries [25]. On the other hand, high-level semantic features might be less responsive to background information and target regions [10], and common downsampling operations result in significant information loss and limit contextual information integration.

To overcome the above shortcomings, some building extraction methods based on encoder–decoder architectures have been proposed [26], [27], [28]. They have effectively reduced network parameters and promoted the fusion of multiscale features by incorporating residual concepts and pyramid pooling. However, the use of simple skip-layer connections for encoder–decoder models can simultaneously increase contextual information and low-level feature transfer [29]. It may lead to inadequate feature representation. Furthermore, it brings challenges in detecting smaller objects in extremely high-resolution images due to the use of dilated convolutions with different dilation rates. Some other approaches [30], [31], [32] have also aimed to enhance network extraction performance through the integration of multiscale input architectures. However, these methods significantly increase the computational complexity and bring difficulties to practical applications [11]. To this end, a series of attention methods have been developed [33], [34], [35]. These methods optimize features from both spatial and channel perspectives, leveraging intraclass similarity to improve the overall feature integrity [36], [37], [38], [39]. This not only enhances the network's ability to handle complex scenes but also reduces the computational complexity associated with multiscale input architectures and feature fusion techniques, making them more suitable for practical applications. Meanwhile, multimodal approaches for cross-city semantic segmentation have opened up new avenues for building extraction [40].

The fusion of contextual information is acknowledged as indispensable in building extraction based on HRSI. But, the incorporation of boundary information is also important in semantic segmentation. Due to complex shapes and diverse lighting conditions, the boundaries of semantic objects often exhibit considerable ambiguity, which is a huge challenge to accurate segmentation. To address this issue, several enhanced building extraction networks have been proposed [41], [42], integrating edge detection mechanisms. These innovative approaches enhance the capacity to handle intricate building edges while maintaining relatively smoother building footprint boundaries through the introduction of constraint terms. Nevertheless, the inclusion of supplementary edge networks often leads to a huge computation burden. Recent studies [43], [44], [45] have integrated structural information about buildings into the workflow by leveraging prior knowledge of building shapes and implementing postprocessing techniques, which have yielded promising outcomes [46]. Moreover, the structural prior information module is combined to refine the building boundaries [47], combined with feature map refinement during training [48], and has contributed to more robust edge detection results.

Although the existing methods have made improvements in determining building boundaries and segmenting building types, these still struggle to resolve internal inconsistencies and discontinuities in building extraction based on HRSI due to the building of distributed discretely, complex characteristics, and vary in scale. Besides, for the blurred difference between the foreground and background in some complex scenes, pixels with similar colors and spatial distances can easily be misjudged as homogeneous pixels, which leads to blurred boundaries. To solve these problems, we combine the priori edge information and propose a low-level feature enhancement and multiscale attention pyramid aggregation network (LFEMAP-Net) based on the low-level feature enhancement model (LFEM) and the multiscale attention pyramid aggregation model (MAPM) for detailed building footprint from HRSI.

The main contributions of this work include the following.

1) This work proposes a novel segmentation architecture, named LFEMAP-Net, characterized by multiscale integration and edge fusion, to achieve the refined extraction of buildings in HRSI.
2) We develop the LFEM by designing a bilateral fusion method to effectively combine prior edges to enhance the expression of network spatial details and provide more details for the decoding results.
3) We proposed MAPM to effectively focus on building feature representations across different scales by building a multiscale mixing attention (MMA) mechanism. Enhance the model's ability to aggregate information across levels.

The rest of this article is organized as follows. Section II presents the proposed LFEMAP-Net and its detailed architecture. Section III includes the descriptions of the dataset, experimental setups, evaluation metrics, as well as detailed analysis and discussions of experimental results. Finally, Section IV concludes this article.

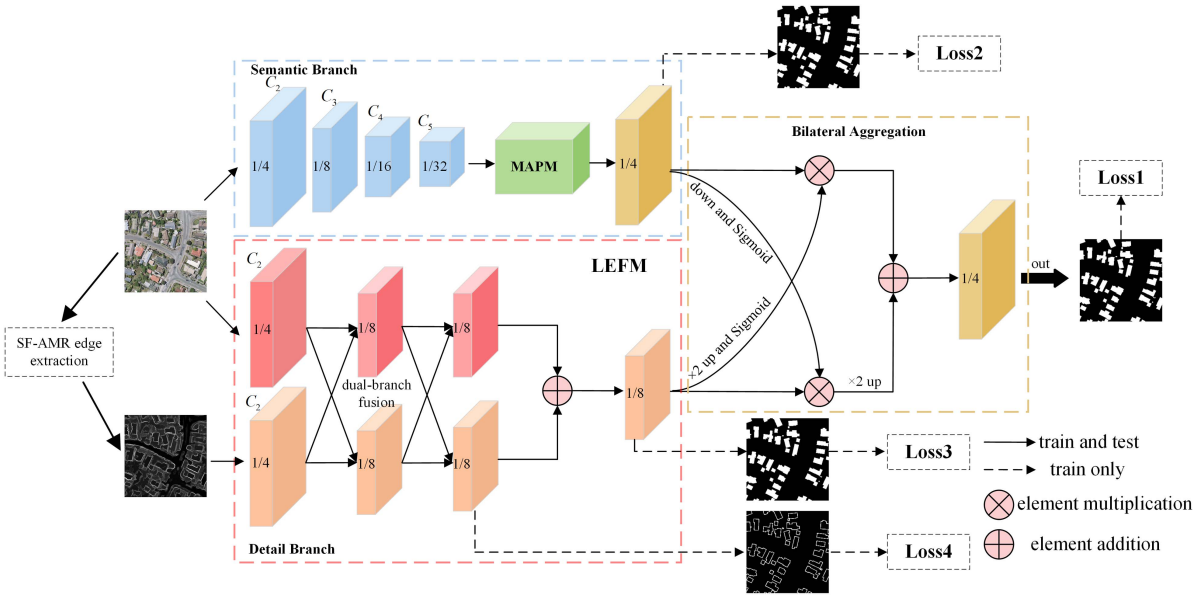

Fig. 1. LFEMAP-Net network architecture.

II. Methodology

A. Overall Framework

For the DCNNs' models, features extracted from deeper layers have a higher level of abstraction, while shallow features contain rich spatial details. Semantic segmentation networks usually use convolutional network models as encoders. As the number of network layers increases, spatial detail information will inevitably be lost, resulting in incomplete content expression or inaccurate segmentation edges during the decoding stage. Most semantic segmentation networks either directly connect multiscale feature maps and decode them into prediction maps or use skip connections to supplement scale features. Although these methods can enhance the feature expression ability of content, they cannot effectively retain accurate edge information. Therefore, LFEMAP-Net based on low-level feature enhancement and multiscale attention pyramid aggregation is designed in this work, as shown in Fig. 1 and Table I. It can integrate both high-level and low-level feature information to construct contextual semantic features and leverage prior edge information to enhance the low-level features associated with objects' edges. First, MAPM is proposed to maximize the utilization of features at various levels and enhance the model's ability to aggregate information. Moreover, we develop the LFEM to further refine boundaries by using prior edge information, enabling the extraction of more discriminative spatial details. Finally, the bidirectional aggregation method [49] is employed to fuse feature representations from both parts. This guidance manner enables efficient communication between both branches, integrating rich high-level semantic information for buildings with spatial detail features to obtain the robust building extraction results.

TABLE I
Architecture of Backbone Network in Encoding Path

ConvNeXt-B Convolutional Layers	Stage Name	Output Feature	Output Scale
$\begin{bmatrix} d7\times7, 128 \\ 1\times1, 512 \\ 1\times1, 128 \end{bmatrix}\times 3$	$conv2$	C_2	1/4
$\begin{bmatrix} d7\times7, 256 \\ 1\times1, 1024 \\ 1\times1, 256 \end{bmatrix}\times 3$	$conv3$	C_3	1/8
$\begin{bmatrix} d7\times7, 512 \\ 1\times1, 2048 \\ 1\times1, 512 \end{bmatrix}\times 27$	$conv4$	C_4	1/16
$\begin{bmatrix} d7\times7, 1024 \\ 1\times1, 4096 \\ 1\times1, 1024 \end{bmatrix}\times 3$	$conv5$	C_5	1/32

B. Multiscale Attention Pyramid Aggregation Model

In semantic segmentation, encoders are usually used to generate feature maps at different scales. Their amalgamation facilitates a more comprehensive capture of global and local context information [50], [51]. Nevertheless, simply concatenating the low-level and high-level features may lead to the underutilization of features across each scale. The lowest resolution branch output features of some popular backbone networks [52], [53], [54] contain the strongest semantic representation. However, the currently popular approach is to construct the feature maps of different dimensions from the lowest scale upward and then fuse them together [55], [56]. This process may not effectively propagate semantic information into higher resolution branches. In addition, generating high-resolution prediction maps through commonly used bilinear upsampling methods may result in the loss of irregular edge detail information. To this end, we propose MAPM, which can be viewed in Fig. 2. The module effectively exploits the spatial and channel dependencies within features across multiple scales to enhance semantic expression. It simultaneously integrates multiple-scale feature maps to form a robust and comprehensive feature representation. The MAPM

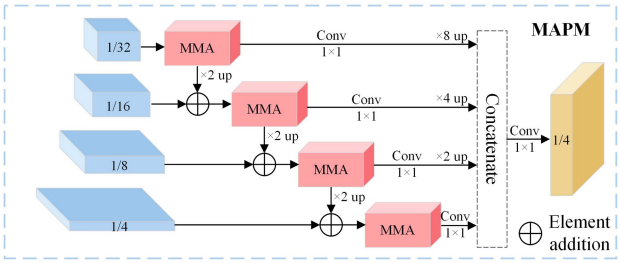

Fig. 2. Structure of MAPM.

process can be expressed as follows:

$$\mathrm{FM}_i = \begin{cases} \mathrm{MMA}\left(F_i\right) & i = 1 \\ \mathrm{MMA}\left(F_i + \mathrm{FM}_{i-1}\right) & 2 \le i \le 4 \end{cases} \tag{1}$$

$$F = \mathrm{Cat}_d\left(\varphi_{1\times1}\left(\mathrm{FM}_i\right)\right) \tag{2}$$

where F_i denotes the output at the i layer, FM_i stands the output processed by the MMA mechanism, $\varphi_{1\times1}$ represents a convolution operation using a 1×1 kernel, and Cat_d denotes the concatenation operation. The MMA module cleverly combines multiple-scale spatial and channel attention mechanisms to extract features more comprehensively from HRSI, thus improving the model's feature representation capabilities.

In order to better aggregate semantic features from high to low layer-by-layer, we have devised an MMA module for optimal utilization of contextual information. Furthermore, it allows the integration of building features with different scales of information. As indicated in Fig. 3, for an input $F \in R^{C\times H\times W}$, we initially construct a multiscale channel attention mechanism by utilizing different-sized pooling windows, followed by concatenation and fusion, resulting in a multiscale channel attention result F_C with dimensions of $C/2\times H\times W$. Subsequently, multiscale spatial attention is designed using different-sized convolution kernels, yielding multiscale spatial attention output F_S. To enrich the feature representation further, we integrate the input feature map features into F_S. Finally, the original image is elementwise added to F_S for fusion, resulting in a feature map with dimensions of $C/2\times H\times W$. The combination of multiscale channel attention and multiscale spatial attention effectively integrates spatial and channel features, resulting in improved semantic segmentation accuracy, which can be expressed as follows:

$$F_{\mathrm{MMA}} = \varphi_{1\times1}\left(F\right) + F_s \tag{3}$$

$$\begin{aligned} F_s = \mathrm{Cat}_d\ \{&\varphi_{1\times1}\left(F\right),\ \varphi_{3\times3}\left(F_c \times \left(S_{3\times3}\left(F_c\right)\right)\right), \\ &\varphi_{3\times3}\left(F_c \times \left(S_{5\times5}\left(F_c\right)\right)\right), \varphi_{3\times3}\left(F_c \times \left(S_{7\times7}\left(F_c\right)\right)\right)\} \end{aligned} \tag{4}$$

$$\begin{aligned} F_c = \mathrm{Cat}_d\ \{&\varphi_{1\times1}\left(F\right) \times \left(P_{\mathrm{global}}\left(F\right)\right), \varphi_{1\times1}\left(F\right) \\ &\times\ \left(P_{3\times3}\left(F\right)\right), \varphi_{1\times1}\left(F\right) \times \left(P_{5\times5}\left(F\right)\right)\} \end{aligned} \tag{5}$$

where φ represents the convolution operation, P is the channel attention operation of different pooling windows, S represents the spatial attention operations at different convolution scales, and F_{MMA} stands for the feature map output from the MMA module.

C. LEFM

We employ the structured forest (SF) combined with the adaptive morphological reconstruction (AMR) for prior edge extraction. SFs [57] use a structured learning method that can fully learn edge features by continuously predicting local segmentation masks for image patches. Specifically, the decision trees are trained to classify image patches as edge or nonedge. For each decision tree, the optimal segmentation parameters are determined based on the principle of maximum information gain.

Given an image $x \in X$ and its corresponding classification result $y \in Y$, the optimization objective function is given as follows:

$$h\ (x, \theta_j) = [x\,(k) < \gamma]\ \in \{0, 1\} \tag{6}$$

where θ_j is the optimal separation parameter, k represents the quantization feature of x, and γ stands for the threshold value of the quantization feature.

At the output stage, the classification result $y \in Y$ of the decision forest is mapped into labels, and Euclidean distance is used to measure whether the image patches with similar labels belong to the same segmentation. The measured results serve as a benchmark for both training and testing.

To mitigate the impact of potential noisy pixels in edge information and optimize subsequent processing, a multiscale and multistructural AMR method [58] is employed to effectively eliminate redundant information in the image and enhance the quality of prior edge.

Given an SF result g, the AMR is performed as follows:

$$\sigma\ (g, m, n) = \max_{m\le i\le n}\ \left\{C_R(g)_{S_i}\right\} \tag{7}$$

where C_R is the morphological closing reconstruction, S_i represents the multiple groups of structural elements, and the scale of structural elements is $i(1 \le i \le n, i \in N^+)$. σ is the AMR operator, which increases with the size of the structural element, and the start and end of the structural element scale selection are represented by m and n. In this study, we set m and n to 1 and 10, respectively.

Fig. 4 presents the results of edge extraction from remote sensing images. Compared with other methods, SF combined with AMR comprehensively represents the information on object boundaries, leading to more robust edge detection results.

To fully leverage the extracted prior edge information, we develop a dual-branch fusion strategy to enhance the network's spatial information representation and improve its boundary discrimination capability. Fig. 5 shows the specific implementation details of the proposed dual-branch fusion strategy. For an input image X_L, where the prior edges are represented as X_E, the final output for detail feature representation is obtained as follows:

$$X_{Li} = F_L(X_{L(i-1)}) + T_{E-L}(F_E(X_{E(i-1)})) \tag{8}$$

$$X_{Ei} = F_E(X_{E(i-1)}) + T_{L-E}(F_L(X_{L(i-1)})) \tag{9}$$

$$X_d = \varphi_{1\times1}(X_{Li} + X_{Ei}) \tag{10}$$

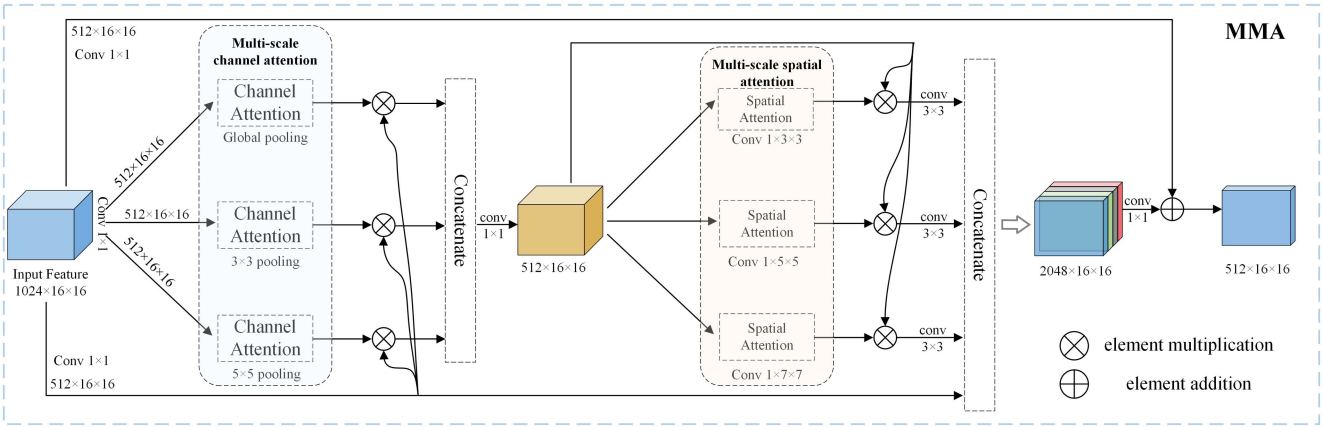

Fig. 3. Structure of MMA.

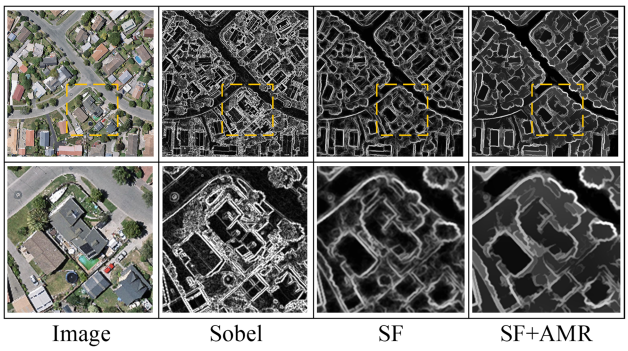

Fig. 4. Edge detection results.

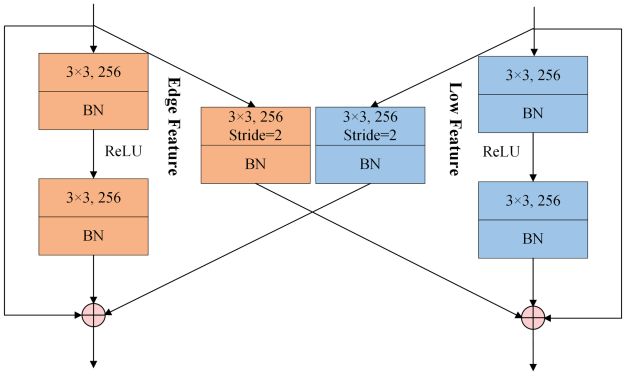

Fig. 5. Dual-branch fusion strategy.

where F_L and F_E correspond to the sequences of residual basic blocks with low resolution and prior edges, T_{E-L} and T_{L-E} refer to the low-to-edge and edge-to-low transformer, and X_d represents the final detail branch output.

D. Loss Function

The loss function is composed of four components: the final output loss ($L1$), pyramid aggregation output loss ($L2$), low-level feature enhancement loss ($L3$), and edge loss ($L4$). The total loss can be defined as follows:

$$\text{Loss} = L1 + \alpha L2 + \beta L3 + L4 \tag{11}$$

where α and β are the hyperparameters that control the weighting between these losses. In this article, it is set to 0.4.

We employ cross-entropy loss with online hard example mining (L_{OHEM}) for $L1$, $L2$, and $L3$, while binary cross-entropy loss (L_{BCE}) is used for edge loss $L4$. For N samples, where y_i denotes the actual category label and p_i stands the predicted probability by the model for the ith sample, then the OHEM cross-entropy loss can be mathematically expressed as follows:

$$L_{\text{OHEM}} = \frac{1}{K} \sum_{i=1}^{K} \text{CE}\left(p_i, y_i\right) \tag{12}$$

where K is the number of difficult examples to mine, which usually selects samples with wrong predictions or the largest loss, and $\text{CE}(p_i, y_i)$ represents the cross-entropy loss.

The $L4$ loss can be expressed as follows:

$$L_{\text{BCE}} = -\frac{1}{N} \sum_{i=0}^{N} \left[y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right)\right]. \tag{13}$$

III. Experiments and Discussion

A. Dataset

In this study, two publicly available datasets, including WHU building dataset [59] and Massachusetts building dataset [60], were chosen as the experimental data to test the proposed method.

The WHU building dataset consists of both aerial and satellite data. For our study, we specifically use aerial imagery, which consists of approximately 220 000 independent buildings in Christchurch, New Zealand. And offers imagery with a ground resolution of 0.3 m over an area of 450 km^2. The dataset is divided into training set, verification set, and test set, containing 4736, 1036, and 2416 images, respectively. Each image and its corresponding label are cropped to the dimensions of 512 × 512 pixels. We employed the default dataset division for our experiments.

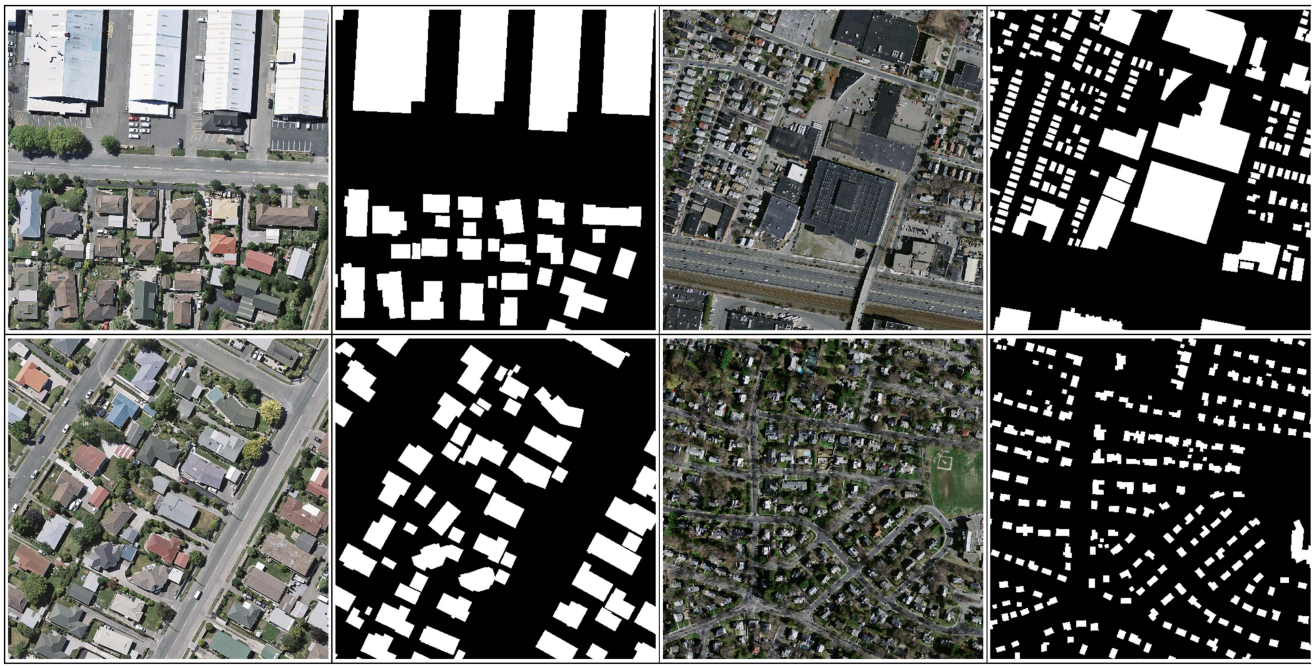

Fig. 6. Images and labels from the WHU dataset and Massachusetts dataset. The two columns on the left are WHU dataset, and the two columns on the right are Massachusetts dataset.

The Massachusetts building dataset comprises 151 aerial images from the Boston area, each with an image resolution of 1500 × 1500 pixels. Covering a vast area of approximately 340 km^2, this dataset offers a spatial resolution of 1 m. It is partitioned into training, validation, and test subsets, consisting of 137, 4, and 10 images, respectively. In our experiments, we resized the dataset to 512 × 512 pixels with a 12-pixel overlap, following the default dataset partitioning.

Fig. 6 shows that the sample images along with their corresponding building labels are presented for the WHU and Massachusetts datasets. The WHU dataset contains high-resolution aerial images that can reveal more detailed representations due to their higher resolution, while the Massachusetts dataset presents challenges with its lower resolution. Moreover, both two datasets exhibit significant variations in building scales, which effectively illustrate the practicality and efficacy of our approach in accurately delineating fine-grained building boundaries across various scales.

B. Implementation Details

All experiments are conducted on Windows 10 with PyTorch 1.12 framework based on Python 3.8. NVIDIA GeForce RTX 3060 GPU for 100 epochs on both two datasets. In addition, the Adam optimizer is employed with an initial learning rate of 0.01 and dynamically adjusted based on the validation accuracy. All compared approaches use the same batch size of 2 and data augmentation, including random scaling, rotation, and flipping.

C. Evaluation Metrics

To effectively assess and compare the models' performance, four commonly used semantic segmentation metrics were employed: Precision (P), Recall (R), Intersection over Union (IoU), and $F1$-score ($F1$). Precision measures the proportion of correctly predicted pixels out of the total. Recall signifies the ratio of predicted pixels to the overall count. The $F1$-score combines both recall and precision, providing a balanced assessment of the model's segmentation performance. Meanwhile, IoU provides a clear indication of the proportion of pixel overlap between the predicted and ground truth masks. The mathematical expressions are given as follows:

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{14}$$

$$R = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{15}$$

$$\mathrm{IoU} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}+\mathrm{FN}) \tag{16}$$

$$F1 = (2\times P\times R)/(P+R) \tag{17}$$

where $\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \text{and } \mathrm{FN}$ represent the numbers of true positive, true negative, false positive, and false negative for pixels.

D. Result and Discussion

1) Ablation Experiments: To demonstrate the effectiveness of different components within LFEMAP-Net, we conducted ablation experiments using ConvNext-B combining feature pyramid networks [55] for scale fusion as the baseline on both the WHU and Massachusetts building dataset. In these experiments, we employed P, R, IoU, and $F1$ scores to evaluate the distinct effects of various modules within LFEMAP-Net through selective exclusion or deactivation.

Fig. 7 presents representative visual results, illustrating variations in building extraction outcomes across different scenarios when the base network is combined with different modules.

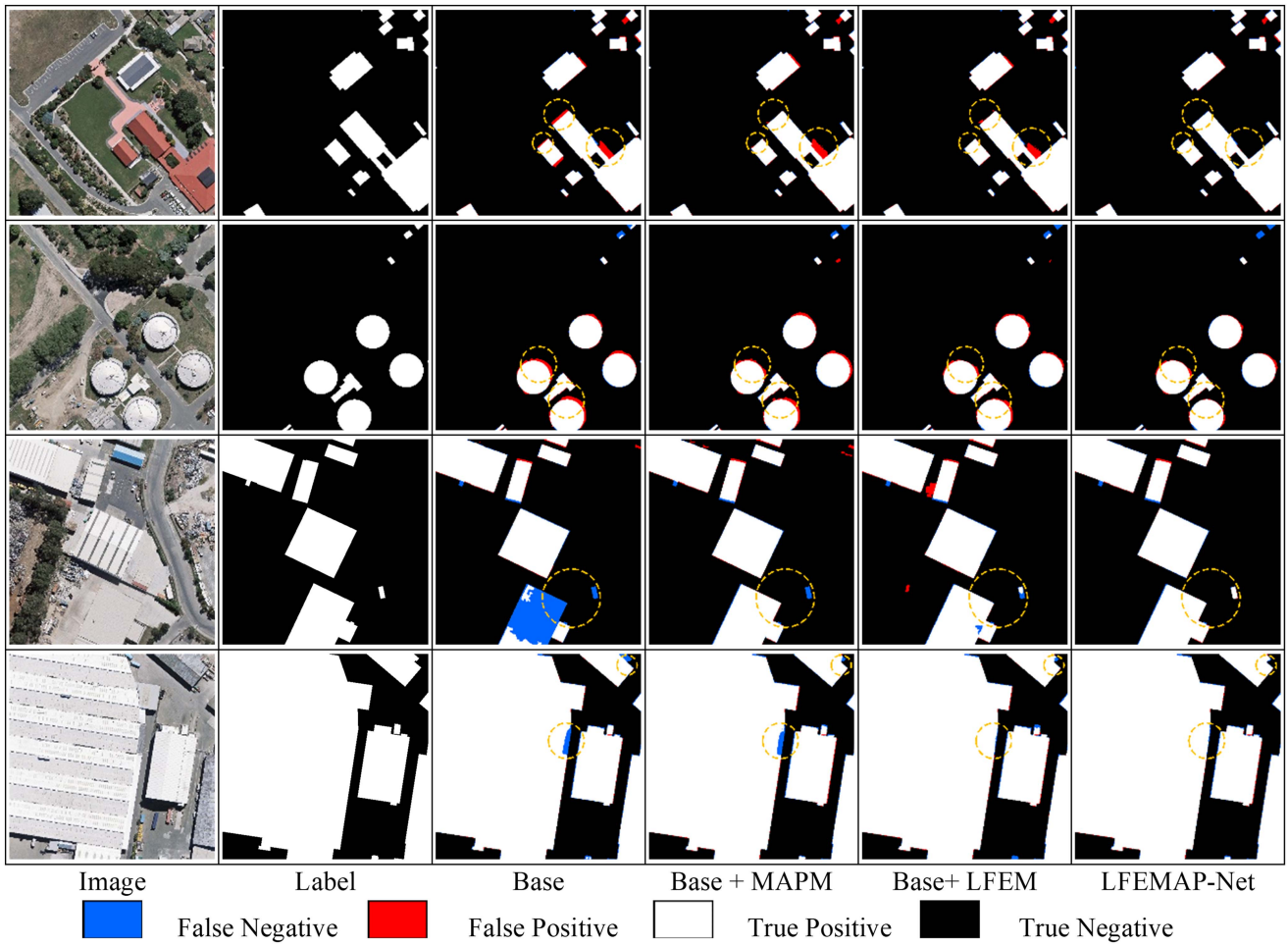

Fig. 7. Building extraction results of different module combinations.

TABLE II
ABLATION EXPERIMENTS OF THE NETWORK STRUCTURE BASED ON WHU DATASET

Method	IoU(%)	F1(%)	Precision(%)	Recall(%)
ConvNext-B	89.22	94.30	94.22	94.39
ConvNext-B +MAPM	90.76(+1.54)	95.16(+0.86)	95.16(+0.94)	**95.38**(+0.99)
ConvNext-B +LFEM	90.27(+1.05)	94.89(+0.59)	94.93(+0.71)	94.84(+0.45)
ConvNext-B + MAPM +LFEM	**91.09**(+1.87)	**95.34**(+1.04)	**95.81**(+1.59)	94.86(+0.47)

The bold entities represent only the highest accuracy, indicating the best performance obtained among all methods.

While the base network exhibits good architectural segmentation capabilities, there is room for improvement in boundary delineation and extraction completeness. In the fourth column, Base + MAPM demonstrates more complete building shapes through multiscale learning. The addition of LFEM to the base network emphasizes building boundaries. Finally, in the sixth column, LFEMAP-Net synthesizes the advantages of both modules, resulting in more complete and accurately delineated building extractions.

Tables II and III indicate that the module we designed significantly improved the performance of the network model. These supplementary modules demonstrated variable levels of the advantageous outcomes. First, optimizing the multiscale attention pyramid aggregation architecture based on the ConvNext-B, denoted as ConvNext-B + MAPM, led to an increase in IoU of 1.54% and 2.18%, $F1$ of 0.86% and 1.5%, Precision of 0.94% and 0.49%, and Recall of 0.99% and 2.39%, respectively. This demonstrates that the MAPM can effectively leverage feature representation at different scales to enhance semantic expression, resulting in more robust extraction results.

ConvNext-B + LFEM, which incorporates prior edge-enhanced low-level features, increased IoU by 1.05% and 1.8%,

TABLE III
ABLATION EXPERIMENTS OF THE NETWORK STRUCTURE BASED ON MASSACHUSETTS DATASET

Method	IoU(%)	F1(%)	Precision(%)	Recall(%)
ConvNext-B	69.16	81.77	85.01	78.77
ConvNext-B +MAPM	71.34(+2.18)	83.27(+1.50)	85.50(+0.49)	81.16(+2.39)
ConvNext-B +LFEM	70.96(+1.80)	83.01(+1.24)	**86.22**(+1.21)	80.03(+1.26)
ConvNext-B + MAPM +LFEM	**72.26**(+3.10)	**83.89**(+2.12)	86.06(+1.05)	**81.84**(+3.07)

The bold entities represent only the highest accuracy, indicating the best performance obtained among all methods.

TABLE IV
COMPARATIVE EXPERIMENTS OF THE NETWORK BASED ON A DIFFERENT BACKBONE

Method	Massachusetts Dataset			
	IoU(%)	F1(%)	Precision(%)	Recall(%)
HRnetv2-18	67.87	80.86	83.60	78.30
HRnetv2-18 + MAPM +LFEM	**69.66**(+1.79)	**82.12**(+1.26)	**84.72**(+1.12)	**79.67**(+1.37)
HRnetv2-48	69.01	81.66	85.41	78.23
HRnetv2-48 + MAPM +LFEM	**71.64**(+2.63)	**83.47**(+1.81)	**85.73**(+0.32)	**81.34**(+3.11)
ConvNext-L	70.11	82.43	84.82	80.17
ConvNext-L + MAPM +LFEM	**72.90**(+2.79)	**84.33**(+1.90)	**85.83**(+1.01)	**82.87**(+2.70)
ConvNext-XL	71.65	83.48	85.45	81.60
ConvNext-XL + MAPM +LFEM	**74.08**(+2.43)	**85.11**(+1.63)	**86.91**(+1.46)	**83.39**(+1.79)

The bold entities represent only the highest accuracy, indicating the best performance obtained among all methods.

$F1$ by 0.59% and 1.24%, Precision by 0.71% and 1.21%, and Recall by 0.45% and 1.26%. The optimization effect was most significant on the Massachusetts dataset, presenting that our proposed model exhibits powerful extraction capabilities for small, densely distributed buildings.

We also conducted comparative experiments using different backbone [54], [61] networks on the Massachusetts dataset. In Table IV, our method consistently demonstrated superior performance with various backbone networks, achieving an average improvement of 2.41% in IoU, 1.65% in $F1$, 0.98% in Precision, and 2.24% in Recall across all metrics.

2) Comparison With Other Methods: So far, many advanced semantic segmentation methods have been proposed, such as the U-Net [62], DeepLabV3+ [63], PSPNet [64], HRNetv2 [61], MaskFormer [65], Mask2Former [66], MAP-Net [67], MSL-Net [68], BOMSC-Net [69], MAFF-HRNet [70], MBR-HRNet [71], and D-LinkNet [21]. Experiments both on the two datasets were conducted to evaluate our LFEMAP-Net, and compared it with these state-of-the-art semantic segmentation approaches, both qualitatively and quantitatively, to validate its performance in building semantic segmentation.

Fig. 8 presents the visual results of our LFEMAP-Net and other common semantic segmentation methods on the WHU dataset. In this comparison, eight representative images were selected to conduct experiments with UNet, DeepLabV3+, PSPNet, HRNetv2, Mask2former, and our LFEMAP-Net. As shown in Fig. 8, when dealing with buildings in complex scenes, our LFEMAP-Net outperforms other tested methods. It produces a more complete building boundary, and it is more sensitive to the scale of both small and large buildings. In the first four rows of Fig. 8, LFMAP-Net significantly reduces misclassification and omissions of other interfering objects, providing more accurate and detailed descriptions of buildings with complex boundary contours. It can accurately distinguish between buildings and background, even in cases where building boundaries are challenging to delineate. The last four rows depict buildings with lower foreground–background contrast and significant scale variations, and LFEMAP-Net can still effectively distinguish buildings from the background.

Leveraging its multiscale advantages, it comprehensively captures information about buildings of different sizes and accurately infers their complete shapes.

Quantitative evaluation results are presented in Table V. For a more comprehensive evaluation of the proposed method, this study further compared building segmentation performance with the latest research, including MAP-Net, MSL-Net, and

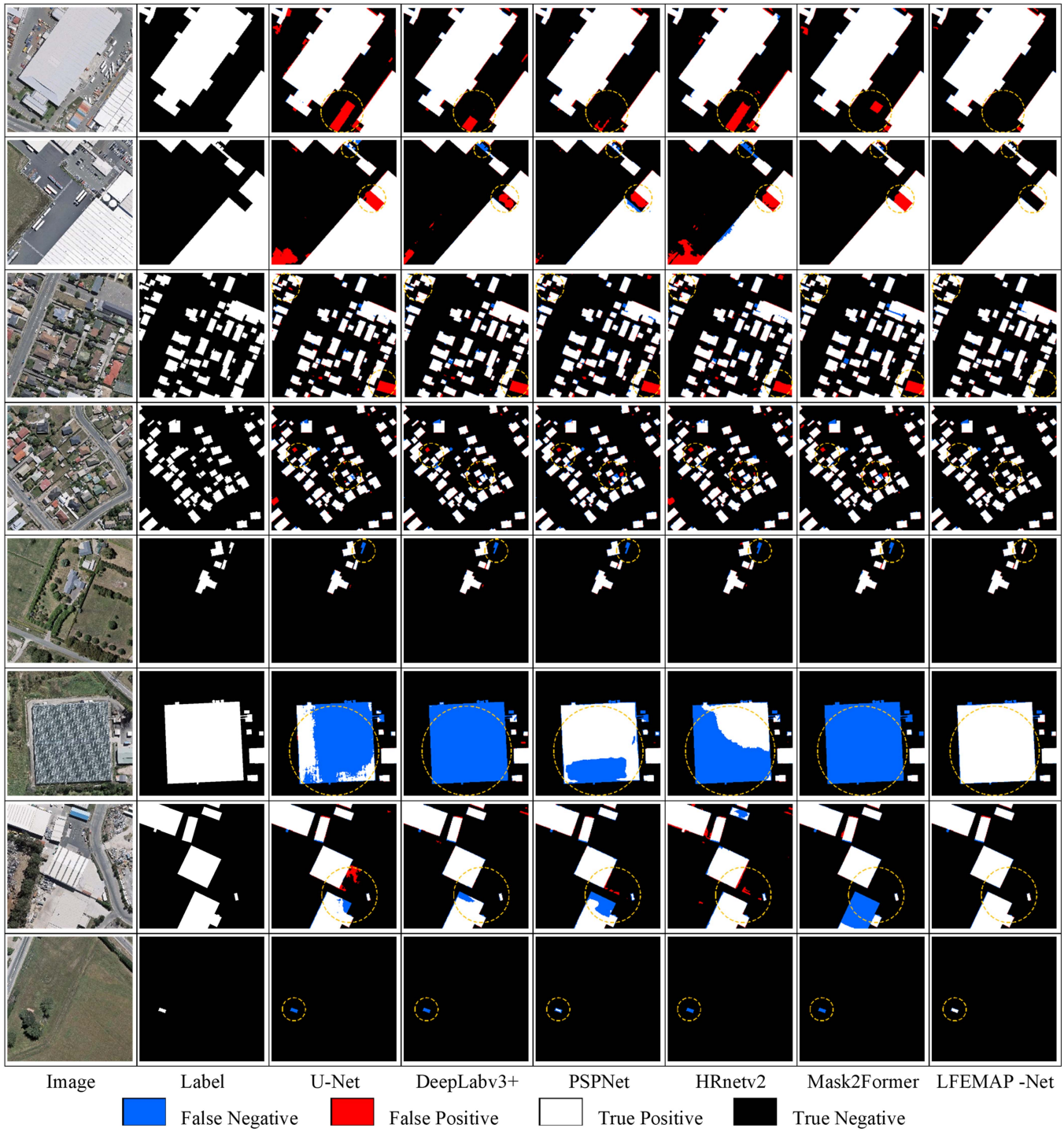

Fig. 8. Visualize results on the WHU aerial building dataset.

BOMSC-Net. LFMAP-Net-L and LFMAP-Net-XL are extensions with increased channel numbers. This enhancement aims to strengthen the model's ability to capture complex features and patterns, ensuring robust performance in highly intricate environments. The rows and columns represent different test methods and evaluation metrics, respectively.

LFMAP-Net achieved the highest precision with 91.09% IoU, 95.34% *F*1, 95.81% Precision, and 94.86% Recall across all metrics. Additionally, when we employed a backbone model, ConvNext-XL, with a larger number of channels, it achieved an accuracy of 91.48% IoU, 95.55% *F*1, 95.65% Precision, and 95.45% Recall with training for 100 epochs, proving the effectiveness of the proposed method.

To further assess LFEMAP-Net's generalization performance in extracting buildings across different datasets, multiple comparative experiments were also carried out on the challenging Massachusetts Dataset. As shown in Fig. 9, due to the limited spatial resolution, building boundary delineation is often

TABLE V
QUANTITATIVE EVALUATION ON THE WHU AERIAL BUILDING DATASET

Method	IoU(%)	F1(%)	Precision(%)	Recall(%)
U-Net	88.80	94.07	93.67	94.47
DeepLabv3+	88.84	94.09	94.04	94.14
PSPNet	87.97	93.60	94.81	92.42
HRnetv2	87.16	93.14	93.01	93.27
Mask2Former	90.31	94.91	95.47	94.36
MAP-Net	90.86	95.21	95.62	94.81
MSL-Net	90.4	95.0	95.1	94.8
BOMSC-Net	90.15	94.80	95.14	94.50
LFMAP -Net	91.09	95.34	**95.81**	94.86
LFMAP -Net-L	91.37	95.49	95.67	95.31
LFMAP -Net-XL	**91.48**	**95.55**	95.65	**95.45**

The bold entities represent only the highest accuracy, indicating the best performance obtained among all methods.

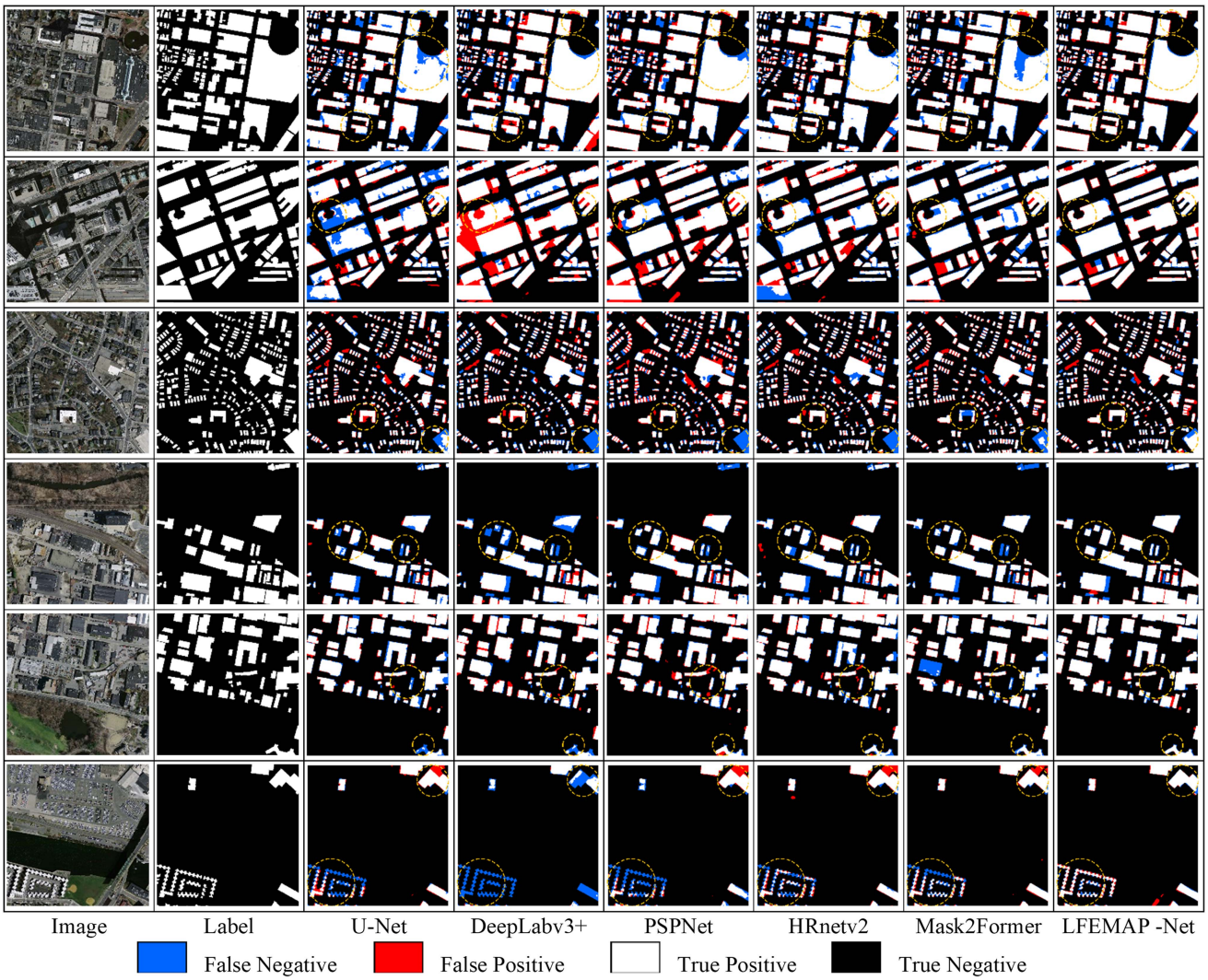

Fig. 9. Visualize results on the Massachusetts building dataset.

TABLE VI
QUANTITATIVE EVALUATION ON THE MASSACHUSETTS BUILDING DATASET

Method	IoU(%)	F1(%)	Precision(%)	Recall(%)
U-Net	68.44	81.26	84.81	78.01
DeepLabv3+	67.19	80.38	83.32	77.63
PSPNet	65.08	78.85	80.02	77.70
HRnetv2	67.87	80.86	83.60	78.30
Mask2Former	70.72	82.85	85.66	80.21
MBR-HRNet	70.97	83.53	86.40	80.85
MAFF-HRNet	68.32	81.17	83.15	79.29
D-LinkNet	70.39	-	73.36	**85.88**
LFMAP -Net	72.26	83.89	86.06	81.84
LFMAP -Net-L	72.90	84.33	85.83	82.87
LFMAP -Net-XL	**74.08**	**85.11**	**86.91**	83.39

The bold entities represent only the highest accuracy, indicating the best performance obtained among all methods.

TABLE VII
QUANTITATIVE EVALUATION OF PARAMETERS OF DIFFERENT MAIN MODULES

	Basic Network	Basic Network + LFEM	Basic Network + MAPM	LFEMAP-Net
Params(M)	94.946	100.35 (+5.404)	121.961 (+27.015)	127.365 (+32.419)

unclear; thus, segmenting small buildings is challenging. Variations in building roof materials and the presence of partial shadows contribute significantly to the challenge of achieving complete building extraction. Compared with other semantic segmentation methods, LFEMAP-Net is capable of achieving more complete building extraction while preserving well-defined boundaries. Fig. 9 presents results for large-scale and small-scale buildings in complex environments. The segmentation results demonstrate that LFEMAP-Net performs better in the challenging scene, resulting in fewer omissions and misclassifications for both large and small buildings, and accurately extracts the outlines of small buildings. Moreover, recent reports were also referenced on MBR-HRNet, MAFF-HRNet, and D-LinkNet for quantitative evaluation in Table VI. Rows represent evaluation metric results for different methods, while columns represent evaluation metrics. In quantitative comparisons, LFEMAP-Net continues to achieve the best performance, demonstrating the advantages of our proposed method in scale learning and building edge optimization, achieving fine-grained building segmentation.

3) Parameter Analysis: In this study, we conducted a comprehensive analysis of model parameters for different main modules to holistically assess the complexity of our model. Specifically, we evaluated the model complexity by calculating the parameters of LFEMAP-Net with various configurations, including the base network, base network + LFEM, base network + MAPM, and base network + LFEM + MAPM (LFEMAP-Net). Table VII reveals that incorporating LFEM into the base network results in a marginal parameter increase of 5.404 (M). On the other hand, introducing MAPM leads to a more substantial parameter augmentation of 27.015 (M). In comparison, LFEM contributes relatively fewer parameters to LFEMAP-Net. The MAPM module, emphasizing scale characteristics and integrating mixing attention across multiple scales, significantly amplifies the model's parameter count.

IV. CONCLUSION

In this study, an improved building extraction approach (LEFMAP-Net) for HRSI has been presented to address the limitations of current methods in boundary accuracy and complete building extraction. Specifically, in order to get more accurate contours, we proposed LFEM, a novel approach incorporating prior edge information through bilateral fusion, which considers more spatial detail information by fusion of prior edge, thereby refining the building boundary details. Moreover, we develop MAPM. Through the designed MMA mechanism, MAPM can effectively capture multiscale and multilevel features and solve the problem of incomplete building extraction and missed detection of small buildings. Experimental results on two publicly available datasets validate the effectiveness of LFEMAP-Net, showcasing its capacity to improve the building boundary and multiscale feature integration. Even in the challenging scene, LFEMAP-Net can make full use of prior edges and multiscale information, extract more accurate building boundaries, and achieve more complete building extraction results.

REFERENCES

[1] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 407.

[2] D. Wierzbicki, O. Matuk, and E. Bielecka, "Polish cadastre modernization with remotely extracted buildings from high-resolution aerial orthoimagery and airborne LiDAR," *Remote Sens.*, vol. 13, no. 4, Feb. 2021, Art. no. 611.

[3] Y. Wang, L. Cui, C. Zhang, W. Chen, Y. Xu, and Q. Zhang, "A two-stage seismic damage assessment method for small, dense, and imbalanced buildings in remote sensing images," *Remote Sens.*, vol. 14, no. 4, Feb. 2022, Art. no. 1012.

[4] C. Qiu et al., "Transferring transformer-based models for cross-area building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4104–4116, 2022.

[5] L. Gu, Q. Cao, and R. Ren, "Building extraction method based on the spectral index for high-resolution remote sensing images over urban areas," *J. Appl. Remote Sens.*, vol. 12, no. 4, Nov. 5, 2018, Art. no. 045501.

[6] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved building detection using texture information," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 143–148, 2013.

[7] L. Wang et al., "Active contours driven by edge entropy fitting energy for image segmentation," *Signal Process.*, vol. 149, pp. 27–35, Aug. 2018.

[8] Z. Zhang, W. Guo, M. Li, and W. Yu, "GIS-supervised building extraction with label noise-adaptive fully convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 12, pp. 2135–2139, Dec. 2020.

[9] D. Chai, "A probabilistic framework for building extraction from airborne color image and DSM," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 948–959, Mar. 2017.

[10] Y. Liu et al., "ARC-net: An efficient network for building extraction from high-resolution aerial images," *IEEE Access*, vol. 8, pp. 154997–155010, 2020.

[11] 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.

[12] M. Zhang, G. Zheng, Z. Jiang, Q. Zhu, L. Wang, and Q. Guan, "Local-aware coupled network for hyperspectral image super-resolution," *GISci. Remote Sens.*, vol. 60, no. 1, Dec. 2023, Art. no. 2233725.

[13] B. Liu, A. Yu, X. Zuo, R. Wang, C. Qiu, and X. Yu, "Deep hierarchical transformer for change detection in high-resolution remote sensing images," *Eur. J. Remote Sens.*, vol. 56, no. 1, Dec. 2023, Art. no. 2196641.

[14] B. Liu, S. Du, L. Bai, S. Ouyang, H. Wang, and X. Zhang, "Water extraction from optical high-resolution remote sensing imagery: A multi-scale feature extraction network with contrastive learning," *GISci. Remote Sens.*, vol. 60, no. 1, Dec. 2023, Art. no. 2166396.

[15] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[16] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, "LRR-Net: An interpretable deep unfolding network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513412.

[17] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[18] 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.

[19] S. Shrestha and L. Vanneschi, "Improved fully convolutional network with conditional random fields for building extraction," *Remote Sens.*, vol. 10, no. 7, Jul. 2018, Art. no. 1135.

[20] W. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder-decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.

[21] M. Chen et al., "DR-net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, no. 2, Jan. 2021, Art. no. 294.

[22] L. Luo, P. Li, and X. Yan, "Deep learning-based building extraction from remote sensing images: A comprehensive review," *Energies*, vol. 14, no. 23, Dec. 2021, Art. no. 7982.

[23] J. Chen et al., "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2021.

[24] Z. Zhang and Y. Wang, "JointNet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, Mar. 2, 2019, Art. no. 696.

[25] M. Guo, H. Liu, Y. Xu, and Y. Huang, "Building extraction based on U-net with an attention block and multiple losses," *Remote Sens.*, vol. 12, no. 9, May 2020, Art. no. 1400.

[26] Q. Tian, Y. Zhao, Y. Li, J. Chen, X. Chen, and K. Qin, "Multiscale building extraction with refined attention pyramid networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8011305.

[27] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, Dec. 31, 2022.

[28] W. Qiu, L. Gu, F. Gao, and T. Jiang, "Building extraction from very high-resolution remote sensing images using refine-UNet," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6002905.

[29] X. Dai, M. Xia, L. Weng, K. Hu, H. Lin, and M. Qian, "Multiscale location attention network for building and water segmentation of remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609519.

[30] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 1350.

[31] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, Feb. 2019, Art. no. 227.

[32] 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, May 3, 2019.

[33] J. Cai and Y. Chen, "MHA-net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5807–5817, 2021.

[34] M. Yu, X. Chen, W. Zhang, and Y. Liu, "AGs-UNet: Building extraction model for high resolution remote sensing images based on Attention Gates U network," *Sensors*, vol. 22, no. 8, Apr. 2022, Art. no. 2932.

[35] D. Zhao, H. Zhao, R. Guan, and C. Yang, "Efficient building extraction for high spatial resolution images based on dual attention network," *Int. J. Comput. Commun. Control*, vol. 16, no. 4, Aug. 2021, Art. no. 4245.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[37] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

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

[39] X. Jin, Y. Xie, X.-S. Wei, B.-R. Zhao, Z.-M. Chen, and X. Tan, "Delving deep into spatial pooling for squeeze-and-excitation networks," *Pattern Recognit.*, vol. 121, 2022, Art. no. 108159.

[40] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[41] 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.

[42] Z. Jiang, Z. Chen, K. Ji, and J. Yang, "Semantic segmentation network combined with edge detection for building extraction in remote sensing images," in *Proc. Int. Symp. Multispect. Image Process. Pattern Recognit.*, 2020, pp. 60–65.

[43] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[44] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 242–2424.

[45] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, 2022.

[46] J. X. Chang, X. J. Gao, Y. W. Yang, and N. Wang, "Object-oriented building contour optimization methodology for image classification results via generalized gradient vector flow snake model," *Remote Sens.*, vol. 13, no. 12, Jun. 2021, Art. no. 2406.

[47] C. Liao et al., "Joint learning of contour and structure for boundary-preserved building extraction," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1049.

[48] Y. Quan et al., "Building extraction from remote sensing images with DoG as prior constraint," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6559–6570, 2022.

[49] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, pp. 3051–3068, 2021.
[50] X. Chen, C. Qiu, W. Guo, A. Yu, X. Tong, and M. Schmitt, "Multiscale feature learning by transformer for building extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2503605.
[51] Z. Wang, N. Xu, B. H. Wang, Y. H. Liu, and S. W. Zhang, "Urban building extraction from high-resolution remote sensing imagery based on multi-scale recurrent conditional generative adversarial network," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 861–884, Dec. 31, 2022.
[52] 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.
[53] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2735–2745.
[54] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.
[55] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6392–6401.
[56] S. Seong and J. Choi, "Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3087.
[57] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2015.
[58] Y. Liu, E. Li, S. Wang, Y. Zhu, and W. Zhu, "Superpixel segmentation of high-resolution remote sensing image based on feature reconstruction method by salient edges," *J. Appl. Remote Sens.*, vol. 17, no. 2, 2023, Art. no. 026516.
[59] 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.
[60] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. of Toronto, 2013.
[61] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5686–5696.
[62] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[63] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[64] 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.
[65] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 17864–17875.
[66] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1280–1289.
[67] 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.
[68] Y. Qiu, F. Wu, J. C. Yin, C. Y. Liu, X. Y. Gong, and A. D. Wang, "MSL-Net: An efficient network for building extraction from aerial imagery," *Remote Sens.*, vol. 14, no. 16, Aug. 2022, Art. no. 3914.
[69] 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.
[70] Z. H. Che et al., "MAFF-HRNet: Multi-attention feature fusion HRNet for building segmentation in remote sensing images," *Remote Sens.*, vol. 15, no. 5, Mar. 2023, Art. no. 1382.
[71] G. D. Yan, H. T. Jing, H. Li, H. C. Guo, and S. He, "Enhancing building segmentation in remote sensing images: Advanced multi-scale boundary refinement with MBR-HRNet," *Remote Sens.*, vol. 15, no. 15, Aug. 2023, Art. no. 3766.

Yu Liu received the B.S. degree in electrical engineering and automation from the School of Xuhai, China University of Mining and Technology, Xuzhou, China, in 2019. He is currently working toward the M.S. degree in electronic information with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China.

His research interests include remote sensing image processing, semantic segmentation, and deep learning.

Erzhu Li received the M.S. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2014, and the Ph.D. degree in cartography and geographic information system from Nanjing University, Nanjing, China, in 2017.

He is currently an Associate Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in urban remote sensing.

Wei Liu received the M.S. and Ph.D. degrees in cartography and geographic information engineering from the China University of Mining and Technology, Xuzhou, China, in 2007 and 2010, respectively.

He is currently an Associate Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include spatial data quality checking, high-resolution remote sensing image processing, and GIS development and applications.

Xing Li received the M.S. degree in cartography and geographic information engineering from the Shandong University of Science and Technology, Qingdao, China, in 2006, and the Ph.D. degree in physical geography from East China Normal University, Shanghai, China, in 2010.

He is currently a Professor with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China. His research interests include high-resolution image processing and computer vision in environmental remote sensing applications.

Yuxuan Zhu received the B.S. degree in detection guidance and control technology from the School of Electronic Engineering and Optoelectronics Technology, Nanjing University of Science and Technology, Nanjing, China, in 2020. He is currently working toward the M.S. degree in photogrammetry and remote sensing with the School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou, China.

His research interests include remote sensing image processing, object detection, and deep learning.