MSB-Net: An End-to-End Network for Extracting Building from High-Resolution Remote Sensing Imagery

Guiwen Lan, Jia Wei, Hanqiang Huang, Fengfan Zou, and Dongbo Li

***Abstract*—Extracting buildings from high-resolution remote sensing imagery (HRSI) is of great significance to emergency management, land resource utilization, and analysis, as well as city planning and construction. However, due to the complex backgrounds and diverse appearances and different sizes of buildings in HRSI, most existing methods for automatic building extraction are difficult to obtain strong building feature representation from low-level and high-level features. Furthermore, existing research mainly focused on regional accuracy, whereas less attention was paid to the description of building boundaries. In this article, MSB-Net, an end-to-end neural network, is proposed to address these issues. A multiscale feature fusion module (MSFFM) is designed to capture and fuse multiscale features. A local branch (LB) constructed by the MSFFM and position attention, is used to obtain long range of context information between different positions and extract the essential features of buildings (e.g., shapes, edges) from low-level features. And a global branch (GB) is designed to use the MSFFM and channel attention to enhance high-level features. Therefore, our method can not only obtain information on building-related attribute categories, but also capture the rich context information in channel dimensions. The boundary enhancement and completion module take the output of the GB and LB as input to search for the missing parts and details of buildings to improve the segmentation accuracy and boundary quality. Our method is tested on two public building datasets and achieves superior classification performance.**

***Index Terms*—Boundary enhancement, building extraction, multiscale feature fusion.**

I. Introduction

EXTRACTING buildings from high-resolution remote sensing imagery (HRSI) can be of great help in various fields, including emergency management, land resource utilization, city planning and construction [1], [2], [3], and so on. With the improvement of HRSI data quality and availability, automatic building extraction from HRSI has become a hot research topic. However, there are still many challenges in building extraction for the following reasons. For example, high-resolution remote sensing images often exhibit small interclass discrimination and high intraclass variance, such as buildings with varying colors, textures, and sizes. In addition, buildings are frequently sheltered by trees and/or obscured by shadows. It is still necessary to design more effective methods for building extraction from HRSI.

Traditionally, building extraction mainly relies on feature operators designed by using features such as the spectra [4], texture [5], geometry [6], edges [7], and shadows [8] of buildings. Some algorithms in common use [9], [10], [11] utilized building features as auxiliary information to extract buildings. However, the abovementioned methods heavily rely on prior knowledge and handcrafted features, so it is difficult for them to achieve high recognition accuracy, and these methods are typically applicable to specific tasks.

Fully convolutional networks (FCNs) can be regarded as a groundbreaking work in image semantic segmentation [12]. By replacing the last fully connected layer of CNN with a transposed convolutional layer, FCN can conduct pixelwise prediction. Many semantic segmentation methods based on FCN were proposed, such as SegNet [13], PSPnet [14], Deeplab [15], and UNet [16], [17], [18] series, which achieve promising performance on some challenging datasets. Some FCN-based methods were then modified and improved to extract buildings from remote sensing images. At the same time, a lot of methods for building artificial neural networks, such as residual connection [19], spatial pyramid pooling [20], capsule feature pyramid network [21], multipath hybrid dilated convolution (HDC) [22], and so on, are applied to fuse or obtain multiscale features among different layers, and exchange information across channels, in order to improve the accuracy of segmentation and extraction. In [19], the features of spatial detail information of the buildings are highlighted with spatial attention units and residual learning, and the contextual information is captured with global features information awareness modules, then the features of different levels are aggregated with cross level feature recalibration modules to bridge the semantic gap between low- and high-level features. CapFPN by Yu et al. [21] uses a set of capsule layers instead of ordinary convolutional layers in its feature pyramid network, in order to represent the relationships between features at different positions. Multipath hybrid attention network [22] adopts a multipath HDC framework to capture building features with varying sizes and styles.

Furthermore, some studies have introduced attention mechanisms to fuse multilevel features, e.g., scene-driven multitask

Manuscript received 16 October 2023; revised 7 December 2023, 26 January 2024, and 22 March 2024; accepted 15 April 2024. Date of publication 23 April 2024; date of current version 23 May 2024. This work was supported by the National Natural Science Foundation of China under Grant 41861050. *(Corresponding author: Jia Wei.)*

The authors are with the College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541006, China (e-mail: 2009043@glut.edu.cn; 1020211827@gult.edu.cn; 2120211860@glut.edu.cn; 1020211829@glut.edu.cn; dongboli@glut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3392757

parallel attention [23] and attention block [24]. In [23] scene-driven multitask parallel attention is proposed to improve the image interpretation based on the scene prior, so as to extract buildings in large areas that cover different scenes. In [24], a multiloss neural network is designed by improving the sensitivity of the model with the attention block and suppress the background effects in irrelevant feature regions.

With the successful application of transformer to computer vision tasks, transformer-based building extraction has attracted a lot of attention, e.g., [25] and [26]. In [25] Swin (shifted windows) transformer is taken as the backbone for building extraction and inserted into a SOTA structure. In [26], a dual-pathway transformer structure is designed to learn the long-term dependency of tokens in both their spatial and channel dimensions. To reduce the computational complexity, buildings are represented as a set of "sparse" feature vectors, based on the assumption that single buildings in remote sensing images usually only occupy a very small part of the image pixels. Especially, the transformer-based segment anything model (SAM) [27] is regarded as a foundational model for image segmentation for its powerful generalization capabilities. However, SAM heavily relies on prior manual guidance, in [28] RSPrompter is proposed for learning how to generate appropriate prompts for SAM so as to produce semantically discernible segmentation results for remote sensing images. With a lightweight feature enhancer to collect features from various intermediate layers of the SAM ViT backbone, RSPrompter can generate prompts with semantic categories to enable SAM to generate multiple instance-level masks with category labels. Comparatively speaking, RingMo [29] is specifically designed as a foundation model developing framework for remote sensing image interpretation, which aims to obtain the foundation models trained in a generative self-supervised manner. The training method has an encoder–decoder structure, which takes ViT and Swin Transformer to extract latent representations of masked images and reconstructing the original images with L1 loss function. The pretraining foundation models can then be used in different downstream tasks, such as object detection and image segmentation. Obviously, transformer-based methods have been proven to be promising in this field, but they usually require a large volume of labeled training samples and computing resources, so it is not suitable to design lightweight building extraction algorithms with transformer. The FCN-based methods are still appealing in downstream tasks such as building extraction from HSRI.

Although the FCN-based approaches mentioned earlier have achieved significant advancements in building extraction, there are still some issues that need to be addressed. First, the buildings in HRSI have different sizes, diverse textures, and varied colors, which makes it difficult for most FCN-based networks to obtain the strong building feature representation. Second, boundaries can effectively describe the morphology and shape of buildings. The abovementioned FCN-based methods primarily focus on regional accuracy while frequently simplifying the description of building boundaries, so that they usually produce blurry boundaries in the extraction results and are unable to preserve the basic morphology and shape of the buildings. Therefore, some postprocessing methods [30], [31] or probability graph models [32] are widely adopted to improve boundaries of building segmentation.

In this article, an end-to-end network, MSB-Net, is designed to tackle the aforementioned issues. Different from the aforementioned approaches, our method enhances both low-level features and high-level features to obtain strong representation of building features, and reverse attention (RA) [38] is introduced to maintain the morphological shapes of the buildings, so as to improve boundary quality and regional segmentation accuracy. The main contributions are as follows.

1) A multiscale feature fusion module (MSFFM) is designed to effectively capture multiscale features. It is used as a decoder and encoder in the local branch (LB) and the global branch (GB). The LB can highlight detailed information (shape, edges, etc.) about buildings in low-level feature and establish long distance dependency between different positions by using position attention [33]. In the GB, channel attention [33] is used to suppress nonbuilding features in the high-level features and extract semantically strong features for building extraction.
2) A boundary enhancement and completion module (BECM) is designed to rectify the inconsistent predictions, and improve the completeness segmented regions and the clarity of building boundaries, by establishing the relationship between regions and boundary cues with RA [34].
3) We compared and tested our method and other methods on the WHU aerial building dataset and INRIA Aerial Image Labeling dataset. We have achieved optimal results and our method can extract buildings of different scales and retain clear boundary.

The rest of this article is organized as follows. In Section II, the methodology of MSB-Net is introduced in detail. The experiment settings and the datasets are introduced in Section III. Section IV covers the comparative experiments, ablation experiments, and the discussion. Finally, Section V concludes this article.

II. Methodology

Fig. 1 presents the structure of MSB-Net. It adopts ResNet50 [35] as the backbone network. ResNet50 is divided into four stages, each of which reduces the spatial resolution of the image by 1/2. The images are fed into ResNet50 to obtain four stages of feature maps. The low-level features, namely, Features 1 and 2, are input into the local branches to capture essential features of the building, such as shape and edges. Moreover, the high-level features, namely, Features 3 and 4, are utilized in the GBs to integrate the semantic information of buildings and obtain attribute information. Thereafter, we employ an aggregation module (AM) to aggregate the feature maps from the two branches, resulting in a global feature map. To further refine the boundary information of buildings and enhance the overall integrity of the extracted outputs, this global feature map, along with the individual feature maps from the two branches, is fed into the BECM. This module employs high-level feature maps as guidance regions to rectify the inconsistent predictions in the lower level feature maps.

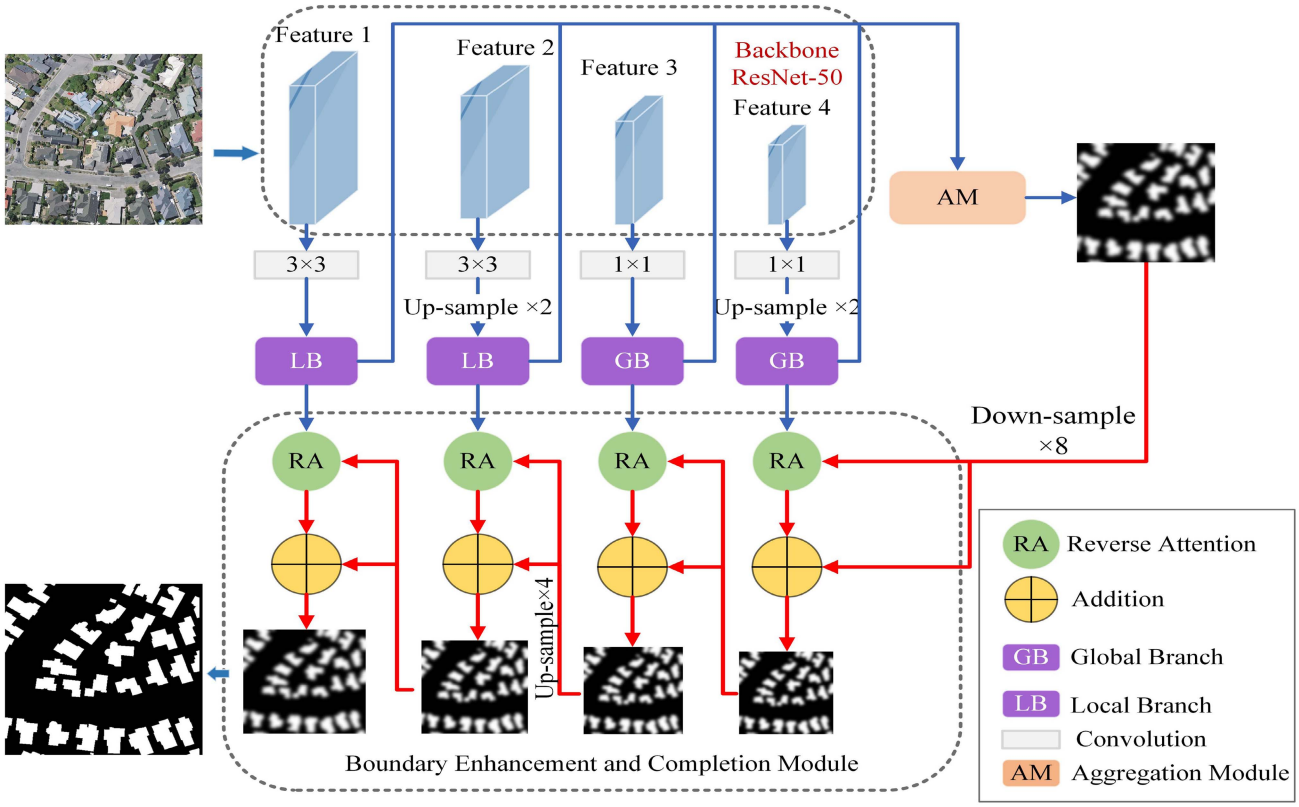

Fig. 1. Framework of MSB-Net.

A. Multiscale Feature Fusion Module

The ordinary 3 × 3 convolution operation is widely used in CNNs for feature learning. However, it requires stacking a lot standard 3 × 3 convolution layers to capture the multiscale features of the objects. On the contrary, the dilated convolution method [36] can increase the receptive field while keeping the size of the feature map unchanged. In this article, we design a MSFFM to capture and aggregate multiscale context information. By stacking dilated convolutions with different dilation rates and ordinary convolutions and then concatenating feature maps with different receptive fields, MSFFM can capture and fuse multiscale features.

MSFFM is designed as a parallel structure with five branches, as shown in Fig. 3, Feature A denotes the input feature maps. To reduce the number of channels of the feature maps, a 1 × 1 convolution layer is used in each branch except branch 1. In branch 1, two 3 × 3 convolutions are stacked. At the same time, to obtain the multiscale feature of the building, branches 2, 3, and 4 add a 3 × 3 dilated convolution layer [dilation rate is 2(k-1)] after the 1 × 1 convolution layer (k is the corresponding branch number) and branch 5 add a 1 × 1 convolution layer. To aggregate multiscale feature, the results of the last four branches are concatenated. After an addition operation between the result of the above 1 × 1 convolution and Feature A, the output of Branch 1 is connected to the addition result to complement and fuse multiscale feature.

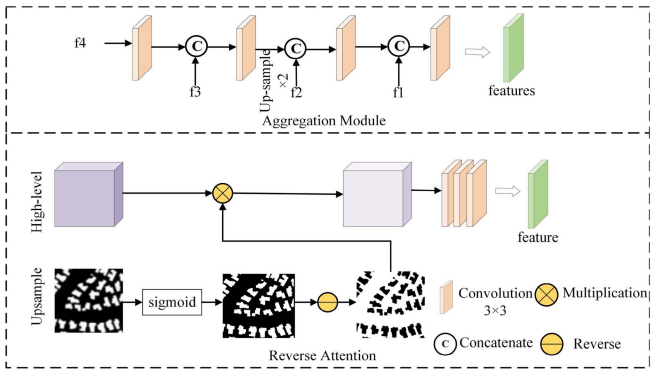

Fig. 2. Reverse attention module and AM.

B. Dual-Branch Structure

The dual-branch structure, which consists of the GB and the LB, is a standard encoder–decoder architecture that enables the acquisition of multilevel feature maps. It exploits skip connections to complement spatial information and facilitate the fusion of deep and shallow features. To capture multiscale building features, MSFFM is used as the encoder and decoder of both the GB and the LB.

1) Local Branch: Though low-level features have more detailed information (color, outline, texture, etc.), they contain a lot of noise and lack of semantic information, which could lead to misclassification of objects. As the position attention can encode a wider range of contextual information into low-level features

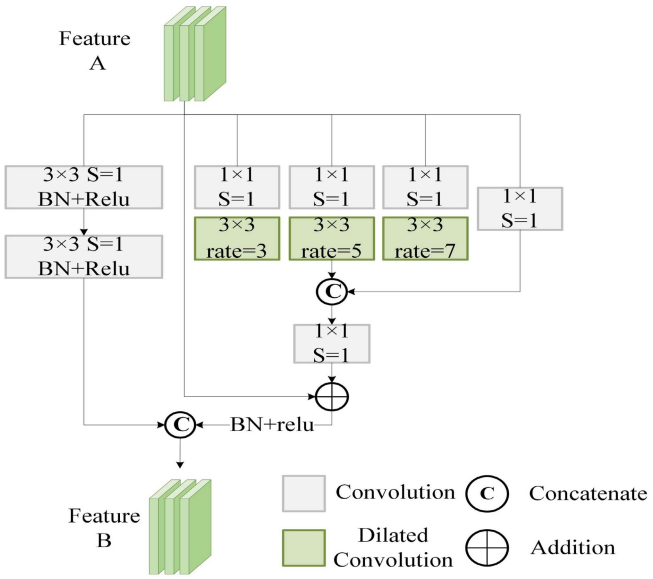

Fig. 3. MultiScale Feature Fusion Module.

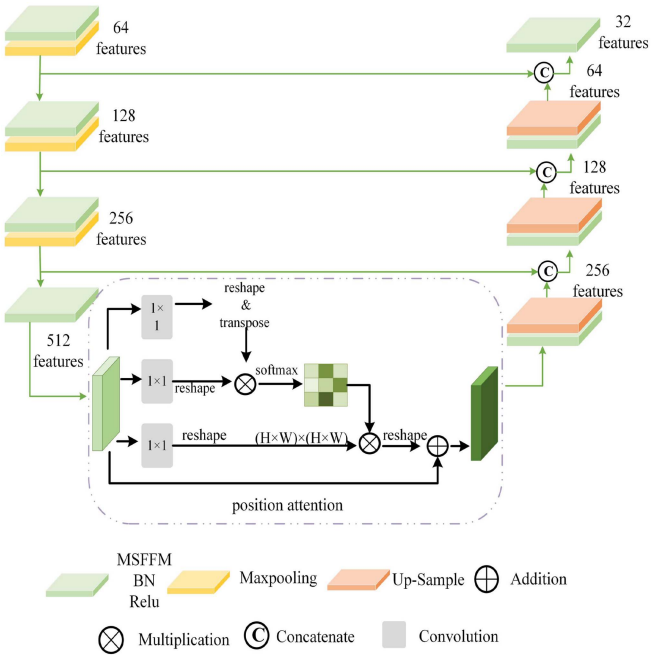

Fig. 4. Local branch.

[33], the LB is designed as a combination of the MSFFM and the position attention module to enhance the representation capability of low-level features and highlight detailed information (shape, edges, etc.) about buildings.

The LB is shown in Fig. 4. First, the feature maps are processed with four encoders with a downsampling stride of 2. Let $\{F_k^L \in R^{C_k \times H_k \times W_k}, k \in \{1,2,3,4\}\}$ denote feature maps generated from the encoder, which are sorted in ascending order from top to bottom of the encoder network. Here K = 4. Second, F_4^L is fed into a position attention module to compute position attention map to obtain the strength of the feature correlation between pixels. The position attention map is then added to F_4^G, resulting in $S(F_4^L)$ which is to capture long-range information between the locations of each building. Finally, $S(F_4^L)$ is passed through four decoders by setting the upsampling factor is 2 to increase the size of the feature maps and obtain multiscale building features and more semantic information from low-level features.

The input of the LB consists of Feature 1 (C = 256 channels, H, W = 128 × 128) and Feature 2 (C = 512 channels, H, W = 64 × 64). Here, C means channels, and H, W means feature map size. The outputs are denoted as f_1 and f_2 (f_1 and f_2 serve as inputs to the AM and BECM). Before being input into the LB, the following operations are performed on the input. Feature 2 is firstly upsampled to match the size of Feature 1. Then, to reduce the computational complexity, a 3 × 3 convolutional layer is used for both Features 1 and 2 to reduce the number of channels to 64.

2) Global Branch: High-level features contain rich semantic information and are very useful and significant for classification. However, due to the potentially complex texture of buildings and background in HRSI, it is necessary to highlight building features and suppress nonbuilding features for effective building extraction. As the channel attention mechanism [33] can emphasize interdependent channel maps by integrating related features between all channel maps, to this end, the GB is designed as a combination of the MSFFM and the channel attention module.

The structure of the GB and the LB is exactly the same, except that the positional attention mechanism is used in the LB, whereas the channel attention mechanism is used in GB. First, the feature maps are processed with four encoders with a downsampling stride of 2. Let $\{F_k^G \in R^{C_k \times H_k \times W_k}, k \in \{1,2,3,4\}\}$ be four feature maps obtained by the encoder from top to bottom. And then, F_4^G is fed into a channel attention module to compute channel attention map to obtain the strength of the feature correlation between channels. The channel attention map is then added to F_4^G, resulting in $C(F_4^G)$, which highlights the building features and clusters rich context information. Finally, $C(F_4^G)$ is passed through four decoders with the up-sampling factor is 2 to increase the size of the feature maps and explore the semantically strong features related to buildings in the high-level features.

The input of the GB consists of Feature 3 (C = 1024 channels, H, W = 32 × 32) and Feature 4 (C = 2048, H, W = 16 × 16). The outputs are denoted as f_3 and f_4 (f_3 and f_4 serve as inputs to the AM and BECM). Before being input into the LB, the following operations are performed on the input features. Feature 4 is firstly upsampled to match the size of Feature 3. Then, to reduce the computational complexity, both feature maps are passed through a 1 × 1 convolution layer to reset the number of channels to 64.

C. Boundary Enhancement and Completion Module

The BECM is designed to refine the boundaries of buildings and maintain their regional integrity. As it is well known, high-level semantic information (such as the coarse spatial locations and shapes) of objects is encoded in the high-level features, whereas low-level semantic information (such as the details of the boundaries) is encoded in the low-level ones. Many studies have shown RA [34] is effective in refining the shapes and the

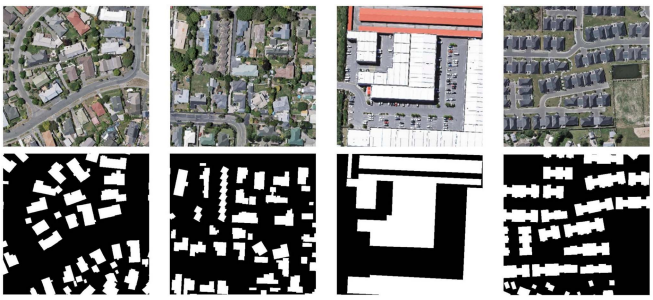

Fig. 5. WHU Building dataset.

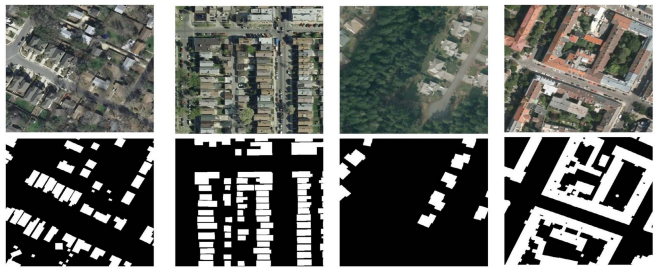

Fig. 6. InriaAIL dataset.

boundaries of detected objects, by taking deep-level features as guidance for shallow-level features to rectify the inconsistent predictions. In this article RA is used as a component of the BECM, to establish relationships between regions and boundaries, to improve boundary quality and regional segmentation accuracy.

In the BECM, shown in Fig. 1, we design four branches to fully use the input feature maps from the GB and LB by establishing relationships between regions and boundaries in top–down manner to rectify the inconsistent predictions. This goal is to improve the boundaries of buildings and explore the missing parts of the target. The purpose of the AM is to aggregate feature maps into higher level layers, resulting in a global feature map that acts as the initial guidance region. The feature map M (32 channels, the size 128 $\times$ 128) is obtained by passing $\{f_i, i \in \{1, 2, 3, 4\}\}$ through the AM. As shown in Fig. 2, initially let $i = 4$, AM is executed as the following steps:

1) The feature f_{i} and M is fed into the RA module to compute the RA feature map R_i;
2) An addition is performed in R_i and M and obtain the rectified feature map P_i;
3) Let $M = P_i$, then $i = i$-1, repeat steps 1) and 2), until $i =$ 1, to obtain the final corrected feature map.

III. Experiments

To evaluate the ability of MSB-Net to extract buildings from HRSI, we trained and tested our model on the WHU building dataset and the INRIA Aerial Image Labeling dataset, where the binary cross-entropy loss function [37] is the loss function. We used evaluation metrics such as precision accuracy (PA), precision (PC), intersection over union (IOU), F1 score (F1), Recall, as well as visualizing the results to assess and analyze the performance of MSB-Net.

A. Dataset

WHU Building dataset (WHU dataset) [38] comprises unmanned aerial vehicle and satellite images with a resolution ranging from 0.0075 m to 0.3 m. The dataset was divided into a training set, validation set, and test set, including 4736, 1036, and 2416 images, respectively, with a size of 512 $\times$ 512 pixels. We followed the official settings in our experiments. Some images and ground truth of the dataset are presented in Fig. 5.

There are 360 remote sensing images in the INRIA Aerial Image Labeling dataset (InriaAIL dataset) [39], with a spatial resolution of 0.3 m/pixel and a size of 5000 $\times$ 5000 pixels. The dataset includes a training set and a test set, each containing 180 images that cover a range of different areas. Only image labels (ground truth) are provided in the training set, whereas they are not provided in the test set. We divided the training set according to the guidelines in [39], selecting the first five images of each region as the test set, and the remaining images as the training set and verification set. We have opted to crop the original images into 512 $\times$ 512 pixels with an overlap of 104 pixels, due to the limitation of GPU memory. To ensure effective training, we removed images without buildings. Fig. 6 shows some samples from the dataset.

B. Experimental Details

In our experiments, all the models were implemented to the PyTorch framework on an NVIDIA Quadro RTX 5000 GPU (16 G). The Adam optimizer [40] was used with an initial learning rate of 2e$-$4 and a batch size of 12. An initial learning rate was subsequently reduced by 0.95 after every 1 epoch. We set weight decay to 1e$-$4 and adopted data augmentation strategies to prevent overfitting. Data augmentation strategies consist of random rotations in increments of 90° from 0° to 270° along both horizontal and vertical directions, random vertical and horizontal flip. Our model was trained for 100 epochs on the GPU.

IV. Results and Discussion

A. Experiments on WHU Building Dataset

We conduct a comparison with seven other methods on the WHU dataset to validate the effectiveness of the proposed method. These methods included three classic semantic segmentation methods and three recent works, namely, U-Net, Link-Net [41], DeepLabV3+ [42], SST [26], and HRNet [43] on the WHU dataset, where U-Net and DeepLabV3+ used ResNet50 as their feature extractor. We used the same settings in the InriaAIL dataset. We also cite recent work to validate the effectiveness of our method, namely, CFENet [44]. Due to the limitation of GPU memory, we chose SST (RS18, S4) and HRNet-w30.

We quantitatively evaluated the performance of these seven methods, and their performances are presented in Table I. In terms of all evaluation metrics, our method surpasses other methods. In the comparison with classic semantic segmentation methods, SST performs the best in building extraction tasks. Compared with SST, our method improves by 0.25% on PA, 1.22% on PC, 1.1% on F1, 0.97% on Recall, and 1.94% on IOU.

TABLE I
PERFORMANCE OF SEVEN METHODS ON THE WHU DATASET

Method	PA (%)	PC (%)	F1(%)	Recall (%)	IOU (%)
U-Net	98.21	91.85	91.97	92.09	85.13
DeepLabV3+	98.00	91.76	90.94	83.99	83.39
Link-Net	97.99	90.09	91.05	92.03	83.57
HRNet	98.18	91.72	91.84	91.96	84.91
SST	98.47	92.79	93.15	93.52	87.19
CEFNet	98.71	91.09	92.62	-	87.22
MSB-Net	**98.72**	**94.01**	**94.25**	**94.49**	**89.13**

"-" Indicates that the data item was not provided in the original literature.
The bold values indicate the best-performing numerical values for each evaluation indicator.

TABLE II
PERFORMANCE OF SIX METHODS ON THE INRIAAIL DATASET

Method	PA (%)	PC (%)	F1(%)	Recall (%)	IOU (%)
U-Net	96.36	87.46	86.68	85.92	76.49
DeepLabV3+	95.90	85.94	84.96	83.99	73.85
Link-Net	95.91	86.98	84.79	82.71	73.60
HRNet	96.11	86.97	85.67	84.40	74.39
SST	96.31	87.40	86.47	85.56	76.16
MSB-Net	**96.63**	**87.67**	**87.81**	**87.94**	**78.26**

The bold values indicate the best-performing numerical values for each evaluation indicator.

Compared with recent work, our method also demonstrates good performance.

As shown in Fig. 7, we qualitatively evaluated the building extraction ability of six methods. Compared with these methods, better visual results in building extraction are obtained by MSB-Net. U-Net can identify most of the building pixels but cannot maintain the shape of the building. Link-Net cannot effectively recognize and overcome the influence of background with similar spectra (row 5). The main problem of DeepLabV3+ is that the boundary of the extraction result is jagged and not smooth enough, and it cannot effectively overcome the influence of building shadows and tree occlusion. HRNet is unable to fully recognize large buildings (row 6). Aside from our method, SST performs best in terms of visual effect. And our method outperforms SST in overcoming shadow and tree occlusion. (row 4 and row 7).

Rows 1–4 correspond to the extraction of small-scale buildings, our method extracted buildings quite well even for very small buildings (row 1 and 2), whereas other methods either did not recognize them or only partially recognized them. Among them, U-Net recognized most of the buildings but failed to maintain their basic shape (row 1). SST identifies fewer correct pixels than our methods (row 2). Regarding the extraction of densely packed small buildings, our method overcame the influence of building shadows and tree occlusion, extracting buildings while also ensuring clear boundaries. Regarding portions heavily covered by trees, our method extracts majority of the buildings (row 4). Rows 5–7 represent the result of multiscale building extraction. Our method can successfully achieve clear boundaries of large buildings (rows 6–8) with complex boundaries. It effectively overcame the influence of the background environment and complex building texture. In row 6, U-Net and SST recognized most of the buildings but with unclear boundaries. Rows 5 and 8 show building extraction results at different scales, and our method almost perfectly eliminates the adhesion problem of adjacent buildings, ensuring the integrity and clear boundaries of small buildings with complex boundaries.

B. Experiments on InriaAIL Dataset

We conduct a comparison with six methods on the InriaAIL dataset to further validate the performance of our method. Table II presents the experimental results of different methods. Our model outperformed other models in all evaluation metrics. Fig. 8 presents the visual extraction results of MSB-Net and five other methods.

For buildings with complex boundaries, the results of rows 1 and 2 indicate that our method can extract complete buildings with clear boundaries. With respect to buildings with complex textures, the results of the rows 7 and 8 demonstrate that we effectively overcome the influence of complex textures on building extraction results, basically ensuring the integrity of buildings. Although U-Net can overcome the influence of complex textures, it cannot ensure the basic shape of buildings relative to our method. The results in rows 3–6 demonstrate that our method successfully mitigates the impact of building shadows and extracts large and small buildings effectively, solving the

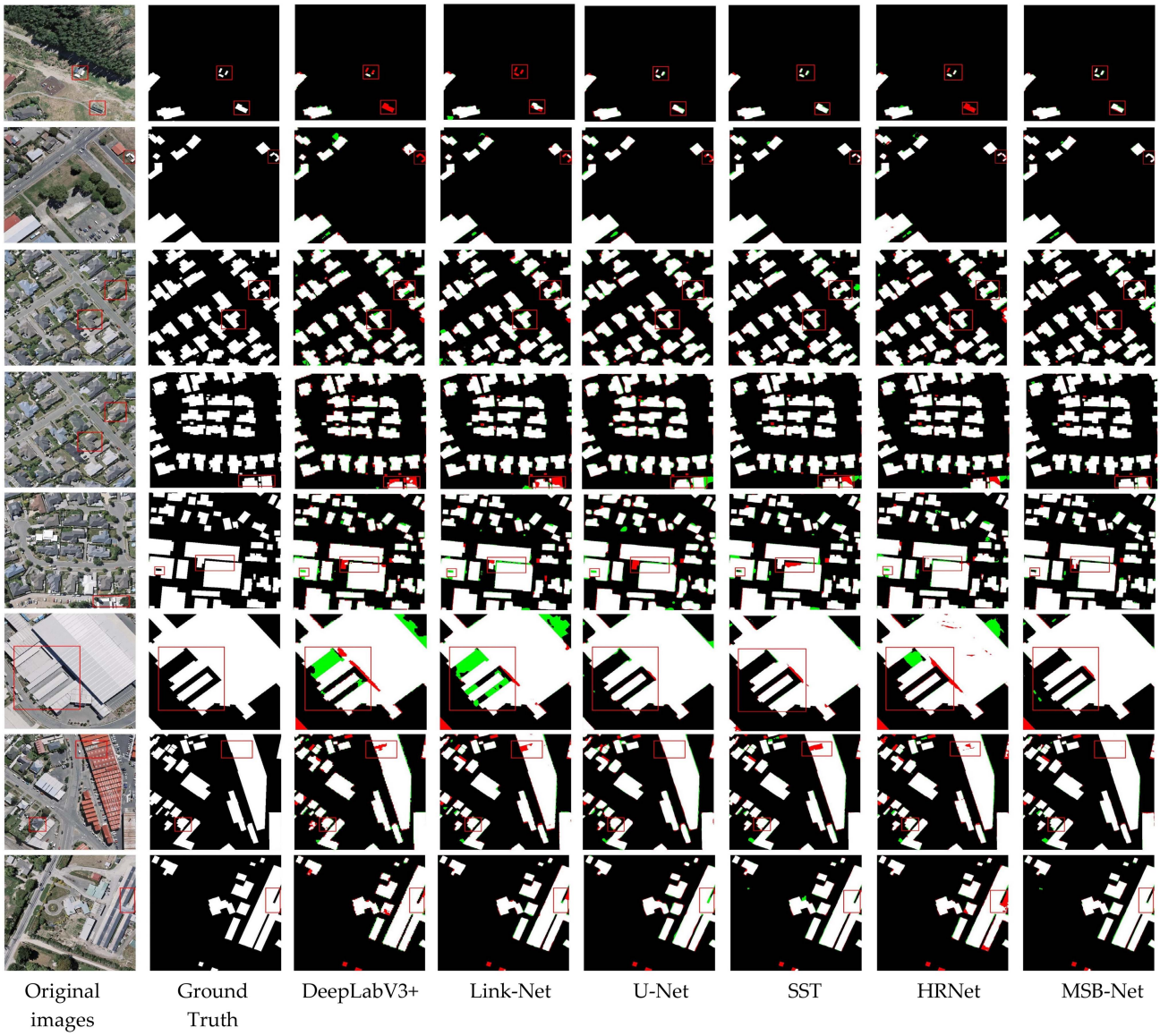

Fig. 7. Performance of six methods on the WHU dataset. White, green, and red pixels on the map denote true positive, false positive, and false negative predictions, respectively. The area in red box shows the regions where the comparative impact is more obvious.

adhesion phenomenon of small dense buildings. Other methods cannot distinguish objects with similar building features well (row 6).

According to the above visualization results and evaluation metric analysis, it can be deduced that MSB-Net effectively enhances the capturing of multiscale features of buildings, aggregates context information, and enhances boundary information. Therefore, MSB-Net demonstrated superior visual extraction capabilities on both datasets, particularly for small-scale, large-scale, and buildings with complex shape. It successfully overcomes the challenges caused by tree occlusion and building shadows.

C. Ablation Experiment

We conducted validation of the three modules in MSB-Net using the WHU dataset, taking FCN (ResNet50) as the baseline model. FCN$*$ denotes the utilization of 3 $\times$ 3 convolutions to decrease the channel dimensions of feature maps from all four stages of ResNet50 to 16 individually. Then, we perform 3 $\times$ 3 convolution to aggregate these feature maps and obtain the final resulting map. The effectiveness of the three modules was quantitatively evaluated using F1 score and IOU, as shown in Table III. Compared with FCN, FCN$*$ improved F1 and IOU by 14.36% and 22.25%, respectively. It can be inferred from this result that the utilization of both high-level and low-level features can make a significant improvement in building segmentation results.

Compared with the baseline network, each individually added module in the experiments has yielded good results. In particular, the addition of LB achieved the best results, with the F1 score and IOU increasing from 79.98% and 65.26% to 93.84% and 88.40%, respectively. The reason for this improvement is that LB, when processing low-level features, not only extracts

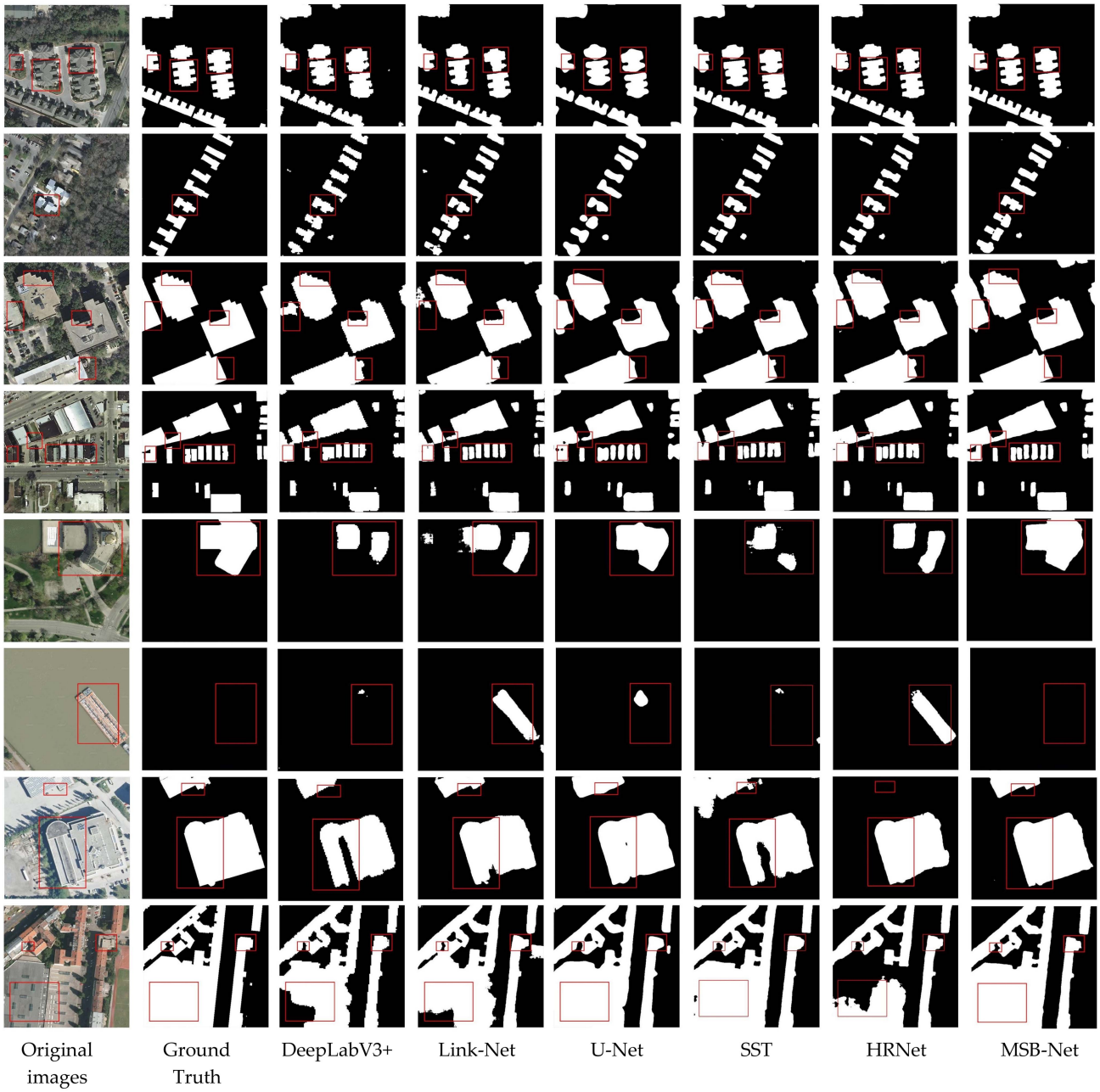

Fig. 8. Results of different methods on the InriaAIL dataset. The area in red box indicates the areas where the comparative effects are more pronounced.

the semantic information of buildings but also emphasizes the spatial information of the buildings, whereas GB primarily emphasize semantic information and the results of BECM are largely dependent on the accuracy of the predicted results of the input feature maps. It can be inferred from this result that spatial information contained in low-level features is crucial for building recognition. Furthermore, the incorporation of both the LB and GB modules further enhanced F1 score and IOU by 14.86% and 23.44%, respectively, illustrating the improved feature extraction ability achieved by leveraging high-level and low-level features based on their respective characteristics.

When all three modules are used in MSB-Net, F1 and IOU are 94.25% and 89.13%, respectively. In summary, the ablation experiments conducted for each module demonstrated the effectiveness of LB, GB, and BECM in improving model performance. Importantly, each component is essential to obtain the best building extraction results. From Table III, it can be observed that the utilization of high-level features is not as effective as that of low-level features. In future work, without compromising the model's runtime, the focus of improvement will be on further exploring and exploiting high-level feature information.

TABLE III
RESULT OF THE ABLATION STUDY

Method	BaseNet	Unit			F1(%)	IOU(%)
		LB	GB	BECM		
FCN	ResNet50				78.98	65.26
FCN*	ResNet50				93.34	87.51
MSB-Net	ResNet50	√			93.84	88.40
MSB-Net	ResNet50		√		90.85	83.24
MSB-Net	ResNet50			√	91.31	84.02
MSB-Net	ResNet50	√	√		94.01	88.70
MSB-Net	ResNet50	√	√	√	94.25	89.13

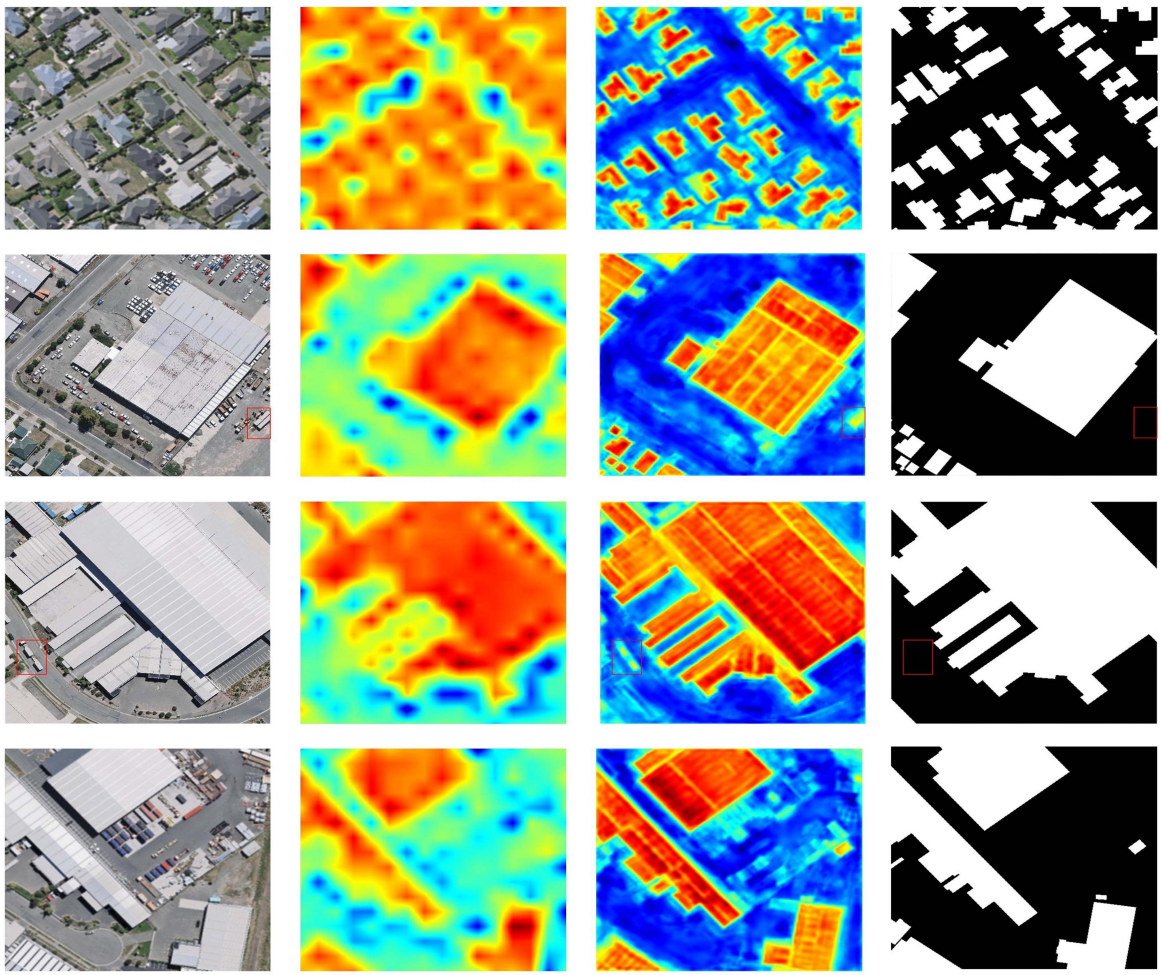

Fig. 9. Heatmaps of different images. ResNet-50 Output is the feature map generated by Res-Net-50, whereas MSB-Net Output is the feature map outputted by MSB-Net.

D. Discussion

1) Model Complexity Comparison: The practical application of a model is significantly influenced by its complexity. In the context of building extraction tasks using CNN methods, models with fewer parameters and floating-point operations (FLOPs) tend to offer faster training and inference speeds. Therefore, models with lower complexity are more suitable for practical applications. To provide an objective assessment of the complexity of each method , we calculated the number of parameters

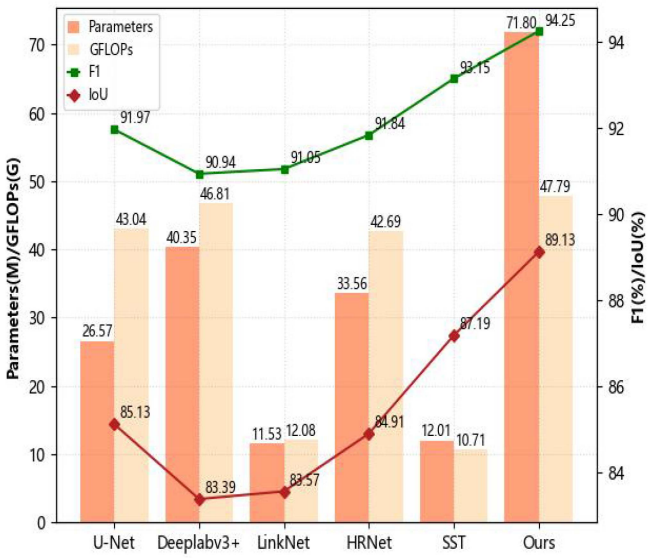

Fig. 10. Parameters/GFLOPs/F1/IOU of different networks.

and GFLOPs of several methods separately, as shown in Fig. 10. GFLOPs were computed from a tensor with dimensions of $1 \times 3 \times 512 \times 512$. Although MSB-Net has a relatively large model parameter (71.8 M), its GFLOPs are slightly larger than those of U-Net and Deeplabv3+ by 0.98 G and 4.75 G, respectively. This result shows that our model demands a relatively larger storage capacity in comparison to the other five approaches, while attaining faster inference speed. Notably, our method exhibits the highest F1 score and IOU among all four methods, indicating its superiority in terms of segmentation outcomes. In addition, the SST achieves good results with relatively low GFLOPs. Hence, it is important for future research to effectively make a tradeoff between computational complexity and precision.

2) Effectiveness of MSB-Net Modules is Verified by Heatmaps: We conduct a heatmap analysis on the feature map of test images to further verify the efficacy of the proposed modules. The heatmap visualizations of these feature maps are shown in Fig. 9. The varying levels of brightness indicate the model's different levels of attention to building features. We compared the feature maps outputted by ResNet50 and MSB-Net. While ResNet50 struggles to effectively suppress background interference and only provides a rough identification of building information, it also faces challenges when it comes to accurately identifying building boundaries. As a result, there is a higher likelihood for misidentification of objects with similar spectra as buildings. In the feature heatmap of MSB-Net buildings are depicted as red, and background is depicted as blue. The result shows that MSB-Net can effectively recognize buildings, and overcome noise interference such as tree obstruction and building shadows. This indicates that our proposed GB, local branch, and BECM can effectively extract buildings of different sizes and make that the extraction results have clear boundaries and complete regions. As shown in the red box in Fig. 10, our method still struggles to accurately identify objects that bear a striking resemblance to the appearance of buildings. Therefore, we will further investigate this issue in our future research.

3) Verification of Boundary Shape Learning: To assess the efficacy of this method in learning boundary shapes, we used two boundary metrics to quantitatively evaluate its performance, including Hausdorff distance (HD) [45] and structural similarity (SSIM) [46].

In total, 95% HD means multiplying HD by 95% to eliminate the influence of very small outlier clusters. The smaller the distance, the more similar the predicted shape is to the shape of the real label. In this article, the ground truth for nonbuilding areas is 0, and the predicted results that do not contain buildings have a 95% HD value of 0. SSIM ranges from -1 to 1. The SSIM is set to 1 when the two images are the same.

The evaluation metrics of the two datasets are shown in Table IV. Our proposed method achieved the highest values in F1, IOU, and SSIM, and the lowest value in 95% HD, for both the InriaAIL dataset and the WHU dataset. This indicates that MSB-Net exhibit the highest shape similarity and resemblance to the ground truth values. This result demonstrates the advantage of our method in learning building boundaries and yields the overall best segmentation results.

4) Generalization Ability of MSB-Net: The Satellite dataset I (global cities) from WHU Building Dataset [WHU (global cities)] was selected to evaluate the generalization ability of our model. This dataset comprises 204 images (512×512 tiles) and encompasses remote sensing images of ten cities, namely, Wuhan, Taiwan, New York, Santiago, Milan, Venice, Los Angeles, Ottawa, Cairo, and Cordoba. The images have varying resolutions ranging from 0.3 m to 2.5 m, and they exhibit diverse architectural styles and distribution patterns. Hence, this dataset is highly suitable for validating the generalization capability of our model.

We conduct transfer learning to validate the generalization ability of MSB-Net. The training results on the WHU dataset were tested on the InriaAIL dataset and WHU (global cities), with results presented in Table V. Similarly, the training results on the InriaAIL dataset were tested on the WHU Building Dataset and WHU (global cities), with results shown in Table VI. Table V shows that our network achieved the best results on the InriaAIL dataset and Link-Net performed well in WHU (global cities). Table VI shows that all methods performed well with transfer learning on the WHU Dataset and WHU (global cities), however, SST achieved the best results in both datasets.

We observed that methods performing well on the InriaAIL dataset exhibit relatively poorer performance on the WHU (global city) dataset. By comparing Tables V and VI, we can observe that the features learned by various methods on the InriaAIL dataset are better suited for the WHU (global city) dataset. This is because compared with the WHU building dataset, the INRIA dataset contains a greater variety of building types.

Overall, our approach has demonstrated relatively stable on different datasets compared with other methods, suggesting a favorable generalizability of our approach. It is noteworthy that the SST, a transformer-based lightweight building extraction model, exhibited excellent performance in the transfer experiments, except for the transfer experiments on the WHU dataset and WHU (global cities) dataset. In future work, we should consider

TABLE IV
PERFORMANCE OF VARIOUS APPROACHES ON THE WHU DATASET AND THE INRIAAIL DATASET IN TERMS OF F1, IOU, 95% HD, AND SSIM

Method	WHU dataset				InriaAIL dataset			
	F1(%)	IOU(%)	95%HD	SSIM(%)	F1(%)	IOU(%)	95%HD	SSIM(%)
U-Net	91.97	85.13	29.00	94.19	86.68	76.49	49.65	91.42
DeepLabV3+	90.94	83.39	35.26	93.72	84.96	73.85	55.94	90.84
Link-Net	91.05	83.57	33.15	94.38	84.79	73.60	55.38	90.84
HRNet	91.84	84.91	38.30	94.62	85.67	74.39	59.49	91.26
SST	93.15	87.19	32.17	95.11	86.47	76.16	58.51	91.49
MSB-Net	**94.25**	**89.13**	**28.33**	**95.59**	**87.81**	**78.26**	**47.42**	**92.00**

The bold values indicate the best-performing numerical values for each evaluation indicator.

TABLE V
RESULT OF TRANSFER LEARNING BY SIX METHODS ON THE INRIAAIL DATASET AND WHU (GLOBAL CITIES)

Method	WHU dataset		Transfer to InriaAIL dataset		Transfer to WHU (global cities)	
	F1(%)	IOU(%)	F1(%)	IOU(%)	F1(%)	IOU(%)
U-Net	91.97	85.13	67.89	51.39	37.90	23.38
DeepLabV3+	90.94	83.39	53.56	36.58	44.35	28.49
Link-Net	91.05	83.57	57.07	39.93	**50.57**	**33.85**
HRNet	91.84	84.91	50.41	33.70	44.99	29.02
SST	93.15	87.19	68.14	51.68	38.33	23.71
MSB-Net	**94.25**	**89.13**	**68.18**	**51.72**	43.69	27.95

The bold values indicate the best-performing numerical values for each evaluation indicator.

TABLE VI
RESULT OF TRANSFER LEARNING BY SIX METHODS ON THE WHU DATASET AND WHU (GLOBAL CITIES)

Method	InriaAIL dataset		Transfer to InriaAIL dataset		Transfer to WHU (global cities)	
	F1(%)	IOU(%)	F1(%)	IOU(%)	F1(%)	IOU(%)
U-Net	86.68	76.49	73.58	58.21	53.68	36.69
DeepLabV3+	84.96	73.85	78.42	64.50	53.87	36.86
Link-Net	84.79	73.60	79.67	66.21	54.46	37.42
HRNet	85.67	74.39	79.71	66.72	54.38	37.34
SST	86.47	76.16	**81.65**	**68.99**	**56.70**	**39.56**
MSB-Net	**87.81**	**78.26**	75.96	61.24	49.56	33.03

The bold values indicate the best-performing numerical values for each evaluation indicator.

collecting more diverse types of building data to enhance model generalization. One of the directions for model improvement is to explore the application of transformers in the model without compromising its operational efficiency.

V. CONCLUSION

In this article, we propose a new method (MSB-Net) on extracting buildings from high-resolution remote sensing images. First, MSB-Net utilizes LB and GB to enhance low-level and high-level features, and obtains rich contextual information and multiscale features of buildings within the remote sensing images. Second, MSB-Net uses BECM to further improve the segmentation accuracy and enhance building boundary information. We test MSB-Net in the WHU dataset and the InriaAIL dataset, achieving IOU scores of 89.13% and 78.26%, respectively. By comparing with five existing methods, the superiority of MSB-Net in accurately extracting building footprints was confirmed. In addition, our method achieved the best scores for boundary accuracy, as indicated by HD and SSIM metrics. This indicates that 1) MSB-Net is an effective and accurate building extraction model, capable of accurately extracting building boundaries, getting a good result in buildings of different scales, and overcoming the impact of building shadows and tree occlusions. 2) MSB-Net exhibits an advantage in learning boundary shapes. There are several potential avenues for further exploration and improvement of our method. 1) The model is limited in situations where there is severe tree shading and shadows. In future work, we will consider paying more attention to auxiliary information to solve these problems. 2) The model is highly dependent on annotated data. In future work, the semisupervised learning will be considered to reduce the reliance on annotated data and improve data diversity.

REFERENCES

[1] 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.

[2] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, "Building extraction from multi-source remote sensing images via deep deconvolution neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1835–1838.

[3] S. Georganos et al., "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 221–242, Mar. 2018.

[4] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *ISPRS J. Photogrammetry Remote Sens.*, vol. 54, no. 1, pp. 50–60, Feb. 1999.

[5] P. S. Tiwari and H. Pande, "Use of laser range and height texture cues for building identification," *J. Indian Soc. Remote Sens.*, vol. 36, no. 3, pp. 227–234, Sep. 2008.

[6] L. Zhang, X. Huang, B. Huang, and P. Li, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.

[7] G. Ferraioli, "Multichannel InSAR building edge detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1224–1231, Mar. 2010.

[8] Y.-T. Liow and T. Pavlidis, "Use of shadows for extracting buildings in aerial images," *Comput. Vis., Graph., Image Process.*, vol. 49, no. 2, pp. 242–277, Feb. 1990, doi: 10.1016/0734-189X(90)90139-M.

[9] J. P. Cohen, W. Ding, C. Kuhlman, A. Chen, and L. Di, "Rapid building detection using machine learning," *Appl. Intell.*, vol. 45, no. 2, pp. 443–457, Sep. 2016.

[10] S. Zhong, J. Huang, and W. Xie, "A new method of building detection from a single aerial photograph," in *Proc. 9th Int. Conf. Signal Process.*, 2008, pp. 1219–1222.

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

[12] 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.

[13] 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.

[14] 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.

[15] L.-C. Chen, G. Papandreou, I. 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.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[18] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1055–1059.

[19] 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.

[20] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019.

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

[22] 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.

[23] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[24] 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, Apr. 2020, Art. no. 1400.

[25] 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.

[26] 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.

[27] A. Kirillov et al., "Segment anything," Apr. 2023. Accessed: May 7, 2024. [Online]. Available: http://arxiv.org/abs/2304.02643

[28] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," Nov. 29, 2023. Accessed: May 8, 2024. [Online]. Available: http://arxiv.org/abs/2306.16269

[29] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612822.

[30] H. Yang, M. Xu, Y. Chen, W. Wu, and W. Dong, "A postprocessing method based on regions and boundaries using convolutional neural networks and a new dataset for building extraction," *Remote Sens.*, vol. 14, no. 3, Jan. 2022, Art. no. 647.

[31] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, Jan. 2018, Art. no. 144.

[32] Q. Zhu, Z. Li, Y. Zhang, and Q. Guan, "Building extraction from high spatial resolution remote sensing images via multiscale-aware and segmentation-prior conditional random fields," *Remote Sens.*, vol. 12, no. 23, Dec. 2020, Art. no. 3983.

[33] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3141–3149.

[34] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection." Apr. 15, 2019. Accessed: Jun. 30, 2023. [Online]. Available: http://arxiv.org/abs/1807.099409

[35] 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.

[36] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," Apr. 30, 2016. Accessed: Jun. 30, 2023. [Online]. Available: http://arxiv.org/abs/1511.07122

[37] P.-T. D. Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Operations Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[38] 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.

[39] 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.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 29, 2017. Accessed: Jun. 30, 2023. [Online]. Available: http://arxiv.org/abs/1412.6980

[41] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. 2017 IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis. – ECCV 2018*, 2018, pp. 833–851.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation." Feb. 25, 2019. Accessed: Nov. 30, 2023. [Online]. Available: http://arxiv.org/abs/1902.09212

[44] J. Chen, D. Zhang, Y. Wu, Y. Chen, and X. Yan, "A context feature enhancement network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 9, May 2022, Art. no. 2276.

[45] 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.

[46] 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.