MAP-Net: Multi Attending Path Neural Network for Building Footprint Extraction from Remote Sensed Imagery

Qing Zhu, Cheng Liao, Han Hu, Xiaoming Mei, Haifeng Li*, Member, IEEE

Abstract—Building footprint extraction is a basic task in the field of surveying, 3D building reconstruction, urban dynamic monitoring, disaster emergency response, etc. Accurately and efficiently extracting building footprints from a wide range of remote sensed imagery remains a challenge due to their complex structure, variety of scales and diverse appearances. Existing convolutional neural network (CNN)-based building extraction methods are complained that they cannot detect the tiny buildings because the spatial information of CNN feature maps are lost during repeated pooling operations of the CNN, and the large buildings still have inaccurate segmentation edges. Moreover, features extracted by a CNN are always partial which restricted by the size of the respective field, and large-scale buildings with low texture are always discontinuous and holey when extracted. This paper proposes a novel multi attending path neural network (MAP-Net) for accurately extracting multiscale building footprints and precise boundaries. MAP-Net learns spatial localization-preserved multiscale features through a multi-parallel path in which each stage is gradually generated to extract high-level semantic features with fixed resolution. Then, an attention module adaptively squeezes channel-wise features from each path for optimization, and a pyramid spatial pooling module captures global dependency for refining discontinuous building footprints. Experimental results show that MAP-Net outperforms state-of-the-art (SOTA) algorithms in boundary localization accuracy as well as continuity of large buildings. Specifically, our method achieved 0.68%, 1.74%, 1.46% precision, and 1.50%, 1.53%, 0.82% IoU score improvement without increasing computational complexity compared with the latest HRNetV2 on the Urban 3D, Deep Globe and WHU datasets, respectively. The TensorFlow implementation is available at https://github.com/lehaifeng/MAPNet.

Index Terms—Building footprint extraction, deep learning, semantic segmentation, attention mechanism, remote sensing imagery.

I. INTRODUCTION

The rapid development of remote sensing technology has made it easier to acquire a large number of high-resolution optical remote sensing images that support the extraction of building footprints in a wide range. Immediate and accurate building footprint information is significant for illegal building monitoring, 3D building reconstruction, urban planning and disaster emergency response. Due to low inter-class variance and high intra-class variance in buildings in optical remote sensing imagery, parking lots, roads and other non-buildings are highly similar to buildings in appearance. With the variety of building materials, scales and illumination, the representation of buildings in remote sensing imagery shows significant differences. Therefore, how to accurately and efficiently extract building footprints from remote sensing imagery remains a challenge.

Over the past two decades, numerous algorithms have been proposed to extract building footprints. They can be divided into two categories: traditional image processing based and CNN-based methods.

Traditional building extraction methods utilize the characteristics of the spectrum, texture, geometry and shadow [1]–[6] to design feature operators for extracting buildings from optical images. Since these features vary under different illumination conditions, sensor types and building architectures, traditional methods can only resolve specific issues on specific data. [7]–[10] combined optical imagery with GIS [11], digital surface models (DSM) obtained from light detection and ranging (Lidar) or synthetic aperture radar interferometry [12] to distinguish non-building areas that are highly similar to buildings, which increased the robustness of building extraction, although a wide range of corresponding multi-source data acquisition is always costly.

Because buildings in remote sensing images are diverse in structure, appearance and scale, building extraction algorithms have evolved from handcrafted feature-based methods to learning feature-based methods, such as deep convolutional neural networks (DCNNs). Moreover, deep networks have been practical in designing CNN models. For instance, ResNet [13] with 152 layers introduced identity mapping in a residual block to solve the problem of gradient explosion in propagation, making it possible to design a deeper network to extract richer semantic features.

Evolving from CNN, fully convolutional network (FCN)-based [12], [14]–[23] building extraction methods achieve incredible results and are often used in building semantic segmentation tasks. The encoder-decoder framework [20], [24]–[29] obtains more accurate building footprints compared with FCN-based methods, particularly on the localization of the boundary since they recover spatial details through skip connections to fuse shallow high-resolution features in the decoder stage. Nevertheless, coarse features introduced by
shallow layers become the main obstacle for accurate building boundaries. [25], [26], [30], [31] used a conditional random field (CRF) for post processing to refine the edge of footprints, which achieved great improvements on building boundaries.

For the problem of multiscale building extraction, [21], [25], [32] integrated hierarchical results extracted from multiple models or design-specific CNN architecture to address multiscale input for accurately extracting multiscale buildings, which improved the performance, yet obviously increased the computational complexity. [33] proposed a pyramid spatial pooling module by introducing several global pooling layers to capture multiscale features without significantly increasing the computational complexity. It is more efficient than [21] and [25] in multiscale building extraction and improves continuously since the global dependency is captured by pooling layers.

The attention mechanism [34]–[40] is another method for capturing global relations with long-range dependencies in spatial or channel, which effectively improves the performance of segmentation. Recently, HRNet [41], [42] proposed a high-resolution CNN to address multiscale feature extraction and achieved a new goal in semantic segmentation.

In previous studies, CNN-based building footprint extraction algorithms have mainly been encoder-decoder-based, which loses the spatial details in the encoder stage and recovers by fusing shallow feature maps during the decoder stage. However, it causes inaccurate localization on building boundaries since the coarse features introduced from shallow layers and small buildings may be unrecognized. Additionally, the extracted features are always partially restricted by the local respective field, and large-scale buildings with low texture are always discontinuous and holey when extracted.

This research proposes a MAP-Net to solve the problems described above. First, a parallel multipath is generated gradually in each stage to extract high-level semantics and preserve localization details through serial convolution blocks with fixed spatial resolution. Then, the attention mechanism-based feature enhancement module is introduced to adaptively squeeze channel-wise feature maps from each path for multiscale feature optimization. Pyramid spatial pooling operations following to extracts global semantic information for continuous building footprints. The main contributions of this study are as follows:

1) We propose a MAP-Net for efficient, accurate and exact multiscale building footprint boundary extraction through parallel localization-preserved convolutional networks.
2) We introduce a channel-wise attention module to adaptively squeeze multiscale features extracted from multipath. These features strengthen the building representation by optimally combining global semantic and spatial localization.
3) We validate the effectiveness of the proposed network and feature enhancement modules through extensive ablation studies.
4) The proposed method outperforms other state-of-the-art algorithms and achieved 1.74%, 0.68%, 1.46% precision, and 1.50%, 1.53%, 0.82% IoU score improvement compared with the latest HRNetv2 [42] on Deep Globe [43] and Urban 3D [44], WHU [45] datasets, respectively.

The rest of the paper is organized as follows. Section II introduces the detailed structure of the proposed network for building extraction. Section III describes the experiments and analyses the results. The discussions and conclusions of this paper are presented in Section IV.

II. METHODOLOGY

A. Overview

![Fig. 1. Structure of the proposed MAP-Net, which is composed of three modules: (A) Detail preserved multipath feature extraction network; (B) Attention-based features adaptive squeeze and global spatial pooling enhancement module; (C) Upsampling and building footprint extraction module. The conv block is composed of a series of residual modules to extract features and is shared with each path. A gen block generates a new parallel path to extract richer semantic features on the basis of conv block.](image)

Repeated pooling layers or stride convolution lose spatial localization during the feature extraction procedure. Existing CNN-based building extraction methods recover spatial localization through skip connections to fuse shallow feature maps or upsampling feature maps extracted from the last layer by interpolation. However, shallow feature maps contain coarse semantics, introducing noise information in building extraction. In addition, convolutions process the local neighbourhood information and cannot capture global dependency for
large buildings. We propose MAP-Net for multiscale building footprint extraction with accurate boundaries and continuous entities. Fig. 1 illustrates the structure of the proposed MAP-Net. It mainly includes three components:

1) A parallel multipath network to extract multiscale high-level semantic features while preserving spatial detail information;
2) An attention-based multiscale features adaptive squeeze and spatial global pooling enhancement module;
3) A building footprint extraction module.

The detail preserved feature extraction network includes three stages, and the parallel path is generated gradually in each stage to extract richer high-level semantic representations with spatial resolution of features fixed to preserve local details. Features are fused among paths at the end of each stage during the feature extraction process for multiscale feature reuse, as shown by the black two-way arrow. Then, multiscale features extracted from each path concatenate fused and squeezed features by an attention-based module for feature optimization. The spatial pooling module extracts global dependency to suppress the holes and obtain continuous building entities in the final extraction module. The numbers of channels (C) and resolution of the feature maps are marked in the figure.

The remainder of this section is arranged as follows. Section B presents a detail preserved multipath feature extraction network. Attention-based multiscale features are adaptively squeezed, and spatial global pooling enhancement is described in section C. Finally, section D describes the basic unit and training strategies involved in this study.

B. Localization-Preserved Multipath Network

Compared with the encoder-decoder-based CNN structure, the advantage of a localization-preserved multipath feature extraction network is that it extracts multiscale features that contain rich high-level semantic representation and accurate spatial localization information rather than recover spatial details by fusing shallow feature maps during the decoding. Multiscale features extracted from different stages are fed into several parallel paths that are gradually generated to extract richer semantics and preserve spatial resolution without increasing the computational complexity of the network.

Fig. 2 illustrates part of the proposed multipath network. There are two parallel path extracted feature maps with different dimensions in the previous stage, and a new path is generated in the next stage to extract high-level semantic features with double-downsampled resolution and double channels. Feature maps extracted in each path maintain spatial resolution, which preserves as many spatial details as possible, providing security for accurate pixel localization in the building segmentation procedure. Features in each parallel path are fused at the end of every stage to make full use of multiscale representation.

To guarantee that the feature maps from different paths have the same dimension, to be fused, specific layers are designed to rescale the dimension of feature maps during the extraction process. The green arrow is composed of a max pooling layer to downsample resolutions and a $1 \times 1$ convolutional layer to increase feature channels. The orange arrow shows the sampling of the resolutions through bilinear interpolation and decreasing channels with a $1 \times 1$ convolutional layer. The blue arrow represents a $3 \times 3$ convolutional layer, and the black arrow indicates each parallel path. Thus, feature maps can be fused by pixel-wise addition or channel-wise concatenation.

In the entire process of feature extraction, spatial resolutions and channels of feature maps in each path are fixed. Features in each path are extracted by a series of convolutional blocks that suppress the coarse semantics in high-resolution feature maps compared with encoder-decoder-based CNN. Because detailed representation is preserved in higher-resolution features, smaller buildings and localization of the boundary can be extracted exactly. The effect of a multipath network that considers preserved localization and high-level semantics is explained in experiment 3.3.1.

Considering the trade-off between complexity and accuracy, the structure of MAP-Net is composed of three parallel paths for extracting multiscale features. The spatial resolutions of feature maps are 1/4, 1/8 and 1/16 of the original image, with corresponding numbers of channels of 64, 128 and 256, respectively.

C. Attention-Based Feature Squeeze

Feature maps extracted from multipaths have different dimensions. Higher-resolution features contain localization details and high-level semantic information, while lower resolution provides richer global features. The features are sampled in up to 1/4 of the original image through bilinear interpolations and fused by concatenation, as shown in Fig. 3. The function of the channel attention squeeze module adaptively measures the significance of each channel for optimizing the multiscale features. A spatial pooling enhancement module is introduced to capture global dependence for continuously extracting building entities, especially for large buildings with low texture. The details are described as follows.
To decrease the computational complexity, a downsampling block is introduced to decrease the resolution of the input before the multipath network, as shown in Fig. 4(a). It consists of two $3 \times 3$ convolutional layers followed by batch normalization (BN) and a rectified linear unit (ReLU) activation function and two max pooling layers to extract feature maps with 64 channels and $1/4$ spatial resolution of the input image. Fig. 4(b) represents the conv block, which includes several residual blocks in the series. The impact of different numbers of blocks on performance are explored in experiment 3.3.2. The residual block consists of a $1 \times 1$ convolutional layer for reducing the dimensions of features, two $3 \times 3$ layers for extracting features, and a $1 \times 1$ convolutional layer for restoring dimensions to the input; a shortcut fuses input to output through element-wise addition, and BN and ReLU execute before the convolutional layers, as illustrated in Fig. 4(c). The building footprint extraction module is shown in Fig. 4(d). The resolution of the features is recovered through bilinear interpolation in two stages, and the convolutional layers are used to decrease the number of channels.

Our research was implemented in TensorFlow using a single 2080Ti GPU with 12 Gigabyte of memory. The am optimizer was chosen with an initial learning rate of 0.001, and beta1 and beta2 were set to default as recommended. All compared methods were trained from scratch for approximately 80 epochs until convergence and randomly rotated and flipped for data augmentation on three building datasets described in section III. Batch size was set to 4 restricted by the GPU memory size and maintained the same hyperparameters to compare the performance with different methods.

Sigmoid cross-entropy loss was selected as the loss function because of the pixel-wise binary classification involved. The computation of loss at position $(i,j)$ was given as $p_{ij}$, logits represent the predicted result and $y_{ij} \in \{0, 1\}$ represents the ground truth; the sigmoid function was applied to logits to ensure that $p \in [0, 1]$, as shown in (2). The loss value is the average of $Loss_{ij}$ at all positions for an input image.

\[
Loss_{ij} = -[y_{ij} \log p_{ij} + (1 - y_{ij}) \log (1 - p_{ij})]
\]  

\[
p_{ij} = \text{Sigmoid}(\text{logits}_{ij}) = \frac{1}{1 + e^{-\text{logits}_{ij}}}
\]

III. EXPERIMENT AND ANALYSIS

A. Dataset

To evaluate the proposed method, we conducted a comprehensive experiment on three open datasets, including the WHU building dataset [45], the Deep Globe Building Extraction Challenge dataset [43], and the USSOCOM Urban 3D Challenge dataset [44]. The details are described as follows.

The WHU building dataset includes both aerial and satellite subsets with corresponding shapefiles and raster images. In our experiment, we selected the aerial subset, which has various appearances and scales of buildings, to evaluate the robustness of the proposed algorithm. It consists of more than 187,000 buildings, covering over a 450 $km^2$ area, with 30 cm ground resolution. Each image has three bands, corresponding to red (R), green (G) and blue (B) wavelengths, with the size of each image being $512 \times 512$ pixels. There are a total of 8,188 tiles of images, including 4,736, 2,416 and 1,036 tiles as training,
The Deep Globe Building Extraction Challenge dataset \cite{43} contains WorldView-3 satellite imagery captured from Vegas, Paris, Shanghai and Khartoum. In this research, the Vegas and Shanghai subsets were selected to evaluate the generalization performance of the proposed algorithm. There were approximately 243,382 buildings with 30 cm ground resolution, covering over 1,216 \text{km}^2, and the size of each image was 650×650 pixels. All images were randomly divided by 6:1:3 as the training set, validation set and test set.

The USSOCOM Urban 3D Challenge dataset \cite{44} contains 208 orthorectified RGB, with corresponding DSM and digital terrain models (DTM) generated from commercial satellite imagery. It contains approximately 157,000 buildings, covering over 360 \text{km}^2 with a ground resolution of 50 cm, and the size of each image is 2048×2048 pixels. DSM and DTM indicate the elevation of buildings, which obviously improves the building extraction performance. We used only the RGB images in our experiment to evaluate the performance of the proposed method. The training, validation and test set include 104, 62, 42 tiles as the original data partition method, and we randomly clipped the images to the size 512×512 of pixels for training and testing.

### B. Evaluation Metric

Generally, evaluation metric methodologies can be divided into two categories: pixel-level metrics and instance-level metrics. The pixel-level method counts the correctly classified and misclassified pixels pixel-wise. In the instance-level method, a building is correctly extracted only when the intersection over union between the prediction and ground truth is larger than a specific threshold. Semantic segmentation-based building footprint extraction aims to classify every pixel, whether or not it belongs to a building, for a specific input image. Therefore, we apply a pixel-level metric including precision, recall, F1-score and intersection over union (IoU) to evaluate the performance of MAP-Net and other different methods.

There are four classifying conditions: true prediction on a positive sample (TP), false prediction on a positive sample (FP), true prediction on a negative sample (TN) and false prediction on a negative sample (FN). Precision represents the percentage of TP in total positive prediction, recall indicates the percentage of TP over the total positive samples, the F1-score is the weighted average of precision and recall, which considers both FP and FN, and IoU is the average value of the intersection of the prediction and ground truth over their union of the whole image set. Equations are given as follows:

\[
Precision = \frac{TP}{TP + FP} \tag{3}
\]

\[
Recall = \frac{TP}{TP + FN} \tag{4}
\]

\[
F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}
\]

\[
IoU = \frac{Precision \times Recall}{Precision + Recall - Precision \times Recall} \tag{6}
\]

### C. Experimental Setup

In this section, we first analysed the significance of the proposed multipath architecture for extracting multiscale buildings with exact localization on boundaries compared with the popular encoder-decoder framework. Second, we explored the impact of different network parameters on the complexity and accuracy of MAP-Net. Third, a contrast experiment was carried out to compare the performance of MAP-Net with four state-of-the-art algorithms on building extraction. Finally, we conducted an ablation experiment to validate the significance of the proposed network and analysed the trade-off between complexity and accuracy among the compared methodologies. Details are described in the following sections.

1) Significance of Multipath: Feature maps extracted from the proposed localization-preserved multipath network are shown in Fig. 5. Columns (b-d) are extracted from path 1 with the same spatial resolution on each stage corresponding to the sample image in column (a). This indicates that feature maps with higher resolution extracted from deeper convolutional layers (larger S) retained richer semantics; in other words, building and background could be distinguished evidently. Columns (d-f) show the extracted feature maps from each
Fig. 5. Feature maps extracted from localization-preserved multipath networks with different paths (P), stages (S) and spatial resolutions (R) referred to in Fig. 1. Column (a) represents two sample images containing multiscale buildings. Columns (b-d) are feature maps extracted from path 1 with the same spatial resolution on each stage. Columns (d-f) are the feature maps extracted from stage 3 but with decreasing spatial resolution in each path.

path at stage 3 with decreasing spatial resolution. It shows that feature maps with lower-resolution blur at the edge of buildings, worse conditions, and small buildings may be lost completely, as shown in column (f), due to the exact localization loss during the feature extraction procedure.

Encoder-decoder-based networks fuse higher-resolution feature maps extracted from shallow layers, such as columns (b) or (c), to recover exact localization through skip connection at the decoder stage, which introduces noise information for the coarse semantic features. In addition, small buildings may be lost in the lowest resolution feature maps, such as column (f), which cannot be refined accurately during the decoder stage. As a result, extracted building footprints were inaccurate on the boundary, or worse, small buildings were unrecognized.

Multipath networks extract multiscale feature maps through parallel paths. The resolutions of feature maps in each path were fixed in the whole feature extraction process. Features with higher spatial resolution preserved exact localization and contained richer semantics, such as column (d), which is beneficial for extracting exact boundaries and small buildings. Additionally, the features with lower spatial resolution captured global semantic representations, which contribute to the extraction of large buildings. Multiscale features extracted from multipath were fused and enhanced to extract buildings with multiscale, which makes up for the shortcoming of the existing network.

2) Network Parameter: The structure of the proposed network is mainly affected by the depth of the convolution network and the number of parallel paths. We designed an experiment to explore the impact of different network parameters on the performance of MAP-Net on the WHU dataset. The depth is represented by the number of residual blocks (N-blocks) in each path; empirically, these were set from 3 to 6 in our experiments. Similarly, the number of paths (N-paths) was chosen from 2 to 4 according to the resolution of the input image. The IoU metric was used to evaluate the accuracy, and the number of trainable parameters (Para.) was counted to represent the complexity of the network.

The experimental result is illustrated in Fig. 6. With the increase in N-blocks, the IoU score increased first, and then decreased after N-blocks were greater than a specific value, which may be explained by the complexity of the network growing with N-blocks increasing while weakening the generalization ability of the model. However, the Para. grow linearly with the increase in N-blocks, while increasing exponentially with the increase in N-paths since the generated path doubles the feature channels, which greatly increases the parameters during the feature extraction and enhancement stage.

Features with specific resolution were extracted from each path. The number of paths impacts the combination of multiscale semantic features fused in MAP-Net. When the N-paths equalled 3, the IoU metric was better than that of 2 or 4, as shown by the solid line marked by the red circle in Fig. 6.

Considering accuracy and complexity, the better structure of the MAP-Net composed of three parallel paths, and each convolutional block consisted of four residual blocks, which
contained fewer parameters and performed better than others, as solid circles marked with red and blue.

| Method      | IoU(%) | Precision(%) | Recall(%) | F1-score(%) |
|-------------|--------|--------------|----------|------------|
| UNet        | 88.75  | 94.85        | 93.25    | 94.04      |
| PSPNet      | 88.87  | 94.28        | 93.93    | 94.10      |
| ResNet101   | 89.18  | 94.47        | 94.09    | 94.28      |
| HRNetv2     | 90.04  | 94.16        | 95.37    | 94.76      |
| Ours        | **90.86** | **95.62**   | **94.81** | **95.21**  |

| Method      | IoU(%) | Precision(%) | Recall(%) | F1-score(%) |
|-------------|--------|--------------|----------|------------|
| UNet        | 76.34  | 89.07        | 84.23    | 86.59      |
| PSPNet      | 78.76  | 87.36        | 88.89    | 88.12      |
| ResNet101   | 79.16  | 89.1         | 87.65    | 88.37      |
| HRNetv2     | 79.13  | 89.55        | 87.17    | 88.35      |
| Ours        | **80.63** | **91.29**   | 87.35    | **89.28**  |

| Method      | IoU(%) | Precision(%) | Recall(%) | F1-score(%) |
|-------------|--------|--------------|----------|------------|
| UNet        | 84.56  | 92.59        | 90.69    | 91.63      |
| PSPNet      | 86.19  | 93.00        | 92.17    | 92.46      |
| ResNet101   | 86.17  | 92.83        | 92.31    | 92.57      |
| HRNetv2     | 86.15  | 92.74        | 92.38    | 92.56      |
| Ours        | **87.68** | **93.42**   | **93.45** | **93.44**  |

3) Performance Evaluation: To evaluate the performance of the proposed network, we conducted contrast experiments to compare MAP-Net with four state-of-the-art methods, including UNet, PSPNet with ResNet50 backbone, ResNet101 and HRNetv2, on datasets [43]–[45]. Experimental results are shown in TABLE I, TABLE II and TABLE III. Our proposed method demonstrates a great improvement compared with other methods on three experimental datasets and obtains approximately 0.82%, 1.50%, 1.53% IoU improvement and 0.45%, 0.93%, 0.88% F1-score improvement on the WHU dataset, Deep Globe dataset and Urban 3D dataset, respectively, compared with the last research HRNetv2. The best records are marked with bold.

To compare different methods, some example results on each dataset are presented in Fig. 7, Fig. 8 and Fig. 9. Fig. 7 shows extracted building footprints on the WHU dataset. There are four examples, including buildings with various appearances and scales. Columns (a) and (g) represent the original image and corresponding ground truth, and columns (b–f) are extracted results from UNet, PSPNet, ResNet101, HRNetv2 and MAP-Net, respectively.

The results show that our proposed method outperforms the other four compared methods obviously, especially by more accurately recognizing small buildings and more completely extracting large buildings, which benefits from the localization-preserved multipath feature extraction network and the multiscale feature enhancement module. The boundary of buildings is more exactly based on the ground truth.

Example results on the Deep Globe dataset and Urban 3D dataset are illustrated in Fig. 8 and Fig. 9. Each column has the same meaning as presented in Fig. 7.

4) Ablation Experiments: To explore the contributions of different modules of the MAP-Net, we conducted ablation experiments on the WHU dataset. The MAP-Net based on HRNetv1 and optimized architecture for building extraction was our baseline.

On the basis of the baseline, fusing localization-preserved multiscale features extracted from multipath represented by (M), in addition, the channel-wise feature squeezes the module based on the attention mechanism called (C) and the spatial pooling enhancement module named (S). We evaluate the performance on IoU, precision, recall and F1-score. Experimental results are recorded in Table IV. The best records are marked with bold.

Compared with the results, the attention-based feature squeeze and spatial pooling enhancement module obviously improved the performance in building extraction. Multipath localization-preserved strategy improved by 0.5% IoU score on a higher baseline.

It is worth noting that our algorithm from No. 3 to No. 6 obtained higher precision measures compared with HRNetv1 and our baseline. A probable explanation is that our methods suppressed false positive prediction, which contributed to accurate multiscale features extracted from localization-preserved multipath networks. The same conclusion can be inferred from other datasets, according to TABLE IV.

5) Complexity of MAP-Net: Our proposed algorithm extracted features with multiscale; specifically, some paths needed to process feature maps with large resolution to preserve exact localization in the whole network, which could lead to large numbers of parameters. To validate the trade-off between the performance and complexity of MAP-Net, we compared the trainable parameter, IoU score and model size of five related methods on the WHU dataset.

The experimental results are presented in Fig. 10. ResNet101 is the most complicated model with poor performance due to the most convolutional layers and the highest number of channels. HRNetv2 has more parameters than HRNetv1 and performs best among the compared methods, except for MAP-Net, which indicates that the multiscale feature is an important factor for building extraction. MAP-Net shows the best performance and requires the fewest parameters compared with other methods. Although it maintains a high-resolution feature map, which may lead to a large number of parameters, the number of channels remains small, allowing it to efficiently extract multiscale features.
Fig. 7. Example of results with the UNet, PSPNet, ResNet101, HRNetv2 and our proposed method on the WHU dataset. (a) Original image. (b) UNet. (c) PSPNet. (d) ResNet101. (e) HRNetv2. (f) Ours. (g) Ground truth.

Fig. 8. Example of results with the UNet, PSPNet, ResNet101, HRNetv2 and our proposed method on the Deep Globe dataset. (a) Original image. (b) UNet. (c) PSPNet. (d) ResNet101. (e) HRNetv2. (f) Ours. (g) Ground truth.
Fig. 9. Example of results with the UNet, PSPNet, ResNet101, HRNetv2 and our proposed method on the Urban 3D dataset. (a) Original image. (b) UNet. (c) PSPNet. (d) ResNet101. (e) HRNetv2. (f) Ours. (g) Ground truth.

| No | Method      | IoU(%) | Precision(%) | Recall(%) | F1-Score (%) | Rise IoU(%) |
|----|-------------|--------|--------------|-----------|-------------|-------------|
| 1  | HRNetv1     | 89.54  | 94.55        | 94.41     | 94.48       | ×           |
| 2  | Baseline    | 89.78  | 94.09        | 95.15     | 94.62       | 0.24        |
| 3  | Baseline+ M | 90.28  | 95.03        | 94.76     | 94.89       | 0.74        |
| 4  | Baseline+ MC| 90.57  | 95.44        | 94.67     | 95.05       | 1.03        |
| 5  | Baseline+ MS| 90.62  | 95.21        | 94.95     | 95.08       | 1.08        |
| 6  | MAP-Net     | 90.86  | 95.62        | 94.81     | 95.21       | 1.32        |

IV. DISCUSSION AND CONCLUSION

To solve the problem of extracted building footprints with inaccurate boundaries and possibly lost small buildings as well as discontinuous buildings for large-scale buildings, in this research, we proposed a novel localization-preserved multipath feature extraction network with a channel and spatial enhancement module for building footprint extraction. Multiscale features extracted from parallel multipaths that contain local details, as well as rich semantic representations allow it to accurately extract building footprints with exact edges and recognize small buildings. The enhanced module further reconstructs and optimizes features in channel and spatial aspects, which suppresses the holes and extracts continuous footprints for large buildings.

The experiments on three different building extraction datasets demonstrate that the MAP-Net outperforms other state-of-the-art algorithms with higher accuracy and lower complexity. In addition, we conducted an ablation experiment to evaluate the significance of the proposed module and proved that localization-preserved multipath network extraction of buildings with higher precision compared with previous methods.

Generally, our research provides a new approach for accurately and efficiently extracting multiscale objects that are commonly in the real world. Currently, our experiments are implemented in building extraction, and we will further study multi-class extraction tasks, such as land cover, to achieve automatic interpretation of remote sensing imagery in future work.

REFERENCES

[1] J. Du, D. Chen, R. Wang, J. Peethambaran, P. T. Mathiopoulos, L. Xie, and T. Yun, “A novel framework for 2.5-d building contouring from
large-scale residential scenes,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 6, pp. 4121–4145, 2019.
[2] Z. Li, W. Shi, Q. Wang, and Z. Miao, “Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions,” IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 2, pp. 883–899, 2014.
[3] N. L. Gavankar and S. K. Ghosh, “Automatic building footprint extraction from high-resolution satellite image using mathematical morphology,” European Journal of Remote Sensing, vol. 51, no. 1, pp. 182–193, 2018.
[4] J.-P. Burochon, B. Vallet, M. Brédif, C. Mallet, T. Brosset, and N. Parpoditis, “Detecting blind building façades from highly overlapping wide angle aerial imagery,” ISPRS journal of photogrammetry and remote sensing, vol. 96, pp. 193–209, 2014.
[5] M. Cote and P. Saeedi, “Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution,” IEEE Transactions on geoscience and remote sensing, vol. 51, no. 1, pp. 313–328, 2012.
[6] G. Zhou and X. Zhou, “Seamless fusion of lidar and aerial imagery for building extraction,” IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 11, pp. 7393–7407, 2014.
[7] M. Arvangieb, C. Zhang, and C. S. Fraser, “Automatic extraction of building roofs using lidar data and multispectral imagery,” ISPRS journal of photogrammetry and remote sensing, vol. 83, pp. 1–18, 2013.
[8] S. Du, Y. Zhang, Z. Zou, S. Xu, X. He, and S. Chen, “Automatic building extraction from lidar data fusion of point and grid-based features,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 130, pp. 294–307, 2017.
[9] S. Gilani, M. Arvangieb, and G. Lu, “An automatic building extraction and regularisation technique using lidar point cloud data and orthoimage,” Remote Sensing, vol. 8, no. 3, p. 258, 2016.
[10] G. Sohn and I. Downman, “Data fusion of high-resolution satellite imagery and lidar data for automatic building extraction,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 62, no. 1, pp. 43–63, 2007.
[11] E. Simonetto, H. Oriot, and R. Garello, “Rectangular building extraction from stereoscopic airborne radar images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 10, pp. 2386–2395, 2005.
[12] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: a review,” 2017.
[13] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in European conference on computer vision, pp. 630–645, Springer, 2016.
[14] Z. Huang, G. Cheng, H. Wang, H. Li, L. Shi, and C. Pan, “Building extraction from multi-source remote sensing images via deep deconvolutional neural networks,” in 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1835–1838, IEEE, 2016.
[15] R. Kemker, C. Salvaggio, and C. Kanan, “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,” ISPRS journal of photogrammetry and remote sensing, vol. 145, pp. 60–77, 2018.
[16] Y. Sun, X. Zhang, Q. Xin, and J. Huang, “Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and lidar data,” ISPRS journal of photogrammetry and remote sensing, vol. 143, pp. 3–14, 2018.
[17] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, “Building extraction from lidar data applying deep convolutional neural networks,” IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 1, pp. 155–159, 2018.
[18] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, “Multiscale object detection in remote sensing imagery with convolutional neural networks,” ISPRS journal of photogrammetry and remote sensing, vol. 145, pp. 3–22, 2018.
[19] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 645–657, 2016.
[20] K. Bittner, S. Cui, and P. Reinartz, “Building extraction from remote sensing data using fully convolutional networks,” International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. 42, 2017.
[21] 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 Sensing, vol. 11, no. 3, p. 227, 2019.
[22] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. Dalla Mura, “Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 130, pp. 139–149, 2017.
[23] R. Zhang, G. Li, M. Li, and L. Wang, “Fusion of images and point clouds for the semantic segmentation of large-scale 3d scenes based on deep learning,” ISPRS journal of photogrammetry and remote sensing, vol. 143, pp. 85–96, 2018.
[24] W. Li, C. He, J. Fang, J. Zheng, H. Fu, and L. Yu, “Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source gis data,” Remote Sensing, vol. 11, no. 4, p. 403, 2019.
[25] S. Ji, S. Wei, and M. Lu, “A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery,” International Journal of Remote Sensing, vol. 40, no. 9, pp. 3308–3322, 2019.
[26] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, “Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network,” Remote Sensing, vol. 11, no. 7, p. 830, 2019.
[27] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, pp. 234–241, Springer, 2015.
[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.
[29] A. Khalef and M. El-Saban, “Automatic pixelwise object labeling for aerial imagery using stacked u-nets,” arXiv preprint arXiv:1803.04953, 2018.
[30] 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 transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2018.
[31] J. Lin, W. Jing, H. Song, and G. Chen, “Esfnet: Efficient network for building extraction from high-resolution aerial images,” IEEE Access, vol. 7, pp. 54285–54294, 2019.
[32] L. Li, J. Liang, M. Weng, and H. Zhu, “A multiple-feature reuse network to extract buildings from remote sensing imagery,” Remote Sensing, vol. 10, no. 9, p. 1350, 2018.
[33] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890, 2017.
[34] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, “Adaptive pyramid context network for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7519–7528, 2019.
[35] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154, 2019.
[36] Y. Yuan and J. Wang, “Ocnet: Object context network for scene parsing,” arXiv preprint arXiv:1809.00916, 2018.

[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Gcnet: Non-local networks meet squeeze-excitation networks and beyond,” arXiv preprint arXiv:1904.11492, 2019.

[38] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803, 2018.

[39] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, “A 2-nets: Double attention networks,” in Advances in Neural Information Processing Systems, pp. 352–361, 2018.

[40] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, 2018.

[41] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” arXiv preprint arXiv:1902.09212, 2019.

[42] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” arXiv preprint arXiv:1904.04514, 2019.

[43] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska, “Deepglobe 2018: A challenge to parse the earth through satellite images,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 172–17209, IEEE, 2018.

[44] H. R. Goldberg, S. Wang, G. A. Christie, and M. Z. Brown, “Urban 3d challenge: building footprint detection using orthorectified imagery and digital surface models from commercial satellites,” in Geospatial Informatics, Motion Imagery, and Network Analytics VIII, vol. 10645, p. 1064503, International Society for Optics and Photonics, 2018.

[45] S. Ji, S. Wei, and M. Lu, “Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 1, pp. 574–586, 2018.