MHA-Net: Multipath Hybrid Attention Network for Building Footprint Extraction From High-Resolution Remote Sensing Imagery

Jihong Cai and Yimin Chen

***Abstract*—Deep learning approaches have been widely applied to building footprint extraction using high-resolution imagery. However, the traditional fully convolution network still has problems in recovering spatial details and discriminating buildings with varying sizes and styles. We propose a novel multipath hybrid attention network (MHA-Net) to address these challenges. We design a separable convolution block attention module and an attention downsampling module as the basic modules with separable convolutions and channel attention. The MHA-Net architecture consists of three components: the encoding network, multipath hybrid dilated convolution (HDC), and dense upsampling convolution (DUC). The encoding network is used to encode the high-level semantic contexts of images. The multipath HDC aggregates multiscale features by combining rich semantic representations extracted by HDCs, which can achieve promising results in extracting tiny buildings. The DUC is capable of recovering precise spatial information of buildings. We evaluate our network on two public datasets: the WHU aerial building dataset and the Massachusetts building dataset. According to the experimental results, MHA-Net outperforms other classical semantic segmentation models and several recent building extraction models. In particular, MHA-Net can improve the extraction accuracy of small buildings and is robust to complicated building roofs.**

***Index Terms*—Building footprint extraction, deep learning, high-resolution remote sensing imagery, semantic segmentation.**

I. Introduction

As THE fundamental entities in urban systems, buildings are the primary carriers of human production and life. Precise building footprint data is essential to the researches of the urban environment [1], [2], energy consumption [3], urban planning [4], urban function [5], building change detection [6], and urban morphology [7]. Remote sensing data have become the primary sources for building footprint extraction. Because of the diversity of buildings' type and scale, it is still challenging to extract buildings accurately from remote sensing imagery [8].

Early studies of building extraction relied on the fusion of LIDAR points and multispectral imagery [9]–[12]. More recent studies, however, aim at extracting building footprints directly from high-resolution imagery [13]. For instance, Huang and Zhang [14], [15] designed a morphological building index (MBI) and a morphological shadow index to automatically extract buildings from high-resolution imagery. Ok [16] developed an automatic building detection method with shadow information and graph cuts. Bi *et al.* [17] proposed a multiscale filtering building index (MFBI) for building extraction. The MFBI can overcome the heavy computation of the MBI morphological operations. The methods mentioned above consistently require image transformation to obtain the complex features of building objects. However, the effectiveness of using these features varies substantially from one case to another because of the varying sensors, building types, and ground conditions [8].

In recent years, deep learning, especially the convolution neural networks (CNNs), has become one of the most prevalent methods in the computer vision field. CNNs can automatically learn rich image features without prior knowledge via deep convolutional architectures. They have been widely used in remote sensing areas for object detection [18], hyperspectral image classification [19], and scene classification [20], [21]. The fully convolutional networks (FCNs), which are developed based on the conventional CNNs, have been used to perform pixel-wise image classification through semantic segmentation [22], [23]. The empirical literature has reported the high performances of several FCN-based models, such as U-Net [24], SegNet [25], PSP-Net [26], and Deeplab series [27]–[30].

With the advances made in FCNs and semantic segmentation, many FCN-based models have been designed to automatically extract building footprints [8], [13], [31]–[42]. For example, Bischke *et al.* [34] developed a multitask building segmentation network to address the problem of preserving building footprint boundaries. Liu *et al.* [35] proposed the SRI-Net model to extract building footprints. In SRI-Net, a modified Res-Net-101 is used as the encoder to capture multilevel context information. The SRI-Net model also incorporates a spatial residual inception module to aggregate multiscale information. However, SRI-Net has higher computation complexity because the convolutional kernel size in SRI-Net is enlarged to broaden the receptive field. The JointNet model proposed by Zhang and Wang [36] uses

Manuscript received March 30, 2021; revised May 12, 2021; accepted May 26, 2021. Date of publication May 28, 2021; date of current version June 16, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2019YFA0607201, in part by the National Natural Science Foundation of China under Grant 41871306, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2021B1515020104, in part by the Fundamental Research Funds for the Central Universities under Grant 20lgzd09, and in part by the Open Fund of Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University under Grant KLGIS2020A02. *(Corresponding author: Yimin Chen.)*

The authors are with the Guangdong Provincial Key Laboratory of Urbanization and Geosimulation, School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: caijh25@mail2.sysu.edu.cn; chenym49@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3084805

the dense connectivity block with atrous convolutions to extract multitype ground objects. Kang *et al.* [37] designed a dense spatial pyramid pooling (DSPP) module based on atrous spatial pyramid pooling (ASPP) to extract dense and multiscale features for buildings. Liu *et al.* [38] used dense upsampling convolution (DUC) and SELU activation functions to strengthen building discrimination and boundary preservation. Guo *et al.* [39] proposed a scene-driven multitask parallel attention convolution network (MTPA-Net) to extract buildings for various scenes. Zhu *et al.* [40] developed the MAP-Net model that can learn multiscale features via a multiparallel path and aggregate them with an attention module. Ding *et al.* [41] argued that traditional CNNs could not capture the shape patterns of buildings and proposed an adversarial shape learning network (ASLNet) to strengthen the performance of building extraction by augmenting building shape features.

Some other studies use prior information that is derived from GIS data to strengthen the performance of building extraction. For instance, Sun and Wang [43] used the digital surface model (DSM) to facilitate building extraction. Li *et al.* [44] proposed several strategies such as data augmentation, post-processing, and data integration to combine multisource GIS data, thereby improving the accuracy of building extraction. Sun *et al.* [45] proposed a conditional GIS-aware network that employs complementary information from GIS data to extract building footprints from a very-high-resolution synthetic aperture radar image. However, although extra auxiliary data can increase the performance of building extraction, they are not always available in practice. It is still important to develop effective models that can directly extract building footprints from images without too much prior information.

According to the literature mentioned above, the major challenge of building extraction is to recover spatial detail and improve the discrimination of buildings with varying sizes and styles. Aiming at addressing this challenge and improving extraction accuracy, we propose a multipath hybrid attention network (MHA-Net) for automatical building footprint extraction. First, we apply the separable depthwise convolutions and the channel attention modules in our network. The separable depthwise convolutions can improve the network's efficiency while the channel attention modules can capture the global relations of the channels, thereby enhancing the segmentation performance. Second, a multipath hybrid dilated convolution (HDC) framework is adopted to aggregate multiscale contexts with large receptive fields, which is capable of capturing building features with varying sizes and styles. Finally, instead of using a conventional encoder-decoder structure, we recover the spatial information by a DUC structure that can maintain the spatial details of building boundaries. The main contribution of this study can be summarized as follows.

1) An effective semantic segmentation model, MHA-Net, is proposed for building footprint extraction. The network can accurately extract buildings of different scales.
2) We introduce the multipath HDC to strengthen the network's ability to detect buildings with varying sizes and styles. By applying the DUC, we can avoid the loss of spatial information caused by encoder-decoder structure and better recover the spatial details of buildings.
3) Our network is tested and compared with other models using two public remote sensing building datasets, i.e., the WHU aerial building dataset and the Massachusetts building dataset. The results illustrate the better performance of the proposed MHA-Net over the conventional building extraction methods.

The rest of this article is organized as follows. Section II introduces the proposed MHA-Net. Section III describes our experiments and the results. Section IV concludes this article.

II. Network

A. Architecture Overview

Fig. 1 demonstrates the architecture overview of MHA-Net, which contains three components.

1) An encoding network to capture high-level semantic features. The basic modules of this component include the separable convolutional block attention modules and the attention downsampling modules (ADMs). Both contain a channel attention module before their outputs to exploit more context information. Section II-B provides the details of the encoding network.
2) A multipath HDC structure. The HDC structure uses several combinations of dilated convolutions with different rates to capture multiscale image context. These features are further aggregated via a channel attention module. Section II-C provides the details of this component.
3) A DUC structure. Here the DUC structure is to restore the spatial information of building footprints. Section II-D provides the details of the DUC structure.

B. Encoding Network

1) Depthwise Separable Convolution: The basic modules of MHA-Net are based on the depthwise separable convolution, which consists of a depthwise convolution and a pointwise convolution [46]. Unlike the standard convolution, the depthwise convolution applies a single filter per each input channel. The pointwise convolution, i.e., a 1×1 standard convolution, can create a linear combination of the depthwise convolution outputs. Suppose that C convolutions ($k \times k$) are applied to an input feature map with C_0 channels. The total number of parameters using standard convolutions would reach C_0Ck^2, while the number of parameters using depthwise separable convolutions reduces to $C_0k^2 + C_0C$. The depthwise separable convolution can significantly save parameters while improving the network performance [46].

2) Channel Attention: According to Woo *et al.* [47], the channel attention module is an improvement of the squeeze and excitation module [48]. The channel attention module focuses on finding channels that are more meaningful in a given feature map. Fig. 2 shows the diagram of the channel attention module. The channel attention module incorporates a global average-pooling and a global max-pooling to aggregate each channel's spatial information, thereby generating two different

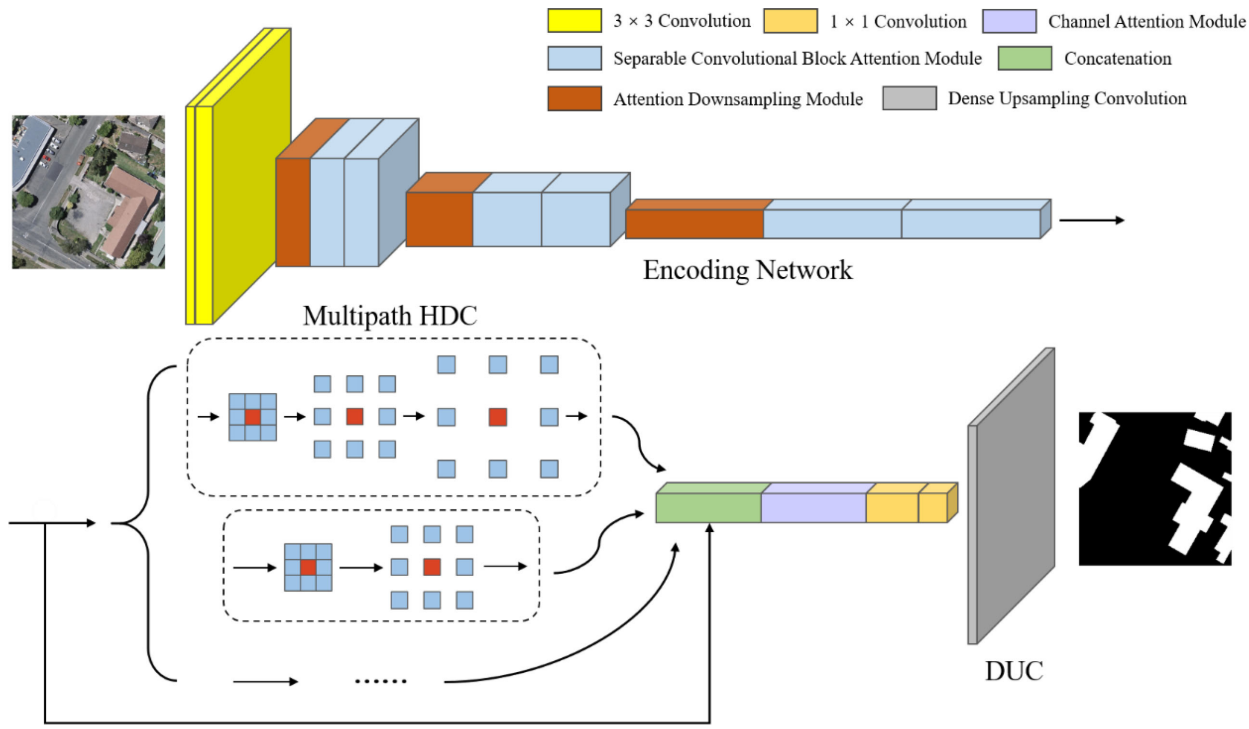


Fig. 1. Overview of the MHA-Net.

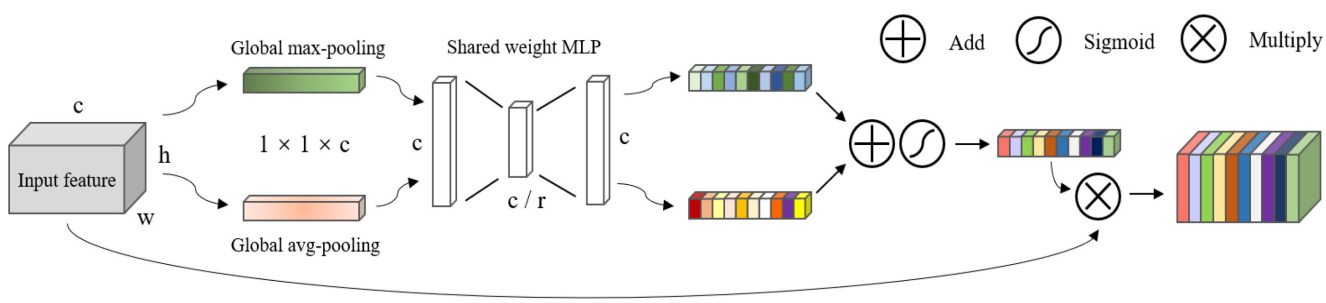


Fig. 2. Diagram of the channel attention module.

spatial context descriptors with a shape of $1 \times 1 \times c$ (where c indicates the channel number of the input feature map). Both of the spatial context descriptors are then passed to a shared weight multilayer perceptron with one hidden layer to produce the channel attention maps. To reduce the number of parameters, the hidden layer's size is set to c/r, in which r denotes the reduction ratio. Afterward, two outputs are merged by element-wise summation. The values of the merged channel attention map are further transformed between 0 to 1 using a sigmoid activation function. Finally, a feature map that highlights the meaningful channels can be obtained by multiplying the input features by the channel attention map.

3) Modules and Architecture Design: Fig. 3(a) and (b) illustrate the separable convolution block attention module (SCBAM) and the ADM, respectively. Based on the residual bottleneck [49], a SCBAM starts with two 1×1 convolution layers and a depthwise convolution layer. The filter number of the first 1×1 convolution layer and the depthwise convolution layer is set to one-fourth of the input channels to enhance the model efficiency. The first 1×1 convolution layer and the depthwise convolution layer are followed by a batch normalization layer and a ReLU layer, while the second 1×1 convolution layer is followed by a batch normalization layer only.

A channel attention module [47] is applied after these three layers in order to highlight the most important channels. The outputs of the channel attention module add input features element-wise via a short connection to avoid gradient vanishment [49]. These outputs are further connected to a ReLU activation layer.

The ADM [Fig.3(b)] is designed for downsampling the feature maps. The ADM includes a max-pooling layer and a depthwise convolution layer with a stride of 2 [38], [50]. Similar to the SCBAM, the ADM has a three-layer convolution architecture, while the number of filters in each layer being the same as the number of input channels. The stride of the depthwise convolution is set to 2. Then, the downsampled features concatenate the output of a 2×2 max-pooling layer with stride 2. The

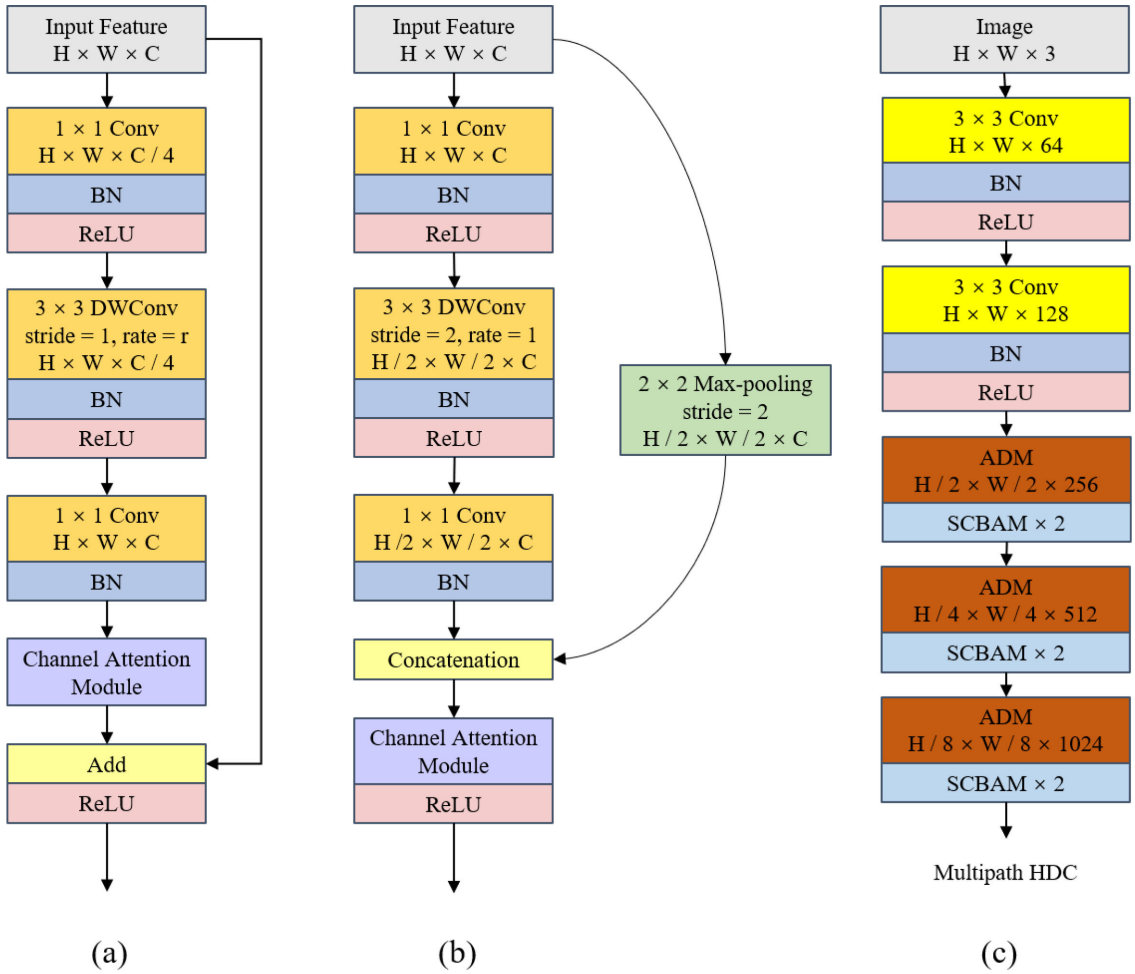


Fig. 3. Primary modules of our proposed model. (a) and (b) are SCBAM and ADM, respectively. (c) shows the encoding part of our network's architecture, designed to capture and downsample the image's semantic information. BN indicates batch normalization; ReLU indicates rectified linear unit.

concatenated features are also forwarded to a channel attention module. A ReLU activation layer is added before the output.

The encoding network is shown in Fig. 3(c). Two 3×3 standard convolutions, each followed by a batch normalization layer and a ReLU activation layer, are first applied to the input images. After that, we use three bottleneck blocks to capture the semantic information. Each bottleneck block consists of an ADM and two SCBAMs. The depths of the bottleneck blocks are 256, 512, and 1024. For an input image with a shape of $256 \times 256 \times 3$, the output features of the encoding network have a size of $32 \times 32 \times 1024$.

C. Multipath Hybrid Dilated Convolution

In semantic segmentation tasks, dilated convolutions are widely used to enlarge the receptive fields of a network and aggregate multiscale context [51]. A dilated convolution is constructed by inserting zeros in the convolutional kernel. However, stacking dilated convolutions may lead to a "gridding" problem [52] [Fig. 4(a)]. Since dilated convolution introduces zeros in the kernel, gaps usually exist between the pixels that participate in the center pixel's computation. These gaps may further cause serious information miss in the input features [52].

The HDC framework is designed to address the "gridding" problem, which has been applied in semantic segmentation by [52]. HDC can make the final receptive field of a series of dilated convolutions fully cover a square region without holes or missing edges. For example, compared with the stacked layers with dilation rates as 2, 2, and 4, grouping succeeding dilated convolution layers with dilation rates like 1, 2, and 5 [Fig. 4(b)] together can capture contexts from a broader range of pixels, despite that they have the same size of the receptive field. Notably, the dilation rate within a layer group should not have a common factor relationship. Layers with dilation rates like 2, 4, and 8 will still lead to the gridding problem on the top layer.

We use the proposed SCBAM to form the HDC framework. As suggested by [52], we set modules' dilation rates to 1, 2, 5, and 9 [Fig. 5(a)]. Inspired by multiscale modules like ASPP [28], [29], Inception [46], [50], and dense prediction cell [53], we developed a multipath HDC framework [Fig. 5(b)] to further aggregate complex multiscale context in building extraction tasks. The multipath HDC consists of four parallel paths to combine the information at different levels. Unlike the ASPP structure,

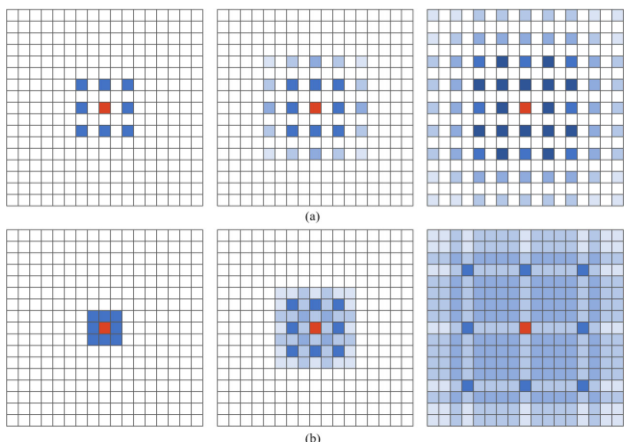


Fig. 4. Gridding problem and the HDC. The blue pixels contribute to the calculation of the red pixel by three 3×3 convolution layers; (a) illustrates the gridding problem. The convolution layers from left to right have dilation rates $r = 2, 2, 4$, respectively. In contrast, the convolution layers in (b) have dilation rates of $r = 1, 2, 5$, respectively; (b) has the same receptive field as (a), but it aggregates more comprehensive contexts.

each path usually has more than one dilation convolution layer. The first path stacks two HDC groups with dilation rates of 1, 2, 5, and 9. The second path has 1/2 depth of the first path and contains four SCBAMs with the same dilation rates of 1, 2, 5, and 9. The third path has two modules with dilation rates of 1 and 2, while the fourth path only contains one module with the dilaton rate of 1. We also add a short connection to concatenate the input features. Combining all these features captured by the four paths can include context information comprehensively from global to local scales. The concatenated features are forwarded to a channel attention module to identify meaningful information. A standard 1×1 convolution layer with 1024 channels is added to aggregate the multiscale context, followed by a batch normalization layer and a ReLU activation layer. Finally, we use a 1×1 convolution with 64 channels to facilitate spatial information restoring through the DUC.

D. Dense Upsampling Convolution

The DUC, which has been termed sub-pixel convolution as well, was first proposed by [54] for image super-resolution tasks. Wang *et al.* [52] applied this operation to better recover spatial details of prediction maps. The DUC has also been applied in building extraction tasks [38] and yielded satisfactory results. As bilinear upsampling and deconvolution would inevitably cause spatial information loss, using these upsampling layers will affect the segmentation precision, especially for building objects that have irregular shapes. The DUC operation, however, does not insert extra values to recover the spatial information. The DUC operation reshapes the input feature map of $H/d \times W/d \times d^2$ to one with a shape of $H \times W \times 1$ (Fig. 6). The number of features is unchanged during the operation. Therefore, the DUC operation can effectively recover tiny objects that may miss in the bilinear interpolation operation. The output feature map is then activated by a sigmoid function. We use a threshold of 0.5 to transform the probability map into the final binary results.

E. Evaluation Metrics

In this article, we use four metrics to evaluate each model's extraction performance, including precision, recall, F1-score, and IoU. Four metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{IoU} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}}. \tag{4}$$

TP, FP, and FN represent the pixel number of true positive, false positive, and false negative.

III. Experimental Results

A. Dataset

WHU Aerial Building Dataset: This dataset is proposed by [31]. It provides highly accurate labels of 187 000 buildings with different colors, sizes, and usage in New Zealand for testing CNN-based methods. The image dataset contains 8189 tiles of 512×512 pixels with 0.3 m ground resolution. Among the samples, the training set consists of 4736 tiles, and the validation set and the test set have 1036 and 1416 tiles, respectively [31].

Massachusetts Building Dataset: This dataset is proposed by [55]. This dataset consists of 151 aerial images of the Boston area. The ground resolution is 1 m. Each of the images has 1500×1500 pixels. The aerial images in this dataset are split into a training set of 137 images, a test set of 10 images, and a validation set of 4 images [55]. Compared with the WHU dataset, the Massachusetts dataset has lower ground resolution and label accuracy [31]. Therefore, using this dataset is appropriate to evaluate the model's ability of segmentation for different sources of images [37].

B. Implementation Setting

Because of the hardware limitation, we did not use the raw images to train our model. The training set and the validation set were cropped into 256×256 pixels in the data preprocessing procedure. Several data augmentation operations were introduced to improve the model robustness, including random flipping, random rotation, and color enhancement. We rescaled all pixel values of each tile to between 0 and 1. We used Adam optimizer [56] with an initial learning rate of 0.0001. The learning rate was updated based on a polynomial decay strategy with a rate of 0.9 per epoch. Four image tiles consisted of a mini-batch. An L2 regularization was applied in all the convolutions to avoid over-fitting. The weight decays were set respectively to 0.00001 and 0.0001 for the WHU dataset and the Massachusetts dataset. We trained 100 epochs for the WHU dataset and 600 epochs for the Massachusetts dataset. In our experiment, we used binary cross-entropy loss as the loss function in our experiments. The

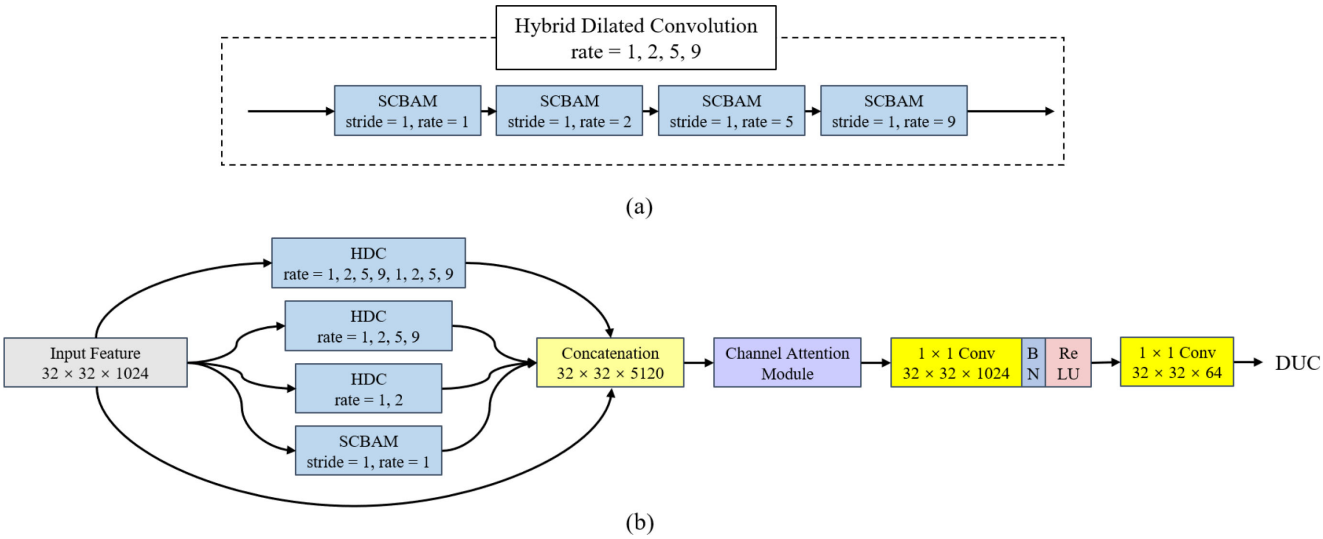


Fig. 5. HDC and the architecture of multipath HDC. (a) Example of HDC. (b) Architecture of multipath HDC.

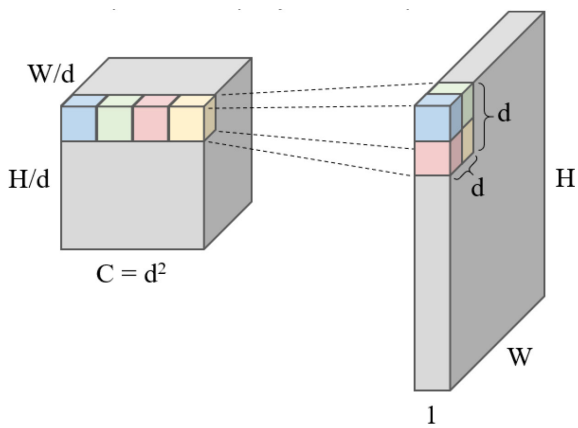


Fig. 6. Illustration of DUC.

TABLE I
QUANTITATIVE COMPARISON OF PRECISION, RECALL, F1-SCORE, AND IOU ON THE TEST SET OF THE WHU AERIAL BUILDING DATASET

Model	Precision	Recall	F1-Score	IoU
SegNet	0.9514	0.9355	0.9434	0.8928
U-Net	0.9327	0.9451	0.9389	0.8848
Deeplab v3+	0.9294	0.9392	0.9343	0.8767
SRI-Net (report)	0.9567	0.9369	0.9451	0.8923
DE-Net (report)	0.9500	0.9460	0.9480	0.9012
EU-Net (report)	0.9498	**0.9510**	0.9504	0.9056
MAP-Net (report)	0.9562	0.9481	0.9521	0.9086
MHA-Net (ours)	**0.9578**	0.9509	**0.9543**	**0.9126**

The best value under each metrics is marked in bold.

experiment was implemented using Keras with a TensorFlow backend.

C. Selected Models for Comparison

We compared the proposed MHA-Net model against three classic semantic segmentation networks, including SegNet, U-Net, and Deeplab v3+. SegNet was proposed by [25] for the semantic segmentation of road scenes. The novelty of SegNet is that the decoder uses the pooling indices to perform nonlinear upsampling. SegNet achieves good segmentation performance and has the advantage of memory efficiency. U-Net [24] was initially developed for biomedical image segmentation. Its architecture consists of a contracting path (encoder) to capture semantic information and a symmetric expanding path (decoder) to obtain precise localization information. Deeplab v3+ [30] has been regarded as one of the best semantic segmentation models. It achieves the state-of-art results on the PASCAL VOC 2012 dataset and the Cityscapes dataset. The innovation of Deeplab v3+ is that it uses the ASPP module to aggregate multiscale contexts captured by an efficient xception-41 encoder.

Besides the networks mentioned above, we also introduce other networks that are specially designed for building footprints extraction, including JointNet, SRI-Net, DE-Net, EU-Net, and MAP-Net. JointNet [36] combines the dense connectivity pattern with atrous convolution layers and can extract large objects efficiently. JointNet had been tested on the Massachusetts dataset and achieved good performance. SRI-Net [35] captures and aggregates multilevel features via spatial residual inception modules and has achieved good performance on WHU aerial building dataset. DE-Net [38] combines the inception-style downsampling modules, SELU activation function, and a densely upsampling module to encode the spatial information contained in the feature maps. EU-Net [37] uses the DSPP module to extract dense and multiscale features and a focal loss function to make the training stage more stable. MAP-Net [40] has an HRNet-like architecture with a multiparallel path to capture spatial multiscale features. We do not reproduce the results of these building extraction networks since most of their source codes are not available. The reported networks' accuracies are shown in Table I, II, and III. The computational complexities of these networks are shown in Table V.

TABLE II
QUANTITATIVE COMPARISON OF PRECISION, RECALL, F1-SCORE, AND IOU ON THE MASSACHUSETTS BUILDING DATASET TEST SET

Model	Precision	Recall	F1-Score	IoU
SegNet	0.8842	0.7732	0.8250	0.7021
U-Net	0.8551	0.8087	0.8312	0.7112
Deeplab v3+	0.8675	0.7877	0.8257	0.7032
JointNet (report)	0.8621	0.8129	0.8368	0.7199
EU-Net (report)	0.8670	**0.8340**	0.8501	0.7393
MHA-Net (ours)	**0.8857**	0.8238	**0.8536**	**0.7446**

The best value under each metrics is marked in bold.

TABLE III
ABLATION STUDY RESULTS OF THE PROPOSED NETWORK ON THE TEST SET OF THE WHU AERIAL BUILDING DATASET

Model	Precision	Recall	F1-Score	IoU
Single HDC	0.9376	**0.9533**	0.9454	0.8964
Single HDC + CA	0.9476	0.9522	0.9499	0.9046
Multipath HDC	0.9471	0.9506	0.9488	0.9026
Multipath Conv + CA	0.9377	0.9489	0.9433	0.8926
Multipath HDC v2 + CA	0.9518	0.952	0.9519	0.9082
Multipath HDC + CA	**0.9578**	0.9509	**0.9543**	**0.9126**

The best value under each metrics is marked in bold.

TABLE IV
QUANTITATIVE COMPARISON OF PRECISION, RECALL, F1-SCORE, AND IOU ON THE TEST SET OF THE WHU AERIAL BUILDING DATASET USING DIFFERENT DILATION RATES

Dilation rates	Precision	Recall	F1-Score	IoU
1, 2, 5, 9 (MHA-Net)	**0.9578**	0.9509	**0.9543**	**0.9126**
1, 2, 4, 8	0.9465	**0.9535**	0.9500	0.9047
1, 2, 3, 5	0.9490	0.9529	0.9509	0.9065
1, 2, 3, 4	0.9462	0.9512	0.9487	0.9025
1, 1, 1, 1	0.9377	0.9489	0.9433	0.8926

The best value under each metrics is marked in bold.

D. Experimental Results Using the WHU Aerial Building Dataset

In the training stage, we cropped the raw images due to the limited GPU memory, making the raw images of the test set have a larger size than images for training. To apply the well-trained model to the test set, as suggested by [35], we predict the raw images with a certain stride. The larger predicted images are obtained by seamlessly stitching many 256×256 tiles with overlap areas. We set the predicting stride as 128 in our experiment. The quantitative comparison of different models on the test set of the WHU aerial building dataset is demonstrated in Table I. MHA-Net outperforms other models on precision, F1-score, and IoU on this dataset. Compared with other building extraction models, our model achieves the F1-score improvements by 1.2%, 0.63%, 0.39%, and 0.22%, and the IoU improvements by 2.17%, 1.14%, 0.6%, and 0.4%. The recall of MHA-Net is only 0.01% less than that of EU-Net.

TABLE V
COMPLEXITY COMPARISON IN TERMS OF FLOPS, TRAINABLE PARAMETERS, AND IOU SCORE ON THE WHU TEST DATA SET

Model	FLOPs (M)	Trainable Parameters (M)	IoU
SegNet	127.16	31.82	0.8928
U-Net	138.03	34.51	0.8848
Deeplab v3+	90.18	41.08	0.8767
SRI-Net (report)	-	-	0.8923
DE-Net (report)	-	9.63	0.9012
EU-Net (report)	> 29.42	> 14.71	0.9056
MAP-Net (report)	48.09	24.00	0.9086
Single HDC	**25.77**	**6.48**	0.8964
Single HDC + CA	38.29	9.62	0.9046
Multipath HDC	87.73	21.99	0.9026
Multipath HDC v2 + CA	91.80	23.01	0.9082
MHA-Net	107.59	26.98	**0.9126**

The best value under each metrics is marked in bold.

The representative examples of building extraction results are shown in Fig. 7. The first two columns are original RGB aerial images and the corresponding ground truth images. The other four columns are the prediction results of SegNet, U-Net, Deeplab v3+, and MHA-Net. Visual inspection reveals that all these models can accurately extract middle-size buildings such as dwellings. Owing to the DUC operation, MHA-Net performs better in recognizing small buildings than the other models. As shown in rows 4 and 6, MHA-Net also outperforms the other three models in the detection of building edges and is more robust to complicated building roofs.

E. Experimental Results Using the Massachusetts Building Dataset

To better illustrate the results of building extraction, we cropped the test set images of this dataset into 512×512 pixels to evaluate the selected models. As in the previous experiment, we set the predicting stride to 128. Table II shows the results of the model comparison. The highest values are highlighted in bold. The results show that building extraction models outperform the classical deep learning models on this dataset. EU-Net has the highest recall of 0.8340, while MHA-Net achieves the best precision, F1-score, and IoU. Our model outperforms other building extraction models by 1.68% and 0.35% for F1-score and by 2.47% and 0.53% for IoU.

Fig. 8 shows the representative examples of building extraction results on the test set of the Massachusetts Building Dataset. From the visual comparison, we can observe that MHA-Net can accurately detect the buildings' internal structure. It also outperforms other models in recognition of small or long and thin buildings. Overall, the proposed MHA-Net has a better performance in extracting the detailed information of urban buildings.

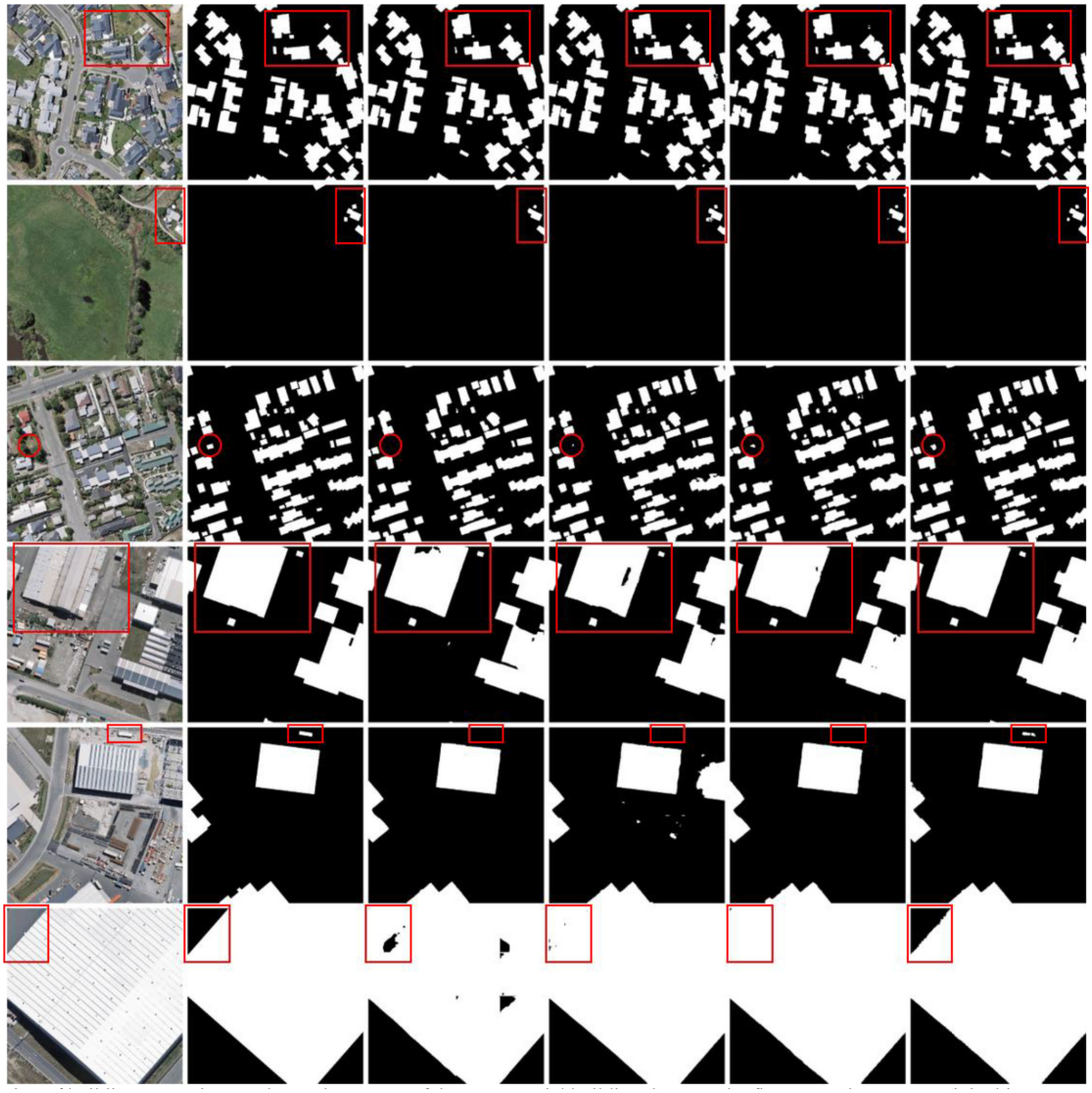


Fig. 7. Visualization of building extraction results on the test set of the WHU aerial building dataset. The first two columns are original images and corresponding ground truth labels. Columns from 3 to 6 are segmentation results of SegNet, U-Net, Deeplab v3+, and MHA-Net, respectively.

F. Ablation Study

We conducted ablation experiments on the WHU aerial building dataset to better understand the influence that each component has in the proposed MHA-Net. The channel attention modules, the multipath HDC framework, and dilated convolutions are removed or replaced in the MHA-Net to show these components' influences on the network. More specific procedures of the experiments are explained below.

First, we test a single HDC network in which the channel attention modules are completely removed from all SCBAMs. Only a single path with two HDC groups remains, and the dilation rates are set as 1, 2, 5, and 9. Second, we evaluate the effectiveness of the channel modules by adding them to the single HDC network. Next, we evaluate the multipath HDC structure of MHA-Net by only removing the channel attention modules from the SCBAMs. Finally, an alternative multipath HDC structure, denoted as multipath HDC v2, is explored. Here the multipath HDC v2 structure is created by cutting the deepest path of the original multipath HDC and replacing it with a three-layer HDC using the dilation rates of 1, 2, and 5.

Table III shows the ablation experiments results of precision, recall, F1-score, and IoU. Surprisingly, the single HDC model, the network with the most straightforward architecture, has the highest recall value in our self-comparison study. This result highlights the merit of the HDC framework with a large receptive field to easily detect tiny buildings and the change inside large buildings. Either adding channel attention modules or multipath structure will reduce the model's recall, but they can significantly improve the model's precision and make the model achieve higher values of F1-score and IoU. Channel attention modules highlight the most meaningful channels of feature maps and restrain invalid information. Compared with the single-path HDC structure, the multipath HDC structure can extract multiscale contexts and aggregate them by a channel attention module. This feature is useful when extracting buildings of various sizes and

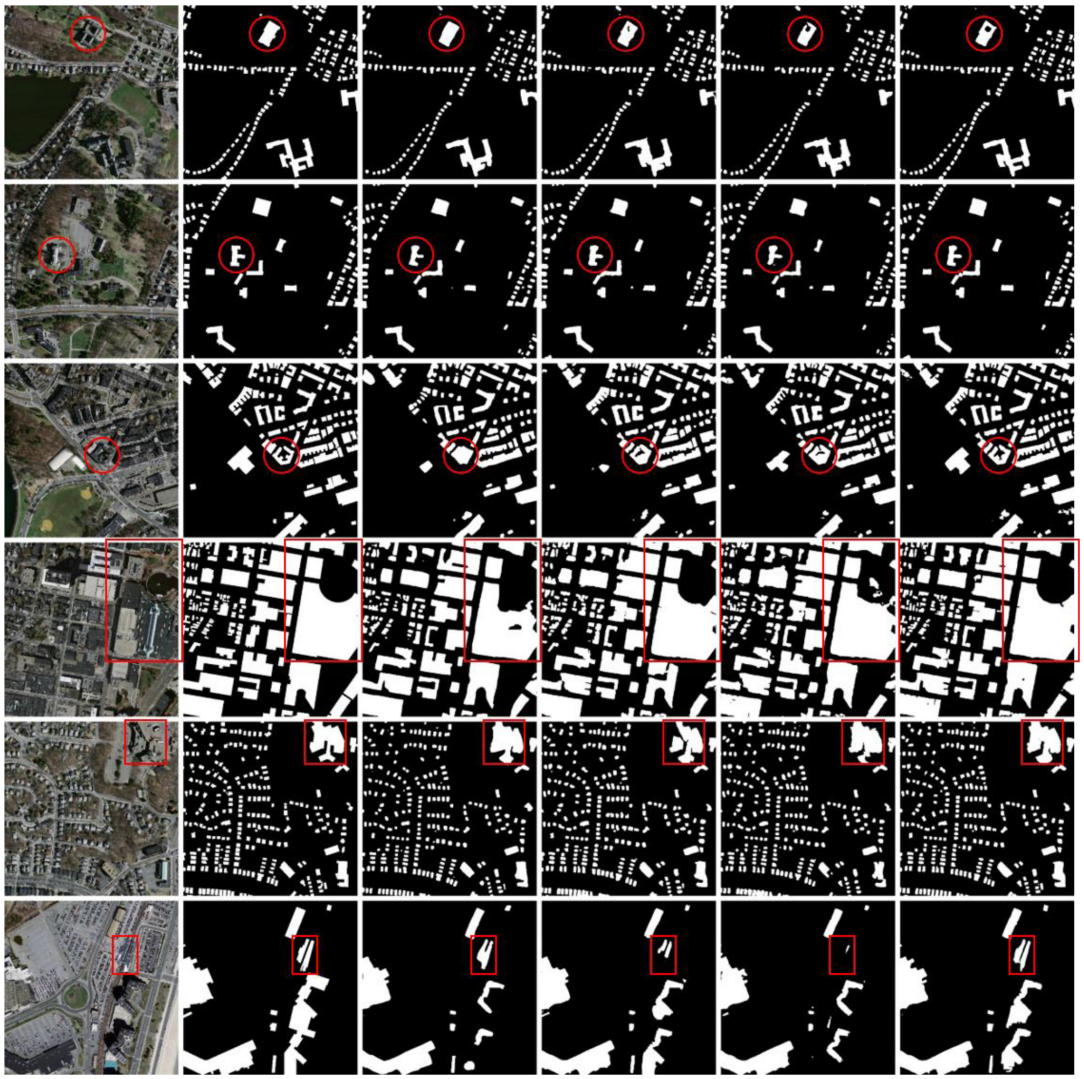


Fig. 8 Visualization of building extraction results on the test set of the Massachusetts building dataset. The first two columns are original images and corresponding ground truth labels. Columns from 3 to 6 are segmentation results of SegNet, U-Net, Deeplab v3+, and MHA-Net, respectively.

styles. These two experiments demonstrate that the network can increase precision by learning more comprehensive information, but at the cost of decreased recall value. As the integrity of shapes is essential for extracted buildings, attaining higher precision values is more reasonable for a building extraction model. The multipath HDC v2 model achieves higher values of F1-score and IoU than the first four models, but it has no advantage over the proposed MHA-Net. Therefore, a deeper network is essential for capturing comprehensive contexts.

G. Experimental Results Using Different Dilation Rates

In our MHA-Net, we set the dilation rates of the multipath HDC as 1, 2, 5, and 9. We also explore the performances of multipath HDC networks using different dilation rates. We tested five groups of dilation rates. Table IV shows their quantitative results. The multipath HDC network with dilation rates of 1, 1, 1, and 1 has the lowest recall, F1-score, and IoU values among all models, since its receptive field is much smaller than others. Our MHA-Net with dilation rates of 1, 2, 5, and 9 has the largest receptive field and obviously outperforms the other four compared models. The network with dilation rates of 1, 2, 4, and 8 has the second-largest receptive field and achieves the best recall value. However, it has lower precision, F1-Score, and IoU performance than the network with dilation rates of 1, 2, 3, and 5 due to the "gridding problem." Overall, these experiments reveal that properly enlarging the receptive field of multipath HDC structure can improve its performance as long as it avoids the "gridding problem."

H. Complexity of MHA-Net

We further compared the computational cost of different models in terms of floating point of operations (FLOPs) and the number of trainable parameters. These two metrics have been frequently used to measure models' computational complexity in the deep learning area [40], [46], [57]. Higher FLOPs and more trainable parameters correspond to greater complexity

of a model. As shown in Table V, U-Net and Deeplab v3+ have the highest FLOPs and the highest number of parameters, respectively, while their performance is relatively poor. For EU-Net, its FLOPs and the number of trainable parameters are quoted from [37]. However, the results of EU-Net only regard the encoder of EU-Net. The actual computational cost of EU-Net should be higher than the results reported by [37]. The proposed single HDC network has the lowest complexity. The single HDC network with channel attention modules has the second-lowest FLOPs and the second-less trainable parameters among all models. However, it achieves a high accuracy of building extraction, suggesting that it is an effective and lightweight model. The proposed MHA-Net has relatively high FLOPs, while the number of required training parameters is moderate compared with other models.

IV. Conclusion

In this article, we proposed an effective fully convolutional network MHA-Net for building extraction tasks using high-resolution remote sensing images. MHA-Net consists of three components: the encoding network, multipath HDC, and DUC. The encoding network encodes the high-level semantic contexts of images. The multipath HDC aggregates multiscale features by combining rich semantic representations extracted by HDCs. The DUC operation is used for upsampling feature maps and recovering spatial details. We designed SCBAM and ADM as the network's basic modules. Separable convolutions replace standard convolutions to improve the efficiency and effectiveness of feature learning. Each module contains a channel attention module that can highlight the meaningful channels of feature maps.

Our network is evaluated on the WHU aerial building dataset and the Massachusetts building dataset. The experimental results show that MHA-Net outperforms other classical semantic segmentation models and building extraction models we tested. The visual extraction results demonstrate that MHA-Net can improve the extraction accuracy of small buildings and is robust to complicated building roofs. The self-comparison results further proved the effectiveness of the proposed architecture.

MHA-Net is specially designed for building extraction on high-resolution remote sensing images. In future works, we will explore its performance on other remote sensing segmentation tasks, such as road detection, land cover classification, classification for multiband images.

References

[1] L. Chen, J. Hang, M. Sandberg, L. Claesson, S. Di Sabatino, and H. Wigo, "The impacts of building height variations and building packing densities on flow adjustment and city breathability in idealized urban models," *Build. Environ.*, vol. 118, pp. 344–361, 2017.

[2] S. Li *et al.*, "Spatially varying impacts of built environment factors on rail transit ridership at station level: A case study in Guangzhou, China," *J. Transp. Geogr.*, vol. 82, 2020, Art. no. 102631.

[3] H. Leng, X. Chen, Y. Ma, N. H. Wong, and T. Ming, "Urban morphology and building heating energy consumption: Evidence from Harbin, a severe cold region city," *Energy Build.*, vol. 224, 2020, Art. no. 110143.

[4] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, 2016.

[5] Y. Chen *et al.*, "Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based k-medoids method," *Landsc. Urban Plann.*, vol. 160, pp. 48–60, 2017.

[6] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[7] X. He, X. Zhang, and Q. Xin, "Recognition of building group patterns in topographic maps based on graph partitioning and random forest," *ISPRS J. Photogramm. Remote Sens.*, vol. 136, pp. 26–40, 2018.

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

[9] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Using the Dempster-Shafer method for the fusion of LIDAR data and multi-spectral images for building detection," *Inf. Fusion*, vol. 6, no. 4, pp. 283–300, 2005.

[10] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Building detection by fusion of airborne laser scanner data and multi-spectral images: Performance evaluation and sensitivity analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 2, pp. 135–149, 2007.

[11] M. Awrangjeb, M. Ravanbakhsh, and C. S. Fraser, "Automatic detection of residential buildings using LIDAR data and multispectral imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 5, pp. 457–467, 2010.

[12] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Automatic extraction of building roofs using LIDAR data and multispectral imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 83, pp. 1–18, 2013.

[13] J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, pp. 2793–2798, Nov. 2018.

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

[15] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[16] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, 2013.

[17] Q. Bi, K. Qin, H. Zhang, Y. Zhang, Z. Li, and K. Xu, "A multi-scale filtering building index for building extraction in very high-resolution satellite imagery," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 482.

[18] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[19] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.

[20] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[21] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[22] 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.

[23] F. Zhou, R. Hang, and Q. Liu, "Class-guided feature decoupling network for airborne image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2245–2255, Mar. 2021.

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

[25] 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.

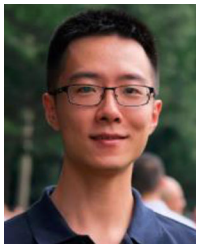
[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFS," 2014. [Online]. Available: arxiv.org/abs/1412.7062

[28] 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.
[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," .2017. [Online]. Available: arxiv.org/abs/1706.05587
[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018. [Online]. Available: arxiv.org/abs/1802.02611
[31] 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.
[32] Y. Shi, Q. Li, and X. X. Zhu, "Building footprint generation using improved generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 603–607, Apr. 2019.
[33] 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, pp. 144, 2018.
[34] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1480–1484.
[35] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens*, vol. 11, no. 7, pp. 830, 2019.
[36] Z. Zhang and Y. Wang, "JointNet: A common neural network for road and building extraction," *Remote Sens.*, vol. 11, no. 6, 2019, Art. no. 696.
[37] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2813.
[38] H. Liu *et al.*, "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 20, 2019,Art. no. 2380.
[39] 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.
[40] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3026051.
[41] L. Ding, H. Tang, Y. Liu, Y. Shi, and L. Bruzzone, "Adversarial shape learning for building extraction in VHR remote sensing images," 2021. [Online]. Available: arxiv.org/abs/2102.11262
[42] W. Deng, Q. Shi, and J. Li, "Attention gate based encoder-decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.
[43] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 474–478, Mar. 2018.
[44] 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 Sens*, vol. 11, no. 4, pp. 403, 2019.
[45] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu, "CG-Net: Conditional GIS-Aware network for individual building segmentation in VHR SAR images," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2020.3043089.
[46] A. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: arxiv.org/abs/1704.04861
[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018. [Online]. Available: arxiv.org/abs/1807.06521
[48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[49] 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.
[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
[51] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015. [Online]. Available: arxiv.org/abs/1511.07122
[52] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.,* 2018, pp. 1451–1460.
[53] L.-C. Chen *et al.*, "Searching for efficient multi-scale architectures for dense image prediction," . 2018. [Online]. Available: arxiv.org/abs/1809.04184
[54] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[55] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, *Dept. Comput. Sci., Univ. Toronto*, Toronto, ON, Canada, 2013.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: arxiv.org/abs/1412.6980
[57] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

Jihong Cai received the B.S. degree in GIScience from Sun Yat-sen University, Guangzhou, China, in 2020. He is currently working toward the M.S. degree in cartography and geographical information system at Sun Yat-sen University.

His research interests include urban remote sensing, image classification, and semantic segmentation.



Yimin Chen received the B.S. degree in GIScience from Sun Yat-sen University, Guangzhou, China, in 2008, and the Ph.D. degree in cartography and GIScience from Sun Yat-sen University, in 2014.

He is currently an Associate Professor with the School of Geography and Planning, Sun Yat-sen University. His research interests include remote sensing and urban analysis, including deep learning, classification, and sematic segmentation.