IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 27, 2020, accepted July 20, 2020, date of publication August 12, 2020, date of current version September 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015701

ARC-Net: An Efficient Network for Building Extraction From High-Resolution Aerial Images

YAOHUI LIU[1,2,3], JIE ZHOU[4], WENHUA QI[2], XIAOLI LI[5], LUTZ GROSS[3], QI SHAO[3], ZHENGGUANG ZHAO[3], LI NI[6], XIWEI FAN[2], AND ZHIQIANG LI[5]

[1]School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China
[2]Institute of Geology, China Earthquake Administration, Beijing 100029, China
[3]School of Earth and Environmental Sciences, The University of Queensland, Brisbane, QLD 4072, Australia
[4]School of Tourism and Geographic Science, Yunnan Normal University, Kunming 650500, China
[5]China Earthquake Networks Center, Beijing 100045, China
[6]Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China

Corresponding authors: Xiwei Fan (fanxiwei@ies.ac.cn) and Zhiqiang Li (lzhq9028@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC1504403 and Grant 2018YFC1504503, in part by the National Natural Science Foundation of China under Grant 41601390 and Grant 41907397, in part by the China Earthquake Administration Special Project Surplus Fund (High-Resolution Rapid Post-Earthquake Assessment Techniques), and in part by the Key Special Fund for the Study on Rapid Assessment of Multi-Source Earthquake Loss under Grant 201308018-5.

ABSTRACT Automatic building extraction based on high-resolution aerial images has important applications in urban planning and environmental management. In recent years advances and performance improvements have been achieved in building extraction through the use of deep learning methods. However, the design of existing models focuses attention to improve accuracy through an overflowing number of parameters and complex structure design, resulting in large computational costs during the learning phase and low inference speed. To address these issues, we propose a new, efficient end-to-end model, called ARC-Net. The model includes residual blocks with asymmetric convolution (RBAC) to reduce the computational cost and to shrink the model size. In addition, dilated convolutions and multi-scale pyramid pooling modules are utilized to enlarge the receptive field and to enhance accuracy. We verify the performance and efficiency of the proposed ARC-Net on the INRIA Aerial Image Labeling dataset and WHU building dataset. Compared to available deep learning models, the proposed ARC-Net demonstrates better segmentation performance with less computational costs. This indicates that the proposed ARC-Net is both effective and efficient in automatic building extraction from high-resolution aerial images.

INDEX TERMS Deep learning, building extraction, high-resolution aerial images, fully convolutional network, asymmetric, separable convolution.

I. INTRODUCTION

Automatic extraction of buildings based on aerial images is of great importance in a broad range of application fields including urban planning, change detection, map services, and disaster management [1]–[5]. Recently, with the continuous advancement of satellite and sensor technology, high-resolution remote sensing products have become the preferred data source for building extraction due to their rich textural, semantic, and spatial details. However, the increasing resolution of aerial images results in an increasing degree of redundant interference information and infernal differences. Moreover, the diversity of building characteristics (color, shape, size, etc.) remains a difficulty and challenge for accurate building extraction. Thus, the efficiency and accuracy of automatic building extraction are still difficult archive and remain a challenging objective which attracts huge research interests [6].

In the past few years, traditional methods including mathematical techniques and morphology approaches have been proposed to address this issue. Many mathematical descriptors have been introduced to extract the spatial and textural

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

features of an image, such as Histogram of Oriented Gradients [7], Haar spaces [8], Grey Level Co-occurrence Matrix [9], and Local Binary Patterns [10]. Furthermore, several machine learning classifiers have been employed for a pixel-by-pixel analysis, including Random Forests [11], Support Vector Machines [12], K-Means [13], Adaptive Boosting [14], and Conditional Random Fields (CRFs) [15]. However, these methods rely heavily on prior knowledge and parameter selections which are leading to limitations as well as significant time and labor costs when applied in real live scenarios.

Recently, with the rapid increase of computational power and available data sources, the use of deep learning technology [16], especially convolutional neural networks (CNNs), has emerged as a powerful tool in computer vision and semantic segmentation [17]. CNNs automatically learn semantic information from the input and generate the classification results through convolutional operations. In the early stages of CNN development, patch-based CNN models, such as VGGNet [18], GoogLeNet [19], ResNet [20], and DenseNet [21], have outperformed traditional machine learning methods on classification applications. Some researchers also utilized patch-based CNN methods to segment buildings in remote sensing images and managed to greatly improve the performance [22]–[25]. However, as patch-based CNN models cannot guarantee spatial continuity and consistency, they are not the best solution for addressing the task of building segmentation [26].

In the fully convolutional network (FCN), proposed by Long *et al.* in 2015 for semantic segmentation [27], fully connected layers are replaced by up-sampling layers so that the output preserves spatial information of contextual features. Over the past five years, various FCN-based variants have been proposed to pursue more accurate segmentation results. The SegNet [28] and U-Net [29] are two classic architectures with symmetric encoder-decoder structures, which were both regarded as effective structures due to their capabilities of recovering semantic details [30]. Some novel FCN-based methods are mainly designed to improve performance by extending the receptive field and by learning multi-scale contextual information. For example, Yu and Koltun [31] utilized dilated convolutions to gather multi-scale contexts. The pyramid pooling module, proposed in PSPNet [32], is applied to capture multi-scale features with different kernel sizes. The DeepLab_v2 [33] employs atrous convolution and atrous spatial pyramid pooling (ASPP) to enlarge the receptive field on different levels. Liu *et al.* [34] merged the spatial pyramid pooling module into the encoder-decoder architecture with a particular focus on building extraction. The JointNet [35] introduced a new, dense atrous convolution block combining a dense connectivity block and atrous convolution to obtain multi-scale features. Ji *et al.* [36] proposed a scale-robust FCN and trained it with five outputs of two ASPP structures. SRI-Net [37] employed large kernel convolution and a spatial residual inception module to preserve details with large receptive fields. Zhang *et al.* [38] proposed the Web-Net with hierarchical dense connections to propagate feature maps among different levels. These novel FCNs have also been successfully applied for land-use detection and are regarded as state-of-the-art methods for semantic segmentation [39].

Some FCN-based models further adopt post-processing approaches to prediction results to optimize the pixel-wise results and to preserve the structure consistency. For example, Shrestha and Vanneschi [40] proposed a novel fully convolutional network using CRFs and exponential linear units for building extraction. Alshehhi *et al.* [41] proposed a post-processing method integrating low-level features of adjacent regions to enhance the performance. Wang *et al.* [42] improved the dense conditional random field (Dense CRF) using a superpixel algorithm in post-processing. However, post-processing methods are only able to improve results within a certain range [37]. The result of semantic segmentation cannot be fundamentally changed.

Although these networks presented before have greatly enhanced the performance of semantic segmentation, their computational cost is high and they require generous training time, which is bringing a heavy burden for the application of deep learning in remote sensing. Therefore, model complexity and computational cost need to be essential indicators to measure the performance of a CNN architecture and should be taken into consideration [43]. One practical way to decrease the number of model parameters is the utilization of efficient structure, such as residual blocks, kernel factorizations, and group convolutions. With these performance considerations in mind but still maintaining high accuracy, a variety of FCN-based architectures have been designed, including ENet [44], ERFNet [45], EDANet [46], the MobileNet family [47], [48], ShuffleNet family [43], [49]. Recent networks such as ICNet [50] and BiSeNet [51] are targeting to compromise performance and efficiency, but these models have still complex structure designs and are difficult to deploy and apply. So, there is still room for further improvement.

To better balance the accuracy and efficiency, we propose a new network for automatic building extraction, named ARC-Net. The basic architecture of the ARC-Net is an asymmetric encoder-decoder structure. We have designed the residual block with an asymmetric convolution (RBAC) module, which incorporates depth-wise separable convolution and asymmetric convolution with the residual connection in order to reduce the computational cost. Dilated convolution is incorporated with the RBAC module to further enlarge the receptive field. Moreover, the advanced atrous pyramid pooling module is added as a connector between the encoder and decoder to aggregate multi-scale contextual information. Experiments on two public building datasets, the INRIA Aerial Image Labeling Dataset [52] and the WHU Building Dataset [53], demonstrate the remarkable performance of the proposed model. Compared to several other FCN-based models, such as SegNet, FCN, U-Net, and ERFNet, higher accuracy with less computational complexity is achieved by the new ARC-Net model when

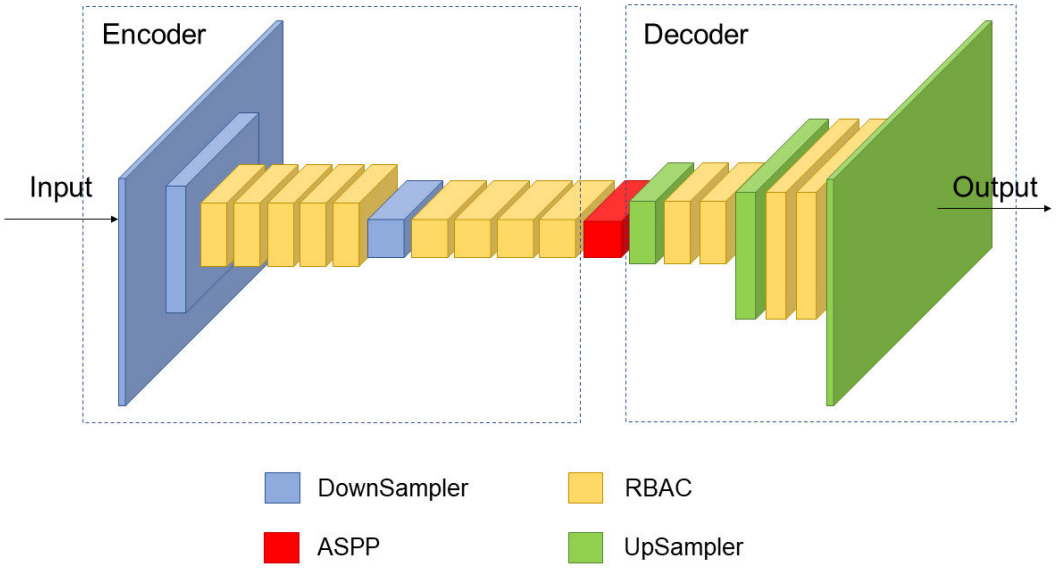


FIGURE 1. The structure of our proposed ARC-Net, consisting of three parts: encoder part (blue & yellow, blocks 1-12), ASPP module (red, block 13), and decoder part (green & yellow, blocks 14-20).

applied to the building extraction from high-resolution aerial images.

The main contributions of this study are summarized as follows: (1) We design a novel efficient network, called ARC-Net, as well as a new residual block with asymmetric convolution module incorporating depth-wise separable convolution to reduce the computational complexity still with sufficient accuracy and (2) we conduct further experiments to provide justifications for some of the design decisions for ARC-Net.

The remainder of this article is organized as follows. The components of the proposed ARC-Net model are introduced in Section II. Section III describes the test datasets and experimental settings. Section IV provides the experimental results of the proposed ARC-Net model including the quantitative and qualitative comparison with other established models. Finally, a discussion and some conclusions from this study are presented in Section V and VI, respectively.

TABLE 1. The detailed blocks of the proposed ARC-Net outlined in figure 1.

Block	Type	Input	Output
1	Downsampler block	256 × 3	128 × 16
2	Downsampler block	128 × 16	64 × 128
3-7	5×RBAC	64 × 128	64 × 128
8	Downsampler block	64 × 128	32 × 512
9	RBAC (dilated 2)	32 × 512	32 × 512
10	RBAC (dilated 4)	32 × 512	32 × 512
11	RBAC (dilated 8)	32 × 512	32 × 512
12	RBAC (dilated 16)	32 × 512	32 × 512
13	ASPP	32 × 512	32 × 832
14	Upsampler block	32 × 832	64 × 64
15-16	2×RBAC	64 × 64	64 × 64
17	Upsampler block	64 × 64	128 × 16
18-19	2×RBAC	128 × 16	128 × 16
20	Upsampler block	128 × 16	256 × 2

II. METHODS

The proposed ARC-Net model follows an asymmetric encoder-decoder architecture, which has already successfully been applied to semantic segmentation. Figure 1 presents the basic structure of the ARC-Net model. In the encoder part (blocks 1-12), several down-sampling blocks and residual blocks with asymmetric convolution (RBAC) modules are employed to extract the feature maps from the inputs and at the same improving computational efficiency. The RBAC modules are also utilized in the decoder phrase with up-sampling operations to recover the details of images in the decoder part (blocks 14-20). The atrous spatial pyramid pooling (ASPP) is employed as a connector in block 13 between the encoder and decoder to further collect the multi-contextual information. The various components of the ARC-Net model are presented in Table 1. In the following, each component will be discussed in detail.

A. ENCODER WITH DOWNSAMPLING BLOCK AND RBAC MODULE

The residual block with the asymmetric convolution (RBAC) module is the fundamental element of the ARC-Net model. It mainly contains two parts: the separable convolution and asymmetric convolution. At the same time, the residual connection is employed to reduce the complexity and to retain dimensions between the input and output.

The depth-wise separable convolution is considered as an efficient tool to reduce the computational cost and the number

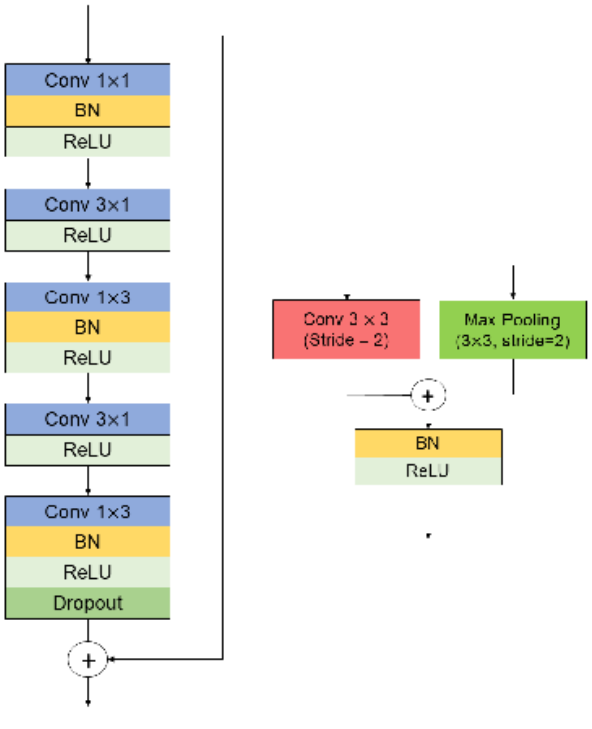


FIGURE 2. The structure of the proposed RBAC module (a) and the down-sampler block (b). Conv: Convolution; BN: Batch Normalization; ReLU: Rectified Linear Unit.

of parameters while achieving similar (or slightly better) performance [54], [55]. It splits the full convolution operations into two independent steps: depth-wise convolution and point-wise convolution [56]. In depth-wise convolution, each kernel has a single feature map in and a single feature map out. As weight kernels are shared the depth-wise convolution requires fewer parameters than the standard version. Point-wise convolution is equivalent to a standard convolution with a kernel size of 1×1 and is aiming to combine the channel-wise independent features from depth-wise convolution. Through such a two-step operation, the number of parameters to be fitted is reduced, speeding up the deep learning computations.

Asymmetric convolutions are widely employed to approximate an existing square-kernel convolutional layer for compression and acceleration [57]. Prior research [58], [59] has shown that a standard $d \times d$ convolutional layer can be factorized as a sequence of two layers with $d \times 1$ and $1 \times d$ kernels. Results of combing a $d \times 1$ and following $1 \times d$ convolution are consistent with the results of a direct $d \times d$ convolution, but the number of a multiplication operation is reduced from $d \times d$ to $2 \times d$ leading to dramatical computational cost saving as d grows. This is the reason why asymmetric convolution performs well in reducing the model parameters and computational work. In this research, we follow this approach and factorize a standard two-dimensional $d \times d$ convolution kernel into two one-dimension $d \times 1$ and $1 \times d$ kernels. As presented in Figure 2 (a), 1×1 point-wise convolution is employed

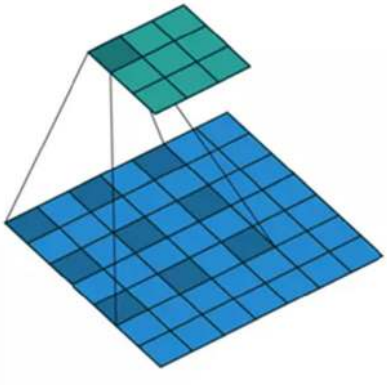


FIGURE 3. Dilated convolution with a 3×3 kernel and a dilation rate of 2.

in the head of the RBAC module. Each 3×1 convolution is then followed by a rectified linear unit (ReLU) while each 1×3 convolution followed by batch normalization and ReLU function.

The down-sampler block in the proposed ARC-Net is inspired by the initial block of ENet [44] and performs down-sampling by concatenating the parallel outputs of a single 3×3 convolution with stride 2 and a MaxPooling operation with stride 2. In contrast to the ENet which used it only as the initial block to perform early down-sampling, we employ it in all down-sampling layers in the ARC-Net. The structure of the down-sampling block is presented in Figure 2 (b).

To improve the accuracy of semantic segmentation for high-resolution aerial images, the models usually need to enlarge the receptive field to gather sufficiently rich contextual information for each individual pixel [60]. The method used in the past is combing the stacking of convolutional layers with down-sampling layers. However, these extra convolution layers substantially increase computational effort during learning. Moreover, over-down-sampling is harmful to the dense pixel-level classifications it leads to a loss of unrecoverable spatial information [61]. Dilated convolution [60] introduces an additional parameter to the convolutional layers named the dilation rate. This rate defined spacing between the values in a kernel, delivering a wider field of view at the same computational cost. Therefore, the dilated convolution is conducive to enlarge the receptive field and to enhance the segmentation performance [62], [63]. Following the suggestion from the literature, we set the dilation rate to the sequence 2, 4, 8, 16 in block 9-12 incorporating the RBAC module to obtain a wide receptive field. The description of the dilated convolution is presented in Figure 3.

B. ATROUS SPATIAL PYRAMID POOLING AS THE CONNECTING MODULE

ASPP, proposed in DeepLab_v2 [33], has several parallel atrous convolutions that maintain the same feature map and fuse the outputs at the end. Comparing with the standard

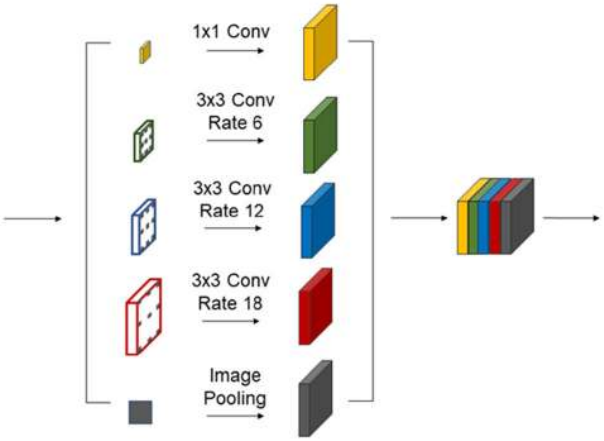


FIGURE 4. Structure of atrous spatial pyramid pooling module in the ARC-Net.

convolutional layer, the atrous convolutions can effectively increase the receptive field of the network without extra down-samplings. In this work, we employ the ASPP module with a 1×1 convolution and three branches of atrous convolution with rate 6, 12, 18 as a connector in block 13 after encoder to effectively capture multi-scale contextual information. Figure 4 presents the detailed structure of the ASPP module in the ARC-Net.

C. DECODER WITH SIMPLE DECONVOLUTIONS

The main task of the decoder phase is to up-sample the feature maps and to recover the input resolution from the encoder phase. Previous works have used heavy-weight decoders [42], [64], which increases computational cost. Inspired by the idea of light-weight and asymmetric decoder, we follow a strategy which similar to ENet [44] and has a small decoder to up-sample the output of the encoder fine-tuning the semantic information. The decoder used in this article is composed of blocks 14 to 20, including the up-sampler block and RBAC module, see Table 1 and Figure 1. In contrast to SegNet and ENet, we utilize simple deconvolution layers with stride 2 as the up-sampling block to reduce computational costs. Two RBAC modules are employed to collect the contextual information after each up-sampling block. This operation is repeated twice (blocks 14-19). Finally, the up-sampling block is utilized in block 20 generating the output segmentation into two classes: building and non-buildings.

III. EXPERIMENTAL DATASETS AND EVALUATION

In this section, we conducted experiments on two building datasets: the INRIA Aerial Image Labeling Dataset and the WHU Building Dataset. Data processing methods and experimental settings are discussed in detail. A standard metric with five values was applied to evaluate the performance and efficiency of the proposed ARC-Net. Other state-of-the-art deep learning models are introduced and their performances are compared to ARC-Net.

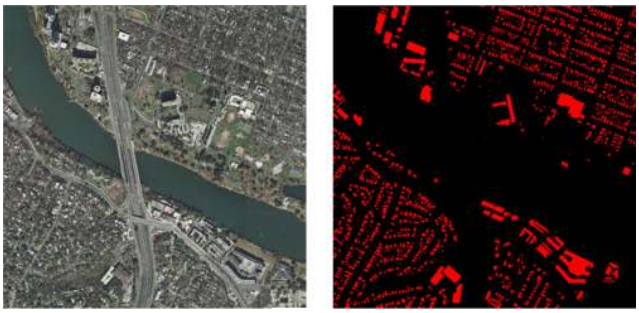


FIGURE 5. Image and label example selected from the INRIA dataset: (a) Aerial Image; (b) Label Image. Black and red pixels mark non-building and building, respectively.

FIGURE 6. Image and label example selected from the WHU dataset: (a) Aerial Image; (b) Label Image. Black and red pixels mark non-building and building, respectively.

A. DATASETS

The first dataset used in this research is the INRIA Aerial Image Labeling Dataset [52]. This dataset covers different cities all over the world, including Austin, Chicago, Kitsap, Western/Eastern Tyrol, Vienna, Bellingham, and San Francisco. The spatial resolution of each image is 0.3 m with a size of 5000×5000 pixels and surface coverage of 1500×1500 m^2. Following previous investigations [29,40], we selected the first five images of each city for validation and the rest for training. Only two semantic classes were considered as the ground truth; buildings, and non-buildings. An example of an input image and its corresponding label are presented in Figure 5. The red color represents the buildings and the black color presents the background.

The WHU Building Dataset is proposed by [53], covering a surface area of about 450 km^2 in Christchurch, New Zealand. The dataset contains 8189 images of 512×512 pixels with a spatial resolution of 0.3 m. This dataset was divided into a training set, a validation set, and a test set, consisting of 4736 images, 1036 images, and 2416 images, respectively. Figure 6 shows an original image and its corresponding label.

B. DATA PROCESSING

Data augmentation is an effective way to enlarge the datasets and to avoid overfitting [35]. In this study, windows were rotated by 90, 180, and 270 degrees. Moreover, horizontal and vertical flipping were randomly applied with a probability

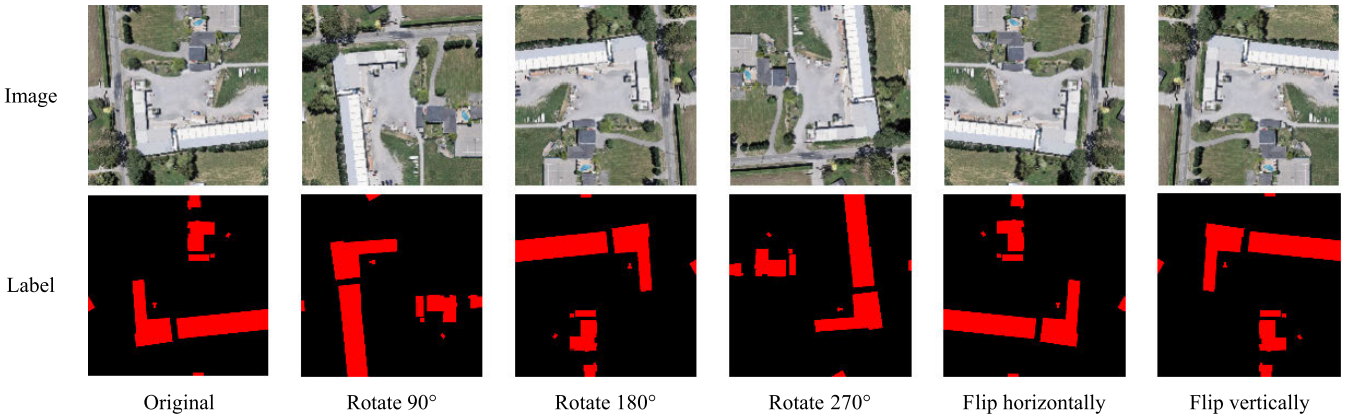


FIGURE 7. An example of data augmentation by rotating and flipping.

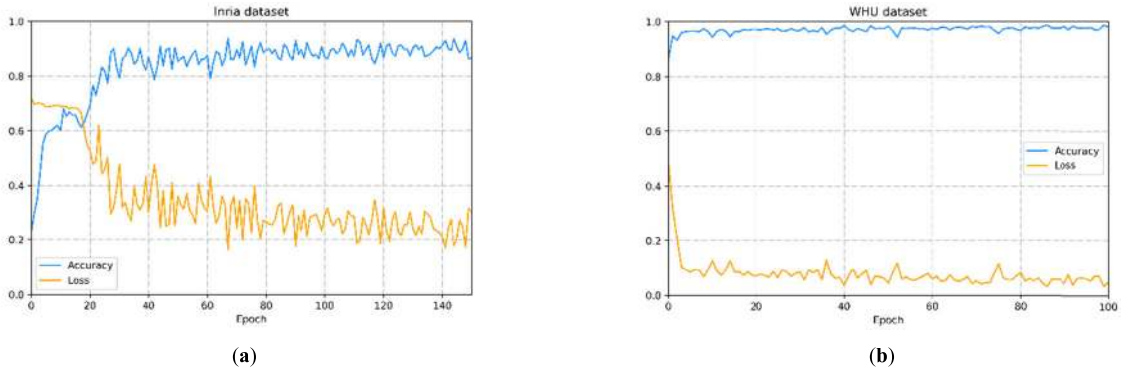


FIGURE 8. The accuracy and loss of the proposed model for training the datasets: (a) INRIA dataset; (b) WHU dataset. Accuracy and loss are plotted as blue and yellow curves respectively.

of 0.5. Figure 7 presents an example of the image after data argumentation by rotating and flipping. For the deep learning process, the pixel values of each image were scaled to the interval [0, 1] by dividing by 255. The sigmoid function was utilized in the final layer to generated outputs within the range [0, 1]. Final segmentation results were produced by further applying a threshold of 0.5. No additional post-processing was performed.

C. EXPERIMENTAL SETTINGS

The building extraction experiments were built on the deep learning framework named PyTorch. The experiments were conducted on computer servers with two NVIDIA GeForce GTX 1080 Ti (11GB). Parallelization was utilized to make full use of the available graphics processing unit (GPU) capability and to accelerate computation. Due to the limitation in GPU memory, we randomly cropped all images in two datasets to be 256×256 pixels for model training and cross-validation of each epoch.

In the process of the experimental setting, we conducted many comparative experiments to finally determine the optimal model parameters. In the training phase, we adopted the ADAM stochastic optimizer [57] with an initial learning rate of 0.0001. To avoid over-fitting, an L2 regularization was introduced with a weight decay of 0.0001 [37]. Models had been trained with 150 epochs for the INRIA dataset and 100 epochs for the WHU dataset, respectively. To overcome the limitation of the GPU memory, the mini-batch size was set as 8. Figure 8 displays the dynamic accuracies and losses of the INRIA and WHU datasets with increasing epochs. It is obvious that the loss gradually decreases while the accuracy increases and retains at a high and stable level.

D. EVALUATION METRICS

The quantitative experiments are based on five evaluation metrics: the 'Overall Accuracy' (OA), 'Precision', 'Recall', 'F1-score', and Intersection-over-Union ('IoU'). 'Overall Accuracy' refers to the number of correctly classified pixels divided by the total number of test pixels. 'Precision' is the fraction of correctly classified positive pixels amongst all predicted positive pixels where 'positive pixel' refers to the pixel of the building. 'Recall' is the proportion of correctly classified positive pixels amongst all true target pixels. 'F1-score' is the weighted average of precision and recall. 'IoU' is the average value of the intersection of the prediction and ground-truth regions over their union. The five metrics are presented as follows:

$$OverallAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives

E. MODEL COMPARISONS

The performance of ARC-Net is compared with the following four FCN-based models:

SegNet: Badrinarayanan *et al.* [28] proposed SegNet for the semantic pixel-wise segmentation. Encoder-decoder structure with MaxPooling operations is employed in SegNet for up-sampling the lower-level information input feature maps. Thus, SegNet is considered as efficient in terms of memory and computational time [17].

U-Net: The U-Net architecture was proposed by Ronneberger *et al.* [29] for biomedical image segmentation. Contracting paths and symmetric expanding paths are used to aggregate contextual information. Multiple skip connections were introduced between the upper and downer layers. Due to its robustness and excellent performance, U-Net and its variants are widely adopted for many semantic segmentation tasks in recent years.

ENet: ENet was proposed by Chaurasia *et al.* in 2017 [44], aiming at performing pixel-wise semantic segmentation with low latency operation. The ENet model is providing an accuracy similar or - in some cases even - better accuracy with far fewer computations achieving a good trade-off between accuracy and processing time of a network.

ERFNet: The ERFNet was proposed by Romera *et al.* in 2017 [45]. The core of the ERFNet architecture is the novel layer that uses residual connections and factorized convolutions to remain computational efficient while delivering remarkable accuracy. The ERFNet model can be applied in real-time while providing accurate semantic segmentation [45].

SRI-Net: Liu *et al.* [37] proposed a novel FCN-based network named SRI-Net in 2019. The spatial residual inception (SRI) module was introduced to capture and aggregate multi-scale contexts for a better semantic representation. Meanwhile, depth-wise separable convolutions were employed to further improve the accuracy and to decrease the number of model parameters.

IV. RESULTS

A. EXPERIMENTAL RESULTS ON THE INRIA DATASET

We first conduct the comparisons on the INRIA dataset between the ARC-Net model and the well-known models including SegNet, U-Net, ENet, and ERFNet. The experiments are implemented on the test dataset with the same experimental settings. Figure 9 presents the qualitative segmentation results for all five models on the INRIA dataset. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative, and true negative, respectively. SegNet and ENet return more false negatives (blue) while U-Net gains more false positives (red) than the other models. ERFNet gets more false positives (red) compared to ENet. By contrast, the proposed ARC-Net shows significantly less false positives (red) and false negatives (blue) than the other models and is able to maintain a high degree of completeness in building segmentation on the INRIA dataset. However, all models have consistently misclassified parts of the built-up area in the top left corner of the first test image.

TABLE 2. Quantitative comparison with the state-of-the-art models on the INRIA dataset. The highest value for each metric is marked as bold.

	OA	Precision	Recall	F1 score	IoU
SegNet	0.880	**0.913**	0.739	0.817	0.692
U-Net	0.899	0.847	0.846	0.844	0.773
ENet	0.909	0.882	0.831	0.855	0.743
ERFNet	0.902	0.826	0.856	0.844	0.733
SRI-Net[37] (report)	-	0.858	0.819	0.831	0.711
ARC-Net	**0.925**	0.896	**0.868**	**0.875**	**0.779**

The quantitative comparison of the networks across the entire test dataset is displayed in Table 2. The 3-digit number is utilized to better differentiate performance. The proposed ARC-Net outperforms the other FCN-based models on all evaluation metrics except for Precision. ARC-Net is the best among all networks on the Overall Accuracy metric with an improvement of 2.0% (0.925 vs. 0.909) over the next best model ENet. As for Precision, SegNet holds the highest value and gains 1.9% (0.913 vs. 0.896) over the proposed ARC-Net. For Recall, U-Net, ERFNet, and the proposed ARC-Net show significantly performance over the other two methods while ARC-Net achieves the highest value being 1.4% (0.868 vs. 0.856) ahead of the ERFNet model. Similarly, ARC-Net reaches the best F1-score where U-Net and ERFNet present the same performance. For the IoU metric, ARC-Net has scored the best value 6.3% ahead of ERFNet (0.779 vs. 0.733) and even 9.6% ahead of SRI-Net (0.779 vs. 0.711).

B. EXPERIMENTAL RESULTS ON THE WHU DATASET

Building segmentation results of different CNN models on the WHU dataset are displayed in Figure 10 for qualitative comparisons. Clearly, all models present quite similar performance in building segmentation except the SegNet model. As for the INRIA dataset, SegNet returns too many false positives (blue) and false negatives (red) indicating the worst performance on the WHU dataset across the five tested models. In contrast, the proposed ARC-Net (last row) performs best among all models accurately detecting an edge

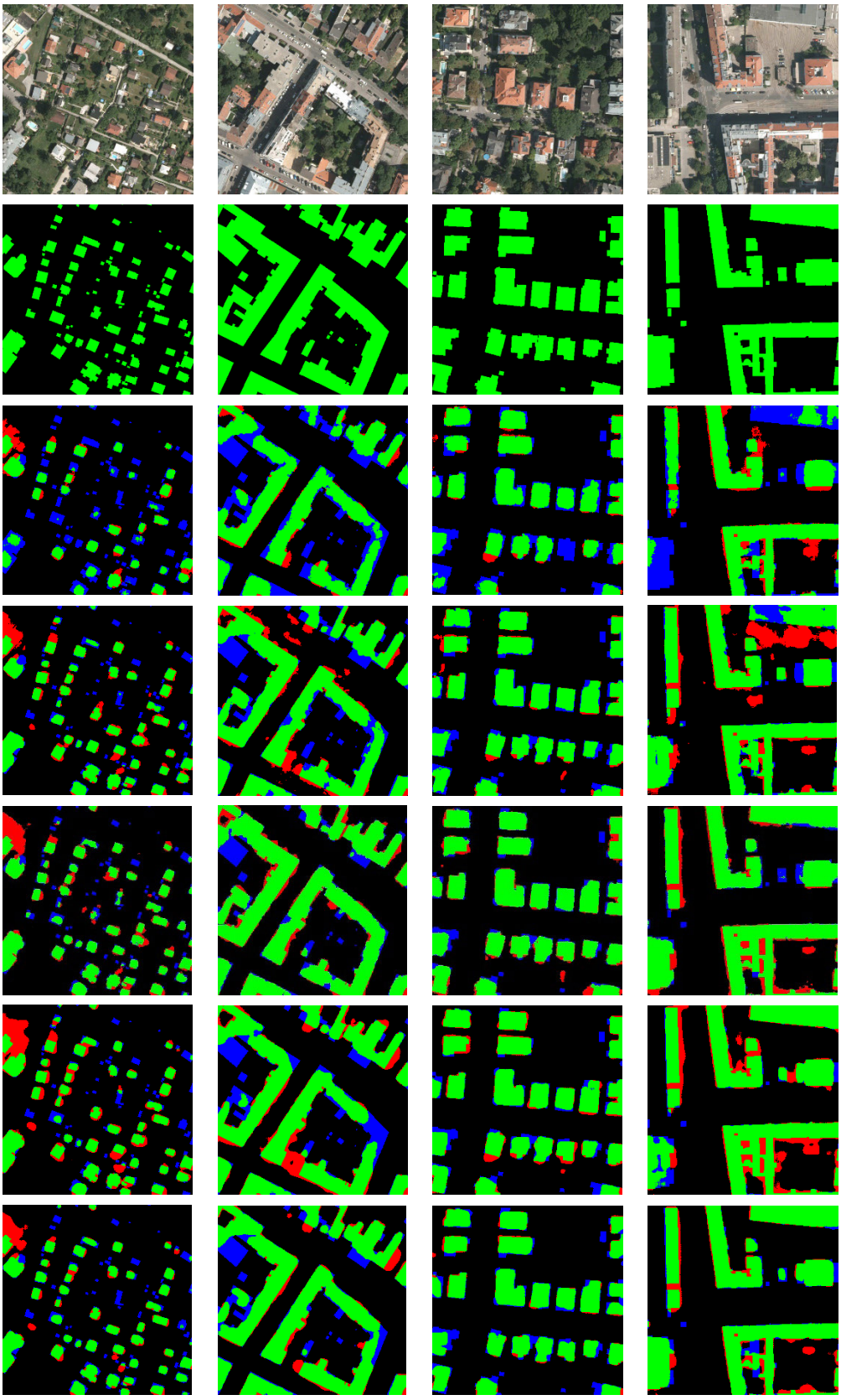


FIGURE 9. Examples of segmentation results by different FCN-based models on the INRIA dataset. The first two rows are aerial images and ground truth, respectively. Rows 3 to 7 are building extraction results of SegNet, U-Net, ENet, ERFNet, and our proposed ARC-Net, respectively. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative, and true negative, respectively.

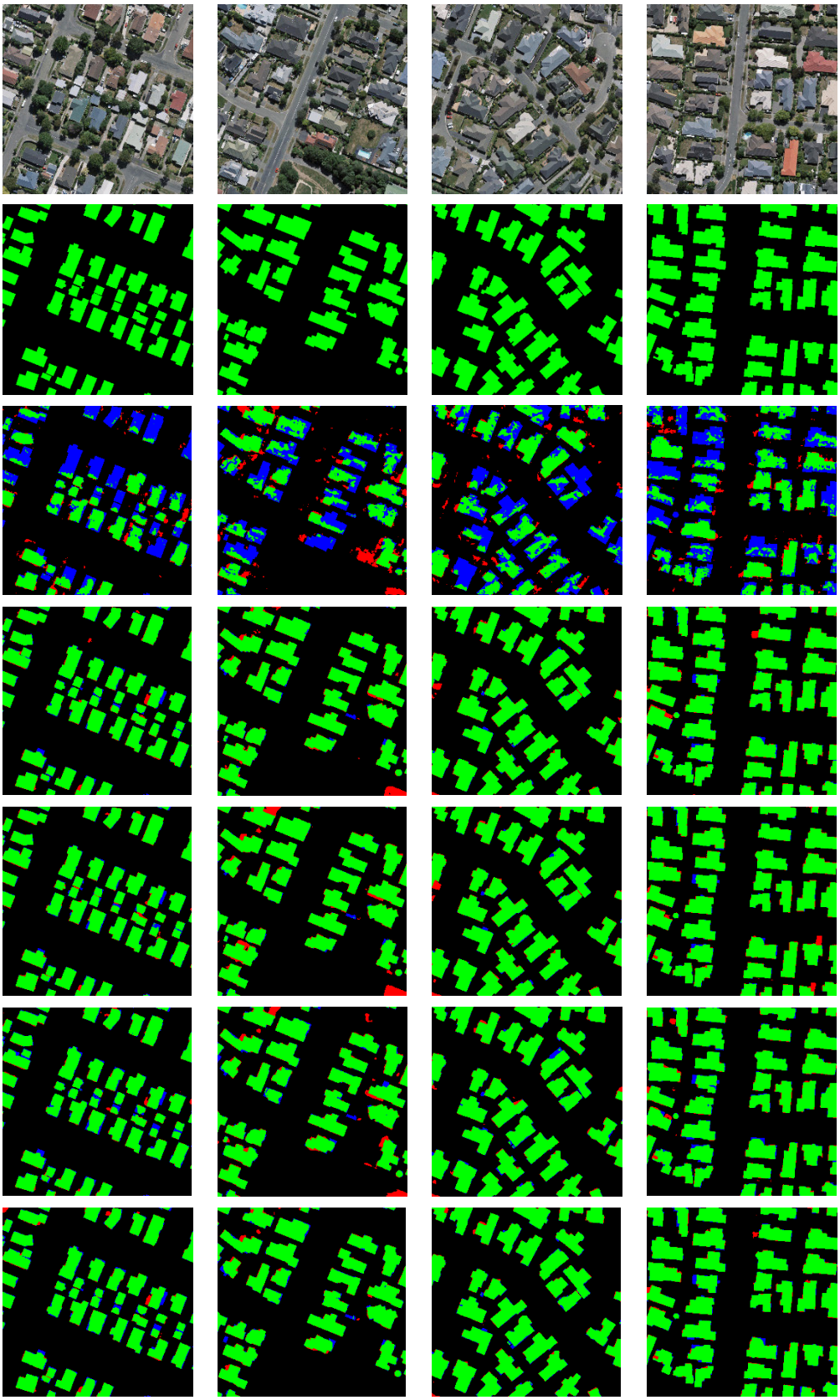


FIGURE 10. Examples of segmentation results by different FCN-based models on the WHU dataset. The first two rows are aerial images and ground truth, respectively. Rows 3 to 7 are building extraction results of SegNet, U-Net, ENet, ERFNet, and our proposed ARC-Net, respectively. The green, red, blue, and black pixels of the maps represent the predictions of true positive, false positive, false negative, and true negative, respectively.

TABLE 3. Quantitative comparison with the state-of-the-art models on the WHU dataset. The highest value for each metric is marked as bold.

	OA	Precision	Recall	F1 score	IoU
SegNet	0.842	0.920	0.684	0.784	0.645
U-Net	0.974	0.949	**0.963**	0.956	0.916
ENet	0.973	0.943	0.948	0.946	0,913
ERFNet	0.968	0.948	0.944	0.946	0.898
SRI-Net[37] (report)	-	0.952	0.933	0.942	0.891
ARC-Net	**0.975**	**0.964**	0.951	**0.957**	**0.918**

in building segmentation. Moreover, for column 2, all deep learning models wrongly classified a building at the bottom of the area except the proposed ARC-Net.

For a quantitative evaluation of the performance, we calculated the individual evaluation metrics presented in Table 3. The proposed ARC-Net model holds the highest scores relative to the established models except for Recall where U-Net performs the better. The performance differences across the deep learning models except SegNet are small, especially for U-Net, ERFNet, and SRI-Net. Compared to the ERFNet, the proposed model still yields a higher F1-score by 1.2% (0.957 vs. 0.946) and a higher IoU by 2.2% (0.918 vs. 0.898).

C. COMPUTATIONAL EFFICIENCY

Computational efficiency is an additional key performance indicator of deep learning models. The computational performance includes the cost and complexity of the model training and testing. As stated in the introduction, the main motivation of ARC-Net is to achieve high prediction accuracy with less computational costs when applied to the building extraction. The model-training for each epoch and testing time of different deep learning models are presented in Table 4.

For the model-training phase, SegNet, and U-Net required about 180 sec while costs for ERFNet and ARC-Net were about 20 more sec on the INRIA dataset. For the WHU dataset, on the other hand, SegNet took the highest training time (455.5 sec) while ARC-Net required only 160.2 sec per epoch. For the model-testing phase, ARC-Net was demanding the least time per epoch on the INRIA dataset (13.4 sec) as well as the WHU dataset (12.4 sec). The SegNet model is the most time-consuming and required 1.2 to 1.5 times more time than the fastest ARC-Net model. It is pointed out that the new ARC-Net took less time than ERFNet for both training and testing on both of the datasets.

To conclude, the new ARC-Net model requires about an extra 11% time more than the fastest SegNet on the model-training of the INRIA dataset, but it returns outstanding performance of building extraction as presented in Section IV-A and IV-B. For the WHU dataset, ARC-Net is the most effective model as well as showing the best computational performance in building extraction. These findings demonstrate that the new ARC-Net implements a good balance between computational performance and segmentation efficiency on the two building datasets.

TABLE 4. Comparison of model-training and model-testing time in seconds of each epoch for different FCN models. The minimums are marked as bold.

	INRIA Training	INRIA Testing	WHU Training	WHU Testing
SegNet	**178.1**	16.8	455.5	19.8
U-Net	180.2	14.6	243.3	14.4
ERFNet	204.3	13.5	209.6	13.3
ARC-Net	198.1	**13.4**	**160.2**	**12.4**

TABLE 5. Comparison of the proposed model with different groups of dilated convolutions on the INRIA dataset. The higher values are marked as bold.

	OA	Precision	Recall	F1 score	IoU
With one group (ARC-Net)	**0.925**	**0.896**	**0.868**	**0.875**	**0.779**
With two	0.917	0.868	0.861	0.871	0.755
With three	0.916	0.879	0.852	0.858	0.741

V. DISCUSSION

A. GOING DEEPER OR NOT

Previous researches have demonstrated that a deeper CNN structure with more convolution operations processes more semantic information during the training phase, which helps to improve the classification accuracy [20]. However, due to the limitation of computational sources and the complexity of structure design, a deeper deep learning neural network requires fitting a larger amount of parameters and can lead to instability introducing gradient explosion and gradient vanishing. In this article, we developed the RBAC module and incorporated it into the ARC-Net in combination with dilated convolutions in order to seek a balance between accuracy and efficiency. As mentioned in Section II-A the dilation rate in the RBAC module in the encoder phase is set as the sequence of 2, 4, 8, 16 which raises the question if a repeated application of dilated convolution in the RBAC module would enhance the performance. To optimize the architecture of the ARC-Net model, we kept other experimental settings unchanged and conducted three comparison experiments with different groups of dilated convolutions for the RBAC module, including one group (as used in ARC-Net), two groups, and three groups, respectively. The results of the comparison on the INRIA dataset are presented in Table 5. Results show that one repeated dilated convolution module (as applied in the proposed ARC-Net) achieves in fact the best score in the five score metrics in comparison to the other two architecture design.

B. THE EFFECT OF ATROUS SPATIAL PYRAMID POOLING

The ASPP module has demonstrated its considerable performance in aggregating multi-scale contextual features, which improve the extraction accuracy of buildings in different

TABLE 6. Comparison of the ARC-Net with or without the ASPP module on the WHU dataset. The higher values are marked as bold.

	OA	Precision	Recall	F1 score	IoU
Without ASPP	0.969	0.953	0.948	0.951	0.903
With ASPP	**0.975**	**0.964**	**0.951**	**0.957**	**0.918**

sizes, especially medium-sized to over-sized buildings [65]. One crucial innovation of the ARC-Net in comparison to the ERFNet is that it employs the ASPP module as a connector between the encoder and the decoder. To test the performance, we conducted a comparison experiment with and without the ASPP module of ARC-Net on the WHU dataset. As presented in Table 6, the model with ASPP shows an obvious improvement over the model without ASPP across all evaluation metrics. The comparison result demonstrates the efficiency and applicability of the ASPP module as a connector for building extraction from high-resolution aerial images [34].

C. ABOUT THE PROPOSED METHOD

Deep learning methods, especially FCN-based models, have been widely applied in automatic building extraction from high-resolution aerial images. Recently, several advanced FCN-based models have delivered improved feature representation capabilities to achieve better classification performance (e.g., USPP [34], EU-Net [65], and MC-FCN [66]). However, most of the existing models focus on improving the accuracy with very little consideration on the computational efficiency, which suffers under large numbers of weight parameters introduced in the model design and high memory costs in the learning phase.

In this article, we designed a novel asymmetric encoder-decoder network with residual connections, named ARC-Net, to pursue good segmentation performance with lower computational cost. The proposed model focuses on three key innovations: (1) The residual block with asymmetric convolutions (RBAC) module is proposed to reduce the model parameters and address the degradation problem (2) Larger convolutional kernels and dilated convolutions are used in the backbone of the architecture to enlarge the receptive field and to obtain rich semantics when detecting objects in complex backgrounds. Moreover, depth-wise separable convolutions are introduced to improve computational efficiency without reducing prediction performance. (3) The ASPP module is utilized as a bridge between the encoder and decoder to further aggregate spatial context information. Through these three innovations, the proposed ARC-Net model implements a good balance between performance and efficiency achieving better predictions with less computational resources.

The training accuracy and loss presented in Section III-C show that the tested CNN models achieved better performance and stability on the WHU dataset than on the INRIA dataset. Compared to the segmentation results on the INRIA dataset, the IoU metric of the different CNN models is all higher than 85%, indicating that the WHU dataset is of higher quality and building and background are easier to distinguish. The INRIA dataset includes wrong labels, high buildings, and shadows, factors that may heavily influence the discriminative ability of CNN models as presented in [37], [53]. For this reason, the differences in the experimental results between the two datasets as shown in this article are reasonable. Moreover, these results verify that the proposed ARC-Net has a beneficial capability for practical application scenarios.

D. LIMITATIONS

Despite the good performance and efficiency achieved, the application of the ARC-Net model is still limited. With the recent progress of remote sensing technology, it is getting much easier to obtain remote sensing images at different scales and spectral bandwidths also to meet different research requirements. However, the datasets used in this article do not contain images from different sensors or different sensor types, such as hyperspectral images and SAR images. Moreover, the buildings have complex morphological characteristics, such as different height, shape, and orientations while with the current approach these attributes cannot directly be determined through deep learning networks. In the future, we will expand to the multi-source training data and integrate multi-disciplinary knowledge to jointly extract the buildings from remote sensing images.

VI. CONCLUSION

The main objective of this research is to propose an efficient FCN-based model for automatic building extraction from high-resolution aerial images that is achieving an outstanding accuracy with less computational resources. To address this issue, we proposed the ARC-Net model which incorporates an asymmetric encoder-decoder structure with the ASPP module as a connector. The RBAC module, the core of the ARC-Net, is designed by incorporating residual connections with depth-wise separable and asymmetric convolutions to reduce the number of model parameters and to accelerate the calculations. In addition, dilated convolutions and the ASPP module are utilized to extend the receptive field for delivering desirable segmentations. Experiments on two public building datasets, the INRIA and WHU datasets, have shown that the proposed ARC-Net outperforms other established FCN-based models with higher metric scores and less computational time. The buildings were extracted successfully by ARC-Net with fewer classification errors and shaper boundaries which demonstrates that the proposed ARC-Net achieves high accuracy and efficiency in building extraction from high-resolution aerial images. In future studies, multi-resources remote sensing data from different sensors will be combined to further improve automatic building extraction.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and an associate editor for their constructive comments and

suggestions. They would also like to thank INRIA for providing the INRIA Aerial Image Labelling Dataset on the website (https://project.inria.fr/aerialimagelabeling/), as well as Shunping Ji for providing the WHU Building Dataset in the website (http://study.rsgis.whu.edu.cn/pages/download/). Also, Yaohui Liu wants to express his great acknowledgment to the China Scholarship Council (CSC) for providing financial support to take a Research Fellow Position at The University of Queensland, Australia.

REFERENCES

[1] R. N. Clark and T. L. Roush, "Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications," *J. Geophys. Res., Solid Earth*, vol. 89, no. B7, pp. 6329–6340, Jul. 1984.

[2] Y. Liu, Z. Li, B. Wei, X. Li, and B. Fu, "Seismic vulnerability assessment at urban scale using data mining and GIScience technology: Application to urumqi (China)," *Geomatics, Natural Hazards Risk*, vol. 10, no. 1, pp. 958–985, Jan. 2019.

[3] X. Li, Z. Li, J. Yang, Y. Liu, B. Fu, W. Qi, and X. Fan, "Spatiotemporal characteristics of earthquake disaster losses in China from 1993 to 2016," *Natural Hazards*, vol. 94, no. 2, pp. 843–865, Nov. 2018, doi: 10.1007/s11069-018-3425-6.

[4] B. Zhang, Z. Chen, D. Peng, J. A. Benediktsson, B. Liu, L. Zou, J. Li, and A. Plaza, "Remotely sensed big data: Evolution in model development for information extraction [point of view]," *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, Dec. 2019.

[5] Y. Liu, E. So, Z. Li, G. Su, L. Gross, X. Li, W. Qi, F. Yang, B. Fu, A. Yalikun, and L. Wu, "Scenario-based seismic vulnerability and hazard analyses to help direct disaster risk reduction in rural weinan, China," *Int. J. Disaster Risk Reduction*, vol. 48, Sep. 2020, Art. no. 101577.

[6] 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, p. 403, Feb. 2019.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[8] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, pp. 511–518.

[9] D. A. Clausi, "An analysis of co-occurrence texture statistics as a function of grey level quantization," *Can. J. Remote Sens.*, vol. 28, no. 1, pp. 45–62, Jan. 2002.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[11] Y. Dong, B. Du, and L. Zhang, "Target detection based on random forest metric learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 4, pp. 1830–1838, Apr. 2015.

[12] J. Inglada, "Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features," *ISPRS J. Photogramm. Remote Sens.*, vol. 62, no. 3, pp. 236–248, Aug. 2007.

[13] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and k-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Oct. 2009.

[14] Ö. Aytekin, U. Zongur, and U. Halici, "Texture-based airport runway detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 471–475, May 2013.

[15] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.

[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[17] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[22] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: Univ. Toronto, 2013.

[23] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 210–223.

[24] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery," *Proc. SPIE*, vol. 9405, Feb. 2015, Art. no. 94050K.

[25] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *J. Imag. Sci. Technol.*, vol. 60, no. 1, pp. 104021–104029, Jan. 2016, doi: 10.2352/J.ImagingSci.Technol.2016.60.1.010402.

[26] X. Wei, K. Fu, X. Gao, M. Yan, X. Sun, K. Chen, and H. Sun, "Semantic pixel labelling in remote sensing images using a deep convolutional encoder-decoder model," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 199–208, Mar. 2018.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[28] 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.

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

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

[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[32] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[33] 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.

[34] Y. Liu, L. Gross, Z. Li, X. Li, X. Fan, and W. Qi, "Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling," *IEEE Access*, vol. 7, pp. 128774–128786, 2019, doi: 10.1109/access.2019.2940527.

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

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

[37] 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 Sens.*, vol. 11, no. 7, p. 830, Apr. 2019.

[38] Y. Zhang, W. Gong, J. Sun, and W. Li, "Web-net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries," *Remote Sens.*, vol. 11, no. 16, p. 1897, Aug. 2019, doi: 10.3390/rs11161897.

[39] B. Yu, L. Yang, and F. Chen, "Semantic segmentation for high spatial resolution remote sensing images based on convolution neural network and pyramid pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3252–3261, Sep. 2018.

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

[41] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 139–149, Aug. 2017.
[42] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 1, p. 20, Dec. 2018, doi: 10.3390/rs11010020.
[43] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.
[44] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: https://arxiv.org/abs/1606.02147
[45] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 263–272, Jan. 2018.
[46] S. Y. Lo, H. M. Hang, S. W. Chan, and J. J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proc. ACM Multimedia Asia*, 2018, pp. 1–6.
[47] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861
[48] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[49] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[50] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.
[51] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.
[52] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229.
[53] 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.
[54] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*. [Online]. Available: http://arxiv.org/abs/1606.02147
[55] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[56] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[57] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[58] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1269–1277.
[59] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," 2014.
[60] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.
[61] 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, doi: 10.1109/access.2019.2912822.
[62] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.
[63] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[64] G. Pasquali, G. C. Iannelli, and F. Dell'Acqua, "Building footprint extraction from multispectral, spaceborne Earth observation datasets using a structurally optimized U-Net convolutional neural network," *Remote Sens.*, vol. 11, no. 23, p. 2803, Nov. 2019.
[65] 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, p. 2813, Nov. 2019.
[66] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, Mar. 2018, doi: 10.3390/rs10030407.

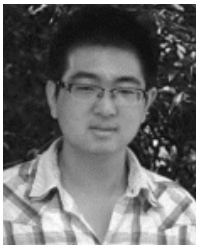
YAOHUI LIU received the Ph.D. degree in cartography and geographical information system from the Institute of Geology, China Earthquake Administration, China, in 2020. He is currently a Lecturer at Shandong Jianzhu University. His research interests include computer vision, remote sensing, deep learning, and risk management.

JIE ZHOU received the B.S. degree in geographic information science from Yunnan Normal University, in 2017, where she is currently pursuing the M.S. degree in geography teaching. Her research interests include geography-education informationization, GIS, and geography teaching.

WENHUA QI received the bachelor's degree in hydrology and water resources engineering from the School of Water Resources and Environment, China University of Geosciences, Beijing, China, in 2008, and the master's degree in tectonics from the Institute of Geology, China Earthquake Administration, Beijing, in 2011. His research interests include remote sensing and citizen science for natural disaster risk assessment and governance.

XIAOLI LI received the master's degree in structural geology from the Institute of Geology, China Earthquake Administration, Beijing, China, in 2008. She is currently a Senior Engineer with the China Earthquake Networks Center, Beijing. Her research interests include earthquake emergency response and management, earthquake disaster risk assessment techniques, and application of GPS, GIS, and RS to earthquake emergency and earthquake resistance and disaster relief.

LUTZ GROSS received the Ph.D. degree in mathematics from the University of Karlsruhe, Karlsruhe, Germany, in 1996. He is currently an Associate Professor with The University of Queensland, Brisbane, QLD, Australia. His research interests include large-scale inversion of geophysical data and numerical modeling.



QI SHAO received the Ph.D. degree in environmental engineering from The University of Queensland, Australia, in 2015. He is currently a Research Fellow with The University of Queensland and The University of Melbourne, Australia. His research interests include numeric simulation of multiphase fluid flow in porous media, geological storage and immobilisation of carbon dioxide, and high performance computing.

ZHENGGUANG ZHAO received the M.S. degree in geodetection and information technology from the China University of Mining and Technology, Beijing, in 2010. He is currently pursuing the Ph.D. degree with The University of Queensland, Australia. He also explored the feasibility of applying machine learning-based methods in micro-earthquake detection and location. His research interests include automated microseismic event detection and location with machine learning.

LI NI, photograph and biography not available at the time of publication.

XIWEI FAN received the Ph.D. degree in cartography and geographical information system from the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, in 2015. He is currently an Associate Research Fellow with the Institute of Geology, China Earthquake Administration. His research interests include the retrieval and validation of land surface temperature/emissivity and earthquake damage estimation.

ZHIQIANG LI received the Ph.D. degree in geodynamics and tectonophysics from the Institute of Geology, China Earthquake Administration, Beijing, China, in 1997. He is currently a Professor with the China Earthquake Networks Center, Beijing. His research interests include earthquake emergency response and management, earthquake emergency basal database technology, earthquake disaster risk assessment techniques, and application of GPS, GIS, and RS to earthquake emergency and earthquake resistance and disaster relief.