IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 17, 2019, accepted January 1, 2020, date of publication January 6, 2020, date of current version January 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964043

Automatic Building Extraction From High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network With Non-Local Block

SHENGSHENG WANG, XIAOWEI HOU, AND XIN ZHAO
College of Computer Science and Technology, Jilin University, Changchun 130012, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

Corresponding author: Xin Zhao (focusxin@outlook.com)

This work was supported in part by the Science and Technology Development Project of Jilin Province, China, under Grant 20190302117GX and Grant 20180101334JC, and in part by the Innovation Capacity Construction Project of Jilin Province Development and Reform Commission under Grant 2019C053-3.

ABSTRACT Extracting buildings automatically from high-resolution aerial images is a significant and fundamental task for various practical applications, such as land-use statistics and urban planning. Recently, various methods based on deep learning, especially the fully convolution networks, achieve impressive scores in this challenging semantic segmentation task. However, the lack of global contextual information and the careless upsampling method limit the further improvement of the performance for building extraction task. To simultaneously address these problems, we propose a novel network named Efficient Non-local Residual U-shape Network(ENRU-Net), which is composed of a well designed U-shape encoder-decoder structure and an improved non-local block named asymmetric pyramid non-local block (APNB). The encoder-decoder structure is adopted to extract and restore the feature maps carefully, and APNB could capture global contextual information by utilizing self-attention mechanism. We evaluate the proposed ENRU-Net and compare it with other state-of-the-art models on two widely-used public aerial building imagery datasets: the Massachusetts Buildings Dataset and the WHU Aerial Imagery Dataset. The experiments show that the accuracy of ENRU-Net on these datasets has remarkable improvement against previous state-of-the-art semantic segmentation models, including FCN-8s, U-Net, SegNet and Deeplab v3. The subsequent analysis also indicates that our ENRU-Net has advantages in efficiency for building extraction from high-resolution aerial images.

INDEX TERMS Deep learning, semantic segmentation, fully convolution network, building extraction, non-local method.

I. INTRODUCTION

Automatic building extraction from high-resolution aerial imagery is a fundamental task for various applications, such as urban planning, economic statistics, disaster monitoring, etc. The target of this task is to distinguish the buildings from background in an aerial image in pixel-wise, as shown in Figure 1. So, it is usually defined as a semantic segmentation task, which is a long-standing topic in computer vision. Extracting buildings accurately from high-resolution aerial imagery is a tough task with several challenges. First, the aerial imagery contains redundant object details, e.g. the building shadow and the trees, especially the high-resolution aerial imagery, which increases the difficulty of building extraction. Another challenge is

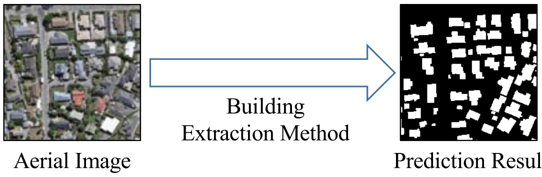


FIGURE 1. Illustration of extracting buildings from aerial images. The white and black pixels in prediction result denote buildings and background respectively.

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

that many objects in high-resolution aerial imagery have low inter-class distance [1]. For instance, the roofs look very similar to the roads in the appearance. Besides, the diverse characteristics of buildings, e.g. size, shape, and color, further increases the hardness of this semantic segmentation task. Consequently, how to extract building accurately from high-resolution aerial images is a challenging task that urgently needs to be solved.

In recent years, owing to the development of the computer hardware and the available large-scale aerial imagery dataset with high quality labels, the data-driven methods based on deep learning have been applied to building extraction. A typical model of deep learning is convolution neural network (CNN), CNN has been successfully applied to image classification [2]–[4], object detection [5], image semantic segmentation [6] and other tasks. The great success of CNNs is mainly due to they can automatically extract hierarchical features by utilizing several successive convolution layers with learnable parameters. Since Long *et al.* [7] transformed the CNN to fully convolution network (FCN) by discarding the fully connected layers and using bilinear interpolation as the upsampling method to conduct a pixel-wise prediction, FCN has been extensively applied to semantic segmentation and outperforms the traditional methods based on the hand-engineered features. Li *et al.* [8] compared some conventional methods to FCN on building extraction and proved that FCN has incomparable advantages over traditional methods in this task.

Although recent FCNs [9]–[15] improve the segmentation accuracy remarkably on various aerial imagery datasets, two challenges of building extraction still exist. First, the employment of pooling layers cause the loss of detailed information, and coarse upsampling layers without the detailed information, e.g. 8× bilinear interpolation directly after the feature extractor, would reduce the recognition accuracy of small buildings, especially the contours. Second, despite the adoption of successive pooling and convolution layers expand the field of view, the FCNs still could not obtain abundant global contextual information due to their local valid receptive field [16], which produces misclassification when dealing the inner pixels in large buildings. As a consequence, the coarse upsampling layers, and the conventional structures of FCNs, provoke numerous misclassification when extracting buildings from aerial images. To address such problems, many novel structures have been proposed, among these structures, the encoder-decoder structure [17], [18] and the non-local block [19] which are carefully designed have been proven to perform well in the previous works.

The encoder-decoder networks adopt several cascaded upsampling layers after the feature extractor to expand the feature maps carefully. And they also deliver the shallow feature maps that contain detailed information to the deep layers by skip-connections, which increases the ability to recognize small buildings from high-resolution aerial imagery. Nevertheless, limited by their simple structures, the ordinary encoder-decoder networks have a feeble ability to capture global contextual information and produce numerous misclassification when classifying the inner pixels of large scale buildings from high-resolution aerial imagery.

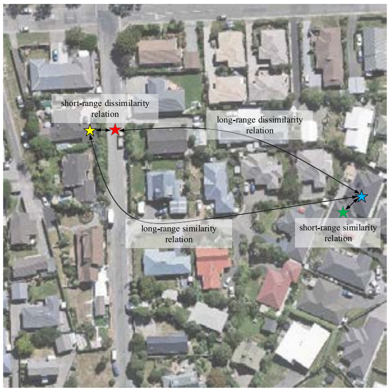

short-range dissimilarity relation
long-range dissimilarity relation
long-range similarity relation
short-range similarity relation

FIGURE 2. Illustration of the spatial similarity relations in aerial imagery. The conventional CNNs could capture the short-range relations (blue-green and yellow-red), but have weak ability to model the long-range relations (blue-red and blue-yellow) directly.

The non-local block [19] is designed to capture global contextual information by utilizing self-attention mechanism. The self-attention mechanism computes the spatial relations between each pair pixels, and the spatial relations can be deemed as semantic similarity among pixels in an image as shown in Figure 2, additionally, the relations introduce global contextual information to neural networks directly [20]. And many improvements of non-local block have been proposed to decrease the computation costs, meanwhile increasing the performance [21]–[23]. However, all of the abovementioned models utilize a coarse 8× upsampling layer to produce the final predictions, which generates numerous misclassification pixels on the boundaries of small buildings as a result of the lack of detail information.

Therefore, it is necessary to build a network that can integrate the abovementioned networks' advantages meanwhile avoid their disadvantages to extract buildings accurately from various high-resolution aerial imagery.

Based on the above analysis, we propose a novel model named ENRU-Net to improve the classification precision of building extraction both on small and large buildings. Specifically, the ENRU-Net contains an U-shape encoder-decoder network as the backbone and an asymmetric pyramid non-local block (APNB) [22] is embedded between the backbone and final classifier. The backbone adopts ResNet-50 [3] as the downsampling path to improve the ability of feature extraction, and employs a simple yet effective upsampling path to expand the feature maps. And the APNB is an improved non-local block in which could capture global contextual information efficiently from high-resolution aerial images. This network restores the feature maps carefully by reusing the detail information from shallow layers, and utilizes the APNB to learn global contextual information as well.

The main highlights of our work can be summed up as follows:

- We propose a novel network, composed of an encoder-decoder structure and the asymmetric pyramid non-local block, for accurate building extraction from high-resolution aerial imagery. The proposed ENRU-Net could efficiently capture global contextual information and at same time sufficiently utilize the detailed information of buildings at various scales.

- We evaluate our ENRU-Net on two public aerial imagery building datasets: the Massachusetts Buildings Dataset [24] and the WHU Aerial Imagery dataset [9], and some state-of-the-art models are also evaluated as comparisons. The experiments demonstrate that the proposed ENRU-Net could achieve higher accuracies on both the two datasets at a relative higher efficiency compared with the established models.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Semantic segmentation is a long-standing foundation challenging task in computer vision, aiming to accurately predict a semantic label for each pixel in an image. With the development of deep learning in computer vision, CNN has been successfully applied to many computer vision tasks, one of which is semantic segmentation. Long *et al.* [7] first proposed the fully convolution network (FCN) that discards the fully connected layers in the last of VGG [25] and used the upsampling operation to restore the feature maps to the same size as the input images, FCN achieves the best result on the PASCAL VOC competition in 2015; In order to fuse features more comprehensively and reduce the loss of detail information that introduced by pooling layers, Ronneberger *et al.* [17] designed a simple yet effective encoder-decoder architecture network named U-Net, which not only obtains the same size output as input, but also concatenate the shallow and the corresponding deep feature maps as feature fusion to harvest more precise segmentation results; Meanwhile, SegNet [18] proposes a novel pooling method that the indices of positions in max pooling layer are saved and passed to decoder, the decoder expands feature maps by using the already saved pooled indices to boost the precise of upsampling operation. Another method to abate the effect of pooling layers is using dilated convolution layers to replace the pooling layers [26]. In addition, some works aggregate contextual information in multi-scales to enhance the ability to recognize multi-scale objects, PSPNet [27] utilizes pyramid pooling module to capture and fuse multi-scale features to increase segmentation accuracy; Deeplab v3 [28] employ the atrous convolution spatial pyramid pooling (ASPP) to aggregate multi-scale contextual information with a larger field of view than [26]; DenseASPP [29] uses densely connected ASPP to obtain multi-scale features that cover scale range densely with cover a larger scale range;

Recently, a self-attention mechanism named non-local block has shown great ability to capture the long-range relations, which increases the performance of various task. The non-local block is proposed by [19], and there are numerous works indicate that the non-local block is also effective in image semantic segmentation. OCNet [30] establishes a multi-scale non-local block to obtain the object contextual information, which finally exhibits robust segmentation performance; DANet [31] proposes two types of non-local blocks that model the semantic interdependencies on spatial and channel dimensions respectively to increase the precision of predictions; Meanwhile, CCNet [21] improves the computation method of self-attention module to obtain appropriate contextual information through a more effective and efficient way; Zhu *et al.* [22] utilizes pooling layers to cut down the costs when compute the relations matrix; A^2-Net [32] optimizes the computing process in the mathematical form to decrease the amount of computation; GCNet [23] analyzes the form of non-local method and combines it with squeeze excitation block [33].

B. BUILDING EXTRACTION

In the past few decades, many approaches have been proposed for building extraction that were based on extracting features through the carefully manual designed descriptors and conventional machine learning classifiers, e.g. support vector machine (SVM). Tuermer *et al.* [34] exploited the histogram of oriented gradients (HOG) feature descriptor for detection vehicles; Yuan and Newsam [35] adopted the scale-invariant feature transform (SIFT) to recognize objects in remote sensing images; Inglada [36] employed SVM to classify man-made objects in high-resolution remote sensing images. Although above methods achieve remarkable scores, these methods extremely rely on the manual designed features, which are always changed with datasets and labor-intensive. In conclusion, these methods are lack of robustness and could not handle various high-resolution aerial imagery effectively.

In the last few years, deep learning has shown incomparable advantages than classic machine learning in various fields, especially in computer vision. For building extraction, there are two prevalent methods for pixel-level classification on remote sensing images by using CNNs. One is training a CNN to classify each pixel a semantic category by inputting a small patch around this pixel which called patch-based method [24], [37], [38]. However, the patch-based method needs overlapping patches to predict each pixel, which causing redundant computations. Another is the pixel to pixel method that training a FCN classify each pixel directly, it has shown more effective and efficient than the patch-based methods without redundant computations. Li *et al.* [10] designed an encoder-decoder network by using dense block in [4] to reuse features excellently while reducing the number of parameters; Xu *et al.* [11] utilizes ResNet [3] with a guided filter to extract buildings from remote sensing images, the guided filter is adopted to refine the prediction map by CNN; Wang *et al.* [12] analyzes feature vector of each position by the entropy maps on a high-level feature maps to control the fusion of shallow and deep feature maps, which ensured the low-level detail information would be delivered to deep layers while producing few noises to the last segmentation results. Zhang *et al.* [13] designed a

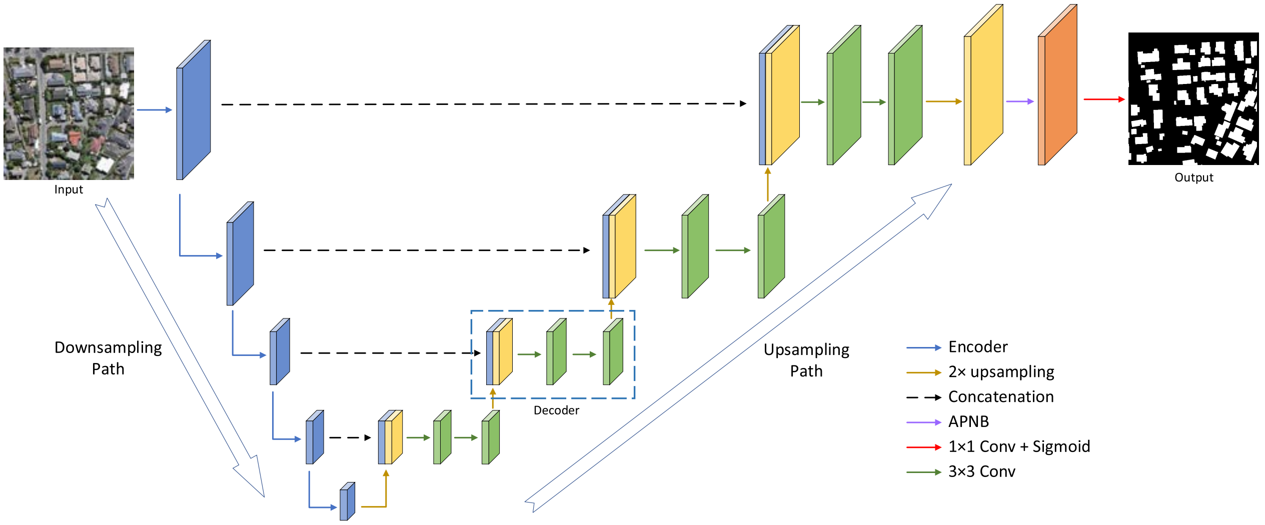


FIGURE 3. Overview of the proposed ENRU-Net. The downsampling path is ResNet-50, and the decoder in the upsampling path is composed of an upsampling layer and two successive 3 × 3 convolution layers. After the backbone, APNB is inserted before the final 1 × 1 convolution layer to capture global contextual information.

nested network architecture with dense hierarchical connection aimed to fuse different level feature maps to recover the structural information properly; Lu *et al.* [14] exploited a dual resolution network to improve the segmentation result of edge areas and large buildings by inputting a large view image and a corresponding small view image at the same time; Liu *et al.* [15] inserted PSPModule [27] into U-Net [17] to aggregate multi-scale contextual information; Mou *et al.* [20] utilized the self-attention mechanism to address the long-dependency issue in building extraction task from aerial imagery.

III. METHODOLOGY

In this section, we explicate the architecture of ENRU-Net, which is illustrated in Figure 3. The ENRU-Net combines an U-shape encoder-decoder network as backbone and a global contextual information computation module APNB. The backbone of ENRU-Net is a typical and widely-used network with some modifications for accurate and efficient building extraction from high-resolution imagery. In addition, the APNB is extracted from [22] to introduce global contextual information to improve the recognition of inner pixels in large buildings. The APNB is embedded between the backbone and the final 1 × 1 convolution layer.

A. BACKBONE OF ENRU-Net

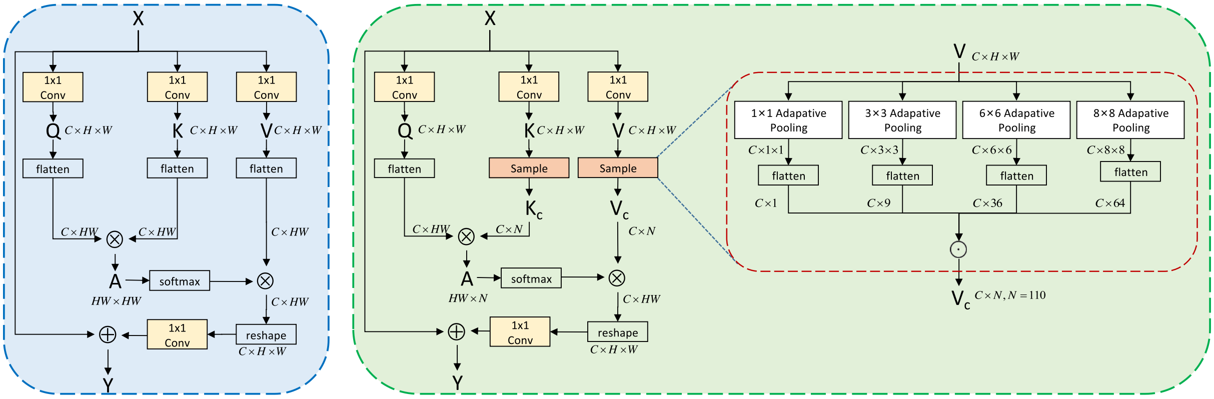
The backbone of our ENRU-Net is a widely-used encoder-decoder structure in previous works for semantic segmentation [10], [12], [15], [17], [18]. The structure could be divided into two parts: the downsampling path and the upsampling path. For the downsampling phase, the ResNet-50 is adopted as the backbone. The downsampling path contains several encoders to extract hierarchical features, and the upsampling path consists of cascaded decoders to stepwise reconstruct the feature maps as the same size as the input image. To carefully utilize the adequate detailed information in the output feature maps of each encoder in the downsampling path when reconstructing the feature maps, we adopt skip-connections to deliver the shallow feature maps to corresponding decoders in the upsampling path. This symmetrical designed structure significantly improves the ability to recognize the boundaries of small buildings when extracting buildings from high-resolution aerial images.

1) DOWNSAMPLING PATH

The downsampling path is beneficial to enlarge the field view so that the deep feature maps could contain rich context information, and it also reduces the computation and memory costs. Through the downsampling path, we can obtain hierarchical feature maps from different encoders, including the low-level feature maps from shallow encoders with detailed information, as well as high-level feature maps from deep encoders with semantic information. To extract features efficiently and effectively, we choose ResNet [3] as the downsampling path in ENRU-Net, which is a widely-used feature extractor among many previous works [12], [19], [27]. According to the sizes of the feature map, ResNet could be divided into five encoders, as shown in Figure 3. Each decoder first halves the size of the input feature map and outputs a new feature map through several convolution layers. Besides, the output of each encoder would be delivered to the corresponding decoder via skip-connections as the black dotted line in Figure 3. As a trade-off between the computation costs and the accuracy, we adopt ResNet-50 as the downsampling path in our implementation.

2) UPSAMPLING PATH

Since the building extraction could be considered as a semantic segmentation task, the upsampling path is employed to restore the output of the downsampling path to the same size as the original input progressively. The upsampling path



(a) (b)

FIGURE 4. Illustration of the original non-local block (a) and asymmetric pyramid non-local block (b), where ⊗, ⊙ and ⊕ denote matrix multiplication, concatenation, pixel-wise sum respectively, X, Q, K, V, K_c, V_c, Y are feature maps, A is the similarity relation matrix, and $N = 1^2 + 3^2 + 6^2 + 8^2 = 110$ in (b).

contains four decoders, and each decoder is a simple light yet capable module like the decoder in U-Net [17]. The blue dashed frame in Figure 3 reveals the structure of encoder, including an upsampling layer and two successive 3 × 3 convolution layers. To be specific, each decoder receives a couple of input features, one of them is a large feature map from a shallow layer with abundant detailed information, the other is a small feature map from the corresponding deep layer with sufficient semantic information. When receiving those two feature maps, the decoder first expands the deep feature map as the same size as the large feature by an upsampling layer. Next, the shallow feature and the expanded deep feature are concatenated on channel dimension as feature fusion. After that, two successive convolution layers is adopted to enhance the nonlinearity of our network. Since the ResNet-50 first uses 2× downsampling, we employ an upsampling layer again at the end of the upsampling path to obtain the same size feature map as the original input.

B. ASYMMETRIC PYRAMID NON-LOCAL BLOCK

The non-local block (NB) [19] could capture global contextual information effectively by computing the relations, i.e. the similarity relations between each pair pixels. However, the complexity of time and space of NB are both $O((H \times W) \times (H \times W))$, where H and W are height and width of a given feature map, see Wang *et al.* [19] for details. To model the relations efficiently, Zhu *et al.* [22] proposed APNB to decrease the computation costs and resource requirements. The structures of NB and APNB are shown in Figure 4. It could be observed from Figure 4 that both NB and APNB adopt matrix multiplication to model the spatial relations. The difference between them is that NB models the relations directly on pixel-level, while APNB first adopts four adaptive pooling layers with different scales to reduce the amount of the pixels in which participate the relation computation. Through these pooling layers, both time and space complexity of APNB are reduced to $O((H \times W) \times N)$, where N is far less than $H \times W$. And the comparison experiment in [22] also demonstrates that APNB outperforms NB both on precision and efficiency.

Specifically, for a given feature map **X**, APNB first feeds it into three convolution layers to reduce the number of channels and then generates three new feature maps named **Q**, **K** and **V**. Next **Q** is flattened. Meanwhile, **K** is fed into the parallel adaptive pooling layers to generate four multi-scale features, and the multi-scale features are also flattened and concatenated to synthesize a new feature $\mathbf{K}_c$. The same operations as **K** are conducted on **V** and produce $\mathbf{V}_c$. After that, APNB performs matrix multiplication between the flattened **Q** and $\mathbf{K}_c$ and a softmax layer is applied after the matrix multiplication, the result of the softmax layer is a huge matrix **A**, which contains the spatial relations, i.e. semantic similarity, between each pixel in **Q** and $\mathbf{V}_c$. Then matrix multiplication is performed again between **A** and $\mathbf{V}_c$ and the result is reshaped to reconstruct each pixel as a weighted sum of all pixels in the $\mathbf{V}_c$, and the spatial relations are weights. In the last, a convolution layer is employed to restore the amount of channels, and pixel-wise sum is conducted between the restored feature map and **X**.

Sufficient global contextual information is obtained via the spatial relations introduced by APNB, it could further improve the ability of our ENRU-Net to accurately classify the inner pixels in large buildings from high-resolution aerial imagery. The subsequent experiments confirm the impressive impact of global contextual information on fully convolution networks for building extraction.

IV. EXPERIMENT

In order to measure the effectiveness of ENRU-Net for building extraction from high-resolution aerial imagery, we conduct numerous experiments on two public datasets: the Massachusetts Buildings Dataset [24] and the WHU Aerial Imagery dataset [9]. And the performance of ENRU-Net is also compared with some state-of-the-art models in semantic segmentation, including FCN-8s [7], U-Net [17], SegNet [18] and Deeplab v3 [28]. All of the models are trained from scratch by using the same datasets and training strategy. And the models are evaluated based on three widely-used metrics: Overall Accuracy(OA), Intersection over Union(IoU) and F1-score(F1).

A. DATASET

1) MASSACHUSETTS BUILDINGS DATASET

Massachusetts Buildings Dataset is proposed by [24], including 155 aerial images of Boston, and the spatial resolution of the images is 1m. The size of each image is 1500 × 1500 and each image covers 2.25km^2 surface. The whole dataset was randomly divided into training set, validation set and testing set. An example image in testing set and the corresponding label are shown in Figure 7(a) and (b). Due to the limitation of GPU memory, we split the large original aerial imagery into small patches. Each image in the training set is randomly cropped into 80 small patches, and the size of each patch is 256 × 256, the same size patches are cropped in the form of sliding window on the original testing set. The number of images in the final cropped training set and testing set images are 10960 and 360 respectively. Figure 5(a) shows some images and corresponding ground truth in the cropped testing set.

2) WHU AERIAL IMAGERY DATASET

WHU Aerial Imagery Dataset is proposed by [9]. This dataset covers a surface of about 450 km^2 and more than 187,000 buildings with different sizes and appearances in Christchurch of New Zealand from the New Zealand Land Information Services website, the spatial resolution of the images is 0.3m, both the aerial image and the corresponding ground truth are provided. The dataset, containing 8189 RGB images of 512 × 512 pixels, is divided into training set, validation set and testing set, the number of the three sub-datasets are 4736, 1036 and 2416 respectively. Figure 5(b) shows some images and corresponding ground truth in the testing set.

B. IMPLEMENTATION DETAILS

The implementations of our ENRU-Net and other models are based on the deep learning framework PyTorch. We train all models for 100 epochs with a mini-batch size of 8 on 2× NVIDIA RTX 2080 Ti. We choose Adam [39] as the optimizer for converging quickly and the learning rate is initialized to 0.001. The learning rate schedule is poly strategy, it could be formulated as:

$$lr = 0.001 \times (1 - \frac{iter}{max\ iter})^{0.9}$$

The loss function of our experiments is binary cross entropy loss function. Moreover, some images would be flipped left to right or up to bottom randomly as data augmentation.

C. METRICS

For better measure the performance of our model, we adopt the Overall Accuracy (OA), Intersection over Union (IoU) and F1-score (F1) as the criteria, all of the metrics are widely-used in semantic segmentation and building extraction, which could be defined as follows:

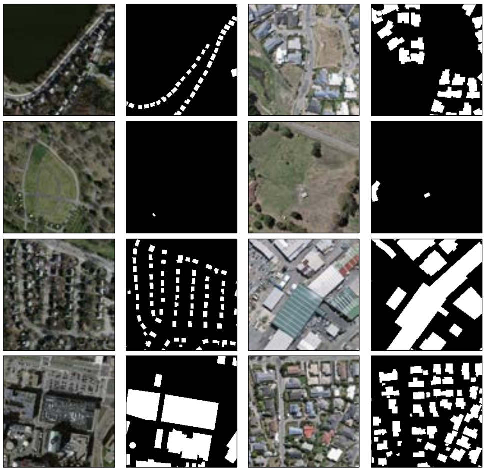


FIGURE 5. (a) Samples in the cropped testing set of Massachusetts Buildings Dataset. (b) Samples in the testing set of WHU Aerial Imagery Dataset.

$$\begin{aligned} OA &= \frac{TP + TN}{TP + TN + FP + FN} \\ IoU &= \frac{TP}{TP + FP + FN} \\ precise &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 \times precise \times recall}{precise + recall} \end{aligned}$$

where the TP, TN, FP, FN denote the true positive, true negative, false positive and false negative respectively.

D. EXPERIMENT RESULT

In this section, we have re-implemented some state-of-the-art semantic segmentation models as comparisons, including FCN-8s [7], U-Net [17], SegNet [18] and Deeplab v3 [28]. We also drop the APNB from ENRU-Net to evaluate the effect of APNB on ENRU-Net. The results are listed in Table 1 and Table 2.

1) COMPARISON EXPERIMENTS ON THE MASSACHUSETTS BUILDINGS DATASET

As summarized in Table1, the quantitative comparison demonstrates that our proposed ENRU-Net is more excellent than these established state-of-the-art semantic segmentation models on all the performance metrics. In the testing set of the Massachusetts dataset, ENRU-Net achieves 94.18%, 73.02%, 84.41% on OA, IoU and F1, which outstrips FCN-8s 0.81%, 3.55%, 2.43% respectively. As for the compared models, FCN-8s reaches 93.37%, 69.47%, 81.98% on OA, IoU and F1 respectively. In addition, the encoder-decoder models

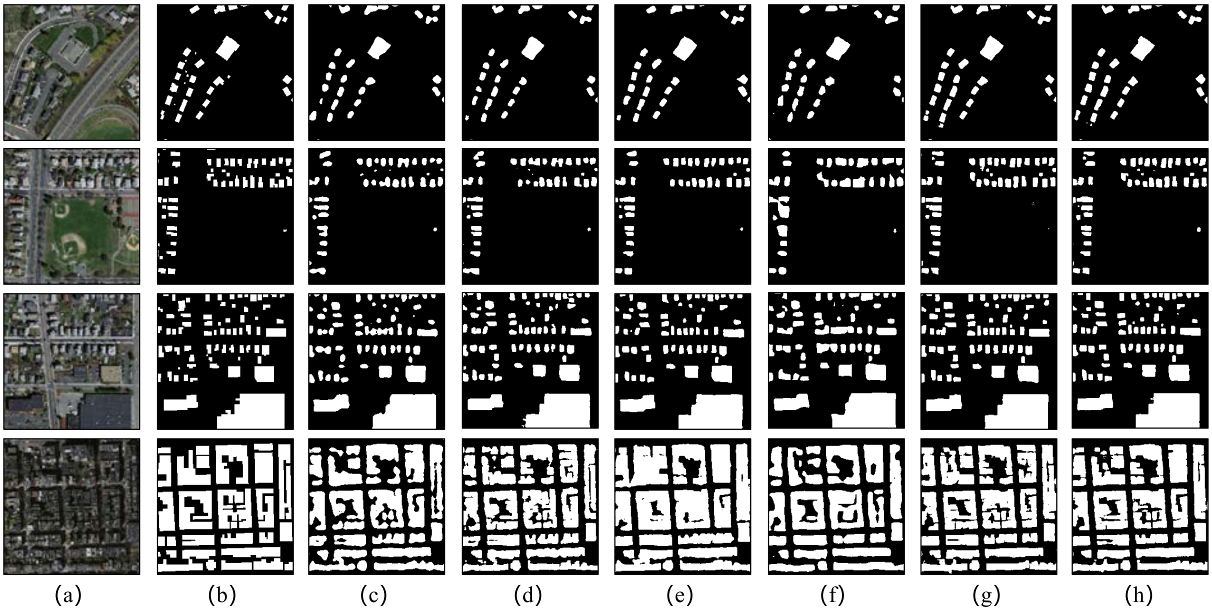


FIGURE 6. Some samples of predictions by ENRU-Net and other models from cropped testing set in Massachusetts Building Dataset. (a)Original Aerial Image. (b)Ground truth. (c)FCN-8s. (d)U-Net. (e)SegNet. (f)Deeplab v3. (g)ENRU-Net without APNB. (h)ENRU-Net.

TABLE 1. Comparison with state-of-the-art models on massachusetts buildings dataset.

Model	OA	IoU	F1
FCN-8s	93.37	69.47	81.98
SegNet	93.84	72.1	83.78
U-Net	93.63	69.97	82.14
Deeplab v3	93.01	68.55	81.34
ENRU-Net without APNB	94.12	72.77	84.24
ENRU-Net	**94.18**	**73.02**	**84.41**

TABLE 2. Comparison with state-of-the-art models on WHU aerial imagery dataset.

Model	OA	IoU	F1
FCN-8s	98.3	85.86	92.39
SegNet	98.53	87.33	93.23
U-Net	98.36	86.8	92.86
Deeplab v3	98.6	88.1	93.67
ENRU-Net without APNB	98.86	90.29	94.9
ENRU-Net	**98.92**	**90.77**	**95.16**

SegNet and U-Net accomplish more precise predictions, which over FCN-8s 1.8 % and 0.16% respectively on F1. Moreover, the SegNet obtains better performance than U-Net since SegNet saves the pooling indices. We also test our ENRU-Net without APNB. As a consequence of the excellent feature extractor and the appropriate feature fusion, the performance of the incomplete ENRU-Net also outstrips all of the established models but is lower than the complete one. However, due to the lack of detailed information that delivered by skip-connections, Deeplab v3 has a lower score than the others, which indicates that the detailed information is more crucial when extracting buildings from high-resolution aerial imagery than the semantic segmentation from natural imagery.

Figure 6 shows some randomly chosen samples from testing dataset and the corresponding prediction maps of these models. We can observe that due to the lack of detailed information and the coarse upsampling layer, FCN-8s and Deeplab v3 generate more misclassification on the boundaries of buildings, particularly on the small buildings. Compared with the two models, SegNet and U-Net have a better ability to extract the precise contours of small buildings as a consequence of their gradual upsampling, but the inner pixels in large buildings could not be adequately classified since they could not obtain sufficient global contextual information. While ENRU-Net has a more remarkable ability to extract buildings accurately both on small buildings and large buildings. As a consequence of the combination of the encoder-decoder network and the improved non-local block, the outlines of small buildings that predicted by ENRU-Net are sharper and more precise, and the pixels located in large buildings are also classified well. Figure 7 shows a whole prediction map in Massachusetts Buildings Dataset by assembling the cropped images.

2) COMPARISON EXPERIMENTS ON THE WHU AERIAL IMAGERY DATASET

The outcomes on the WHU Aerial Imagery Dataset are listed in Table 2. It can be observed that all models achieve higher scores on all of the metrics, even the simple FCN-8s could reach 98.3%, 85.86%, 92.39% on OA, IoU and F1 respectively, which profits by the lower image complexity, higher labeling precision and spatial resolution.

Due to the larger spatial resolution, Deeplab v3 could maintain sufficient detailed information even without

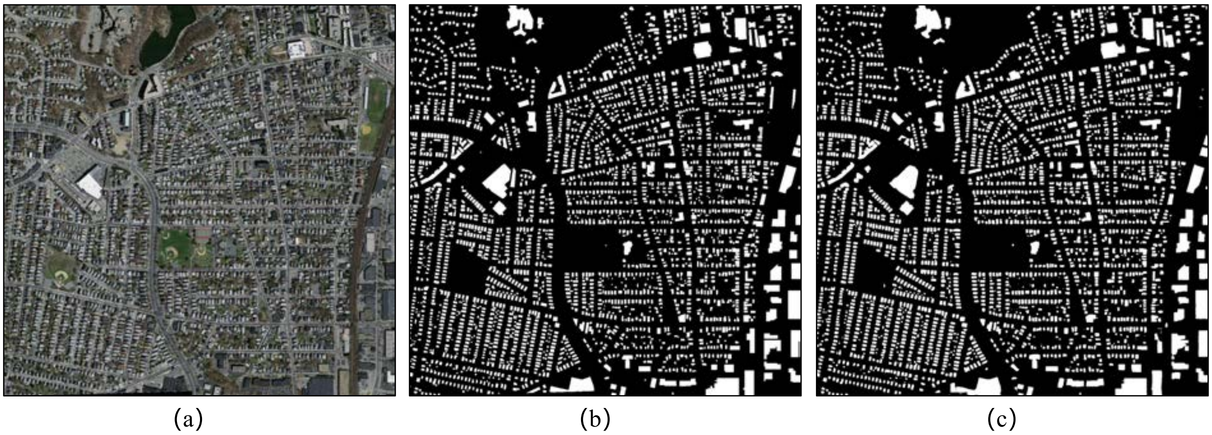


FIGURE 7. (a) An image in original testing set of Massachusetts Buildings Dataset. (b) The corresponding ground truth. (c) The binary predictions by ENRU-Net.

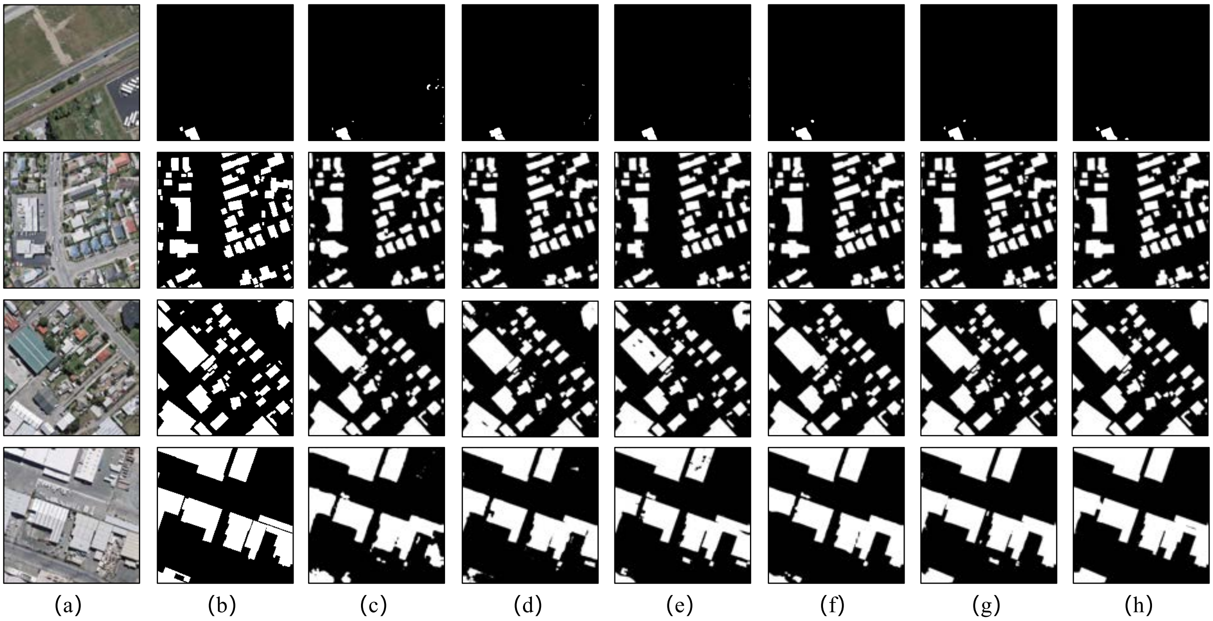


FIGURE 8. Some samples of predictions by ENRU-Net and other models from testing set in WHU Aerial Imagery Dataset. (a)Original Aerial Image. (b)Ground truth. (c)FCN-8s. (d)U-Net. (e)SegNet. (f)Deeplab v3. (g)ENRU-Net without APNB. (h)ENRU-Net.

skip-connections. Thus Deeplab v3 defeats all other established models benefits from the excellent feature extractor ResNet-101 [3], which is contrary to the Massachusetts Dataset. Deeplab v3 gains 1.28% higher scores than FCN-8s on F1. SegNet and U-Net also work well and respective achieve improvement of 0.84% and 0.47% on F1 against FCN-8s. Additionally, the ENRU-Net without APNB also performs well, which obtains further improvement when compared with the next best model Deeplab v3 on OA, IoU and F1. When compared with the aforementioned models, the proposed ENRU-Net shows the best ability for building extraction, where the OA, IoU and F1 is 98.92%, 90.77% and 95.16% respectively. The scores of ENRU-Net is 0.32%, 2.67% and 1.49% higher than the Deeplab v3, despite the deeper feature extractor ResNet-101 and the more complicated structure of Deeplab v3, e.g. the ASPP and dilated convolution layers. The results of Deeplab v3 and ENRU-Net also illustrate that the detailed information and global contextual information play a significant role in building extraction.

Figure 8 lists some randomly chosen prediction results in the testing set of WHU Dataset. It could be observed that all of the models could predict more accurate results than the Massachusetts Dataset. The edges of buildings predicted by FCN-8s are sharper than the Massachusetts Dataset, but the corners are still smooth. Besides, the encoder-decoder models, i.e. U-Net and SegNet, address this problem to a certain degree by using skip-connections and the upsampling path. Nevertheless, the lack of global information makes these models have poor ability to classify the inner pixels in large buildings accurately. Since Deeplab v3 adopts dilated convolution layers to expand the field of view, the ability to classify inner pixels in large scale buildings achieves a little

TABLE 3. **Comparison of FLOPs and parameters between ENRU-Net and other state-of-the-art models.**

Model	FLOPs(G)	Parameters(M)
FCN-8s	73.49	134.27
U-Net	16.59	31.06
SegNet	79.89	39.87
Deeplab v3	121.06	60.99
ENRU-Net	51.87	73.71

improvement. Additionally, owing to the excellent backbone and the employment of APNB, ENRU-Net outperforms all of the abovementioned models. It is obviously from Figure 8 that ENRU-Net predicts more precise and sharper contours, and the pixels that inside buildings, especially large scale buildings, are also recognized with higher accuracy.

E. MODEL EFFICIENCY

The complexity of a model also influences the practical applications. In deep learning, the complexity of networks could be measured by the computation overhead and parameter amount. In this section, we summarize these two indicators of our ENRU-Net and other state-of-the-art semantic segmentation models, the amount of computation consumption of all models are calculated on a 256×256 aerial image. The statistics results are listed in Table 3. It could be seen that U-Net has the smallest amount of floating point operations (FLOPs) and parameters because of its simple structure. FCN-8s and SegNet in which employ VGG [25] as backbone have similar computation costs. As a result of adopting ResNet-50 as feature extractor, ENRU-Net has less than half FLOPs compared with Deeplab v3. However, due to the extra parameters in the upsampling path, ENRU-Net has more parameters than Deeplab v3 but still fewer than FCN-8s.

The results in Tabel 3 indicate that ENRU-Net could extract buildings from high-resolution aerial imagery at a relative lower complexity when compared with the established models.

V. CONCLUSION

In this paper, we propose an efficient and effective model called ENRU-Net for extracting building from high-resolution aerial imagery. ENRU-Net adopts an U-shape encoder-decoder structure as the backbone to adequately utilize the detailed information to improve the recognition of small buildings. Meanwhile, to further decrease the misclassification of the inner pixels in large scale buildings, an improved non-local block named APNB is applied in ENRU-Net between the backbone and the final classifier. APNB could capture sufficient global contextual information via computing the spatial relations among each pair pixels. The significant contributions of this work are that it first analyzes the existing two key challenges in building extraction and then combines two outstanding structures to address these challenges.

To validate the effectiveness of the proposed ENRU-Net, we conduct several experiments on two public high-resolution aerial building imagery datasets: the Massachusetts Buildings Dataset and the WHU Aerial Imagery Dataset. Both on the two datasets, the proposed ENRU-Net achieves impressive scores, which proves ENRU-Net is robust for aerial imagery. Moreover, the quantitative comparison with other state-of-the-art models demonstrates that ENRU-Net outstrips the established segmentation models for building extraction task. In addition, the qualitative comparison also indicates that ENRU-Net has more accurate and sharper boundaries of small buildings, and the inner pixels in large buildings are classified more precisely.

Nevertheless, building extraction based on RGB aerial imagery do not make use of other type information, such as the digital surface model and multi-spectral information. How to utilize these extra information efficiently and effectively in deep learning models for building extraction needs to be further investigated in our future works.

REFERENCES

[1] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

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

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

[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

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

[8] Y. Li, B. He, T. Long, and X. Bai, "Evaluation the performance of fully convolutional networks for building extraction compared with shallow models," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 850–853.

[9] 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.

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

[11] 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, p. 144, Jan. 2018.

[12] H. Wang, Y. Wang, Q. Zhang, S. Xiang, and C. Pan, "Gated convolutional neural network for semantic segmentation in high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 446, 2017.

[13] 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, 2019.

[14] K. Lu, Y. Sun, and S.-H. Ong, "Dual-resolution u-net: Building extraction from aerial images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 489–494.

[15] 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.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," 2014, *arXiv:1412.6856*. [Online]. Available: https://arxiv.org/abs/1412.6856

[17] 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.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] 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.

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

[20] L. Mou, Y. Hua, and X. X. Zhu, ''A relation-augmented fully convolutional network for semantic segmentation in aerial scenes,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12416–12425.

[21] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, ''CCNet: Criss-cross attention for semantic segmentation,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 603–612.

[22] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, ''Asymmetric non-local neural networks for semantic segmentation,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 593–602.

[23] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, ''GCNet: Non-local networks meet squeeze-excitation networks and beyond,'' 2019, *arXiv:1904.11492*. [Online]. Available: https://arxiv.org/abs/1904.11492

[24] V. Mnih, *Machine Learning for Aerial Image Labeling*. Citeseer, 2013.

[25] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[26] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''Semantic image segmentation with deep convolutional nets and fully connected CRFs,'' in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015. [Online]. Available: http://arxiv.org/abs/1412.7062

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

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' *Comput. Vis. Pattern Recognit.*, vol. 22, no. 7, pp. 1182–1189, 2017.

[29] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, ''Denseaspp for semantic segmentation in street scenes,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.

[30] Y. Yuan and J. Wang, ''OCNet: Object context network for scene parsing,'' 2018, *arXiv:1809.00916*. [Online]. Available: https://arxiv.org/abs/1809.00916

[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, ''Dual attention network for scene segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[32] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, ''A^2-nets: Double attention networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 352–361.

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

[34] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, ''Airborne vehicle detection in dense urban areas using hog features and disparity maps,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.

[35] Y. Yang and S. Newsam, ''Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery,'' in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1852–1855.

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

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

[38] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, ''Convolutional neural networks for large-scale remote-sensing image classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[39] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

SHENGSHENG WANG received the B.S., M.S., and Ph.D. degrees in computer science from Jilin University, in 1997, 2000, and 2003, respectively. He is currently a Professor with the College of Computer Science and Technology, Jilin University. His current research interests are in the areas of computer vision, deep learning, and data mining.

XIAOWEI HOU received the B.S. degree from the College of Computer Science and Technology, Jilin University, in 2018, where he is currently pursuing the M.S degree. His current research interests include deep learning and remote sensing imagery semantic segmentation.

XIN ZHAO received the B.S. degree from the College of Computer Science and Technology, Jilin University, in 2016, where he is currently pursuing the Ph.D. degree. His current research interests include deep learning, transfer learning, and image processing.