


Article

BARNet: Boundary-Aware Refined Network for Automatic Building Extraction in Very High-Resolution Urban Aerial Images

Yuwei Jin ¹ , Wenbo Xu ^{1,2*}, Ce Zhang ^{3,4}, Xin Luo ¹, Haitao Jia ¹

¹ Yangtze Delta Region Institute (HuZhou) & School of Resource and Environment, University of Electronic Science and Technology of China, Huzhou 313001, P. R. China.; yuwei_jin@163.com (Y.J.); xuwenbo@uestc.edu.cn (W.X.); luoxin@uestc.edu.cn (X.L.); jhtao@uestc.edu.cn (H.J.)

² Center for Information Geoscience, University of Electronic Science and Technology, Qingshuihe Campus, Chengdu 611731, China; xuwenbo@uestc.edu.cn (W.X.)

³ Lancaster Environment Centre, Lancaster University, LA1 4YQ, U.K.; c.zhang9@lancaster.ac.uk (C.Z.)

⁴ UK Centre for Ecology & Hydrology, Lancaster, LA1 4AP, U.K.; c.zhang9@lancaster.ac.uk (C.Z.)

* Correspondence: xuwenbo@uestc.edu.cn

Received: date; Accepted: date; Published: date

Abstract: The convolutional neural networks (CNNs), such as U-Net, have shown competitive performance in automatic extraction of buildings from very high-resolution (VHR) aerial images. However, due to the unstable multi-scale context aggregation, the insufficient combination of multi-level features, and the lack of consideration about semantic boundary, most existing CNNs produce incomplete segmentation for large-scale buildings and result in predictions with huge uncertainty at building boundaries. This paper presents a novel network embedded a special boundary-aware loss, called Boundary-aware Refined Network (BARNet), to address the gap above. The unique properties of the proposed BARNet are the gated-attention refined fusion unit, the denser atrous spatial pyramid pooling module, and the boundary-aware loss. The performance of BARNet is tested on two popular data sets that include various urban scenes and diverse patterns of buildings. Experimental results demonstrate that the proposed method outperforms several state-of-the-art approaches in both visual interpretation and quantitative evaluations.

Keywords: VHR aerial images; building extraction; convolutional neural network; feature fusion; context aggregation; boundary

1. Introduction

Automatic building extraction from VHR aerial images has been a hot topic in the field of photogrammetry and remote sensing for decades. The end product is of paramount importance for various applications such as urban planning, regional administration [1,2], and disaster management [3]. However, the heterogeneous spectral and structural characteristics among different buildings coupled with highly complex urban scene pose enormous challenges to extract buildings precisely from VHR aerial images in an automatic fashion. Therefore, developing an advanced method for automated building extraction is essential and urgently needed.

The existing building extraction methods can be divided into three main categories based on different data sources, including optical image-based [4], non-optical image-based [5,6], and data fusion-based approaches [7,8]. Furthermore, in terms of the adopted algorithms, these methods can be categorized into two major groups: non-learning-based and learning-based.

For non-learning-based methods, buildings are extracted by: (1) thresholding buildings using specific characteristics such as spectral [9], shadow [10], and texture [11]; and (2) detecting building

edges [12]. For learning-based approaches, supervised classification methods, such as support vector machine (SVM) [13], are applied to acquire building extraction maps at pixel level or object level [14,15]. However, it is rather difficult for the conventional methods to realize really automatic building extraction, particularly when handling complex VHR aerial images, because the empirically designed features vary across different building structures, imaging conditions, and roof materials. Recently, with the rapid development of deep learning methods [16–18], a significant breakthrough has been made in mapping buildings using CNNs [19,20]. CNNs have the potential to address the challenge by closing the semantic gap between different semantic levels, and the feature representation can be learned autonomously from data itself in a hierarchy. Simultaneously, remote sensing has entered into a big data era, with massive amounts of aerial images being captured, providing the fuel for deep learning methods to learn for automatic building extraction. Thus, the latest research paid much attention to exploiting CNNs for automatic building. In general, under the support of massive training data with high-quality annotations, the performance of CNN-based approaches is superior to the other learning-based methods in terms of generalization and precision. As a result, CNN-based learning-based methods are widely utilized and represent an exciting program of research [21–25].

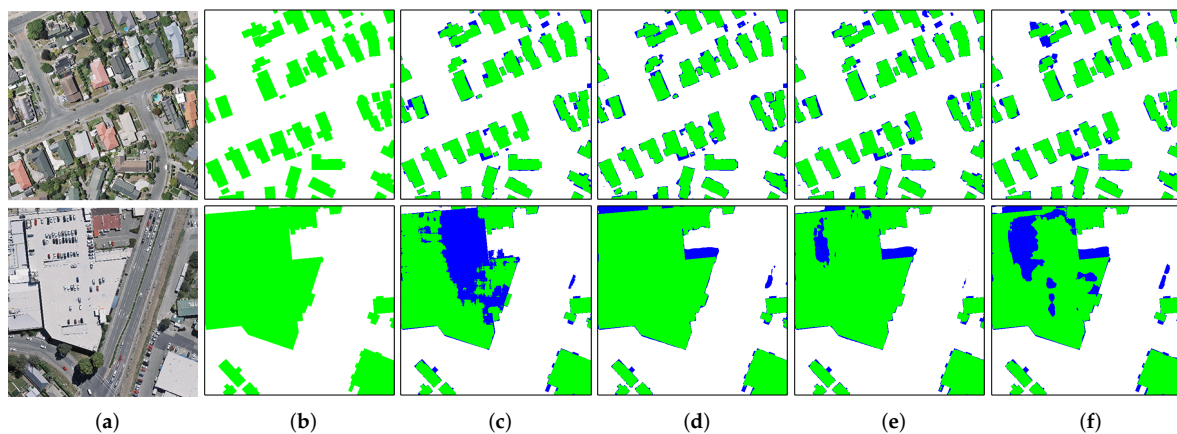


Figure 1. Examples of segmentation error map for several existing state-of-the-art methods performed on WHU aerial building dataset [21]. Column (a), original images. Column (b), reference labels. Columns (c–f), results obtained by U-Net, PSPNet, DeepLab-v3+ , and MA-FCN, respectively. In (b–f), green, white, and blue indicate building pixels, non-building pixels, and misidentified building pixels, respectively.

Building extraction aims to estimate a mask, where each pixel represents a specific category (i.e., building or non-building). Based on the fully convolutional network (FCN), previous works have achieved great success as reflected by numerous variants of FCNs, such as the encoder-decoder based U-Net [26] and SegNet [27]. In previous methods, the encoder-decoder structure [21,23] and some off-the-shelf context modeling approaches [22,25] were adopted to perform building extraction without considering the multi-scale problem and the boundary segmentation accuracy. There are some early attempts to tackle the multi-scale problem [28], they are not yet adequate to cope with the scale variations in buildings and often involve insufficient multi-scale context aggregation, which leads to incomplete segmentation, particularly for large-scale building extraction. In addition, each pixel is treated equally for a standard FCN. Boundary estimation is extremely challenging since the spatial details are lost during down-sampling process. As a consequence, the boundary accuracy of the building mask is limited. To support these statements, an instance of segmentation error map acquired by U-Net, PSPNet [29], DeepLab-v3+ [30], and MA-FCN [23] is demonstrated in Figure 1, from which all four state-of-the-art methods present errors, with missed holes in the large buildings and inconsistent boundary segmentation. Therefore, a novel method needs to be developed to address these issues mentioned above and further enhance the performance.

Based on U-Net and DeepLab-v3+, BARNet is proposed in this paper, with the aim to refine the building extraction with accuracy, particularly for those large buildings and boundary regions. Different from other studies, our BARNet method has a significant novelty to learn the boundary structure information in an end-to-end manner as the boundary-aware refined network. The performance of our method is compared with several state-of-the-art methods comprehensively through extensive experiments. The three major contributions of this work are summarized as follows:

- (1) we developed the gated-attention refined fusion unit (GARFU), which realized a better fusion of cross-level features in skip connection;
- (2) we proposed a denser atrous spatial pyramid pooling (DASPP) module to capture dense multi-scale building features; and
- (3) we designed a boundary-enhanced loss that allowed models to pay attention to boundary pixels.

The remainder of this article is organized as follows. Section 2 reviews the relevant works for this study. Section 3 detail presents the proposed method. Experiments and results are provided in Section 4. Section 5 gives some discussion of this work, and this paper is concluded in Section 6.

2. Related Work

In this section, we briefly review the relevant works of this study, i.e., CNNs for semantic segmentation in Section 2.1, multi-level feature fusion in Section 2.2, aggregation of multi-scale context in Section 2.3, and boundary refinement in Section 2.4.

2.1. CNNs for Semantic Segmentation

Following the pioneering study of FCN [31], numerous FCN schemes have been put forward in semantic segmentation domain, including dilated FCN, encoder-decoder, and multi-path. For dilated FCNs, the last convolution layers of the backbone network (e.g., ResNet [16]) were usually replaced with dilated convolutions for maintaining the resolution of feature maps, and the transposed convolution and bilinear interpolation were embedded after the backbone network as segmentation heads. The encoder-decoder structure derived from U-Net [26] and SegNet [27] is composed of two parts: encoder and decoder. Because the low-level features containing rich details can be reused through multiple skip connections, this scheme enables the network to better restore the spatial details. The multi-path architecture often has multi-paths such as Bilateral Segmentation Network (BiSeNet) [32], which has a spatial and a context path. The highlight of this structure is that it is capable of constructing a lightweight network.

2.2. Multi-level Feature Fusion

In general, there are rich spatial details such as edges in the high-resolution feature maps from shallow layers. By contrast, the abstract global representation is learned while the spatial details are lost with the successive convolution and down-sampling operations. Maintaining the resolution of feature maps is the key to semantic segmentation, while it is laborious for a standard CNN to balance the high resolution and abstract semantic representation. Accordingly, it is natural to combine the high-level features from the bottom layers and the high-resolution features from the top layers for exploiting the complementary features. FCN combined the low-level features and high-level features via element-wise addition operation. U-Net employed channel-wised concatenation in the lateral shortcuts between encoder and decoder to reuse the low-level features. Based on the structure of U-Net, Feature Pyramid Network (FPN) [28] integrated each same level low- and high-level features to make predictions. These fusion approaches fuse the different-level features directly without awareness of the usefulness of all features, limiting the propagation for beneficial features.

2.3. Aggregation of Multi-scale Context

The context is corresponding to the receptive field size of CNN. Small objects need small RF, and vice versa. Owing to the fixed RF size, it is rather complicated for a CNN to suit the objects with diverse sizes. Common attempts to handle this issue are to append a well-designed sub-module after the backbone network for modeling multi-scale context. PSPNet [29] introduced the pyramid pooling module (PPM) to capture the multi-scale information. Based on the dilated convolution, Deeplab series [30,33] proposed the atrous spatial pyramid pooling (ASPP) to obtain multi-scale context. DenseASPP [34] combined dense skip connection with ASPP, which effectively enlarges the receptive field size of network. Recently, inspired by the success of attention mechanism in natural language processing, the self-attention mechanism is also applied to aggregate the dense pixel-wise context [18,35–37]. The major drawback of self-attention is that it is with excessive computation and memory consumption.

2.4. Boundary Refinement

Due to the inevitable degradation of spatial details caused by the down-sampling process, the boundary accuracy of segmentation mask is usually limited for most existing CNN-based methods. To resolve this problem, painstaking efforts have been made. One way is to employ many post-processing operations with high computational costs such as the dense conditional random filed (DenseCRF) [38]. Another way relies on combining the boundary prior information and the extra sub-network that is responsible for detecting edges [39]. For example, Gated-SCNN [40] refined the boundary predictions through exploiting the duality between semantic segmentation and boundary segmentation with two branches and a regularizer. Evidently, this way increases the complexity and the parameter amount of the model. The others focus on exploiting online hard example mining strategy [29,37] and perceptual loss [41], which both require careful re-training or fine-tuning the hyper-parameters.

3. Methodology

In this section, the proposed method is presented in detail. We first overview the architecture of the proposed BARNet briefly. Then, the proposed gated-attention refined fusion unit, denser atrous spatial pyramid pooling module, boundary-aware loss, and training loss are elaborated.

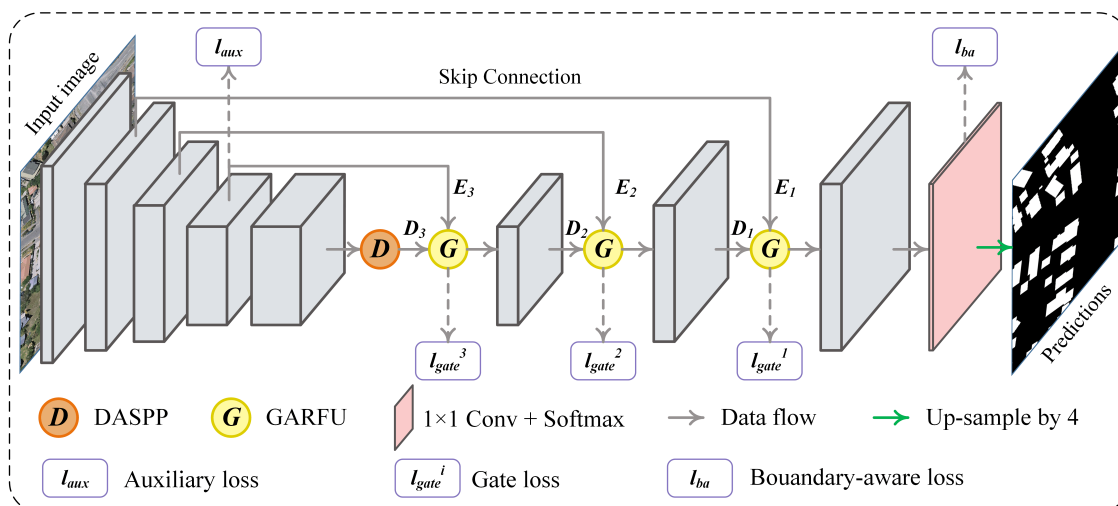


Figure 2. Overall structure of the proposed BARNet.

3.1. Model Overview

BARNet takes VHR aerial images as input and performs pixel-level building extraction in an end-to-end manner. As illustrated in Figure 2, BARNet is a standard encoder-decoder structure composed of three parts: encoder, context aggregation module, and decoder. ResNet-101 [16] is

adopted as the backbone network to encode the basic features of buildings. The fully connected layer and the last global average pool layer in ResNet-101 are removed. To retain more details in the final feature map, for the last stage of ResNet-101, all 3×3 convolutions are modified with dilated convolutions with dilated rates of $\{1, 2, 4\}$ and the stride in the down-sampling module is set to 1. After the encoder, DASPP is appended to capture the dense global context semantics. Before harvesting the low-level features into the decoder, each low-level feature map from the encoder is reduced to 256 channels with a 3×3 convolution layer followed by batch normalization (BN) and ReLU layers, for less computational cost. The reduced low-level feature map and the corresponding high-level feature map from the decoder are fused in GARFU. Then, the fused features are fed into the decoder block to restore the details. Each decoder block in BARNet is equipped with two cascaded Conv-BN-ReLU blocks, the same as U-Net and DeepLab-v3+. At the end of BARNet, a 1×1 convolution layer and a softmax layer are applied to output the final predictions. To get the same size as the original input, the final predictions are further up-sampled by 4 times using bilinear interpolation.

3.2. Gated-Attention Refined Fusion Unit

The significant highlight of U-Net and FPN is integrating the cross-level features via simple concatenation and addition operations. Given a low-level feature map $E_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ and a high-level feature map $D_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, in which i , C , H , and W denote the level order, number of channels, height and width of feature map, respectively, the two basic fusion strategies can be formulated as

$$\text{Concatenation: } F_i = \text{concat}(E_i, \text{Upsample}(D_i)), \quad (1)$$

$$\text{Addition: } F_i = E_i + \text{Upsample}(D_i), \quad (2)$$

where $\text{concat}(\cdot)$ denotes concatenation operation, $\text{Upsample}(\cdot)$ means up-sampling operation, and F_i represents the fused feature map. It can be evidently observed from the above two equations that all feature maps are fused directly without considering the contribution of each feature. As described earlier in Section 2.2, different level features are complementary, which is beneficial for building extraction. However, to our best knowledge, the informative features in a feature map are mixed with massive redundant information [42]. As a result, the cross-level features should be re-calibrated before combining them for the best exploitation of beneficial features. To achieve this goal, we developed GARFU. As presented in Figure 2, GARFU is embedded at each shortcut, which only increases a small extra computational cost.

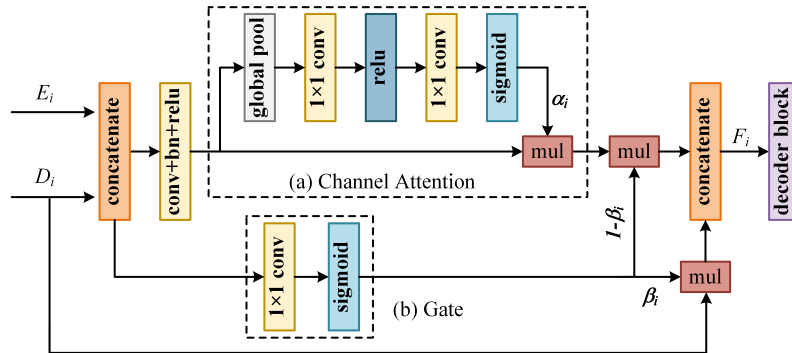


Figure 3. Structure of the proposed gated-attention refined fusion unit.

GARFU consists of two major components: channel-attention module and spatial gate module. As displayed in Figure 3, the low-level feature map E_i is first re-calibrated by a channel-wise attention vector $\alpha_i \in \mathbb{R}^{C_i \times 1 \times 1}$. Then, E_i' and D_i are re-calibrated by multiplication a gated map β_i and $(1 - \beta_i)$, respectively, in which $\beta_i \in [0, 1]^{1 \times H_i \times W_i}$. The mechanism of GARFU can be defined as

$$E_i' = f_S(f_C(E_i, \alpha_i), (1 - \beta_i)), \quad (3)$$

$$D'_i = f_S(D_i, \beta_i), \quad (4)$$

$$E_i = \text{concat}(E'_i, D_i), \quad (5)$$

where $f_C(\cdot)$ represents the multiplication in channel axis, and $f_S(\cdot)$ represents the multiplication on spatial dimension.

(1) Channel attention: The latest works have demonstrated the effectiveness of modeling the contribution of each channel using channel attention mechanism [1,43]. Therefore, we adopt the channel attention module reported in SENet [43] to exploit channel-wise useful features in low-level feature map, because the low-level features usually carry more noises. The attention vector α_i is generated based on the concatenated result of D_i and E_i . Before concatenating them, D_i is first up-sampled with bilinear interpolation to make D_i have the same spatial shape with E_i . Then, the concatenated features are passed through a 1×1 convolution layer for less parameters. Let $x = \text{GAP}(c)$, where c is concatenated features, and $\text{GAP}(\cdot)$ is the channel-wise global average pooling operation. The attention vector α_i is obtained by

$$\alpha_i = \sigma_2(W_2(\sigma_1(W_1(x)))) , \quad (6)$$

in which $W_1 \in \mathbb{R}^{\frac{C_i}{2} \times 1 \times 1}$ and $W_2 \in \mathbb{R}^{C_i \times 1 \times 1}$ are two linear transformations, σ_1 denotes relu activation function, and σ_2 is sigmoid activation. α_i is multiplied with E_i to enable the network to learn the salient channels that contribute to distinguish building.

(2) Gate: Many studies [26,27,30] suggest that introducing the low-level information can improve the accuracy of predictions on boundary and details, whereas lacking global semantic may lead to confusions in other regions due to limited local receptive field size. On the other hand, there exists a semantic gap for low- and high-level features, that is, not all features benefit for building extraction. With this motivation, we adopt the gate mechanism to generate a gate map β_i , which serves as a guide to enhance the informative regions and suppress the useless regions both in low- and level features. Gates are widely used in deep neural network to control the information propagation [44]. For example, the GRU in the LSTM network is a typical gate [45]. In this work, the gate β_i is generated by

$$\beta_i = \gamma \sigma(W(c_i)) \quad (7)$$

where W is a linear transform parameterized with $\mathbb{R}^{C_i \times 1 \times 1}$, σ denotes sigmoid activation for normalizing the value into $[0, 1]$, and γ is a trainable scale factor to prevent the minima occurs during the initial training. The gate map is learned under the supervision of ground-truth during training, and the pixel value in it measures the degree of importance for each pixel. The feature at position (x, y) would be highlighted when the value of $g_i(x, y)$ is large, and vice versa. In this manner, the useless information is suppressed and only useful features can be harvested to the following decoder block, thus obtaining better cross-level feature fusion. Different from the self-attention mechanism, the gate is learned with the explicit supervision of ground-truth.

3.3. Denser Atrous Spatial Pyramid Pooling

As is well known, building scale variance frequently occurs in complex urban scenes, resulting in non-unified extraction scales. Thus, an ideal context modeling unit should capture the dense multi-scale features as much as possible. To achieve this goal, a new ASPP module is developed. As it is inspired by ASPP and the main idea is to capture the denser image pyramid feature, we name it denser atrous spatial pyramid pooling. As illustrated in Figure 4, the DASPP module consists of a skip connection, a cascaded atrous spatial pyramid block (CASPB), and a global context aggregation block. The skip pathway is only composed of a simple 1×1 convolution layer, aiming at reusing the high-level feature and accelerating network convergence [16]. In CASPB, we cascade the hybrid multiple dilated 3×3 depthwise separable convolution [46] layers with different dilation rates and connect them with dense connections. Here, the depthwise separable convolution is utilized for

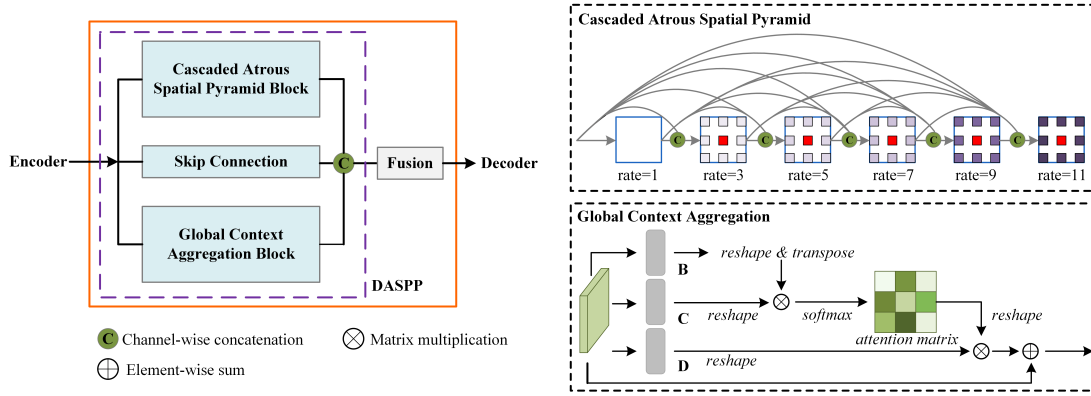


Figure 4. Architecture of the proposed denser atrous spatial pyramid pooling module.

reducing the parameters of DASPP, and the negative effect is almost negligible. CASPB can be formulated as $L_i = Conv_{i,d_i}(concat(L_0, L_1, \dots, L_{i-1}))$, where d_i represents the dilation rate of the i th layer L , $Conv(\cdot)$ means convolution operation, and $concat(\cdot)$ denotes the concatenation operation. In this study, $d = \{1, 3, 5, 7, 9, 11\}$. Compared to ASPP, this change brings us two main benefits: a denser feature pyramid and a larger receptive field. The sequence of receptive field size in the original ASPP is 13, 25, and 37, respectively, when the output stride of encoder is 16. However, the max receptive field of each layer in CASPB is 3, 8, 19, 36, 55, and 78, which is denser and larger than ASPP. This means CASPB is more robust with the building scale variations. In addition, the position-attention module of DANet [36] is also introduced to replace the image pooling branch in ASPP to generate denser pixel-wise global context representation. Unlike the global average pooling used in PPM and ASPP, the self-attention can generate global representation and capture the long-range dependence between each pixel. As presented in Figure 4, the position attention module re-weights each pixel according to the degree of correlation between any pixels.

3.4. Boundary-aware Loss

Although the re-calibrated low-level features contribute to refining the segmentation results [30], it is not still sufficient enough to locate accurate building boundaries. As mentioned earlier, the commonly used per-pixel cross-entropy loss treats each pixel equally. In fact, depicting boundaries is more challenging than locating semantic bodies because of the inevitable spatial details degradation. Consequently, an individual loss should be applied to force the model to pay more attention to boundary pixels explicitly. The key here is how to decouple the building edges from the final predicted maps. If the corresponding boundary maps are obtained, we could use the binary cross-entropy loss to reinforce the boundary prediction. Herein, the Laplacian operator, defined by Equation 8, is applied both on the final prediction maps and ground-truths to produce the boundary predictions and corresponding boundary labels.

$$f = \frac{\partial^2 f}{\partial^2 x} + \frac{\partial^2 f}{\partial^2 y}, \quad (8)$$

where f is the an 2-D gray-scale image, and x, y are the two coordinate directions of f . The output of Equation 8 is a termed gradient information map, where a higher value stands for that the probability of a pixel locates at the boundary, and vice versa. We extend the Laplacian operator to process the multidimensional tensor. An instance of using the Laplacian operator to obtain building boundary is given in Figure 5. With the yielded boundary maps $\tilde{B} \in \mathbb{R}^{N \times 1 \times H \times W}$ and boundary labels $B \in \mathbb{R}^{N \times 1 \times H \times W}$, where N, H , and W are batch size, image height, and image width, respectively, the boundary refinement is defined as a cost minimum problem expressed by

$$\tilde{\theta} = \operatorname{argmin}(\tilde{B}, B|\theta), \quad (9)$$

where θ denotes the trainable parameters of BARNet. For every single image, the weighted binary cross-entropy loss [26] is employed to compute the boundary-enhanced loss \mathcal{L}_{be} as:

$$\mathcal{L}_{be} = -\frac{1}{H \times W} \sum_i^{H \times W} \left[\frac{1}{Z^+} B_i \log(\tilde{B}_i) + \frac{1}{Z^-} (1 - B_i) \log(1 - \tilde{B}_i) \right], \quad (10)$$

where Z^+ and Z^- represent the number of pixels in boundary and non-boundary regions, respectively.

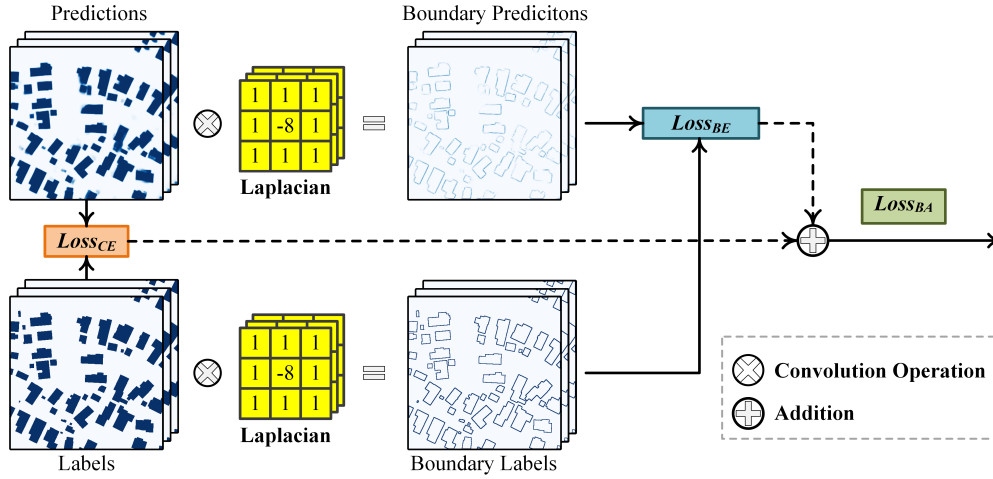


Figure 5. Principle of the proposed boundary-aware loss.

The final boundary-aware (BA) loss \mathcal{L}_{ba} is defined as the addition of two losses, i.e.,

$$\mathcal{L}_{ba} = \lambda_{be} \mathcal{L}_{be} + \mathcal{L}_{ce}, \quad (11)$$

$$\mathcal{L}_{ce} = -\sum_{i=0}^1 w_i \tilde{y}_i \log(y_i), \quad (12)$$

in which λ_{be} is empirically set to 1 to balance the contribution of boundary-enhanced loss, w_i is the class weight calculated using the median frequency balance strategy [1], \tilde{y}_i and y_i denote the model predictions and corresponding labels, respectively.

3.5. Training Loss

The cross-entropy loss expressed in Equation 12 is utilized to supervise each learned gate. It should be noted that each gate is up-sampled to the same size with ground-truth for computing loss. In addition, in order to facilitate the training process, an auxiliary loss with a weight of 0.4 is set on the output feature map at the third stage of ResNet-101 [29]. Thus, the final total loss of our network is

$$\mathcal{L} = \sum_{i=1}^3 \lambda_i \mathcal{L}_{gate}^i + 0.4 \times \mathcal{L}_{aux} + \mathcal{L}_{ba} \quad (13)$$

where $\lambda_i \in \{0.8, 0.6, 0.4\}$ denotes the balanced weight parameter for different gate loss. Following the works [29,36], the online hard example mining (OHEM) strategy [29] is adopted to compute BA loss \mathcal{L}_{ba} during training, to boost the performance of BARNet.

4. Experiments and Results

4.1. Datasets

Two standard open-source VHR aerial datasets were used to verify the effectiveness of the proposed method. All the images are collected in the complex urban scene by airborne sensors and are with very-high resolution. Two close-ups of the datasets are shown in Figure 6.



Figure 6. Examples of the images and corresponding labels for the employed two datasets. (a) and (b) represent the WHU dataset and Potsdam dataset, respectively. The white regions in the two reference maps stand for buildings.

WHU Aerial Building Dataset (WHU) [21]: This dataset consists of more than 22000 independent buildings extracted from the aerial images with 0.075m spatial resolution and 450 km² covering in Christchurch, New Zealand. The structure and roof materials of these buildings vary in different locations, ranging from low-raised urban residential settlements to homogeneous industry areas. In previous literature, this dataset is the most popular benchmark for building extraction. Due to the size of original images with 0.075m ground resolution is very large, the organizer down-sampled them to 0.3m ground resolution and seamlessly cropped them into 8189 tiles with 512×512 pixels. The dataset is officially divided into three parts: 4736 tiles for training, 1036 tiles for validation, and 2416 tiles for testing.

ISPRS Potsdam Dataset (Potsdam): This dataset, consisting of 38 true orthophotos (TOP) aerial images, is provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) and is widely used to evaluate the algorithm for urban remote sensing semantic labeling. The size of each image is 6000×6000 pixels. Compared to the WHU dataset, it is more challenging due to the finer spatial resolution of 0.05m. The dense residual buildings with different shapes and roof materials are dominated in this dataset, making it hard to accurately separate buildings from background objects. Following the official suggestion, 14 images were set as the test set, and the remainder 24 images were randomly split into 7 images for validating and 17 images for training.

4.2. Experimental Settings

Experimental Configurations: The proposed BARNet was implemented based on Pytorch-1.6 framework in Ubuntu 20.04 environment. All experiments were conducted on an Nvidia GeForce RTX 2080ti GPU with 11GB RAM. The shape of the original input data for a batch is 3×512×512.

Training Settings: The encoder was initialized with the weight of ResNet-101 trained on ImageNet [47], and the rest was initialized using Kaiming uniform [48]. The Adam algorithm [49], where the initial learning rate is set to 0.0002 and the weight decay is set to 0.0005, was selected to optimize the network. The warmup strategy [50] with a base learning rate of 5e-8 and exponential weight-decay strategy with a decay rate $\gamma = 0.9$ was employed to adjust the learning rate for each epoch. We set the warmup period to 5 epochs. During the warmup period, the learning rate was linearly increased. The number of epochs was set to 100 for WHU and Potsdam dataset. We set the batch size to 8 to make full use of the GPU memory. Note, all experiments were done using mixed-precision training [51].

Dataset Settings: For Potsdam dataset, the images used for training and validation were cropped into 512×512 pixels without overlap, and the tested images were cropped into 512×512 pixels with an overlap of 128 pixels. To avoid the risk of over-fitting, some commonly used data augmentation approaches, including random horizontal-vertical flipping, random cropping with the crop size of 512×512 , random scaling within the range of $\{0.75, 1.0, 1.25, 1.5, 1.75\}$, and random Gaussian smooth, were applied on each training image.

Test Settings: Following the works [29,30,36,37], the multi-scale (MS) inference strategy was employed. When using multi-scale inference, the final results were generated by averaging the all predictions with scales $\{0.75, 1.0, 1.25, 1.50, 1.75\}$.

4.3. Comparison of State-of-the-art Studies

We evaluated the proposed method on the WHU and Potsdam dataset. To reveal whether the proposed method gains an advantage over other recent state-of-the-art (SOTA) studies, several remarkable CNNs for semantic segmentation and building extraction were chosen as comparative methods, namely, U-Net [26], DeepLab-v3+ [30], and DANet [36]. Moreover, MA-FCN [23], which achieved a very high IoU score of 90.7% on WHU dataset [23], was also chosen to verify the superiority of our method. Among these methods, DANet is a representative dilated FCN, where the self-attention mechanism was applied to aggregate the holistic context. The robustness of DeepLab-v3+ equipped with a strong decoder head has been proved in previous studies. MA-FCN, one of the variants of U-Net, focuses on the effects of building scale. Except the mini-batch size and the input size were changed to suit the GPU memory used in this study, all the comparative methods were reproduced using the default settings given by the authors.

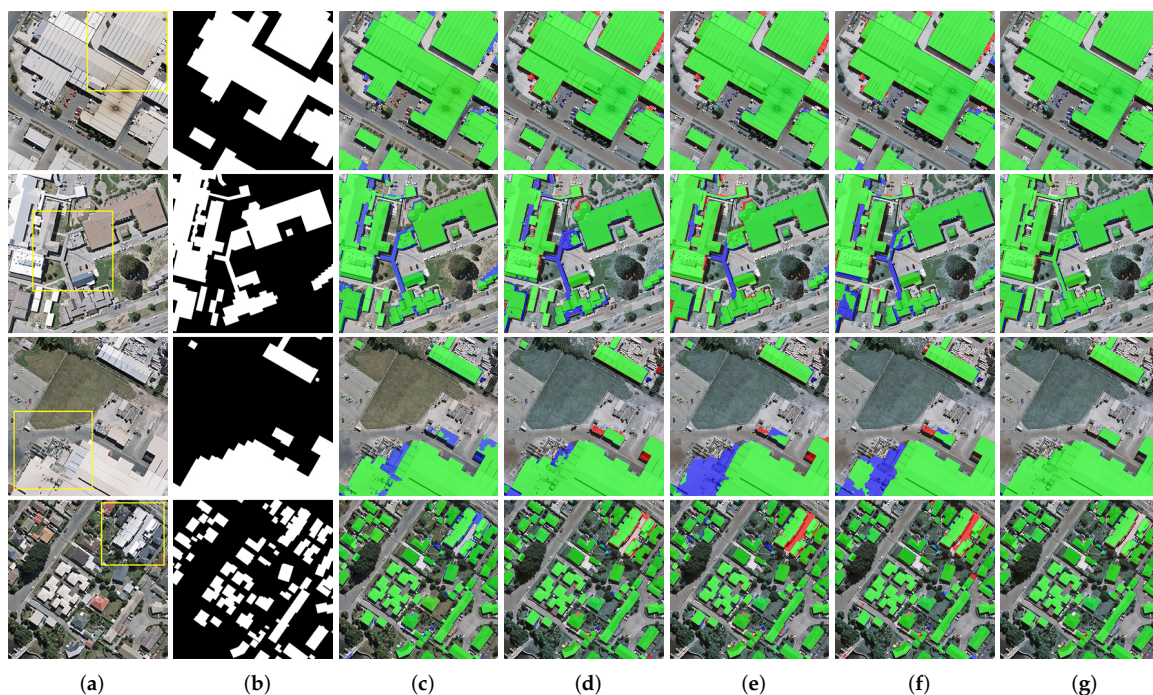


Figure 7. Examples of building extraction results obtained by different methods on the WHU dataset. (a) Original image. (b) Ground-truth. (c) U-Net. (d) DeepLab-v3+. (e) DANet. (f) MA-FCN. (g) BARNet. Note, in columns (c)–(g), green, blue, and red indicate true-positive, false-positive, and false-negative, respectively. The yellow rectangles in (a) are the selected regions for close-up inspection in Figure 8.

4.3.1. Visualization Results

The comparisons with different models are elaborated as follows:

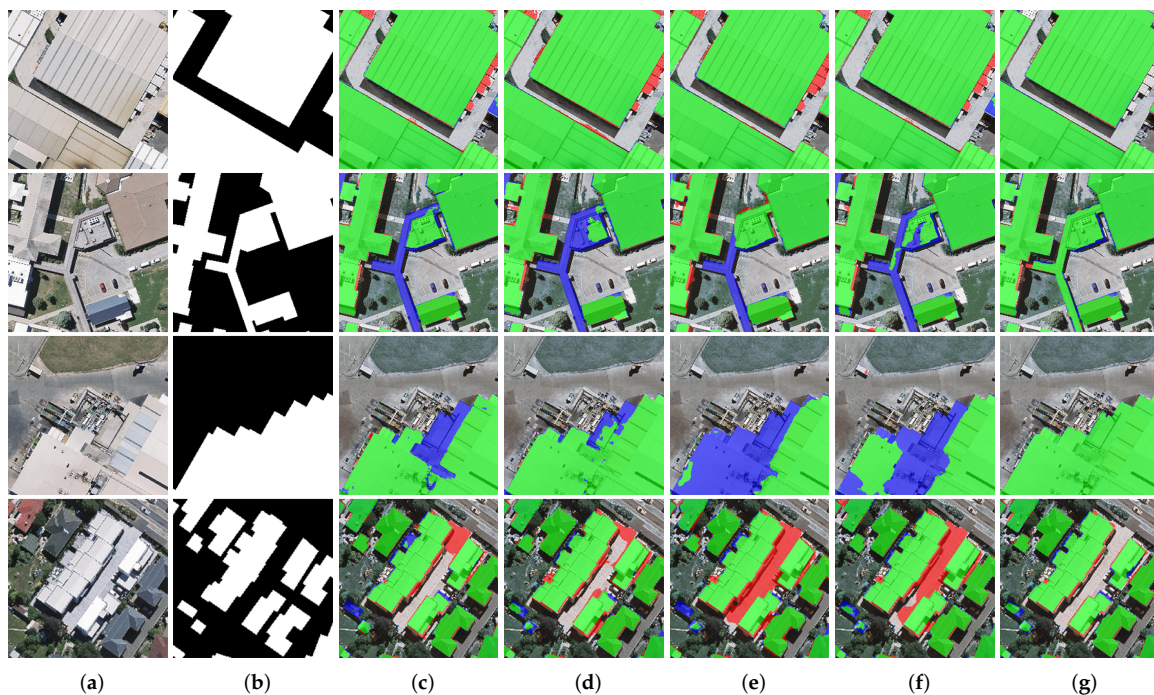


Figure 8. Close-up views of the results obtained by different methods on the WHU dataset. Images and results shown in (a)–(g) are the subset from the selected regions marked in Figure 7a. (a) Original image. (b) Ground-truth. (c) U-Net. (d) DeepLab-v3+. (e) DANet. (f) MA-FCN. (g) BARNet.

(1) WHU dataset. The results produced by different methods on the WHU dataset are illustrated in Figure 7. Visually, the proposed BARNet obtained the best global extraction results compared with other SOTA methods. As displayed in the first row of Figure 7a, the reasonable performance was achieved by U-Net, DeepLab-v3+, DANet, and MA-FCN in a simple scene. However, with the increase in complexity and structure for buildings, a dramatically decreased performance was clearly observed in the second and third rows of Figure 7, where parts of buildings were missed, implying that they have difficulty in accurately recognizing the buildings with irregular structure and large scale. This phenomenon seriously affects the visualizations. Conversely, almost all buildings were identified correctly and completely by BARNet, and there are only a few errors according to the results presented in Figure 7g. This is mainly because our model can better aggregate contextual information. Comparing U-Net, PSPNet, and MA-FCN, DeepLab-v3+ could maintain the completeness of final predictions for relatively large buildings to some extent, whereas it also failed to handle the large buildings with complex shapes. A striking illustration for this can be seen in the third row of Figure 7d, where problems like high missed rate occurred. Even though MA-FCN is the improved version of U-Net, its performance is similar to U-Net, sensitive to the variation in building scale and structure. To better clarify the detailed inspection, the close-ups of the selected regions (as marked in yellow rectangles in Figure 7) in tested images are displayed in Figure 8. From the close-up views, we can observe that these SOTA methods exhibited limited ability to separate the confusing non-building objects adjacent to buildings, yielding inaccurate boundary predictions. Nevertheless, owing to reinforce the boundary and refine the multi-level feature fusion, our method performed well by generating only a small number of misclassified pixels in boundary regions.

(2) Potsdam Dataset. Figure 9 provides the experimental results on the Potsdam dataset for different methods. According to the results displayed in Figure 9, BARNet always obtained the most consistent results with the reference building maps visually. Compared with other methods, our model is robust to cope with buildings in different complex scenes. In contrast, the performance of the other methods looks unstable. A striking illustration of close-up views for different methods is given in

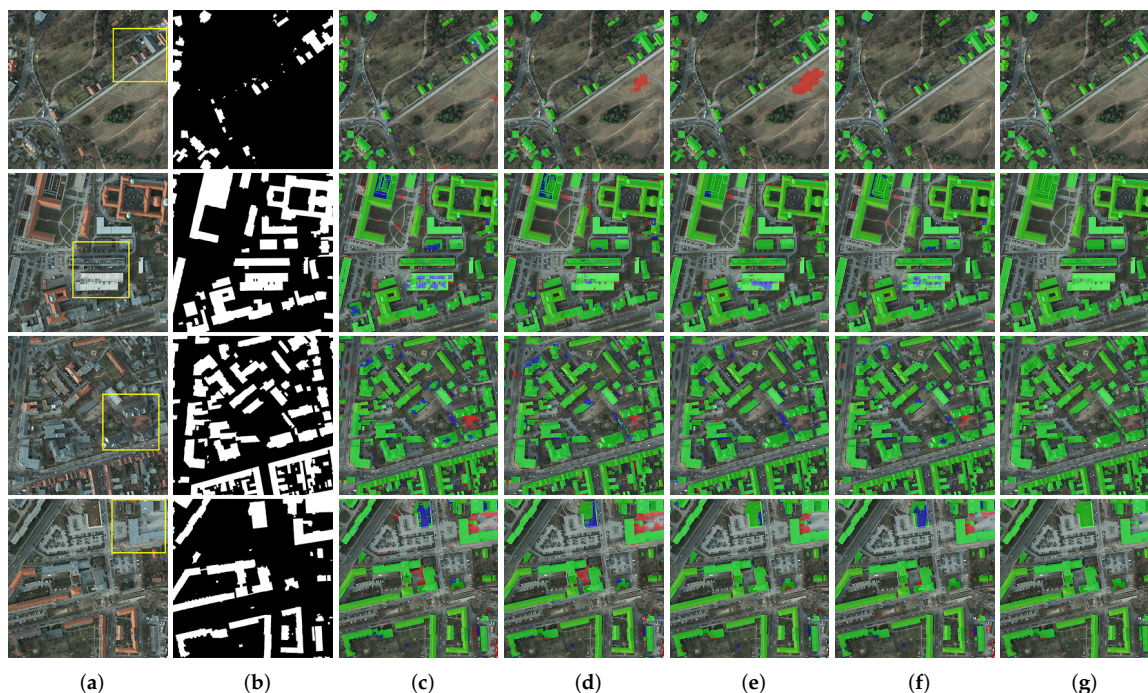


Figure 9. Examples of building extraction results obtained by different methods on the Potsdam dataset. (a) Original image. (b) Ground-truth. (c) U-Net. (d) DeepLab-v3+. (e) DANet. (f) MA-FCN. (g) BARNet. Note, in columns (c)–(g), green, blue, and red indicate true-positive, false-positive, and false-negative, respectively. The yellow rectangles in (a) are the selected regions for close-up inspection in Figure 10.

Figure 10, from which we can see that there are only a few errors in Figure 10g. Among the SOTA methods, U-Net produced building extraction results with incomplete predictions, suggesting that it is with a poor ability to aggregate multi-scale contextual information. Meanwhile, the boundary location is not precise enough. Similar behaviors are observed in Figure 9 and Figure 10 for DANet and DeepLab-v3+. Although DANet and DeepLab-v3+ are both equipped with a strong context modeling module, they still suffer incomplete detection for large buildings. Another prominent problem for DANet and DeepLab-v3+ is that there exists evident grid effect (see column (d) and column (e) in Figure 9) resulted from the atrous convolution, bringing unstable performance. Benefiting from the attention on the aggregation of multi-scale representation, MA-FCN could extract most of the buildings stably, but its performance is not as good enough. It can be clearly observed in Figure 10f, some buildings were not detected completely, and many non-building pixels are misclassified as buildings. For boundary refinement, as illustrated in the last row of Figure 10, the other four methods failed to accurately separate the cement square from the buildings that have similar characteristics to it, resulting in over-extracting.

According to the above analysis, we can conclude that the improvements of our method lie in recognizing the multi-scale buildings, especially the buildings with large scale, and locating building boundaries more accurately among different scenes, which demonstrates the effectiveness of the proposed method for automatic building extraction in urban VHR aerial images.

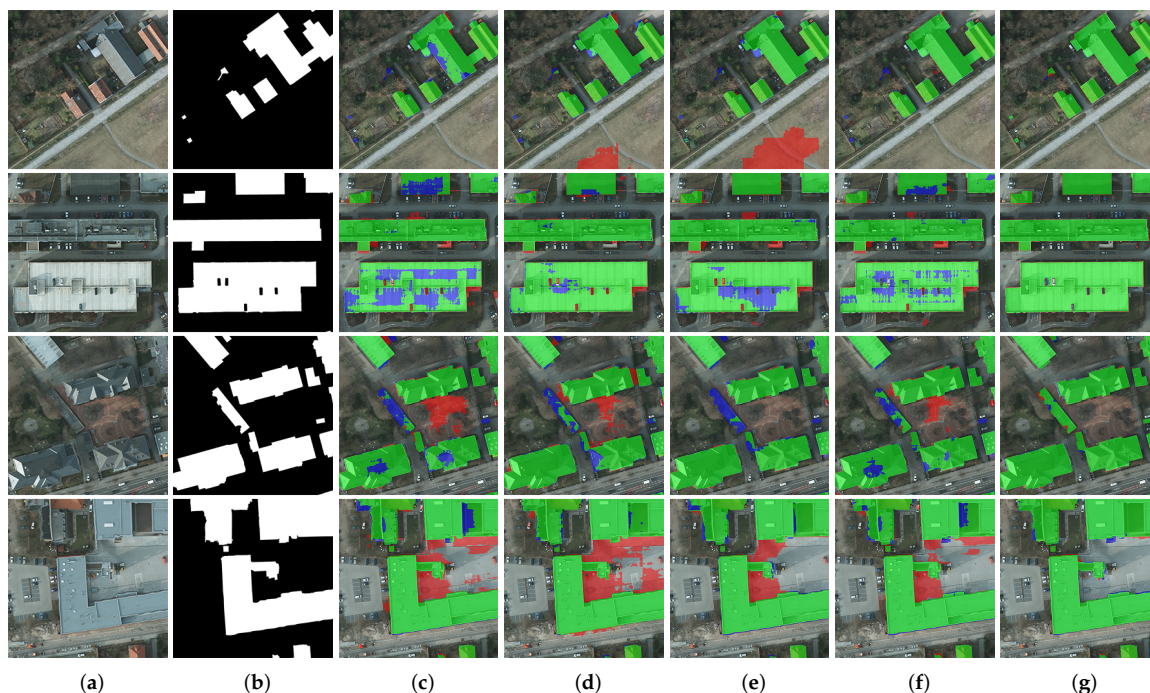


Figure 10. Close-up views of the results obtained by different methods on the Potsdam dataset. Images and results shown in (a)–(g) are the subset from the selected regions marked in Figure 9a. (a) Original image. (b) Ground-truth. (c) U-Net. (d) DeepLab-v3+. (e) DANet. (f) MA-FCN. (g) BARNet.

4.3.2. Quantitative Comparisons

To objectively assess the performance of the proposed method, following the work in [1,21,23], four commonly used metrics, i.e., precision, recall, F_1 score, and Intersection over Union (IoU), were adopted in the following experiments. These metrics are expressed as

$$Precision = \frac{TP}{TP + FP} \times 100\%, Recall = \frac{TP}{TP + FN} \times 100\%, \quad (14)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\%, \quad (15)$$

$$IoU = \frac{TP}{TP + FN + FP} \times 100\%, \quad (16)$$

where TP (true positive), number of correctly identified building pixels, FP (false positive), number of missed building pixels, TN (true negative), number of correctly classified non-building pixels, and FN (false negative), number of non-detected non-building pixels. Precision reports the ratio of TP in the whole positive predictions, and recall assesses the proportion of TP over entire building pixels in ground-truth. F_1 score is the weighted numerical assessment by taking both precision and recall into consideration. IoU measures the ratio that building pixels are correctly identified as building category.

The quantitative comparison results of different methods are listed in Table 1, where the best entries are in bold. According to Table 1, BARNet achieved an outstanding performance which is over than other SOTA methods. For the WHU dataset, we can see that the proposed method performs favorably against the SOTA methods in terms of all metrics. Compared with the second-best method (i.e., MA-FCN), BARNet improved the IoU score from an already very high IoU (90.70%) to a new highest score of 91.51%. In particular, we can see that the precision of BARNet is boosted by 2.01 points than MA-FCN, which is a near-perfect performance on this dataset. The near-saturated performance confirms that BARNet is capable of extracting buildings in VHR aerial images completely. Even in

Table 1. Quantitative results of different methods on the two datasets. The entries in bold denote the best on the corresponding dataset. The short line in this table is the unknown results that are not given by the authors.

Datasets	Methods ¹	Precision (%)	Recall (%)	F ₁ (%)	IoU(%)
WHU	U-Net	92.88	93.18	91.61	86.95
	DeepLab-v3+	95.07	90.75	92.85	88.05
	DANet	94.25	93.93	94.09	88.87
	MA-FCN (our implement)	94.20	94.47	94.34	89.21
	MA-FCN (overlap & vote) [23]	95.20	95.10	-	90.70
	SiU-Net [21]	93.80	93.90	-	88.40
	BARNet (ours)	97.21	95.32	96.26	91.51
Potsdam	U-Net	93.90	93.61	93.75	86.70
	DeepLab-v3+	95.70	93.95	94.81	88.21
	DANet	95.63	94.30	94.97	88.19
	MA-FCN	93.70	93.20	93.42	87.69
	DAN [52]	-	-	92.56	90.56
	Wang et al. [1]	94.90	96.50	95.70	-
	BARNet (ours)	98.64	95.12	96.84	92.24

¹ This table incorporates the numerical results reported by other works and the results obtained by our method.

the face of the challenging Potsdam dataset, where the images are with very high resolution, BARNet also achieved a remarkable performance with the precision score of 98.64%, the F₁ score of 96.84%, and the IoU score of 92.24%. Here, the recall score of BARNet is lower than the method reported by Wang et al. [1] is that the additional normalized digital surface model (nDSM) data was not utilized to enhance the model performance. Under the condition of only using single R-G-B data, the proposed method improved the precision, F₁, and IoU by 2.94%, 1.14%, and 1.68%, respectively, compared against the second-best metrics. The improvements indicate that BARNet is robust enough to cope with the building extraction in VHR aerial images with complex urban scenes.

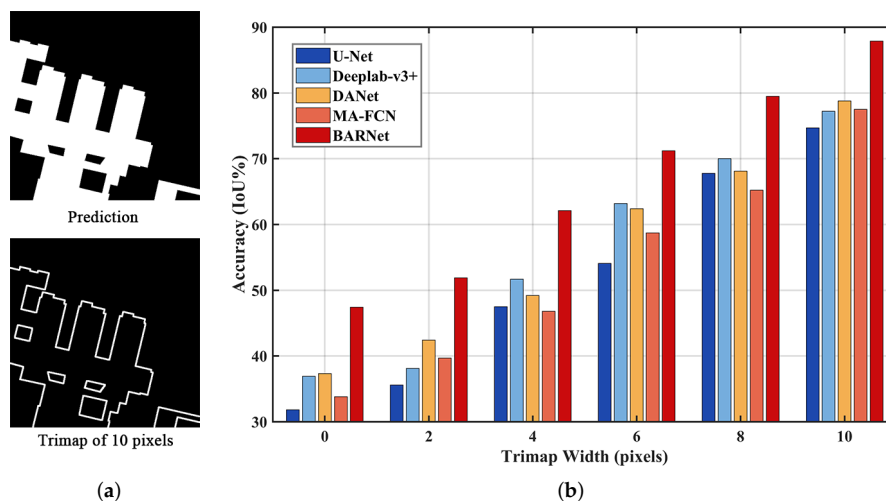


Figure 11. Instance of trimap and comparisons of boundary refinement. (a) Boundary trimap. (b) Mean IoU Results for trimap experiments with band width of {2, 4, 6, 8, 10, 12}.

To quantify the boundary segmentation quality of different methods, the comparison result of trimap boundary experiments performed on the WHU dataset are also reported, as illustrated in Figure 11. Notably, we did not conduct the trimap experiment on the Potsdam dataset owing to the evident texture distortion in building boundary regions for tested images, leading to that the ground truth is not precisely corresponding to the actual building boundary. Especially, as presented

in Figure 11a, the eroded and dilated band along the boundary with a given width (pixels) is called trimap. We utilized the morphological dilation and erosion operations to generate the boundary trimap with the width of $\{2, 4, 6, 8, 10, 12\}$. After obtaining the predicted trimap and reference trimap, we computed the mean IoU between them. The higher the mean IoU, the better performance for boundary segmentation. As shown in Figure 11, when the bandwidth of trimap is lower than 8 pixels, the comparative methods exhibited poor performance, suggesting that massive boundary pixels were classified wrongly. Nevertheless, it can be clearly observed that our method achieved significant performance on refining the boundary extraction, which verifies the positive effect of GARFU and BA loss on enhancing model performance.

5. Discussion

5.1. Ablation Studies

The ablation experiments include two parts: (a) an explore for further investigating the contribution of each sub-module introduced in Section 3 and the improvement strategies adopted for training and inference; (b) several quantitative comparisons for the GARFU, the DASPP, BA loss, and boundary-enhanced loss with other corresponding SOTA methods. Unless otherwise stated, all the involved experiments were conducted under the same conditions and settings given in for fairness.

5.1.1. Network Design Evaluation

To verify the design of BARNet, U-Net was chosen as the baseline model, and IoU was adopted to assess the effectiveness quantitatively. The detailed evaluation results were summarized in Table 2. We first replaced the encoder part in the baseline with ResNet-101. Meanwhile, all the encoder feature maps were reduced to 256 channels to keep consistent with BARNet, which is different from the baseline. Due to the strong feature encoding ability of ResNet-101, these changes bring an improvement of 1.49% IoU. After inserting the GARFU module in the lateral skip connections, IoU was improved by 0.72% point, implying that selectively fusing cross-level features is better than the direct concatenation fusion. Adding the DASPP module to capture the multi-scale context obtained a significant improvement with IoU of 1.12%, indicating that the proposed GARFU is robust to handle the critical scale issue for building objects in VHR aerial images. As expected, BA loss notably boosted the performance with an increment of 0.53% IoU, compared with the commonly used cross-entropy loss. Additionally, the OHEM strategy further improved the IoU by 0.53% point. With the help of MS inference, our model achieved 91.51% IoU, which significantly outperforms the previous SOTA model MA-FCN that achieved 90.7% IoU on the WHU dataset with multi-model voting and refined-overlap strategies.

Table 2. Evaluation results for network design, where U-Net serves as the baseline model. ✓ indicates the corresponding component is adopted. ↑ stands for the improvement of IoU for using the adopted component, compared to the previous step.

Baseline	ResNet-101	GARFU	DASPP	CE loss	BA loss	OHEM	MS	IoU (%)
✓				✓				86.95
✓	✓			✓				88.44 (1.49↑)
✓	✓	✓		✓				89.16 (0.72↑)
✓	✓	✓	✓	✓				90.28 (1.12↑)
✓	✓	✓	✓		✓			90.84 (0.56↑)
✓	✓	✓	✓		✓	✓		91.37 (0.53↑)
✓	✓	✓	✓		✓	✓	✓	91.51 (0.14↑)

5.1.2. Comparison with Multi-level Feature Fusion Strategy

To test the proposed GARFU, herein, the BARNet served as the baseline model. We utilized the addition operation and the concatenation operation to replace the GARFU, respectively. As described

in Section 5.1.1, all the feature maps from the encoder were also reduced to 256 channels using 3×3 convolutions, and the others remained unchanged. Table 3 shows the comparison results, from which we can observe that IoU reduces by 0.89% and 0.78% respectively, without GARFU. The decreased performance for addition and concatenation can be attributed to ignoring the semantic gap between high- and low-level features. We rethink the contribution of different level features and fill the gap well. GARFU can adaptively harvest the useful information to fuse them by learning a spatial gated map and a channel attention vector from two adjacent high- and low-level features. This simple yet efficient fusion strategy contributes to making full use of different level features for building extraction.

Table 3. Comparison of different fusion strategies on the WHU dataset. ↓ stands for the decrement of IoU compared with the baseline. Note, OHEM and MS are not used for this comparison.

Fusion Strategies	IoU (%)
GARFU (baseline)	90.84
Addition	89.94 (0.89↓)
Concatenation	90.06 (0.78↓)

5.1.3. Comparison with Multi-scale Context Scheme

The idea of DASPP is to make the network more stable for coping with the buildings in VHR remote sensing images with complex scenes by capturing denser multi-scale semantic context. It has been proved that it is non-trivial for semantic segmentation task to append an additional module after the backbone network for enlarging the receptive field size of the network. We compared the IoU performance of DASPP with several well verified context modeling approaches, i.e., Pyramid Pooling Module (PPM) in PSPNet [29], ASPP in DeepLab-v3+ [30], Self-Attention in Non-local Net [35], and Dual-Attention in DANet [36]. Additionally, complexity is also a key factor to a context modeling method, thus complexity also is reported. The experimental results are summarized in Table 4, from which we can find that DASPP outperforms other multi-scale context aggregation schemes both in terms of IoU and floating-point operations (FLOPs). Self-Attention-based methods can establish dense pixel-wise relations and achieve a reasonable performance of IoU, but the practical application is restricted by high computational cost. Despite ASPP has the maximum trainable parameters, its performance looks insufficient. In contrast, DASPP can acquire the best performance when maintaining efficiency, which demonstrates the robustness of DASPP for coping with the building scale variation.

Table 4. Comparison of different multi-scale contest schemes, where the best is in bold. We report the IoU results performed on the WHU dataset. FLOPs is calculated when processing the input with a size of $1 \times 2048 \times 128 \times 128$. Note, OHEM and MS are not used for this comparison.

Mthods	IoU (%)	Paramters (M)	FLOPs (G)
PPM	89.79	22	619
ASPP	90.34	15.1	503
Self-Attention	90.42	10.5	619
Dual-Attention	90.45	10.6	1110
DASPP (Ours)	90.84	10.6	172

5.1.4. Analysis of the Generality and Effectiveness of BE loss

Apparently, the traditional cross-entropy loss function only considers the pixel-level similarities between predictions and labels, resulting in that this loss is not sensitive when tackling boundary pixels and non-boundary pixels. In this study, we developed the boundary-enhanced (BE) loss to strengthen boundary segmentation explicitly. Since the importance has been proved in Section 5.1.1, we only compared BE loss with the well verified conventional DenseCRF introduced in DeepLab-v1 [38].

We fine-tuned the hyper-parameters in DenseCRF for yielding the optimal results. According to the result listed in Table 5, the proposed BE loss outperforms DenseCRF no matter what baseline model it is embedded. DenseCRF only brings a slight improvement in IoU. On the other hand, in the experiments, we observed that embedding DenseCRF is with a high computational cost. The comparisons powerfully verify the generality and robustness of BE loss is better than the SOTA DenseCRF.

Table 5. Comparison of BE loss and DenseCRF, where the best is in bold. We report the results performed on the WHU dataset. ✓ means that the corresponding method is adopted. Note, OHEM and MS are not used for this comparison.

Methods	CE loss	CE loss + DenseCRF	CE loss + BE loss	IoU (%)
U-Net	✓			86.95
U-Net		✓		87.03
U-Net			✓	87.76
BARNet (Ours)	✓			90.28
BARNet (Ours)		✓		90.39
BARNet (Ours)			✓	90.84

5.2. Limitations and Future Works

Although the proposed approach has filled the left gap that the performance of most of the existing methods is insufficient to recognize large-scale buildings and locate building boundaries well accurately, there are still some inherent pending problems that should not be ignored. First, the number of the total trainable parameters in BARNet is 67.49M, which is greatly larger than some medium-scale networks, such as U-Net (about 28.95M). On the other hand, the efficiency of our method is relatively slow. It took about 9 hours for training and about 3 minutes to inference on the WHU dataset, making the proposed method impractical to be deployed on mobile platforms such as the UAV. Thus, future work should pay attention to achieve real-time extraction. Then, in the experiments, we found that extracting small buildings is insufficient (54.29% IoU on WHU dataset for small buildings with an area less than 2000 pixels). So, a stronger high-resolution network should be developed to handle this problem. Last, we observed that there are many latent mistaken labels in several existing open-source benchmarks, making it quite hard to improve the performance further based on these datasets. For this reason, semi-supervised learning and few-shot learning should be given enough attention to reducing the dependence on mass high-quality labeled data.

6. Conclusions

Even though tremendous efforts have been made in automatic building extraction from VHR images using CNNs, extracting large-scale buildings completely and locating building boundaries precisely remains a challenging issue due to the limited multi-scale context and the lack of boundary consideration. With such motivation, BARNet was proposed to address the issues. Within BARNet, GARFU was introduced to make full use of multi-level features by re-calibrating the information contribution in the channel and the spatial dimensions. Besides, DASPP was developed to encode the multi-scale context better. In particular, the BE loss was embedded into the network to force the model to pay attention to the boundary. Comprehensive experiments performed on WHU and ISPRS benchmarks indicate that BARNet is suitable for processing building extraction in VHR aerial remotely sensed images over complex urban scenes. Compared with several SOTA models, our method exhibits the best performance with the highest accuracy constantly. A light-weighted network and semi-supervised learning will be developed to improve computational efficiency and extraction accuracy in our future research.

Author Contributions: Conceptualization, Y.J.; Methodology, Y.J.; Software, Y.J.; Validation, X.L. and H.J.; Resources, W.X.; Data curation, Y.J.; Writing—original draft preparation, Y.J.; Writing—review and editing, C.Z.; Visualization, Y.J.; Supervision, W.X.; Funding acquisition, W.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sichuan Science and Technology Project (No. 2020YFG0306, No. 2020YFG0055, and No. 2020YFG0327) and Science and Technology Program of Hebei (No. 19255901D and No. 20355901D).

Acknowledgments: The authors thank Wuhan University and ISPRS for providing the open-access and free aerial image dataset. The authors would also like to thank anonymous reviewers and the editors for their insightful comments and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
FCN	Fully Convolutional Network
GARFU	Gated-Attention Refined Unit
FPN	Feature Pyramid Network
ASPP	Atrous Spatial Pyramid Pooling
DASPP	Denser Atrous Spatial Pyramid Pooling
BN	Batch Normalization
ReLU	Rectified Linear Unit
CASPB	Cascaded Atrous Spatial Pyramid Block
CE	Cross-Entropy
OHEM	Online Hard Example Mining
ISPRS	International Society for Photogrammetry and Remote Sensing
GPU	Graphics Processing Unit
PPM	Pyramid Pooling Module
DenseCRF	Dense Conditional Random Filed
MS	Multi-scale Inference

References

1. Wang, Y.; Chen, C.; Ding, M.; Li, J. Real-time dense semantic labeling with dual-Path framework for high-resolution remote sensing image. *Remote Sens.* **2019**, *11*, 3020.
2. Chaudhuri, D.; Kushwaha, N.; Samal, A.; Agarwal, R. Automatic building detection from high-resolution satellite images based on morphology and internal gray variance. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2015**, *9*, 1767–1779.
3. Wang, X.; Li, P. Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data. *ISPRS-J. Photogramm. Remote Sens.* **2020**, *159*, 322–336.
4. Huang, X.; Zhang, L. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 721–732.
5. Du, S.; Zhang, Y.; Zou, Z.; Xu, S.; He, X.; Chen, S. Automatic building extraction from LiDAR data fusion of point and grid-based features. *ISPRS-J. Photogramm. Remote Sens.* **2017**, *130*, 294–307.
6. Awrangjeb, M.; Fraser, C.S. Automatic segmentation of raw LiDAR data for extraction of building roofs. *Remote Sens.* **2014**, *6*, 3716–3751.
7. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS-J. Photogramm. Remote Sens.* **2019**, *151*, 91–105.
8. Awrangjeb, M.; Ravanbakhsh, M.; Fraser, C.S. Automatic detection of residential buildings using LIDAR data and multispectral imagery. *ISPRS-J. Photogramm. Remote Sens.* **2010**, *65*, 457–467.
9. You, Y.; Wang, S.; Ma, Y.; Chen, G.; Wang, B.; Shen, M.; Liu, W. Building detection from VHR remote sensing imagery based on the morphological building index. *Remote Sens.* **2018**, *10*, 1287.

10. Huang, X.; Yuan, W.; Li, J.; Zhang, L. A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2016**, *10*, 654–668.
11. Zhai, W.; Shen, H.; Huang, C.; Pei, W. Fusion of polarimetric and texture information for urban building extraction from fully polarimetric SAR imagery. *Remote Sens Lett.* **2016**, *7*, 31–40.
12. Qin, X.; He, S.; Yang, X.; Dehghan, M.; Qin, Q.; Martin, J. Accurate outline extraction of individual building from very high-resolution optical images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1775–1779.
13. Turker, M.; Koc-San, D. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 58–69.
14. Inglada, J. 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.* **2007**, *62*, 236–248.
15. Miao, Z.; Shi, W.; Gamba, P.; Li, Z. An object-based method for road network extraction in VHR satellite images. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2015**, *8*, 4853–4862.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 27–30 June, 2016, pp. 770–778.
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137.
18. Jin, Y.; Xu, W.; Hu, Z.; Jia, H.; Luo, X.; Shao, D. GSCA-UNet: Towards Automatic Shadow Detection in Urban Aerial Imagery with Global-Spatial-Context Attention Module. *Remote Sensing* **2020**, *12*, 2864.
19. Wu, G.; Guo, Z.; Shi, X.; Chen, Q.; Xu, Y.; Shibasaki, R.; Shao, X. A boundary regulated network for accurate roof segmentation and outline extraction. *Remote Sens.* **2018**, *10*, 1195.
20. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 2380.
21. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sensing* **2018**, *57*, 574–586.
22. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322.
23. Wei, S.; Ji, S.; Lu, M. Toward automatic building footprint delineation from aerial images using cnn and regularization. *IEEE Trans. Geosci. Remote Sensing* **2019**, *58*, 2178–2189.
24. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813.
25. Xie, Y.; Zhu, J.; Cao, Y.; Feng, D.; Hu, M.; Li, W.; Zhang, Y.; Fu, L. Refined Extraction Of Building Outlines From High-Resolution Remote Sensing Imagery Based on a Multifeature Convolutional Neural Network and Morphological Filtering. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 1842–1855.
26. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N.; Hornegger, J.; Wells, W.M.; Frangi, A.F., Eds.; Springer International Publishing: Cham, 2015; pp. 234–241.
27. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495.
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA, 22–26 July, 2017, pp. 2117–2125.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern. Honolulu, HI, USA, 21–26 July, 2017, pp. 2881–2890.
30. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 8–14 September, 2018, pp. 801–818.
31. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 7–20 June, 2015, pp. 3431–3440.

32. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 8–14 September, 2018, pp. 325–341.
33. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* **2017**.
34. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Densenaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 18–22 June, 2018, pp. 3684–3692.
35. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 18–22 June, 2018, pp. 7794–7803.
36. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA, 16–20 June, 2019, pp. 3146–3154.
37. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065* **2019**.
38. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* **2014**.
39. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a discriminative feature network for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 18–22 June, 2018, pp. 1857–1866.
40. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 27 October–2 November, 2019, pp. 5229–5238.
41. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; others. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 7–20 June, 2017, pp. 4681–4690.
42. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA, 14–19 June, 2020, pp. 1580–1589.
43. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 18–22 June, 2018, pp. 7132–7141.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European conference on computer vision. Springer, Amsterdam, Netherlands, 8–16 October, 2016, pp. 630–645.
45. Wu, Y.; Yuan, M.; Dong, S.; Lin, L.; Liu, Y. Remaining useful life estimation of engineered systems using vanilla LSTM neural networks. *Neurocomputing* **2018**, *275*, 167–179.
46. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Mobilenets, H.A. Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* **2017**.
47. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, Miami, FL, USA, 20–25 June, 2009, pp. 248–255.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA, 7–20 June, 2015, pp. 1026–1034.
49. Da, K. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
50. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* **2017**.
51. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; others. Mixed precision training. *arXiv preprint arXiv:1710.03740* **2017**.
52. Yang, H.; Wu, P.; Yao, X.; Wu, Y.; Wang, B.; Xu, Y. Building extraction in very high resolution imagery by dense-attention networks. *Remote Sens.* **2018**, *10*, 1768.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).