remote sensing

MDPI

Article

Enhancing Building Segmentation in Remote Sensing Images: Advanced Multi-Scale Boundary Refinement with MBR-HRNet

Geding Yan, Haitao Jing *, Hui Li, Huanchao Guo and Shi He

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China; 212104020020@home.hpu.edu.cn (G.Y.)
* Correspondence: jht@hpu.edu.cn

Abstract: Deep learning algorithms offer an effective solution to the inefficiencies and poor results of traditional methods for building a footprint extraction from high-resolution remote sensing imagery. However, the heterogeneous shapes and sizes of buildings render local extraction vulnerable to the influence of intricate backgrounds or scenes, culminating in intra-class inconsistency and inaccurate segmentation outcomes. Moreover, the methods for extracting buildings from very high-resolution (VHR) images at present often lose spatial texture information during down-sampling, leading to problems, such as blurry image boundaries or object sticking. To solve these problems, we propose the multi-scale boundary-refined HRNet (MBR-HRNet) model, which preserves detailed boundary features for accurate building segmentation. The boundary refinement module (BRM) enhances the accuracy of small buildings and boundary extraction in the building segmentation network by integrating edge information learning into a separate branch. Additionally, the multi-scale context fusion module integrates feature information of different scales, enhancing the accuracy of the final predicted image. Experiments on WHU and Massachusetts building datasets have shown that MBR-HRNet outperforms other advanced semantic segmentation models, achieving the highest intersection over union results of 91.31% and 70.97%, respectively.

Keywords: building footprint extraction; remote sensing imagery; boundary refinement; multi-scale context fusion; intra-class inconsistency

Citation: Yan, G.; Jing, H.; Li, H.; Guo, H.; He, S. Enhancing Building Segmentation in Remote Sensing Images: Advanced Multi-Scale Boundary Refinement with MBR-HRNet. *Remote Sens.* **2023**, *15*, 3766. https://doi.org/10.3390/rs15153766

Academic Editors: Wen Liu and Yonas Zewdu Ayele

Received: 5 July 2023
Revised: 27 July 2023
Accepted: 28 July 2023
Published: 28 July 2023

1. Introduction

Buildings constitute an essential element of urban areas. Accurate information about buildings is crucial for various applications, including city planning [1,2], environmental monitoring [3], real estate management [4], population estimation [5], and disaster risk assessment [6].

Rapid advancements in aerospace technology, photography, and remote sensing techniques have made it possible to rapidly obtain high-resolution spectral and spatial information of objects over large regions. The detailed texture and semantic information provided by this VHR imagery is useful for building extraction and land-cover classification [7], making building extraction an exciting but challenging research topic [8]. Traditional building extraction methods primarily rely on manual and expert interpretations of the statistical features of building shape and texture derived from remote sensing images. However, these approaches suffer from limited accuracy, making it challenging to meet the demands of large-scale datasets and intelligent automatic updates. In contrast, object-based extraction methods that use spectral, textural, and geometric features of buildings, and auxiliary information, such as DSM, LiDAR, and building shadows, can yield more accurate results [9–11]. For example, Jiang et al. [12] used an object-oriented method to extract trees and buildings with DSM data, then removed the trees to retrieve the building, while Vu et al. [13] leveraged LiDAR to obtain building structural information through height and spectral data to remove vegetation affecting building extraction. However, handcrafted

feature-based object extraction methods have limitations. They cannot extract semantic information about complex land features from VHR images. Additionally, these methods suffer from low accuracy, limited completeness, and require a considerable amount of manpower and resources. Additionally, they cannot meet the demands of industrial big data and intelligent automatic updates [14,15]. Therefore, the imperative lies in developing deep learning-based techniques for extracting building features with heightened precision and accuracy, thus fulfilling the demands of research and industrial applications.

In recent years, with the expeditious advancements in deep learning computer vision technology, the application of deep convolutional neural networks (DCNNs) for automated building extraction has been extensively adopted. Deep learning architectures are characterized by their ability to learn features specifically from the data, without the need for domain-specific knowledge to design features, which avoids the problem of feature dependence on specialized field knowledge [16]. The integration of remote sensing techniques with deep learning has emerged as a prevalent methodology for the extraction of buildings, demonstrating its widespread adoption [17–21]. Classical CNN models, such as VGGNet [22], ResNet [23], DenseNet [24], AlexNet [25], and InceptionNet [26], have become baseline networks for many segmentation models. Shrestha et al. [27] replaced the last three FC layers of the original VGGNet to construct an FCN model and applied conditional random fields to extract buildings from remote sensing images. Deng et al. [28] applied an improved ResNet50 encoder to extract building features and used ASPP blocks to capture multi-scale features of objects between the encoder and decoder. Chen et al. [29] employed a modified version of the Xception network as their backbone feature extraction network to identify buildings. This approach effectively reduced computational requirements, accelerated model convergence, and minimized training time.

Despite the substantial advancements achieved in building extraction accuracy through deep learning-based methods, certain challenges persist [30–32]. In the process of extracting buildings from VHR images, the incorporation of contextual information assumes a critical role and is deemed indispensable [33,34]. Although the DCNN has strong semantic feature extraction ability, common down-sampling operations can lead to the loss of spatial details, resulting in blurred images [35]. Tian et al. [36] utilized dilated convolutions with different dilation rates to expand the receptive field and a densely connected refine feature pyramid structure for the decoder to fuse multi-scale features with more spatial and semantic information. However, dilated convolutions may miss small objects in VHR images. Wang et al. [37] used an improved residual U-Net for building extraction, with residual modules to reduce the network's parameters and degradation. The skip connection of U-Net can transmit low-level features and increase the context information [38,39]; however, the simple concatenation of features from the low and high levels can lead to insufficient feature exploitation and inaccurate extraction of boundaries or small buildings. The boundary information of semantic segmentation in remote sensing images is also important for performance. Due to the complex shape, large-scale changes, and different lighting conditions of the targets, the boundaries between semantic objects are often ambiguous, which poses a challenge to segmentation. A common method to improve edge accuracy is to combine edge detection algorithms and add constraint terms to the loss function [40]. However, this method is susceptible to false detection of edges due to noise and different angles. When boundaries cross multiple regions, it becomes challenging to reflect the situation accurately and enhance the accuracy of segmentation results.

To further deal with these problems, we propose an MBR-HRNet network with boundary optimization for building extraction, which can automatically extract buildings from VHR images. The main contributions of this study are as follows:

(1) In the encoding stage of MBR-HRNet, we propose a boundary refinement module (BRM) that utilizes deep semantic information to generate a weight to filter out irrelevant boundary information from shallow layers, focusing on edge information and enhancing edge-recognition ability. The boundary extraction task is coordinated

with the building body extraction task, increasing the accuracy of building body boundaries;

(2) In the decoding stage of MBR-HRNet, a multi-scale context fusion module (MCFM) is applied to optimize the semantic information in the feature map. By incorporating extracted boundary data and effectively retaining intricate contextual nuances, this module successfully tackles the integration of both global and local contextual information across various levels, resulting in enhanced segmentation precision.

The rest of this paper is organized as follows. Section 2 provides a comprehensive overview of the methodology employed for refining a boundary extraction buildings network. Section 3 presents the dataset description, experimental configuration, and evaluation metrics. Section 4 describes the experimental results and discussions. Finally, Section 5 concludes this paper.

2. Materials and Methods

This section begins with an overview of the model's architecture outlined in Section 2.1. Subsequently, detailed explanations of the BRM and MCFM are presented in Sections 2.2 and 2.3, respectively. Concluding this section, we introduce our proposed loss function in Section 2.4.

2.1. Architecture of the Proposed Framework

Previous semantic segmentation tasks mostly used an encoder–decoder architecture and convolutions to continuously down-sample the image, process contextual semantic information on the resulting low-resolution image, and then restore the original high-resolution output [41]. However, as the number of convolutional layers increases, this operation cannot maintain the high-resolution feature information. Moreover, this feature extraction through serially connected encoders causes information loss and resource waste. When information is transmitted across multiple layers, each subsequent layer can only receive limited information from the previous layer. This means that, as the number of layers increases, the amount of information received by each subsequent layer decreases, which may result in the loss of important boundary contour information. In contrast, HRNet parallelizes the serial structure and replaces the operation of reducing the resolution with an operation that maintains the resolution [42]. Furthermore, the high-resolution and low-resolution feature maps continuously exchange information and advance synchronously. The existence of high-resolution maps makes the spatial resolution more accurate, while the existence of low-resolution maps makes semantic information more comprehensive. Furthermore, we proposed the BRM to further refine boundaries, using a separate branch to capture building boundary information. Moreover, to fully utilize features at different levels and improve the model's ability to extract fine contextual details, we proposed an MCFM. The pipeline of MBR-HRNet is shown in Figure 1.

2.2. BRM

Although many semantic segmentation models can effectively extract buildings, they are not accurate enough in capturing the details and edge structures of buildings, especially when dealing with closely adjacent ones. Furthermore, for networks trained only with building label loss, the loss contributions of edge regions are often relatively small. Because the smaller loss values equal smaller gradients in backpropagation, the network tends to focus on constructing the body of the building rather than the edges. To solve this problem, we proposed a BRM, as shown in Figure 2.

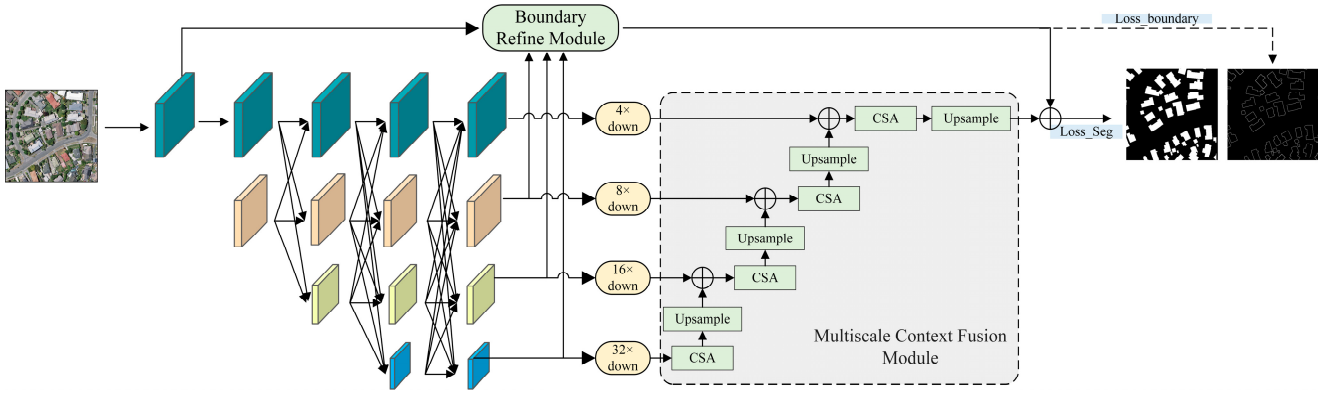

Figure 1. The structure of multi-scale boundary-refined HRNet (MBR-HRNet). The MBR-HRNet consists of two main components: the boundary-refinement module (BRM) and multi-scale context fusion module (MCFM). The CSA module represents channel–spatial attention.

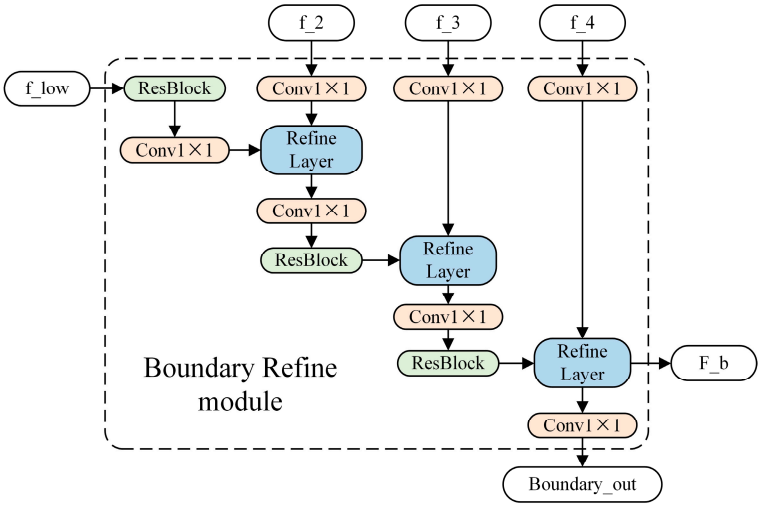

Figure 2. Boundary-refinement module (BRM).

The BRM uses a separate stream to process boundary information, which can be considered as an auxiliary to the main feature extraction flow. The main feature extraction flow forms a higher-level semantic understanding of the scene, while the boundary processing flow focuses solely on the parts related to boundaries. This allows the boundary flow to employ an efficient shallow structure for image processing at a very high resolution, avoiding down-sampling and the loss of fine detail information. It takes the shallow feature f_low, which contains rich detail information in the encoder, as the input and calculates weights by jointly considering three different scales of deep features, f_2, f_3, and f_4, output by the encoder, in order to filter out the unrelated boundary information from the shallow layer and obtain F_b that focuses only on the boundary information. The binary cross-entropy loss function is used to predict the boundary map Boundary_out. Specifically, because f_2, f_3, and f_4 correspond to features of different sizes, they are first up-sampled to match the size of f_low. Subsequently, a 1×1 convolution is applied to transform their channel dimensions. Similarly, f_low is initially subjected to a 1×1 convolution to modify its channel dimension. Subsequently, it undergoes calculations with the refine layer, along with the aforementioned features, until arriving at the final outputs F_b and Boundary_out.

The refine layer plays a vital role in filtering out irrelevant information regarding boundaries. In the refine layer, the features containing boundary information from shallow layers are first combined with the features capturing the main body of buildings from deep layers. Weights are then derived through network adaptation and applied to the input shallow features. Finally, the refined boundary features, which have been selected

based on weight adaptation through a residual connection, are outputted. For a detailed illustration, refer to Figure 3. First, IF_i and RF_i are connected, followed by a normalized 1×1 convolution layer $Conv_{1\times1}$ and a sigmoid function σ, resulting in an attention map $\alpha_i \in R^{H\times W}$.

$$\alpha_i = \sigma(Conv_{1\times1}(IF_i \oplus RF_i)) \tag{1}$$

$$OF_i = (\alpha_i \odot IF_i) + IF_i \tag{2}$$

where $\oplus$ denotes the concatenation of the feature; $\odot$ denotes the Hadamard product of the feature. For the obtained feature map α_i, it is applied to IF_i through an element-wise multiplication. Finally, the output OF_i is obtained through the residual connection and transmitted to the subsequent layer for further processing. The computations of the attention map and refine layer are differentiable, allowing for end-to-end backpropagation. Intuitively, α is an attention map that focuses on important boundary information and assigns greater weight to boundary areas. In the proposed BRM, three refine layers are used to connect the features of the encoder as refine features.

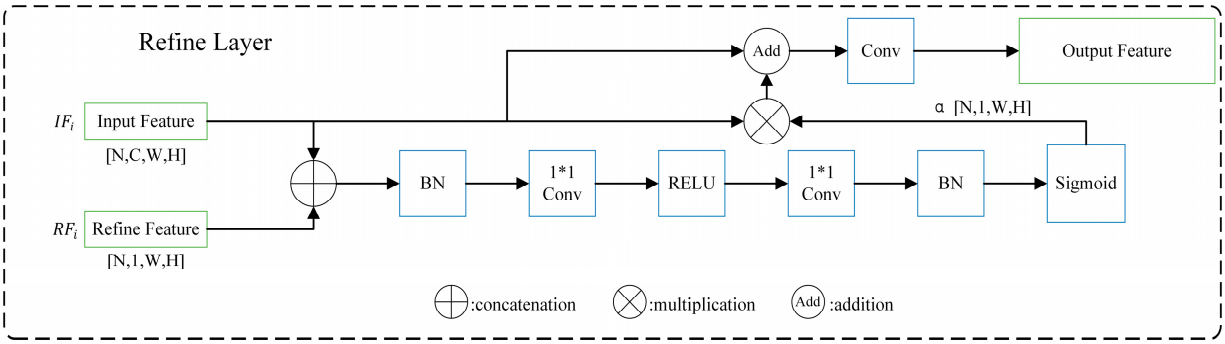

Figure 3. Structure diagram of the refine layer in the BRM.

2.3. MCFM

In semantic segmentation, the encoder usually generates feature maps at multiple scales. These feature maps correspond to semantic information at different levels and aggregating them can better capture global and local contextual information, improving segmentation accuracy [43,44]. However, simply concatenating low-level and high-level features can result in the inadequate utilization of different level features. HRNet's lowest-resolution branch output features have the strongest semantic representation [45]; however, in the original segmentation network, they are directly concatenated with features of other scales, and output, which cannot be fully propagated to higher-resolution branches, and semantic information are not fully utilized. Restoring higher-resolution prediction maps based on simple bilinear up-sampling may result in a loss of fine details. As shown in Figures 4 and 5, we proposed an up-sampling module that combined spatial and channel attention to fully utilize the spatial and channel dependencies of features to improve the semantic reconstruction ability and gradually restore them to the size of the prediction map. By effectively capturing the relevant features in the input data and removing useless information, this module can maintain a lightweight and efficient model. Moreover, using this module to combine regional features with boundary features can output refined semantic segmentation results. The formula is as follows:

$$F = F_{i+1} + CSA(F_i) \tag{3}$$

where F_i represents the feature map from the i-th layer, and F represents the feature map processed by the channel–spatial attention (CSA) module. The CSA module to is used to efficiently integrate multi-scale information; spatial and channel attention mechanisms are employed, thereby enhancing the feature representation capability. This module's design

minimizes the number of learnable parameters, resulting in a substantial reduction in model complexity. Additionally, it enables a better adaptation to scenarios that require an advanced semantic analysis.

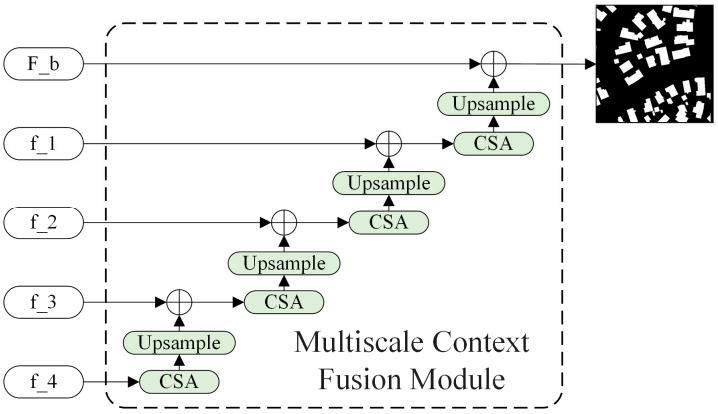

Figure 4. Multi-scale context fusion module (MCFM).

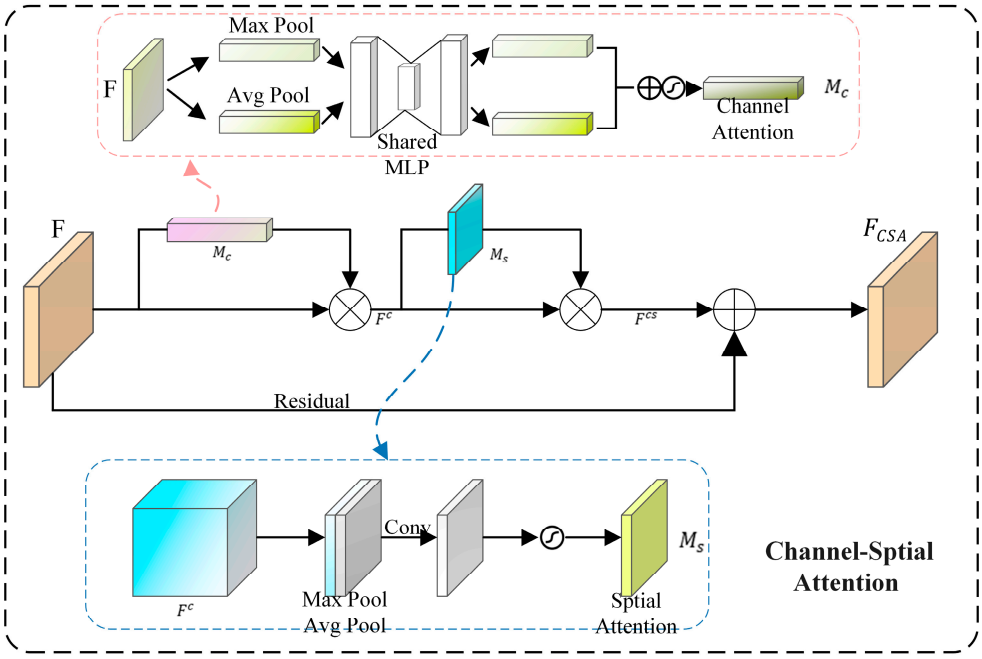

Figure 5. Structure of channel–spatial attention (CSA). F represents the input, while F_{CSA} represents the feature map obtained after passing through the CSA layer.

In the CSA module, for the input feature map, $F \in R^{C \times H \times W}$ undergoes parallel transformations through MaxPool and AvgPool layers, resulting in a feature map size of C × 1 × 1. Through the Share MLP module, the two output results are then added element-wise, and a sigmoid activation function is applied to obtain the output weights M_c of the channel attention. The original image is then multiplied by these output weights to restore the size of the feature map to C × H × W. The result of channel attention is then obtained using max pooling and average pooling to obtain two 1 × H × W feature maps, followed by a concatenation operation to combine the two feature maps into a single channel feature map using a 7 × 7 convolutional layer. Finally, a sigmoid is applied to obtain the spatial attention feature map, M_s. The output is multiplied by the original image to restore its size to C × H × W. The weights M_c and M_s for the channel attention and spatial attention can be calculated using the following equations. Residual connections

are employed to integrate the spatial and channel features effectively, thereby enhancing the semantic segmentation accuracy. The formulas are as follows:

$$F_{CSA} = F + F \times M_c \times M_s \tag{4}$$

$$M_c = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{5}$$

$$M_s = \sigma(Conv_{7\times7}(AvgPool(F) \oplus MaxPool(F))) \tag{6}$$

where σ represents the sigmoid function. *MLP* refers to a fully connected layer, and *AvgPool* and *MaxPool* refer to global average pooling and global maximum pooling, respectively. $\oplus$ denotes the concatenation of feature maps, and F_{CSA} is the feature map output by the CSA module.

2.4. Loss Function

The semantic segmentation task for buildings encounters a notable disparity in sample quantities between foreground (building) and background (non-building regions) classes. When training the convolutional neural network using traditional loss functions, the network tends to predict all pixels as a background class since it is easier to predict the background class that occupies most of the pixels. This phenomenon leads to insufficient feature representation and discriminative learning of the foreground class during training, thereby affecting the accuracy of the final segmentation results. Accordingly, we put forth a modified cross-entropy (CE) loss function with weights and assigned a higher weight to the foreground class to balance the sample quantities between the two classes. In the proposed framework, MBR-HRNet outputs two main results that aim to generate the building segmentation masks and building boundaries, respectively. The formula is as follows:

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i log(\hat{y}_i) + w_2(1-y_i)log(1-\hat{y}_i)] \tag{7}$$

where N is the number of samples, y_i represents the true label of the i_{th} sample (0 or 1), $\hat{y}_i$ represents the prediction for the i_{th} sample (ranging from 0 to 1), and w_1 and w_2 are the weight parameters for positive and negative classes, respectively. In this loss function, the cross-entropy error of the belonging class is calculated for each sample and multiplied by a corresponding weight. Different weight parameters can be set to adjust the impact of misclassification of different classes when applying the CE loss function, thus solving class imbalance in binary classifications.

When constructing boundary prediction models, the CE loss function is employed.

$$L_{boundary} = -\sum_i [y_i log(\hat{y}_i) + (1-y_i)log(1-\hat{y}_i)] \tag{8}$$

Therefore, the overall training loss is:

$$L = L_{seg} + L_{boundary} \tag{9}$$

3. Experiment

Within this section, we begin by providing an overview of the dataset used in the experiment in Section 3.1. Then, in Sections 3.2 and 3.3, we present the experiment configuration and evaluation metrics used, respectively.

3.1. Datasets

3.1.1. The WHU Dataset

For our experiment, we utilized the WHU building dataset, which consisted of aerial images [46]. Spanning across a 450 km^2 region within Christchurch, New Zealand, the dataset encompassed 187,000 buildings that exhibited diverse textures, shapes, and colors.

The geographical region under consideration was partitioned into 8189 tif format images with dimensions of 512 × 512 pixels, possessing a spatial resolution of 0.3 m. For the purpose of our experiment, this dataset was arbitrarily split into three sets: a training set consisting of 4736 images (approximately 60% of the dataset), a validation set comprising 1036 images (approximately 10% of the dataset), and a test set containing 2416 images (approximately 30% of the dataset). Figure 6 illustrates the original images along with their corresponding labels as depicted in the WHU building dataset.

Figure 6. Images and labels selected from the WHU dataset.

3.1.2. The Massachusetts Building Dataset

The Massachusetts building dataset is a comprehensive collection [47], encompassing 151 aerial images of the Boston area, portraying diverse urban and suburban regions adorned with buildings of varying sizes, including independent houses and garages. With an image resolution of 1500 × 1500 pixels, this dataset covered an expansive area of approximately 340 square kilometers, offering a spatial resolution of 1 m. Firstly, the training set comprised 137 images, which were further cropped to dimensions of 500 × 500 pixels, resulting in 1233 images. Secondly, the validation set encompassed four images that were subsequently cropped, yielding 36 images. Lastly, the test set contained 10 images, resulting in 90 cropped images. The images and their corresponding labels are shown in Figure 7.

3.2. Experimental Settings

Table 1 presents a comprehensive depiction of the primary software and hardware configurations implemented in this experiment, providing essential details and specifications. The primary emphasis of this experiment revolved around the extraction of architectural structures. Preceding the model training phase, a crucial preprocessing step entailed the rescaling of annotated images. Specifically, the pixel values were subjected to a normalization process, wherein their original range spanning from 0 to 255 was transformed into a standardized range of 0 to 1. In this representation, pixels possessing a value of 1 denoted the presence of buildings, while pixels with a value of 0 symbolized the background region. This paper compared the performance of four models, U-Net [48], PSPNet [49], HRNetV2 [50], and DeepLabv3+ [51], which are widely used for image segmentation tasks, using ResNet50 as their backbone. Considering that the optimal optimization of the models may not be achievable due to hardware limitations, we set the same standards for all experimental models to ensure fairness under our experimental conditions. We trained, validated, and tested the models on the same dataset, and used Adam optimizer for training with an initial learning rate of 0.0001, batch size of 8, and decayed it to 0.1 of

the current learning rate every 50 epochs. Our network was trained for 200 epochs on a GPU.

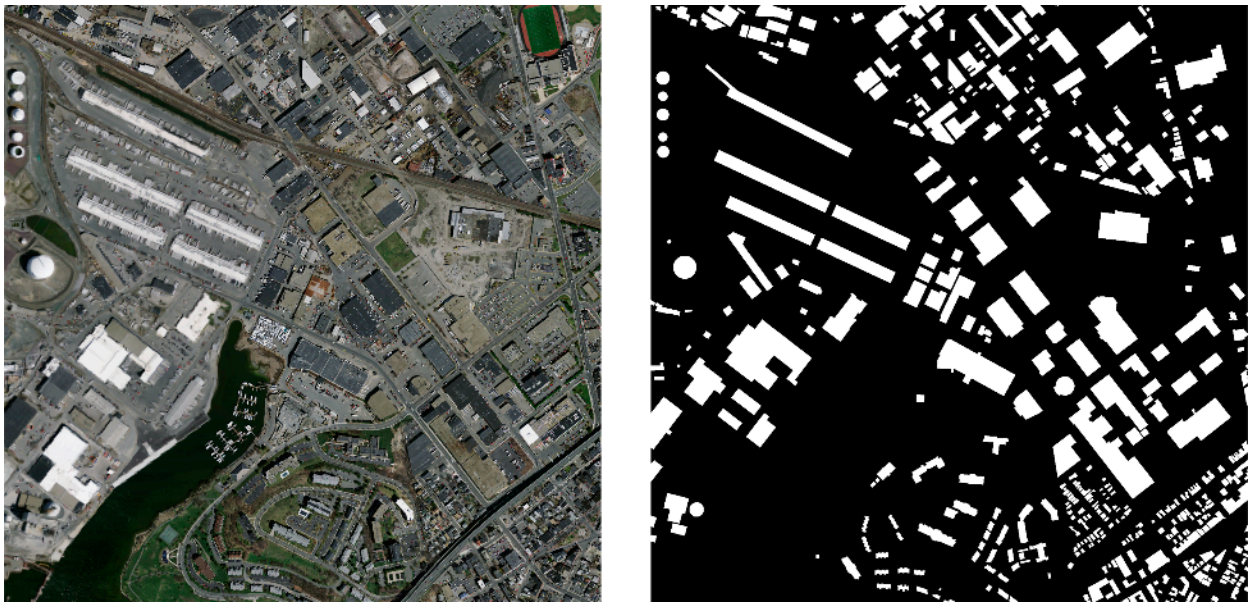

Figure 7. Images and labels selected from the Massachusetts dataset.

Table 1. Details of the employed hardware and software.

Item	Details
CPU	Intel i5-13600 3.50 GHz
GPU	GeForce RTX 3080
OS	Windows 10
Language	Python 3.8
Framework	PyTorch 1.7.1

During the training phase, the data augmentation of input images is an effective method to obtain more feature information, especially in cases where the dataset is limited [52]. Since buildings have rich geometric features, we first used spatial data augmentation techniques to enhance the data. This included random horizontal flipping with a probability of 50%, random rotation within a range of –10° to 10° with a probability of 50%, and random scale. To improve the network's generalization ability and reduce the influence of imaging condition differences, we adopted spectral data augmentation strategies. To augment the spectral data, we applied a range of techniques that included enhancing the hue and saturation, as well as introducing random Gaussian blur. By incorporating these data augmentation methods, we expanded the diversity and richness of the spectral information, thereby improving the overall quality and variability of the dataset.

3.3. Evaluation Metrics

To verify and compare the performance of the model, we adopted commonly used evaluation metrics, including precision, recall, intersection over union (IoU) ratio, and F1 score. These metrics can help measure the algorithm's performance. Precision is a metric that quantifies the proportion of accurately predicted pixels among all predicted pixels. By directly reflecting the model's segmentation proficiency and efficacy, accuracy serves as an insightful measure of its performance. In essence, it gauges the model's ability to precisely identify and classify pixels within the given data. Recall represents the ratio of the number of samples correctly classified as a certain category by the classifier to the number of true samples belonging to that category. IoU reflects the ratio of the common region between

pixel labels and actual pixel labels to the joint region between the two, and is not affected by the number of categories. F1 score is the harmonic mean of precision and recall and can be used to measure the model's performance in the presence of class imbalance. The formulas are shown in (10) to (13):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{F1 score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

4. Results and Discussion

In this section, we first evaluate the proposed model on the results of building extraction capability using the two public datasets in Section 4.1. Then, in Section 4.2, we design ablation experiments to verify the contributions of different modules in the network and demonstrate the visualized results of the BRM edge module. Finally, in Section 4.3, we analyze the existing shortcomings and future work.

4.1. Comparative Experiments

4.1.1. Quantitative Comparison

Based on Figures 8 and 9, it can be observed that our model exhibits a higher convergence speed compared to other algorithms, and the final convergence value is lower than that of other algorithms. Therefore, the MBR-HRNet model demonstrated the fastest convergence speed and the best convergence value. Tables 2 and 3 show the quantitative evaluation results of MBR-HRNet. For ease of reading, the highest score is shown in bold, and the second-highest score is underlined. Our model consistently outperformed other models on both datasets, achieving the highest IoU, accuracy, recall, and F1 scores. Specifically, our model achieved IoU scores of 70.97% and 91.31%, as well as F1 scores of 83.53% and 95.18% on the Massachusetts and WHU datasets, respectively. Compared to the second-best performing HRNet on the WHU dataset, MBR-HRNet improved the IoU and F1 scores by 1.98% and 1.02%, respectively. On the Massachusetts dataset, compared to the second-best performing U-net, MBR-HRNet improved the IoU and F1 scores by 2.01% and 2.13%, respectively. The data show that our model's improvement on the Massachusetts dataset compared to other models is higher, indicating that our model has strong capabilities for identifying buildings with low spatial resolutions and limited details provided.

Table 2. Comparative results from the selected models on the WHU dataset. Bold represents the best result, underline represents the second-best result.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
U-net	87.75	93.75	93.15	92.41
HRNetv2	<u>89.33</u>	94.51	<u>94.22</u>	<u>94.16</u>
Deeplabv3+	88.42	<u>94.78</u>	92.96	93.70
PSPNet	86.62	92.52	93.14	92.64
Our model	**91.31**	**95.48**	**94.88**	**95.18**

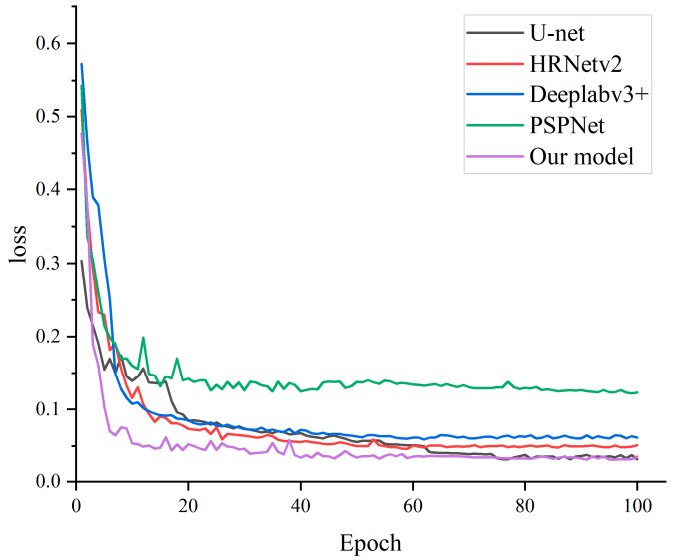

Figure 8. Curves of loss values of different models as the number of iterations increases in the training process on the WHU dataset.

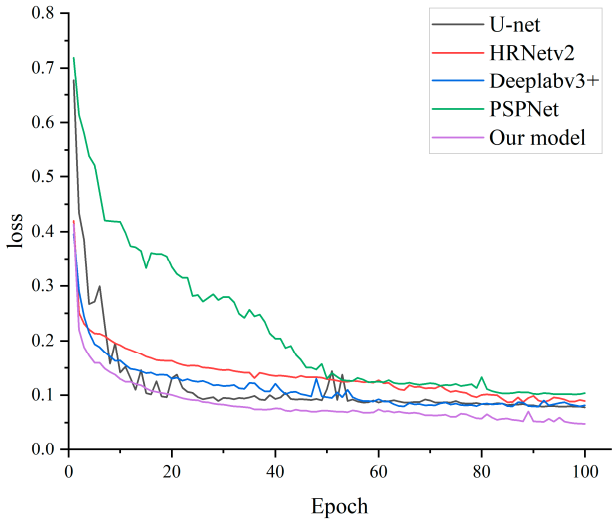

Figure 9. Curves of loss values of different models as the number of iterations increases in the training process on the Massachusetts dataset.

Table 3. Comparative results from the selected models on the Massachusetts dataset. Bold represents the best result, underline represents the second-best result.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
U-net	<u>68.96</u>	<u>85.32</u>	<u>76.68</u>	<u>81.40</u>
HRNetv2	67.71	82.08	79.60	80.59
Deeplabv3+	64.54	81.06	76.11	77.98
PSPNet	61.17	81.24	71.22	75.60
Our model	**70.97**	**86.40**	**80.85**	**83.53**

Regarding the study's comparison, four different models were utilized; HRNetV2 exhibited the best performance on the WHU dataset. This result can be attributed to the fact that U-Net and DeepLabv3+ lack semantic information in the up-sampling process, as they only fuse low-resolution feature maps. Conversely, HRNetV2 adopts a parallel computational architecture to extract features from four distinct scales. This approach

ensures that the output features at different scales contain richer semantic information. On the Massachusetts dataset, U-Net demonstrated a better performance. This can be attributed to the dataset's lower spatial resolution and prevalence of small-scale buildings. In such cases, U-Net benefits from incorporating skip connections during up-sampling, which helps preserve the correlation between high-level semantic information and low-level edge texture information, thereby enhancing the model's segmentation capability. The MCFM of the MBR-HRNet proposed by us has the capability to integrate multi-scale features. Additionally, the BRM can accurately identify building boundaries and small-scale structures.

4.1.2. Qualitative Comparison

Figure 10 shows the test images, corresponding labels, and building extraction results of four sample areas in the WHU dataset using the selected model. In Figure 10a, U-Net, PSPNet, HRNetV2, and DeepLabv3+ all failed to detect buildings that were obstructed by trees to some extent. Our method showed a better performance in overcoming the occlusion problem, and our model was almost unaffected by tree and shadow obstructions, as marked in the green box. In Figure 10b, U-Net, PSPNet, and Deeplabv3+ all had edge sticking problems with densely connected buildings, while MBR-HRNet accurately extracted the boundaries of the buildings. This was because the proposed BRM could improve the ability to recognize building edges and construct finer boundary parts of the extracted buildings. Figure 10c illustrates that the remaining four approaches exhibit different levels of false detection, which can be attributed to the similarity between the color and texture of building roofs and the ground. However, our model can accurately extract them, solving the problem of "intra-class spectral heterogeneity" remarkably well. In Figure 10d, none of the other methods correctly detect buildings with complex texture colors on the roof.

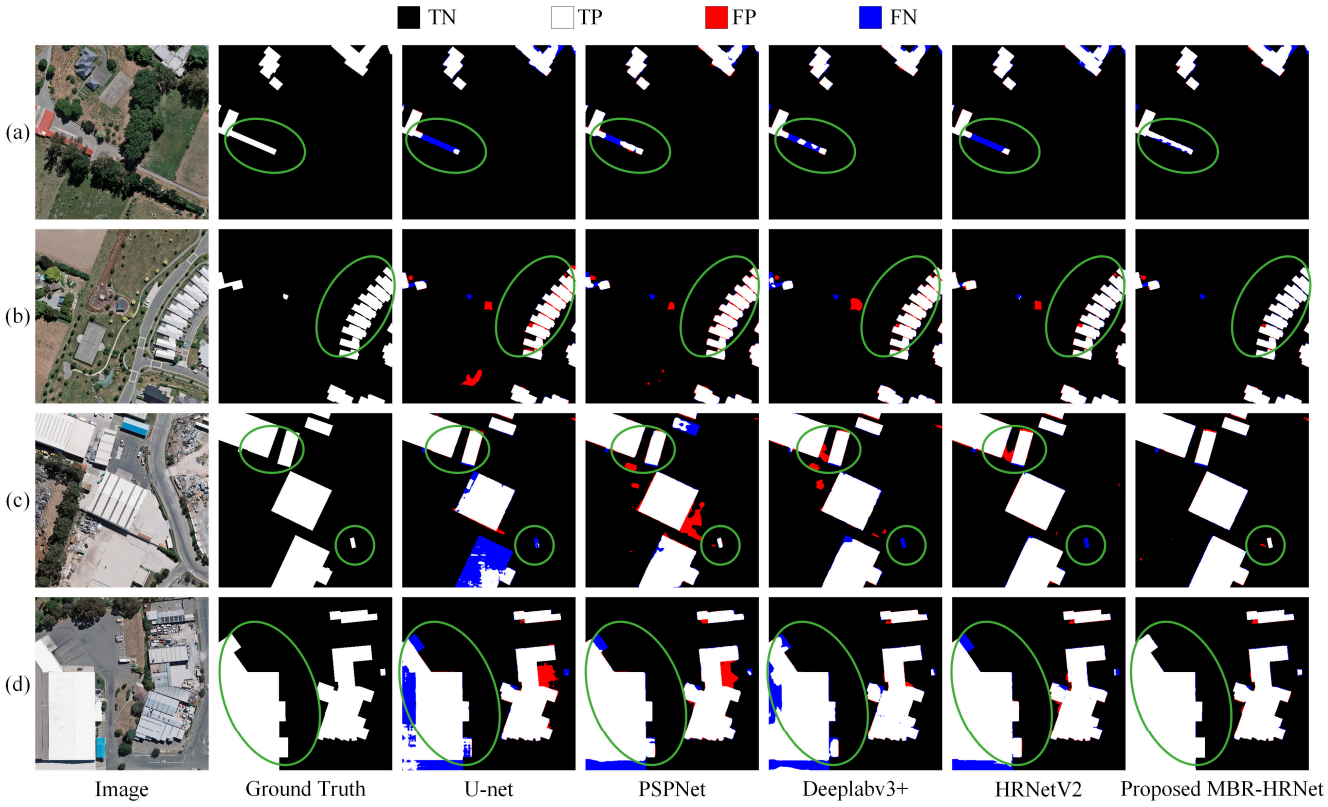

Figure 10. (**a**–**d**) The building extraction results of the proposed model on the WHU dataset. TN indicates the background, TP indicates the accurately detected building area, FP indicates the wrongly detected building area, and FN represents the parts of the building that were missed.

Figure 11 shows the building extraction results for the Massachusetts dataset using the selected models. The resolution of this dataset is relatively low, and most buildings appear as scattered patches, which poses a certain difficulty for extracting small buildings. In Figure 11a,c, darker-colored buildings are easily misidentified as ground color, and all compared models show some degree of missed detections. In Figure 9b, only MBR-HRNet accurately extracts two small buildings located on the water surface, while other models fail to detect them. In Figure 11d, U-Net, PSPNet, and Deeplabv3+ all have problems with the extracted building boundaries being stuck together in areas with densely distributed small buildings. Our proposed method can more accurately extract buildings, especially identifying the edges between buildings with dense and adjacent areas.

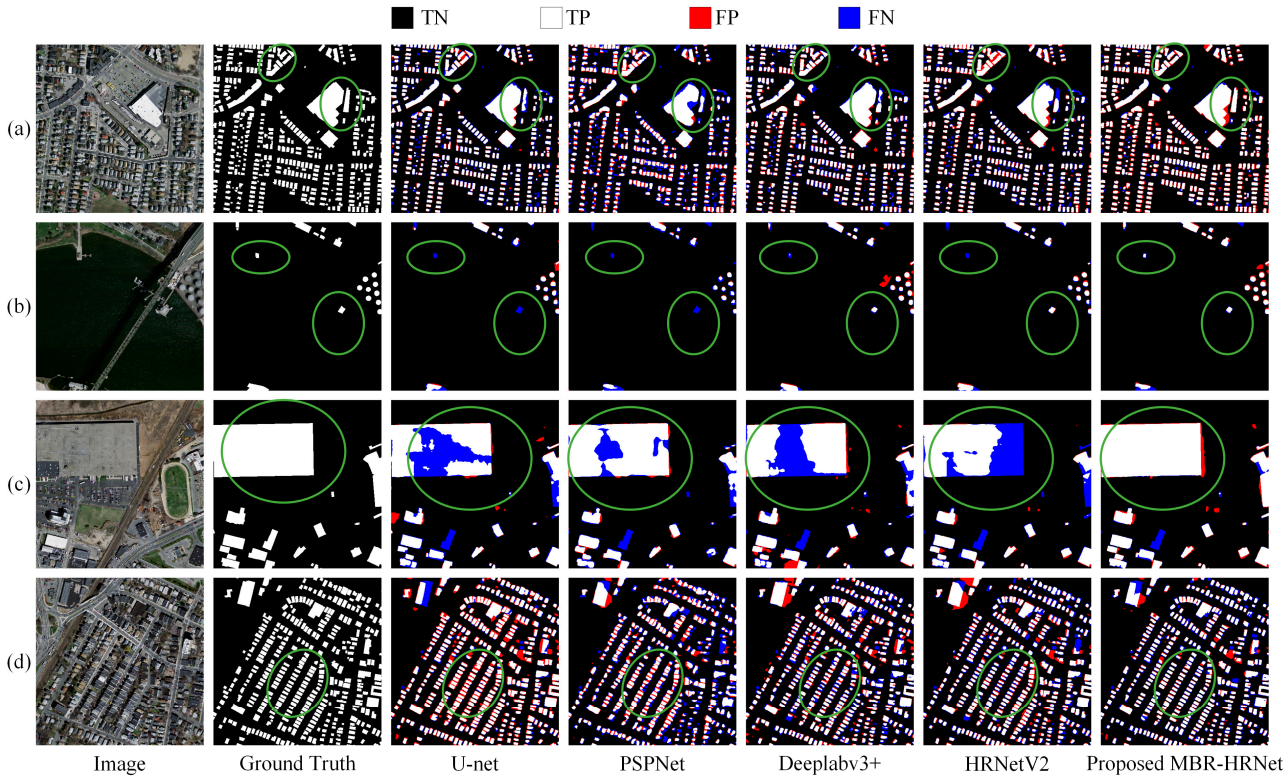

Figure 11. (**a**–**d**) The building extraction results of the proposed model on the Massachusetts dataset.

4.1.3. Comparison with State-of-the-Art Methods

In order to assess the efficacy of the proposed network, MBR-HRNet was compared with SOTA, including AGs-Unet, MAP-Net, MSL-Net, CFENet, DR-Net, and BOMSC-Net on the WHU dataset. The quantitative comparison results are shown in Table 4. We also compared MBR-HRNet with state-of-the-art building extraction methods on the Inria and Massachusetts datasets, as shown in Tables 5 and 6.

Table 4. Comparison of state-of-the-art methods and our model on the WHU dataset. The "-" symbol indicates the absence of a result for this method in its respective paper.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
AGs-Unet [53]	85.50	93.70	-	-
MAP-Net [54]	90.86	95.62	94.81	95.21
MSL-Net [55]	90.40	95.10	94.80	95.00
CFENet [56]	87.22	93.70	-	92.62
DR-Net [29]	86.00	92.70	92.20	92.50
BOMSC-Net [57]	90.15	95.14	94.50	94.80
Our model	91.31	95.48	94.88	95.18

Table 5. Comparison of state-of-the-art methods and our model on the Inria dataset. The "-" symbol indicates the absence of a result for this method in its respective paper.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
DeepLabV3 + ResNet-50 [58]	80.75	-	-	95.87
U-uet ResNet-50 [59]	69.29	84.38	78.77	80.20
LRAD-Net [60]	79.82	89.39	88.79	88.37
LCS-Net [61]	78.82	89.58	86.77	88.15
BOMSC-Net	78.18	87.93	87.58	87.75
CGSANet [62]	80.90	90.22	88.68	89.44
Our model	81.39	90.98	89.56	90.01

Table 6. Comparison of state-of-the-art methods and our model on the Massachusetts dataset. The "-" symbol indicates the absence of a result for this method in its respective paper.

Method	IoU (%)	Precision (%)	Recall (%)	F1 (%)
DeepLabV3 [63]	68.55	-	-	81.34
Res-Unet [64]	66.21	76.97	82.58	79.67
MultiBuildNet [65]	70.72	80.08	85.82	82.85
MAFF-HRNet [66]	68.32	83.15	79.29	81.17
D-LinkNet [67]	70.39	73.36	85.88	-
Our model	70.97	86.40	80.85	83.53

As shown in Table 4, our model performs the best in terms of IoU, reaching 91.31%, compared to all test methods developed in recent studies on the WHU dataset. At the same time, it also performed well in precision and recall, approaching or surpassing the performance of other models. MAP-Net slightly outperformed our model in the F1 indicator. In the Inria and Massachusetts datasets, our model also performed exceptionally well. Overall, our model performed excellently in multiple evaluation metrics, demonstrating the state-of-the-art performance of the proposed method.

4.2. Ablation Experiments

In order to investigate the roles played by various components of MBR-HRNet, we performed ablation experiments on the WHU and Massachusetts datasets, using HRNetV2 as the baseline, and evaluated the accuracy of the IoU and F1 scores. These experiments aimed to assess the individual contributions of different modules within MBR-HRNet by selectively removing or disabling them. As shown in Table 7, the incorporation of additional modules into the baseline yielded varying degrees of positive impacts. These supplementary modules exhibited the beneficial effects of different magnitudes. Firstly, the architecture of multi-scale feature fusion was optimized based on the HRNetV2 baseline, and Baseline + MCFM improved the baseline's IoU by 0.48% and 0.34%, respectively, on the two datasets. This demonstrated the effectiveness of using spatial and channel attention mechanisms to fully utilize the spatial and channel dependencies of features to improve the semantic reconstruction capability. Baseline + BRM improved the baseline's IoU by 0.69% and 1.96%, respectively, on the two datasets, proving the effectiveness of using independent boundary processing streams. The optimization effect on the Massachusetts dataset was the most significant, indicating that our model had a strong extraction ability for small, dense buildings. We visualized the edge features of the last two refine layers in the BRM for the sample images of the two datasets, as shown in Figure 12. As the feature maps were passed into the BRM, it could be seen that the network's attention to the edge increased significantly, gradually obtaining accurate edge texture information, which helped to segment buildings more accurately.

Table 7. The results of the ablation study on two datasets.

Method	WHU Dataset		Massachusetts Dataset	
	IoU (%)	F1 (%)	IoU (%)	F1 (%)
Baseline	89.33	94.16	67.71	80.59
Baseline + MCFM	89.81	94.38	68.05	81.60
Baseline + BRM	90.02	94.61	69.67	82.29
Baseline + MCFM + BRM	91.31	95.18	70.97	83.53

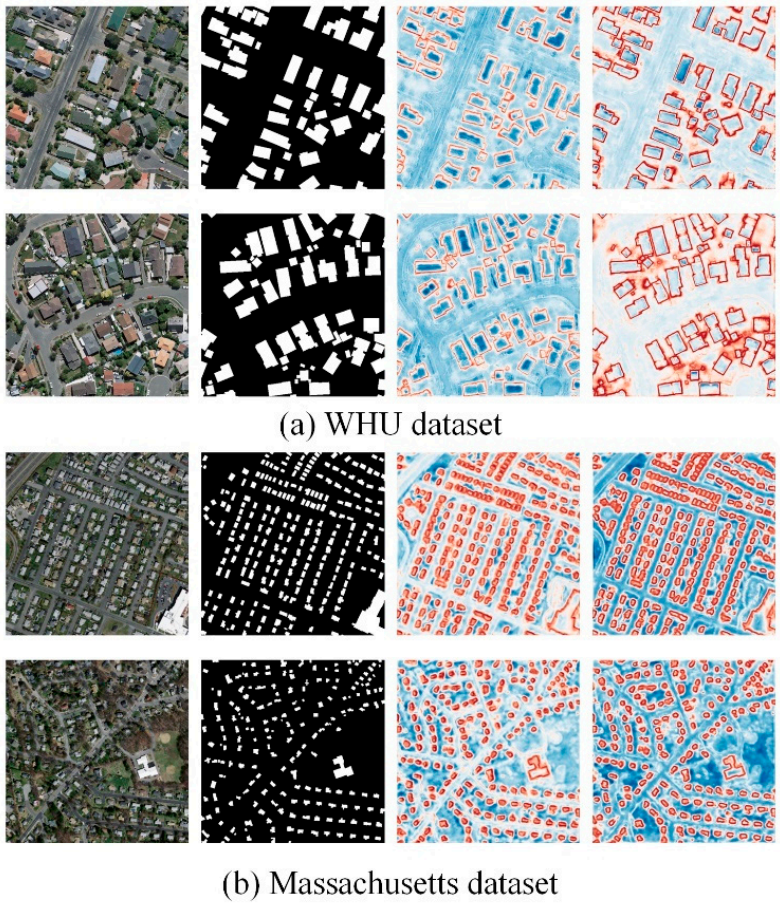

Figure 12. Visualization of edge feature maps in the refine layer of BRM. The color red indicates that the model pays higher attention to this area, blue indicates lower attention.

4.3. Limitations and Future Work

Despite the exceptional performance achieved by MBR-HRNet in building extraction accuracy on two datasets, there is still room for improvement. We conducted comparative experiments on the number of trainable parameters and floating-point operations (FLOPs) for each model, as shown in Table 8. Table 9 presents the inference time of the model for a single image slice.

Table 8. The efficiency comparison on the Massachusetts dataset.

Methods	IoU (%)	FLOPs (G)	Parameters (M)
U-net	68.96	262.09	34.53
HRnetv2	67.71	45.46	29.54
Deeplabv3+	64.54	164.12	39.63
PSPNet	61.17	11.84	2.24
Our model	70.97	68.71	31.02

Table 9. The inference time of the model for a single image.

Methods	PSPNet	U-Net	Deeplabv3+	HRNetv2	Our Model
Inference time	0.011 s	0.026 s	0.031 s	0.024 s	0.028 s

While PSPNet offered the highest efficiency, its accuracy was significantly lower compared to our model's performance. Deeplabv3+ had a similar number of parameters to our model but exhibited much greater computational complexity. In comparison to HRNetV2, our approach demonstrated a negligible increase in the number of trainable network parameters. Additionally, when compared to U-net, our proposed method exhibited a significant improvement in building recognition IoU accuracy on the Massachusetts test dataset without introducing additional computational complexities. These observations underscore the capabilities and advantages of our method over the existing alternatives. In summary, although the proposed method had good extraction accuracy, the computation complexity of BRM was not considered a lightweight module. In future work, we will try to introduce prior knowledge, such as building structural information, into the network, and further explore lightweight boundary-refinement modules.

5. Conclusions

In this study, we introduced the utilization of MBR-HRNet to tackle various difficulties encountered in extracting buildings from remote sensing images. These challenges encompassed issues, such as the potential oversight of small-scale buildings, indistinct boundaries, and non-uniform building shapes. By employing MBR-HRNet, we aimed to overcome these obstacles and improve the accuracy and robustness of building extraction in remote sensing applications. Within the MBR-HRNet architecture, the MCFM module plays a crucial role in capturing and amplifying diverse global multi-scale features. It effectively harnesses intricate representations of spatial dimensions found within both high-level and low-level features. This capability proves highly valuable in extracting building boundaries that exhibits discontinuities, as well as accurately representing the appearance of irregular buildings. By incorporating the MCFM module, MBR-HRNet enhances its capacity to capture nuanced details and improve the overall performance of building extraction tasks. BRM progressively refines the building boundary information within the features, resulting in more accurate and refined boundary features, so that the extracted buildings have more accurate contours and prevent the extraction results of adjacent buildings from being sticky. The quantitative experimental outcomes highlight the superior performance of the MBR-HRNet model compared to other methods. This advantage was particularly evident in its remarkable achievements in both WHU and Massachusetts building datasets, as it attained the highest IoU results of 91.31% and 70.97%, respectively. From the qualitative experimental results, it can be seen that MBR-HRNet has a clear advantage in extracting small, dense buildings and accurately constructing building boundaries. In the future research, there are two main directions for further investigation. Firstly, integrating prior knowledge into the network to enable a faster and more accurate extraction of buildings in the network. Secondly, leveraging the precise boundary extraction advantages of MBR-HRNet for buildings, we aim to design a module that directly converts grid prediction maps into regular vector boundaries and integrate it into an end-to-end deep learning architecture. This approach would enable the direct acquisition of building boundaries and provide more comprehensive applications.

Author Contributions: Conceptualization, G.Y., H.J. and S.H.; data curation, G.Y., H.L. and H.G.; funding acquisition, S.H.; methodology, G.Y.; software, G.Y.; supervision, H.J.; validation, H.L.; visualization, H.G.; writing—original draft, G.Y.; writing—review and editing, G.Y. and H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Scientific Research Project of Henan higher education institutions (Grant No. 22B420004) and the Scientific and Technological Research Project of Henan Province (Grant No. 232102210043).

Data Availability Statement: In the aforementioned passage, it is mentioned that two distinct datasets pertaining to public building semantic labeling were employed in the study. These datasets include the WHU building dataset and the Massachusetts buildings dataset. Individuals can access these datasets through the respective sources: http://gpcv.whu.edu.cn/data/ (accessed on 20 October 2020) for the WHU building dataset and https://www.kaggle.com/datasets/balraj98/massachusetts-buildings-dataset (accessed on 13 August 2021) for the Massachusetts buildings dataset.

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

References

1. Huang, X.; Wen, D.; Li, J.; Qin, R. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote Sens. Environ.* **2017**, *196*, 56–75.
2. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sens.* **2018**, *10*, 407.
3. Zhang, Y.; Sun, L. Spatial-temporal impacts of urban land use land cover on land surface temperature: Case studies of two Canadian urban areas. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *75*, 171–181.
4. Wierzbicki, D.; Matuk, O.; Bielecka, E. Polish cadastre modernization with remotely extracted buildings from high-resolution aerial orthoimagery and airborne LiDAR. *Remote Sens.* **2021**, *13*, 611.
5. Song, J.; Tong, X.; Wang, L.; Zhao, C.; Prishchepov, A.V. Monitoring finer-scale population density in urban functional zones: A remote sensing data fusion approach. *Landsc. Urban Plan.* **2019**, *190*, 103580.
6. Xiong, C.; Li, Q.; Lu, X. Automated regional seismic damage assessment of buildings using an unmanned aerial vehicle and a convolutional neural network. *Autom. Constr.* **2020**, *109*, 102994.
7. Hanan, N.P.; Anchang, J.Y. Satellites could soon map every tree on Earth. *Nature* **2020**, *587*, 42–43.
8. Hu, Q.; Zhen, L.; Mao, Y.; Zhou, X.; Zhou, G. Automated building extraction using satellite remote sensing imagery. *Autom. Constr.* **2021**, *123*, 103509.
9. Liu, Z.; Wang, J.; Liu, W. Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform. In Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium, Seoul, Republic of Korea, 29 July 2005; pp. 2250–2253.
10. Chen, Y.; Su, W.; Li, J.; Sun, Z. Hierarchical object oriented classification using very high resolution imagery and LIDAR data over urban areas. *Adv. Space Res.* **2009**, *43*, 1101–1110.
11. Su, W.; Li, J.; Chen, Y.; Liu, Z.; Zhang, J.; Low, T.M.; Suppiah, I.; Hashim, S.A.M. Textural and local spatial statistics for the object-oriented classification of urban areas using high resolution imagery. *Int. J. Remote Sens.* **2008**, *29*, 3105–3117.
12. Jiang, N.; Zhang, J.; Li, H.; Lin, X. Object-oriented building extraction by DSM and very highresolution orthoimages. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2008**, *37*, 441–446.
13. Vu, T.T.; Yamazaki, F.; Matsuoka, M. Multi-scale solution for building extraction from LiDAR and image data. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 281–289.
14. Tarantino, E.; Figorito, B. Extracting buildings from true color stereo aerial images using a decision making strategy. *Remote Sens.* **2011**, *3*, 1553–1567.
15. Wen, Q.; Jiang, K.; Wang, W.; Liu, Q.; Guo, Q.; Li, L.; Wang, P. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. *Sensors* **2019**, *19*, 333.
16. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204.
17. Zhang, G.; Lei, T.; Cui, Y.; Jiang, P. A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 582.
18. 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. [CrossRef]
19. Guo, H.; Shi, Q.; Marinoni, A.; Du, B.; Zhang, L. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sens. Environ.* **2021**, *264*, 112589. [CrossRef]
20. Liu, T.; Yao, L.; Qin, J.; Lu, N.; Jiang, H.; Zhang, F.; Zhou, C. Multi-scale attention integrated hierarchical networks for high-resolution building footprint extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102768.
21. Wei, S.; Zhang, T.; Ji, S.; Luo, M.; Gong, J. BuildMapper: A fully learnable framework for vectorized building contour extraction. *ISPRS J. Photogramm. Remote Sens.* **2023**, *197*, 87–104. [CrossRef]
22. 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, 7–12 June 2015; pp. 3431–3440.
23. 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.

24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
25. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
26. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
27. Shrestha, S.; Vanneschi, L. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
28. Deng, W.; Shi, Q.; Li, J. Attention-gate-based encoder–decoder network for automatical building extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2611–2620. [CrossRef]
29. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An improved network for building extraction from high resolution remote sensing image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
30. Luo, L.; Li, P.; Yan, X. Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies* **2021**, *14*, 7982. [CrossRef]
31. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1194–1206. [CrossRef]
32. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
33. Liu, M.; Chai, Z.; Deng, H.; Liu, R. A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4297–4306. [CrossRef]
34. Hu, K.; Li, M.; Xia, M.; Lin, H. Multi-scale feature aggregation network for water area segmentation. *Remote Sens.* **2022**, *14*, 206. [CrossRef]
35. Liu, Q.; Dong, Y.; Li, X. Multi-stage context refinement network for semantic segmentation. *Neurocomputing* **2023**, *535*, 53–63. [CrossRef]
36. Tian, Q.; Zhao, Y.; Li, Y.; Chen, J.; Chen, X.; Qin, K. Multiscale building extraction with refined attention pyramid networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
37. Wang, H.; Miao, F. Building extraction from remote sensing images using deep residual U-Net. *Eur. J. Remote Sens.* **2022**, *55*, 71–85.
38. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* **2020**, *12*, 1400.
39. Dai, X.; Xia, M.; Weng, L.; Hu, K.; Lin, H.; Qian, M. Multi-Scale Location Attention Network for Building and Water Segmentation of Remote Sensing Image. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5609519.
40. Jiang, Z.; Chen, Z.; Ji, K.; Yang, J. Semantic segmentation network combined with edge detection for building extraction in remote sensing images. In Proceedings of the MIPPR 2019: Pattern Recognition and Computer Vision, Wuhan, China, 14 February 2020; pp. 60–65.
41. Yi, Y.; Zhang, Z.; Zhang, W.; Zhang, C.; Li, W.; Zhao, T. Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. *Remote Sens.* **2019**, *11*, 1774.
42. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
43. Chen, X.; Qiu, C.; Guo, W.; Yu, A.; Tong, X.; Schmitt, M. Multiscale feature learning by transformer for building extraction from satellite images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
44. Wang, Z.; Xu, N.; Wang, B.; Liu, Y.; Zhang, S. Urban building extraction from high-resolution remote sensing imagery based on multi-scale recurrent conditional generative adversarial network. *GIScience Remote Sens.* **2022**, *59*, 861–884. [CrossRef]
45. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
46. 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 Sens.* **2018**, *57*, 574–586.
47. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Toronto, ON, Canada, 2013.
48. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
49. 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 Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
50. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
51. 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.

52. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]
53. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. *Sensors* **2022**, *22*, 2932. [CrossRef]
54. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6169–6181. [CrossRef]
55. Qiu, Y.; Wu, F.; Yin, J.; Liu, C.; Gong, X.; Wang, A. MSL-Net: An efficient network for building extraction from aerial imagery. *Remote Sens.* **2022**, *14*, 3914. [CrossRef]
56. Chen, J.; Zhang, D.; Wu, Y.; Chen, Y.; Yan, X. A context feature enhancement network for building extraction from high-resolution remote sensing imagery. *Remote Sens.* **2022**, *14*, 2276. [CrossRef]
57. Zhou, Y.; Chen, Z.; Wang, B.; Li, S.; Liu, H.; Xu, D.; Ma, C. BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
58. Atik, S.O.; Atik, M.E.; Ipbuker, C. Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. *J. Appl. Remote Sens.* **2022**, *16*, 024510. [CrossRef]
59. Sariturk, B.; Seker, D.Z. Comparison of residual and dense neural network approaches for building extraction from high-resolution aerial images. *Adv. Space Res.* **2023**, *71*, 3076–3089. [CrossRef]
60. Liu, J.; Huang, H.; Sun, H.; Wu, Z.; Luo, R. LRAD-Net: An Improved Lightweight Network for Building Extraction From Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *16*, 675–687. [CrossRef]
61. Liu, Z.; Shi, Q.; Ou, J. LCS: A collaborative optimization framework of vector extraction and semantic segmentation for building extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
62. Chen, S.; Shi, W.; Zhou, M.; Zhang, M.; Xuan, Z. CGSANet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 1526–1542. [CrossRef]
63. Alsabhan, W.; Alotaiby, T. Automatic building extraction on satellite images using Unet and ResNet50. *Comput. Intell. Neurosci.* **2022**, *2022*, 5008854. [CrossRef]
64. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
65. Yin, J.; Wu, F.; Qiu, Y.; Li, A.; Liu, C.; Gong, X. A multiscale and multitask deep learning framework for automatic building extraction. *Remote Sens.* **2022**, *14*, 4744. [CrossRef]
66. Che, Z.; Shen, L.; Huo, L.; Hu, C.; Wang, Y.; Lu, Y.; Bi, F. MAFF-HRNet: Multi-Attention Feature Fusion HRNet for Building Segmentation in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1382. [CrossRef]
67. Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An efficient building extraction method from high spatial resolution remote sensing images based on improved mask R-CNN. *Sensors* **2020**, *20*, 1465. [CrossRef]