remote sensing

MDPI

Article

Recurrent Residual Deformable Conv Unit and Multi-Head with Channel Self-Attention Based on U-Net for Building Extraction from Remote Sensing Images

Wenling Yu [1,2], Bo Liu [1,2,3,*], Hua Liu [1,2,3] and Guohua Gou [4]

[1] School of Surveying and Geoinformation Engineering, East China University of Technology, Nanchang 330013, China; yuwenling@ecut.edu.cn (W.Y.); liuhua@ecut.edu.cn (H.L.)
[2] Key Laboratory of Mine Environmental Monitoring and Improving around Poyang Lake of Ministry of Natural Resources, East China University of Technology, Nanchang 330013, China
[3] Jiangxi Province Engineering Research Center of Surveying, Mapping and Geographic Information, Nanchang 330025, China
[4] State Key Laboratory Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430070, China; guohua.gou@whu.edu.cn
* Correspondence: liubo@ecut.edu.cn

Abstract: Considering the challenges associated with accurately identifying building shape features and distinguishing between building and non-building features during the extraction of buildings from remote sensing images using deep learning, we propose a novel method for building extraction based on U-Net, incorporating a recurrent residual deformable convolution unit (RDCU) module and augmented multi-head self-attention (AMSA). By replacing conventional convolution modules with an RDCU, which adopts a deformable convolutional neural network within a residual network structure, the proposed method enhances the module's capacity to learn intricate details such as building shapes. Furthermore, AMSA is introduced into the skip connection function to enhance feature expression and positions through content–position enhancement operations and content–content enhancement operations. Moreover, AMSA integrates an additional fusion channel attention mechanism to aid in identifying cross-channel feature expression Intersection over Union (IoU) score differences. For the Massachusetts dataset, the proposed method achieves an Intersection over Union (IoU) score of 89.99%, PA (Pixel Accuracy) score of 93.62%, and Recall score of 89.22%. For the WHU Satellite dataset I, the proposed method achieves an IoU score of 86.47%, PA score of 92.45%, and Recall score of 91.62%, For the INRIA dataset, the proposed method achieves an IoU score of 80.47%, PA score of 90.15%, and Recall score of 85.42%.

Keywords: building extraction; remote sensing; recurrent residual convolution; U-Net; multi-head self-attention

Citation: Yu, W.; Liu, B.; Liu, H.; Gou, G. Recurrent Residual Deformable Conv Unit and Multi-Head with Channel Self-Attention Based on U-Net for Building Extraction from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5048. https://doi.org/10.3390/rs15205048

Academic Editors: Lefei Zhang, Tao Lei, Asoke K. Nandi and Tao Chen

Received: 31 July 2023
Revised: 18 October 2023
Accepted: 19 October 2023
Published: 20 October 2023

1. Introduction

Building extraction from remote sensing images constitutes a paramount application in the realm of remote sensing image analysis. In the spheres of urban planning, environmental monitoring, and natural disaster assessment, among others, edifices embody the most fundamental spatial entities, thereby endowing the accurate extraction of building information with profound significance for these diverse applications. In its nascent phase, building extraction from remote sensing images relied upon manual visual interpretation, albeit manifesting inefficiency and excessive reliance on human experiential knowledge. Subsequently, traditional building extraction methods hinged upon color, texture, and other salient attributes in the original image, facilitating image segmentation. The primary approaches encompassed the edge-based segmentation algorithm [1–3], region-based segmentation algorithm [4], and texture-based extraction algorithm [5]. Yet, with the

unremitting advancement of society, the exigency for timely updates in geographic information systems necessitates the pursuit of proficient and automatic building extraction from remote sensing images, a formidable challenge. The advent of deep learning technology, in particular, convolutional neural networks, has witnessed widespread adoption in building extraction methods, yielding commendable outcomes [6–9]. Owing to its capacity for autonomous feature learning, sans manual intervention, this approach exhibits remarkable adaptability across diverse scenarios and datasets.

Convolutional neural networks have made some strides in the domain of building extraction from remote sensing images, albeit being encumbered by certain drawbacks, notably the unwieldy computational parameters and suboptimal outcomes. The advent of FCNs (fully convolutional networks) [10–12] has ushered in a new realm of possibilities for leveraging deep learning in the context of building extraction from remote sensing images. This innovative approach employs a deconvolution layer, encompassing upsampling and convolutional operations, to supplant the conventional fully connected layer within CNNs, thereby enabling pixel-wise predictions across an image. The advent of FCNs establishes a fundamental framework for semantic segmentation methods grounded in deep learning, delineated by the encoding–decoding structure. Notably, Emmanuel et al. [13] introduced FCNs as a means of extracting buildings from remote sensing images, yielding superior results compared to prior algorithms, thereby underscoring the feasibility of FCNs in the realm of remote sensing image applications. Moreover, the exclusion of the fully connected layer in an FCN contributes to a reduction in the computational parameters. However, despite these advancements, the building extraction results attained using FCNs still leave room for enhancement, inciting ongoing research to bolster FCNs' efficacy in this domain. This quest has led to the emergence of a plethora of akin encoding–decoding structures. For instance, SegNet [14] leverages the initial convolutional layer of the VGG network for its encoding structure, coupled with a multi-layer upsampling architecture in the decoder, culminating in a finely detailed pixel-wise classification map. Likewise, U-Net [15] capitalizes on an encoding structure to extract features layer by layer, followed by the restoration of feature maps in a progressive manner. Diverging from FCNs, U-Net integrates feature maps from the encoding structure into the decoder using skip connections, and employs deconvolution to progressively enlarge feature maps within the decoding structure. The outcome is a feature map that encompasses both low-dimensional feature details and high-dimensional feature abstractions, thereby harnessing spatial context information to enhance feature extraction accuracy and demonstrating remarkable performance in the task of building extraction from remote sensing images. It also performs well in other similar image segmentation tasks.

Recently, most deep-learning-based building extraction from remote sensing images has been improved based on the encoder–decoder structure. Some methods replace the convolutional module for feature extraction to capture more building feature information. DeepLabV3 [16–18] replaces ordinary convolution with atrous convolution and uses the ASPP (Atrous Spatial Pyramid Pooling) [19] module to capture multi-scale information. Wang et al. [20] introduced multi-scale recursive residual convolution into the encoding–decoding layer and used an attention mechanism to enhance the information interaction between features. Dixit et al. [21] replaced the convolutional module of U-Net with Dilated-ResUnet, and through extensive experiments, studied deep learning hyperparameters such as optimizers and activation functions, and applied image enhancement using high-boost filters to input satellite images to further improve the fine details of building boundaries. Chen et al. [22] used the encoding–decoding structure as the backbone network and used a dense connected convolutional neural network (DCNN) [23] and residual network (ResNet) [24] to obtain the global features and local detail features contained in remote sensing images. EfficientUNet+ [25] uses UNet as the basic architecture, adopts EfficientNet-b0 as the encoder, and embeds spatial and channel squeeze excitation (scSE) [26] into the decoder to further improve the accuracy and speed of model extraction. Guo et al. [27]

designed an edge preservation neural network (EPUNet) that combines edge detection with contextual aggregation in the proposed SG-EPUNet framework.

In addition to changing the convolutional module used for feature extraction, some methods also make improvements at the time of feature output. Ji et al. [28] combined the feature pyramid network [29] with the encoding–decoding structure model to reduce the impact of inconsistent building scales in remote sensing images on the extraction results and improve the building extraction accuracy. Shi et al. [30] integrated a graph convolutional network (GCN) [31] and deep structured feature embedding (DSFE) [32] into an end-to-end workflow and performed well on building extraction tasks. Chen et al. [33] designed an efficient dual-pathway transformer structure (Sparse Token Transformers, STT) that learns the long-term dependency of tokens in both their spatial and channel dimensions and achieves state-of-the-art accuracy on benchmark building extraction datasets. Unlike existing multi-scale feature extraction strategies, MAP-Net [34] learns multi-scale features with spatial localization preservation using multiple parallel paths, where high-level semantic features are extracted at a fixed resolution, gradually generated at each stage. Then, the attention module adaptively squeezes the channel features extracted from each path for optimized multi-scale fusion, and the pyramid spatial pooling module captures global dependencies to refine discontinuous building outlines. Although encoder–decoder-based methods have achieved many successes in building extraction from remote sensing images, when applied to automatic building extraction from high-resolution remote sensing images, there are still problems in accurately identifying the shape features of buildings, and the acquired building features cannot be fully utilized, resulting in confusion between building features and non-building features, so they are ultimately unable to accurately extract building features from remote sensing images.

In this paper, we introduce a pioneering building extraction network, presenting an innovative approach to extracting buildings from remote sensing images. The proposed method leverages a recurrent residual deformable convolution unit (RDCU), eschewing the conventional convolution module, thereby heightening the model's acumen in discerning building shape features. Additionally, we deploy augmented multi-head self-attention (AMSA), integrated into the skip connection module, to yield feature maps of the encoding layer. This ingenious incorporation empowers the model with enhanced expression capabilities for building features, facilitated by the global context information of features. Moreover, the channel feature enhancement component augments the model's sensitivity to building features, resulting in a more precise identification of buildings amidst intricate landform environments. The main contributions of this paper are as follows:

1. We present a novel neural network tailored to building information extraction, drawing upon a recurrent residual deformable convolution unit and multi-head channel self-attention, incorporated into the U-Net architecture. This composite structure adeptly extracts and aggregates building features, enriches the representation of building detail features, and significantly elevates pixel-level building extraction accuracy.
2. We introduce a pioneering network module, the recurrent residual deformable convolution module. This module seamlessly integrates the merits of a recurrent residual structure and deformable convolution operations, effectively capturing comprehensive building features and amplifying the model's aptitude for recognizing building shape features.
3. A multi-head channel self-attention module is meticulously crafted to fine-tune the weightage of each feature information during the feature fusion process. This results in the retention of more pertinent information conducive to the segregation of buildings from the background, while effectively filtering out irrelevant noise information. By doing so, we ensure the model's efficient utilization of feature information, thus enhancing the coherency of building segmentation.

The rest of this paper is organized as follows: Section 2 introduces the architecture of the proposed method. Section 3 reports experimental results. Section 4 gives conclusive opinions.

2. Methodology

The architecture of the proposed method is illustrated in Figure 1. It constitutes a marked enhancement over the commonly utilized U-Net. Within this novel approach, the encoding layers acquire profound features by progressively downsizing the resolution from high to low. At each encoding layer, the feature map is halved, paving the way for its subsequent restoration to high resolution and output through the decoding layers. Instead of traditional convolutional modules, the RDCU modules are embraced, exuding the capability to enhance the preservation of geometric features while ameliorating challenges like gradient vanishing during deep learning model training. Thus, the loss of architectural feature information is mitigated, bolstering the model's adeptness at preserving architectural features during network learning. Furthermore, to further bolster the accuracy of building extraction, the AMSA module is introduced during the fusion of deep high-level features and shallow low-level features within the skip connection linking the encoding and decoding layers. The AMSA module adroitly addresses the issue of erroneous building feature output during skip connections, harnessing the global contextual information of features to enrich content–position and content–content relationships. Additionally, it incorporates a channel attention mechanism, elevating feature expression and capturing pivotal channel features, thereby facilitating the discrimination of feature expression disparities among channels. Consequently, this module empowers the model with heightened sensitivity to building features and amplifies its capacity to effectively represent them. The feature map within the encoding structure seamlessly converges with the feature map in the decoding structure after traversing through the AMSA module. The amalgamated feature map is then upsampled using the RDCU module, in an iterative process, culminating in the complete restoration of the image to its original size. Lastly, the output is artfully obtained through the sigmoid activation function

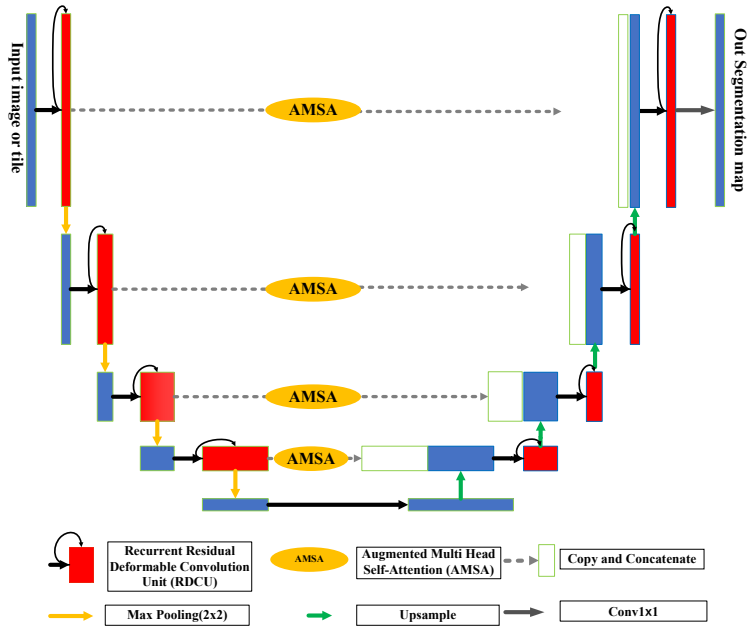


Figure 1. The framework of the proposed method. The encoder–decoder uses RDCU module for feature extraction. The skip connection part is fused after passing through AMSA module, and outputs building segmentation mask after the classifier.

2.1. RADU Module

The conventional convolutional modules typically consist of convolution and ReLU activation functions. However, when the network structure becomes increasingly complex, network training often encounters challenges like gradient explosion. To address this issue, ResNet emerges as an effective solution, significantly ameliorating the problem of network

degradation. ResNet adopts a residual structure, incorporating skip connections within its internal residual blocks. This strategic approach effectively deepens the network structure while circumventing issues such as gradient vanishing. In this regard, the accuracy of building semantic segmentation is heightened by reformulating the convolution module in both the encoding and decoding layers. Notably, ResU-Net [35] adopts a transformative approach by replacing the traditional convolution module with ResNet. Figure 2a exemplifies the fundamental convolution module of U-Net's encoding and decoding layers, while Figure 2b illustrates the convolution unit employed by ResU-Net.

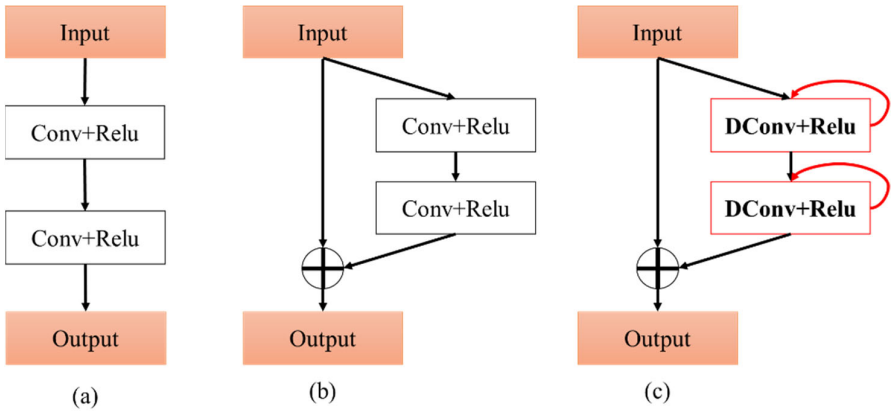


Figure 2. Different variants of the convolutional units including (**a**) the forward convolutional unit, (**b**) the ResNet block, and (**c**) the RDCU.

The fundamental convolution module and ResNet module utilize feature stacking to further expand the depth of the network structure. However, conventional convolutional neural networks struggle to precisely capture intricate edge details and other essential building features. Taking inspiration from networks like ResNet, we introduce the RDCU module. This innovative module incorporates features into the input features following two residual sub-modules, facilitating the effective accumulation of features and ensuring a more robust and potent feature representation.

The architecture of the RDCU sub-module is illustrated in Figure 3. The RDCU sub-module can help the RDCU module improve the expression ability of the building shape and highlight the detail features of the building. In the sub-module of the RDCU, the input feature map x is fused with the original input feature map after the deformable convolution operation, and the feature map x_1 is obtained. x_1 is fused with x after the deformable convolution operation to obtain x_2. The sub-module calculation formula of the RDCU is expressed:

$$x_t = x + D_t(x_{t-1}, w_t), \tag{1}$$

where $D_t(x_{t-1}, w_t)$ denotes the feature output after t times of deformable convolution operations. The output sub-module of the RDCU is fed to the standard ReLU activation function f and is expressed:

$$F(x_l, w_l) = max(0, x_l), \tag{2}$$

where $F(x_l, w_l)$ represents the outputs from of l layer of the RDCU unit. Let us consider that the output of the RDCU is X_l and can be calculated as follows:

$$X_l = x + F(x_l, w_l), \tag{3}$$

where x represents the input samples of the RDCU. The X_l represents the the input feature map for the immediate succeeding sub-sampling or up-sampling layers in the encoding and decoding convolutional units of the proposed method.

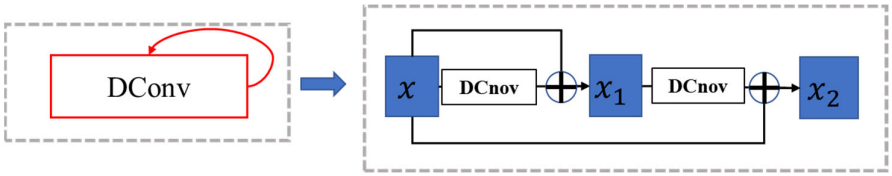


Figure 3. The submodule of RDCU. The DCnov in the figure is denoted as deformable convolution.

2.2. AMSA Module

To maximize the effective utilization of acquired features and bolster the model's proficiency in discriminating between building and non-building features, we introduce an AMSA module after the cyclic residual convolution module in each layer of the feature map within the encoding layer. The AMSA module harnesses the global context information of features to execute content–position enhancement operations and content–content enhancement operations between features [36]. It additionally integrates a channel attention mechanism, thereby endowing the model with the capability to not only enhance feature expression and feature positions, facilitating mutual enhancement of similar features at different positions, but also discern disparities in feature expression across channels. This intelligent mechanism automatically prioritizes crucial channel features, acquiring effective building feature information and harnessing the full potential of the acquired features. This strategic augmentation elevates the model's acumen in detecting building features, while fortifying its ability to express them coherently. The specific structure of the AMSA module is visually presented in Figure 4.

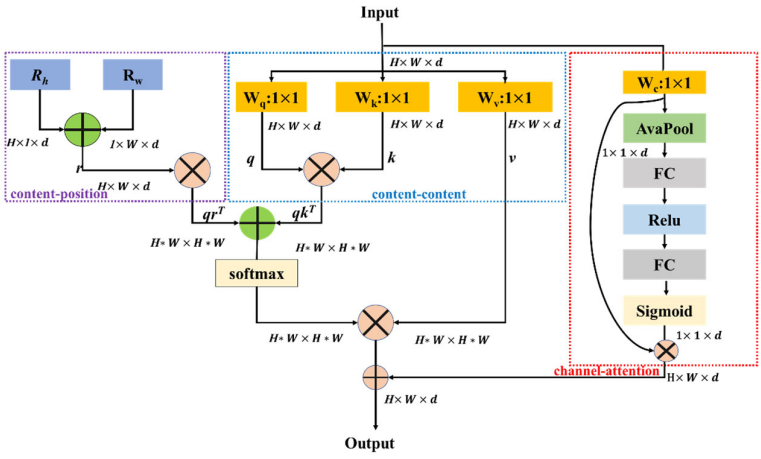


Figure 4. The structure of AMSA. Attentional weights are calculated at the channel and spatial dimensions.

In Figure 4, H represents the length of the feature map, W represents the width of the feature map, and d represents the number of channels in the feature map. The specific operations of AMSA are as follows: (1) The enhancement operation between feature positions is as follows: After the position encoding of the length (R_h) and width (R_w) of the input image, channel expansion is performed to obtain the feature maps $H \times 1 \times d$ and $1 \times W \times d$, and the features are added to obtain a feature map r of size $H \times w \times d$. After transposing the feature map r, it is multiplied by the feature map q to enhance the expression of features at different positions. (2) The specific operation of enhancing between features is as follows: After the feature map is input, 1×1 convolution operations are performed to obtain feature maps q, k, and v. The transpose of feature map q and feature map k are multiplied by the features to enhance the expression between similar features. (3) Then, the feature map qr^T output from the operation between feature positions is added to the feature map qk^T output from the enhancement operation between features, activated with a soft-max function, and multiplied by the feature map to reduce the loss of original feature

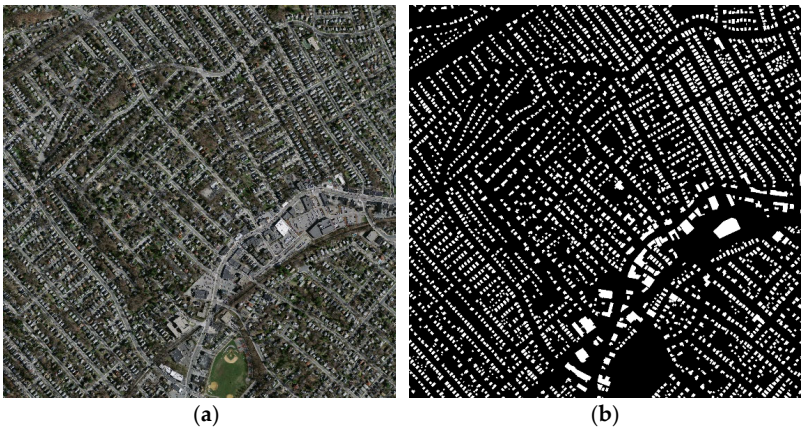
information. (4) The channel attention operation is as follows: After inputting the feature map and performing convolution, it is input into a Relu activation function after global average pooling and fully connected operations, and then the fully connected operations and Sigmoid activation are performed again to obtain the channel feature weights. Finally, the channel feature weights are multiplied by the convolved feature map. (5) Feature fusion module: the feature map obtained from the channel attention operation is fused with the other feature enhancement module's feature maps, and a feature map with the same size as the original input's feature map is obtained.

3. Experiments and Results

3.1. Dataset Details

3.1.1. Massachusetts Dataset

The Massachusetts dataset was proposed by Dr. Mnih [37], covering an area of about 340 square kilometers, including the urban, suburban, and dock areas of Boston, with a spatial resolution of 1 m. It consists of 151 three-channel TIFF format remote sensing images with pixel sizes of 1500 × 1500 and their corresponding label images, of which 137 remote sensing images are used for training, 4 remote sensing images are used for validation, and 10 remote sensing images are used for testing. The building density in the Massachusetts dataset is large, the roof colors are different, each building occupies not many pixels, and some images are shown as shown. In order to better adapt to the experimental environment, first, remove the images with blank spaces, convert the images into single-band grayscale images, and convert the data format into JPG format. Due to the large size of the remote sensing images and limited computer memory, it is necessary to perform data cropping, rotation, mirroring along the y-axis, mean filtering, salt and pepper noise enhancement, Gaussian noise enhancement, and other preprocessing operations on the Massachusetts dataset. After preprocessing operations, the training set in the processed dataset has 8352 images and their label maps, the validation set has 288 images and their label maps, and the test set has 720 images and their label maps. The pixel size is all 512 × 512. A sample case from the Massachusetts dataset is shown in Figure 5.



(**a**) (**b**)

Figure 5. An example of the Massachusetts building dataset. (**a**) is the example of the original image, and (**b**) is the label of (**a**).

3.1.2. WHU Satellite Dataset I

The WHU Satellite dataset I [28], open-sourced by the team of Ji Shunping from Wuhan University, contains a collection of 204 remote sensing image pairs, each consisting of a 512 × 512 image and its corresponding label image. It includes images from different sensors of ZY-3, IKONOS, and Worldview series satellites with spatial resolutions ranging from 0.3 m to 2.3 m. The images cover different urban areas in Europe, China, North and South America, and Africa. The remote sensing image types are diverse and can effectively test the robustness of building extraction algorithms, making it very challenging. Some

examples of dataset I are shown. After preprocessing the dataset, we obtained a total of 20,094 images, each with dimensions of 256 × 256. Finally, the processed dataset is divided into training, validation, and test sets in an 8:1:1 ratio, with 16,076 images in the training set, 2009 images in the validation set, and 2009 images in the test set. Building images from four different regions are shown in Figure 6.

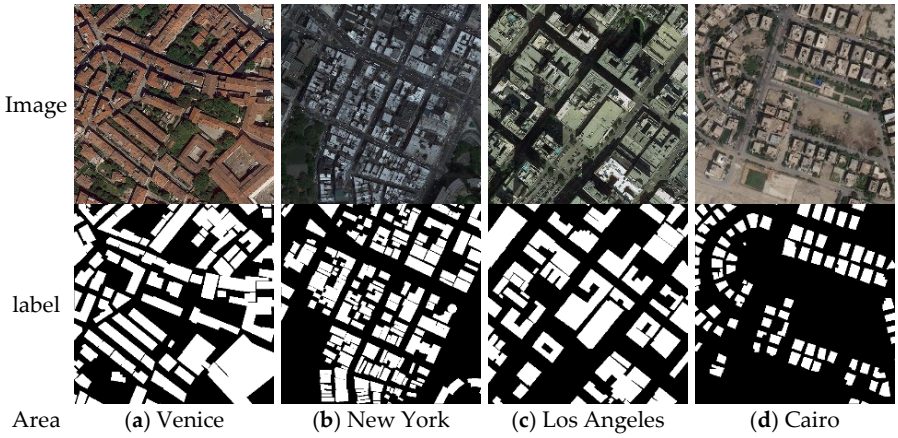


Figure 6. Building images from four different regions in the Massachusetts building dataset: (**a**) a building image from Venice, (**b**) a building image from New York, (**c**) the building image from Los Angeles, and (**d**) a building image from Cairo.

3.1.3. INRIA Dataset

The INRIA dataset [38] consists of 360 aerial images covering a wide range of regions, from densely populated areas (such as the financial district in San Francisco) to mountain towns (such as Lienz in Tyrol, Austria). Each image is 5000 × 5000 visible spectrum data with a ground resolution of 0.3 m. The dataset contains 180 images for training and 180 for testing. Since the reference annotations for the test set are not publicly available, we cropped the original images into 512 × 512 patches and divided them into training, validation, and test sets with a ratio of 6:2:2. The final dataset partition consisted of a training set with 12,960 images, a validation set of 4320 images, and a test set of 4320 images. Building images from five different regions are shown in Figure 7.

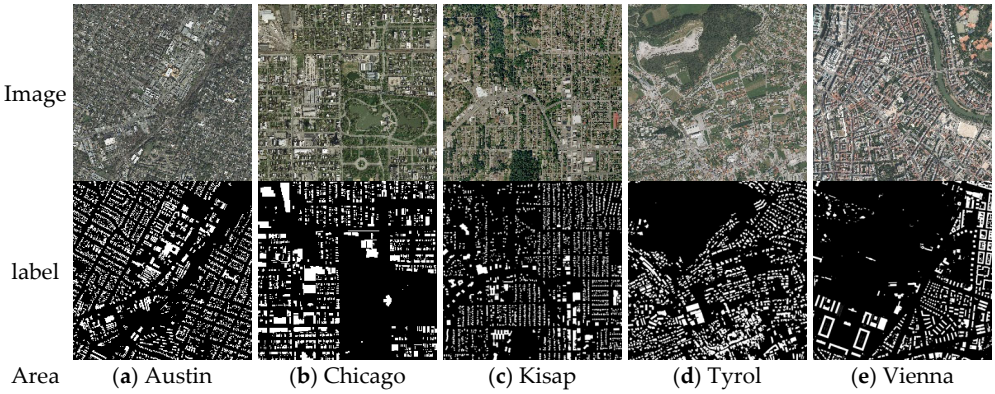


Figure 7. Building images from five different regions in the Massachusetts building dataset: (**a**) a building image from Austin, (**b**) a building image from Chicago, (**c**) a building image from Los Kisap, (**d**) a building image from Tyrol, and (**e**) a building image from Vienna.

3.2. Experimental Settings

In the experiment, the models used for comparison include U-Net, DeepLab v3+, ResU-Net, EPUNet, and STT. Both ResU-Net and EPUNet are semantic segmentation models improved based on U-Net and have shown superior performance in many remote

sensing images. DeepLab v3+ is a relatively classic semantic segmentation model. SST is a promising model because of using transformers for efficient building extraction. To ensure the fairness of the experiment, the proposed method was trained in the same experimental environment as the comparison models and used the same training parameters. According to our existing experimental environment, based on the scale of the dataset and repeated experimental results, the final number of epochs in the experiment with the Massachusetts dataset is 18, with a batch size of 2, the number of epochs in the experiment with the WHU Satellite dataset I is 15, with a batch size of 2, and the number of epochs in the experiment with the INRIA dataset I is 25, with a batch size of 2. All experiments use the common BCELoss as the loss function, with the Adam optimizer and a learning rate set to 0.0001. All experiments were conducted under the Windows 10 operating system using the Pytorch deep learning framework, with the GPU being NVIDIA Quadro RTX and the CPU being Intel® Core™ i7-10700. The experimental environment parameters are shown in Table 1.

Table 1. Configuration of the experiment.

Name	Configuration	Versions
OS	Windows 10	N/A
GPU	NVIDIA Quadro RTX	Quadro RTX 4000
VRAM	8 G	N/A
RAM	64 G	N/A
Programming Language	Python	3.5
Deep Learning Framework	PyTorch	3.6
LIBS	NumPy, PIL, torch, sklearn, opencv, tqdm, torch, torchvision, tensorboardX	N/A
IDE	PyCharm	PyCharm2019

Note: "N/A: not applicable".

In order to more intuitively display the experimental results of our method, we have selected four relatively common accuracy evaluation indicators, namely Intersection over Union (IOU), Pixel Accuracy (PA), Recall, and mean Pixel Accuracy (mPA). These three classic evaluation indicators are used to comprehensively analyze the experimental results, with the IOU and PA as the main evaluation indicators and mPA and Recall as the auxiliary evaluation indicators. The IOU is a standard measure of semantic segmentation and is widely used. It calculates the ratio of the intersection and union of two sets, the ground truth, and the predicted segmentation (Intersection over Union, IOU). PA is a common simple measure in semantic segmentation, which represents the proportion of correctly predicted pixels to the total number of pixels. mPA is an improved semantic segmentation measure based on PA, which calculates the average value of PA for each category. Recall represents the proportion of samples with a predicted value of 1 and a true value of 1 among all samples with a true value of 1. The specific calculation formula is as follows:

$$\mathrm{IoU} = \frac{p_{ii}}{\left(\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n}\left(p_{ij} - p_{ii}\right)\right)}, \tag{4}$$

$$\mathrm{PA} = \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} P_{ij}}, \tag{5}$$

$$\mathrm{mPA} = \frac{PA}{n+1} = \frac{\sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij}}}{n+1}, \tag{6}$$

$$\mathrm{Recall} = \frac{p_{ii}}{p_{ii} + p_{ij}}, \tag{7}$$

$n + 1$ represents the number of categories of n target objects, with 1 background class, p_{ii} represents the total number of pixels with actual type i and predicted type i, and p_{ij} represents the total number of pixels with actual type i and predicted type j.

3.3. Comparisons and Analysis

3.3.1. Quantitative Results

The accuracy evaluation results of the Massachusetts dataset experiment results are shown in Table 2. It is obvious from Table 2 the proposed method model has significantly improved in all accuracy indicators compared to DeepLab v3+, U-Net, ResU-Net, EPUNet, and STT. For the IOU, it is 4.96%, 5.06%, 3.04%, 1.27%, and 1.04% higher, respectively; for PA, it is 5.50%, 6.4%, 4.48%, 2.55%, and 1.51% higher, respectively; for mPA, it is 5.35%, 6.55%, 3.12%, 1.63%, and 1.27% higher, respectively; and for Recall, it is 6.1%, 3.89%, 2.57%, 1.49%, and 0.73% higher, respectively. Table 3 shows the accuracy evaluation of the WHU Satellite dataset I experiment results. It is obvious from Table 3 the proposed method model has significantly improved in all accuracy indicators compared to DeepLab v3+, U-Net, ResU-Net, EPUNet, and STT. For the IOU, it is 3.44%, 4.94%, 2.52%, 1.26%, and 0.84% higher, respectively; for PA, it is 3.22%, 2.94%, 1.74%, 1.12%, and 0.73% higher, respectively; for mPA, it is 4.58%, 4.74%, 3.35%, 2.06%, and 1.84% higher, respectively; and for Recall, it is 3.5%, 3.73%, 1.97%, 1.56%, and 0.95% higher, respectively. Table 4 shows the accuracy evaluation of the INRIA dataset experiment results. It is obvious from Table 4 the proposed method model has also significantly improved in all accuracy indicators compared to DeepLab v3+, U-Net, ResU-Net, EPUNet, and STT. For the IOU, it is 6.20%, 6.89%, 3.44%, 1.52%, and 1.34% higher, respectively; for PA, it is 6.76%, 7.12%, 5.03%, 3.78%, and 1.83% higher, respectively; for mPA, it is 6.00%, 6.94%, 4.58%, 2.72%, and 1.07% higher, respectively; and for Recall, it is 7.03%, 7.94%, 4.47%, 3.39%, and 1.07% higher, respectively

Table 2. The accuracy of the different building detection methods in the Massachusetts dataset.

Method	IOU	PA	mPA	Recall
DeepLapv3+	85.03%	88.12%	86.24%	83.12%
U-Net	84.93%	87.22%	85.04%	85.33%
ResU-Net	86.95%	89.14%	88.47%	86.65%
EPUNet	88.72%	91.07%	89.96%	87.73%
STT	88.95%	92.01%	90.32%	88.49%
The proposed method	**89.99%**	**93.62%**	**91.59%**	**89.22%**

Table 3. The accuracy of the different building detection methods in the WHU Satellite dataset I.

Method	IOU	PA	mPA	Recall
DeepLapv3+	83.03%	89.23%	87.24%	88.12%
U-Net	81.53%	89.54%	87.08%	87.89%
ResU-Net	83.95%	90.71%	88.47%	89.65%
EPUNet	85.21%	91.33%	89.76%	90.06%
STT	88.95%	92.01%	90.32%	88.49%
The proposed method	**86.47%**	**92.45%**	**91.82%**	**91.62%**

Table 4. The accuracy of the different building detection methods in the INRIA dataset.

Method	IOU	PA	mPA	Recall
DeepLapv3+	74.27%	83.39%	80.26%	78.39%
U-Net	73.58%	83.03%	79.32%	77.48%
ResU-Net	77.03%	85.12%	81.68%	80.95%
EPUNet	78.92%	86.37%	83.54%	82.03%
STT	79.31%	88.32%	85.19%	84.35%
The proposed method	**80.47%**	**90.15%**	**86.26%**	**85.42%**

3.3.2. Qualitative Results

In order to reflect the advantages of the proposed method, we selected some extracted results from the test set in the experiment for local feature comparison, as shown in Figure 8. The first column is the suburban boundary image, which not only contains small residential building areas but also irregular suburban buildings and some forest areas. From the result map, it can be seen that other comparison methods cannot accurately identify and there are cases of missed detection and missed detection. The proposed method performs well on irregular buildings and can completely identify the irregular shape of the building. The second column image is a residential industrial mixed area with a variety of buildings, most of which are medium and large buildings, and a small part of which are small buildings. Although the comparison method can identify most buildings, it can not accurately identify the detailed features of the building. The proposed method can not only accurately extract the detailed features of the building but can also extract smoother building edges and less noise. The third column image shows that buildings are distributed around water bodies, with some water bodies and a relatively small proportion of buildings. Due to the similarity between the texture features of the dock ground and the top surface features of the building, there is a large interference from non-building features during building extraction, so most of the model building extraction effects are not ideal. Compared with the comparison methods, the proposed method can identify the most buildings, some buildings close to water bodies can be accurately identified, and the interior of the building is more complete. The fourth column image is a suburban remote sensing image, mostly forest land, and the rest are roads and houses. When the texture information of the house and the surrounding environmental features (such as forest land) are low in similarity, all building extraction methods can detect buildings well, but only the proposed method can perfectly display the detailed parts of a building. After comparative analysis, the proposed method has a good overall recognition of buildings and recognition of building shape features.

Figure 9 shows the local feature comparison of some image results of the test set. The proposed method model has a high sensitivity to building features in various complex environments, and the situation of missed detection of buildings has been improved. The specific performance is as follows: (1) In the first column, the buildings are arranged neatly and are of the same size. Compared with other comparison methods, the proposed method model can accurately identify buildings and reduce the missed detection of buildings. (2) The buildings in the second column are of varied sizes and have texture features similar to the non-building features on the ground. Compared with other methods, the proposed method model can identify buildings in a noisy background well. (3) The building features in the third column are blurred, and the building shapes are diverse. Compared with other methods, the proposed method model can identify the shape of the building and clearly express the relationship between buildings. (4) Most of the buildings contained in the fourth column are medium and large buildings, which are distributed chaotically and have similar texture features to non-building features. Therefore, most methods have large noise and other situations, while the proposed method model can not only identify the outline features of buildings but also has relatively less noise.

Figure 10 presents the partial test results of the INRIA dataset: (1) In the first column, the original remote sensing image depicts buildings of various sizes and arrangements; compared to the contrastive methods, the proposed method accurately captures the shapes of the buildings, and successfully identified smaller buildings with minimal interference from non-building features. (2) The second column's original remote sensing image contains very large buildings, with some non-building objects (e.g., ground and cars) sharing similar features. Many contrastive methods fail to identify these buildings entirely. In contrast, the proposed method can effectively extract the majority of these buildings, ensuring high completeness in the extracted structures despite the challenging conditions. (3) The images in the third column feature diverse types of buildings; the proposed method shows minimal sensitivity to noise interference during building extraction. (4) In the fourth column, the original remote sensing image contains numerous irregularly shaped buildings.

In comparison to the contrastive methods, the proposed method has high performance in extracting highly realistic building shapes and intricate details.

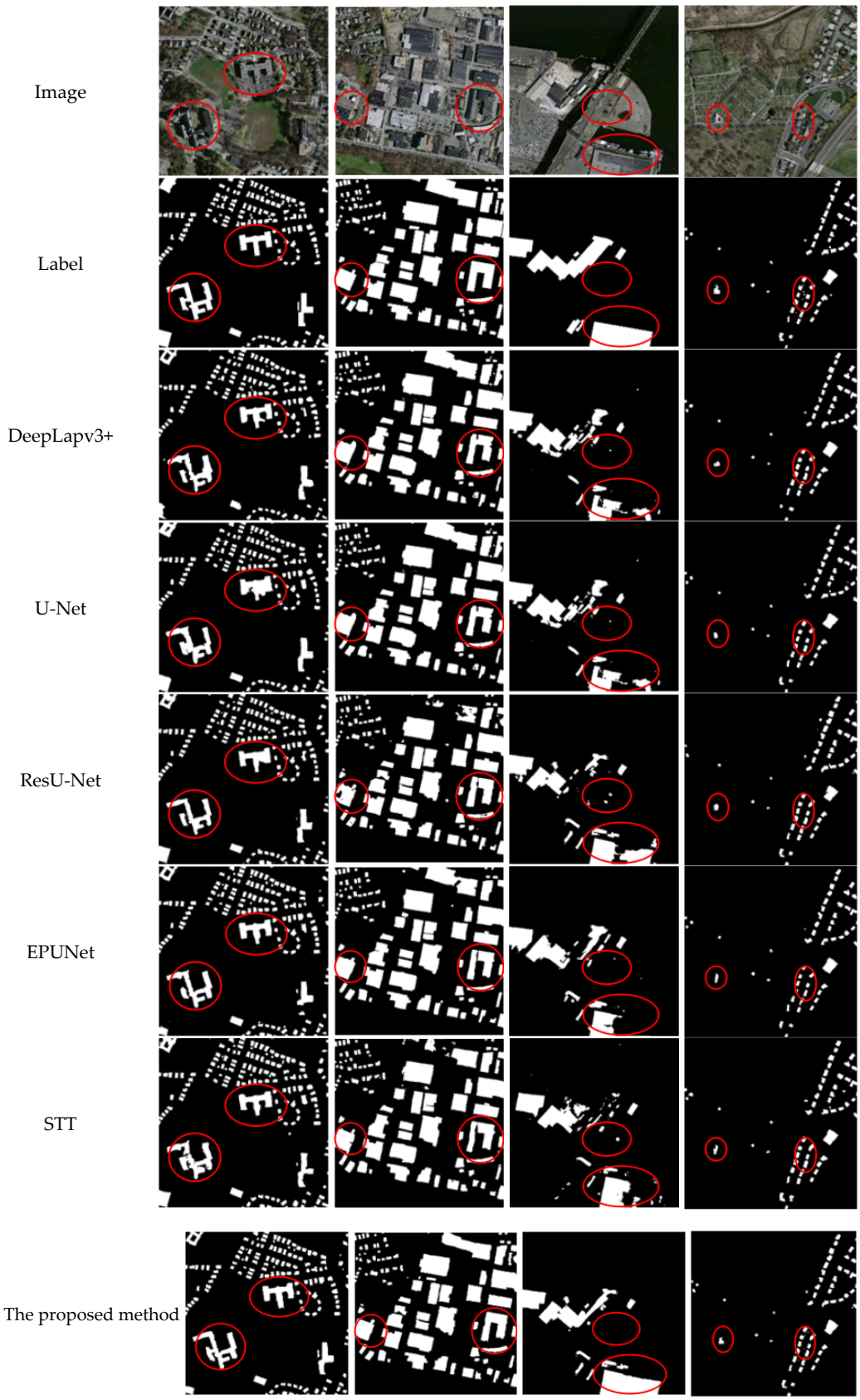


Figure 8. The local results of building detection using different methods in the Massachusetts dataset.

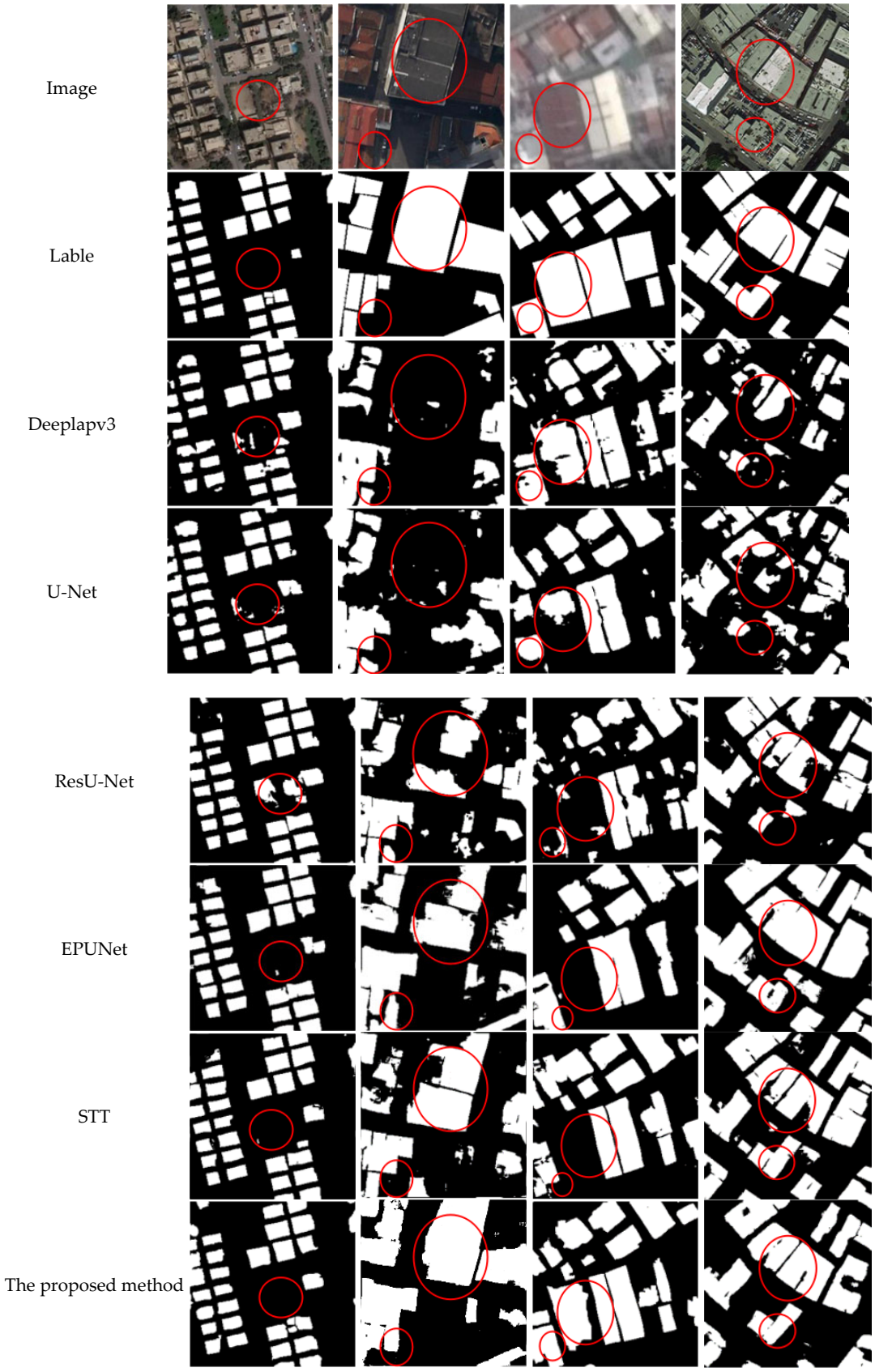


Figure 9. The local results of building detection using different methods in the WHU Satellite dataset I.

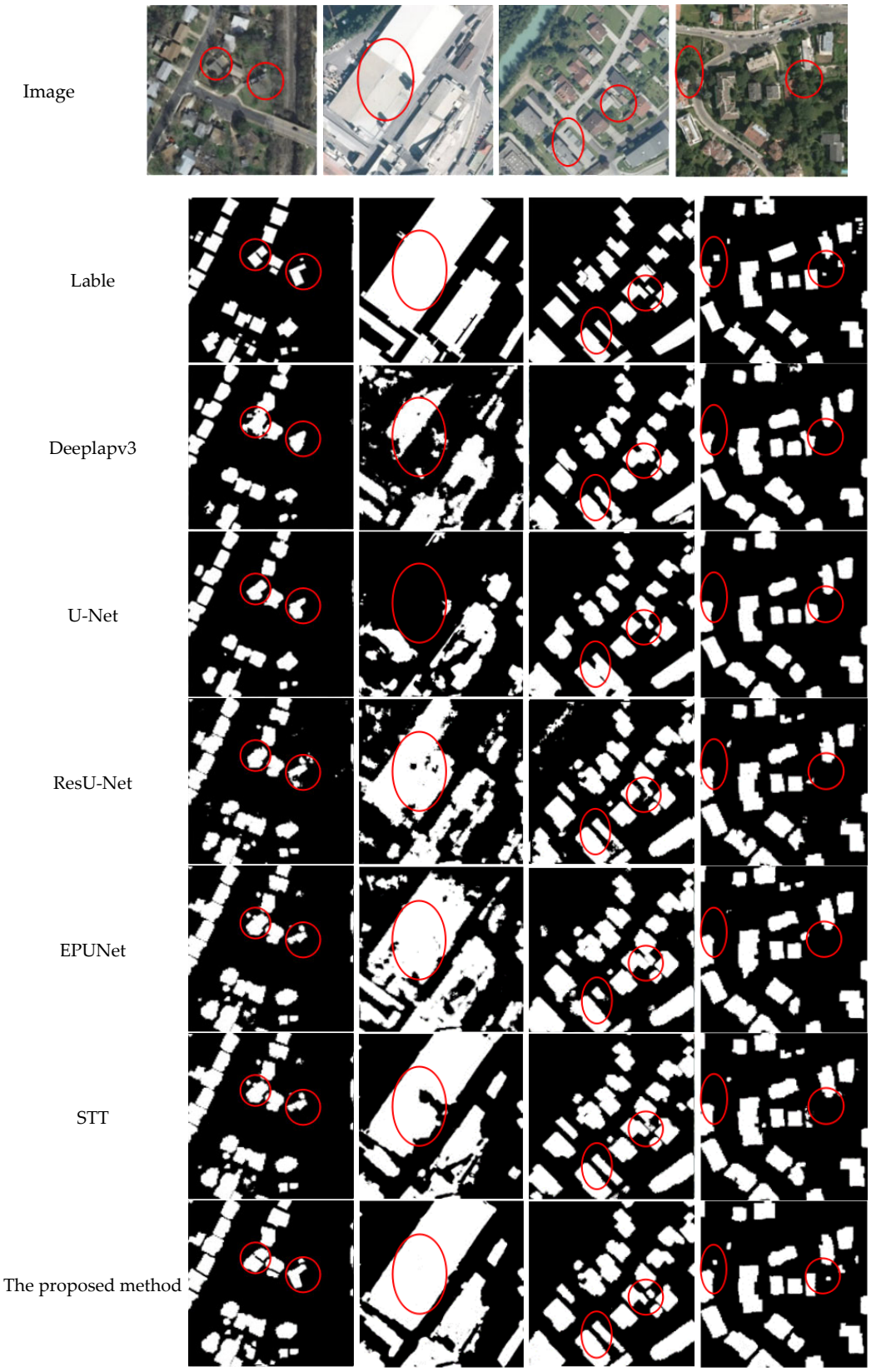


Figure 10. The local results of building detection using different methods in the INRIA dataset.

It can be seen from the results of the three datasets shown in Figures 8–10 that this proposed method can accurately extract the building shapes and detailed features, with less interference from non-building objects. The strong performance of the proposed

method benefited from the RDCU module and the AMSA module, which strengthens the representation of building shapes and details while suppressing the influence of non-building features.

3.4. Ablation Study

3.4.1. Dataset Details and Experimental Settings

In order to demonstrate the impact of different modules in the proposed method model on overall network performance, we chose the WHU Aerial imagery dataset [28] for ablation experiments, as shown in Figure 11. The data are located in Christchurch, New Zealand. As shown in Figure 11, the training area in this dataset is located in area ①, the validation area is located in area ②, and the test area is located in areas ③ and ④, covering a total of 220,000 buildings. The spatial resolution of the image is 0.075 m. The data are later downsampled to 0.3 m and the image is cropped to a size of 512 × 512. After preprocessing, all images are 256 × 256 in size, with 18,944 training images, 4144 validation images, and 9664 test images. The number of training iterations for the ablation experiment is 18, with a batch size of 2. The commonly used BCELoss is used as the loss function, and the Adam optimizer is used. The learning rate is set to 0.0001.

Figure 11. WHU Aerial imagery dataset: ① is the training area in this dataset, ② is the validation area in this dataset, and ③ and ④ are the test area in this dataset.

3.4.2. Quantitative Analysis

U-Net was selected as the baseline for experiments to study the impact of the AMSA and RDCU modules on overall network performance. The IOU, PA, and Recall are used for accuracy evaluation. The evaluation results are shown in Table 5. It can be seen from Table 4 that, in terms of the accuracy evaluation indicators of the IOU, PA, and Recall, compared with the baseline, the addition of the RDCU module increased them by 2.5%, 1.98%, and 1.77%, respectively. Adding the AMSA module increased them by 2.92%, 2.28%, 2.83%, and 0.36%, respectively. Combining the RDCU and AMSA modules increased them by 3.28%, 2.62%, and 3.3%, respectively. Therefore, it can be seen that when the RDCU and AMSA modules are used separately, they can improve the performance of the model, and the combined use of the two is better than using them independently.

Table 5. The accuracy of the ablation study experiments.

Method	IOU	PA	Recall
Baseline	87.03%	89.83%	85.12%
Baseline + RDCU	89.53%	91.81%	86.89%
Baseline + AMSA	89.95%	92.11%	87.95%
Baseline + RDCU + AMSA	**90.31%**	**92.45%**	**88.42%**

3.4.3. Qualitative Analysis

Figure 12 shows the results of the ablation experiments. From Figure 12, it can be seen that the RDCU module and the AMSA module can effectively improve the model's sensitivity to building features. After adding the RDCU module, the extracted building shape is more complete and better reflects the real building outline, indicating that the RDCU module can help the model greatly improve its ability to extract building shape features. After adding the AMSA module, the noise interference of the obtained buildings is small and also reduces the false rate, indicating that the AMSA module can strengthen the expression of building features and solve the problem of confusion between building features and non-building features. The combination of the two perfectly shows each other's advantages, achieving high-precision building feature extraction while avoiding interference from non-building features, and can achieve the goal of high-precision building extraction.

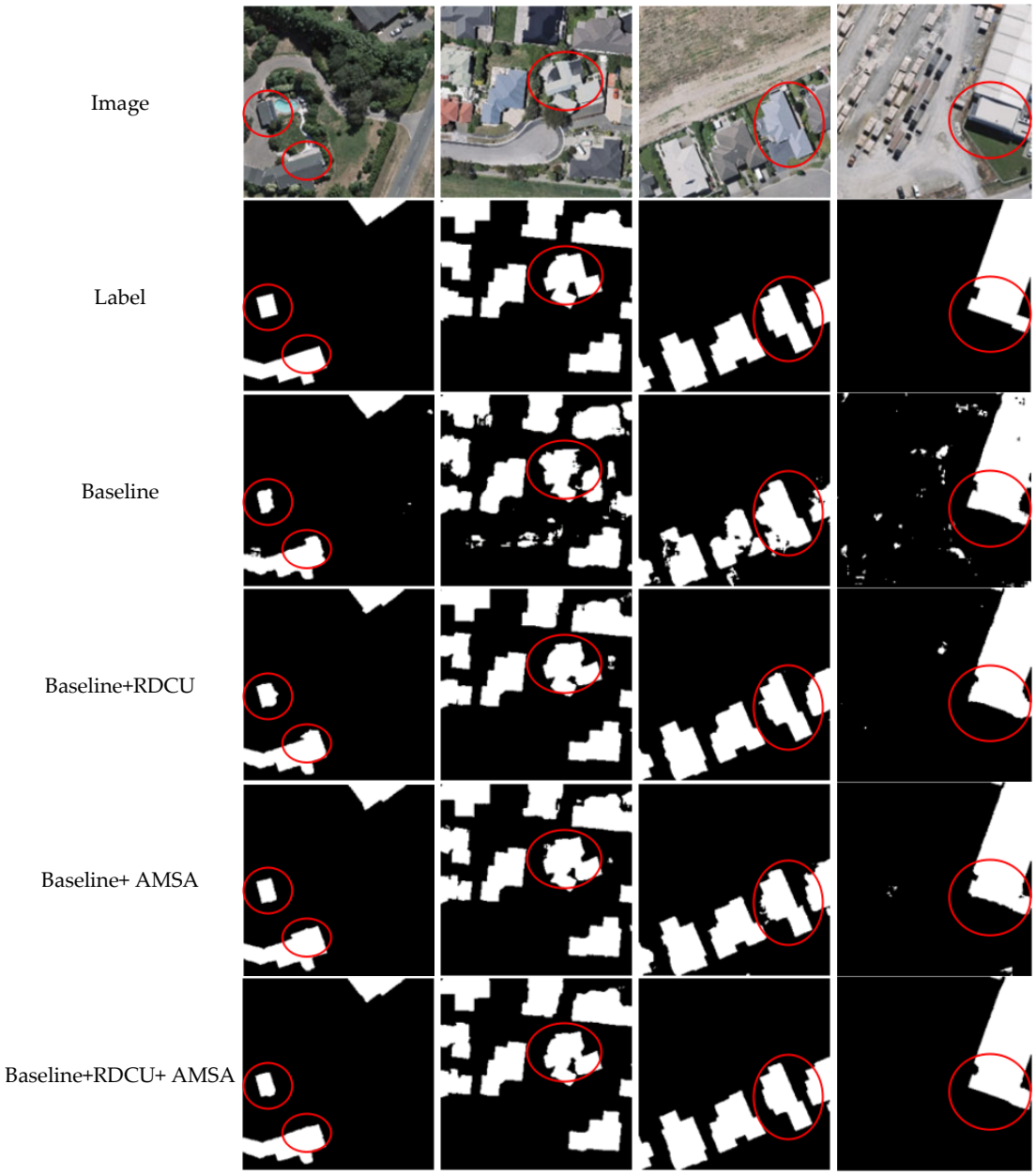


Figure 12. Samples of building extraction results using different models with the WHU Aerial imagery dataset (ablation study).

4. Conclusions

In this study, we have introduced a novel neural network, amalgamating the recurrent residual deformable convolution unit (RDCU) and multi-head with channel self-attention (AMSA) for building extraction. Through meticulous experimental analyses conducted on the Massachusetts dataset, the WHU dataset I, and the INRIA dataset, we have drawn the following significant conclusions:

(1) The residual structure in the proposed RDCU module has been shown to alleviate challenges with gradient vanishing and loss of feature information during deep learning model training. Deformable convolutional neural networks can enhance the module's ability to learn detailed features such as building shapes. By doing so, it substantially reinforces consistent learning of feature shapes throughout the network.

(2) The AMSA module profoundly augments the expression ability of building features by harnessing the global contextual information of features. Consequently, it enhances the model's sensitivity to building shape features, fortifies building features, and effectively suppresses other irrelevant features, culminating in a remarkable enhancement in building extraction accuracy.

(3) The proposed method has exhibited a superlative performance in the conducted experiments. For the Massachusetts dataset, the proposed method achieves an IoU score of 89.99%, PA score of 93.62%, and Recall score of 89.22%. For the WHU Satellite dataset I, the proposed method achieves an IoU score of 86.47%, PA score of 92.45%, and Recall score of 91.62%. For the INRIA dataset, the proposed method achieves an IoU score of 80.47%, PA score of 90.15%, and Recall score of 85.42% Ablation experiments have additionally corroborated the individual efficacy of using the RDCU module or the AMSA module, each contributing to improved building extraction accuracy. When both modules are synergistically combined, their respective strengths are harmonized, culminating in even higher accuracy.

Author Contributions: Conceptualization, W.Y. and B.L.; methodology, W.Y.; software, W.Y.; validation, W.Y. and H.L.; formal analysis, W.Y.; investigation, W.Y. and H.L.; resources, W.Y.; data curation, W.Y.; writing—original draft preparation, W.Y.; writing—review and editing, W.Y., G.G. and B.L.; visualization, W.Y. and H.L.; supervision, W.Y.; project administration, B.L. and H.L.; funding acquisition, B.L. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (No. 42161064, 42001411), supported by Jiangxi Provincial Natural Science Foundation (No. 20232ACB204032; 20212BAB204003). Graduate Innovation Foundation of East China University of Technology (No. DHYC-202302).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Corbane, C.; Lemoine, G.; Pesaresi, M.; Kemper, T.; Sabo, F.; Ferri, S.; Syrris, V. Enhanced automatic detection of human settlements using Sentinel-1 interferometric coherence. *Int. J. Remote Sens.* **2018**, *39*, 842–853. [CrossRef]
2. Zhou, R.-G.; Yu, H.; Cheng, Y.; Li, F.-X. Quantum image edge extraction based on improved Prewitt operator. *Quantum Inf. Process.* **2019**, *18*, 261. [CrossRef]
3. Kavzoglu, T.; Tonbul, H. A comparative study of segmentation quality for multi-resolution segmentation and watershed transform. In Proceedings of the 2017 8th International Conference on Recent Advances in Space Technologies (RAST), Istanbul, Turkey, 19–22 June 2017; pp. 113–117.
4. Yu, H.; Zhang, Y.; Cheng, G.; Ge, X. Rural residential building extraction from laser scanning data and aerophotograph based on quadtree segmentation. In Proceedings of the 2011 International Conference on Remote Sensing, Environment and Transportation Engineering, Nanjing, China, 24–26 June 2011; pp. 8476–8479.
5. Futagami, T.; Hayasaka, N. Automatic extraction of building regions by using color clustering. In Proceedings of the 2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), Hiroshima, Japan, 10–13 September 2019; pp. 415–419.

6. Jiang, B.; An, X.; Xu, S.; Chen, Z. Intelligent Image Semantic Segmentation: A Review through Deep Learning Techniques for Remote Sensing Image Analysis. *J. Indian Soc. Remote Sens.* **2022**. [CrossRef]
7. Tejeswari, B.; Sharma, S.K.; Kumar, M.; Gupta, K. Building footprint extraction from space-borne imagery using deep neural networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *XLIII-B2-2022*, 641–647. [CrossRef]
8. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
9. Li, W.; Sun, K.; Zhao, H.; Li, W.; Wei, J.; Gao, S. Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102970. [CrossRef]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Sarıtürk, B.; Bayram, B.; Duran, Z.; Seker, D.Z. Feature extraction from satellite images using segnet and fully convolutional networks (FCN). *Int. J. Eng. Geosci.* **2020**, *5*, 138–143. [CrossRef]
12. He, C.; Li, S.; Xiong, D.; Fang, P.; Liao, M. Remote sensing image semantic segmentation based on edge information guidance. *Remote Sens.* **2020**, *12*, 1501. [CrossRef]
13. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]
14. 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. [CrossRef]
15. 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; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
16. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
17. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:170605587.
18. Yurtkulu, S.C.; Şahin, Y.H.; Unal, G. Semantic segmentation with extended DeepLabv3 architecture. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; pp. 1–4.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
20. 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. *GISci. Remote Sens.* **2022**, *59*, 861–884. [CrossRef]
21. Dixit, M.; Chaurasia, K.; Mishra, V.K. Dilated-ResUnet: A novel deep learning architecture for building extraction from medium resolution multi-spectral satellite imagery. *Expert Syst. Appl.* **2021**, *184*, 115530. [CrossRef]
22. 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]
23. 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 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
24. 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 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. You, D.; Wang, S.; Wang, F.; Zhou, Y.; Wang, Z.; Wang, J.; Xiong, Y. EfficientUNet+: A Building Extraction Method for Emergency Shelters Based on Deep Learning. *Remote Sens.* **2022**, *14*, 2207. [CrossRef]
26. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018; Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2018; pp. 421–429.
27. 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]
28. 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. [CrossRef]
29. 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 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
30. Shi, Y.; Li, Q.; Zhu, X.X. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 184–197. [CrossRef] [PubMed]
31. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:160902907.
32. Song, H.O.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep Metric Learning via Lifted Structured Feature Embedding. *arXiv* **2015**, arXiv:1511.06452.
33. Chen, K.; Zou, Z.; Shi, Z. Building extraction from remote sensing images with sparse token transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]

34. 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]
35. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762v5.
37. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013; ISBN 0-494-96184-8.
38. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.