remote sensing

MDPI

Article

A Multi-Scale Edge Constraint Network for the Fine Extraction of Buildings from Remote Sensing Images

Zhenqing Wang [1,2,†], Yi Zhou [1,†], Futao Wang [1,2,3,*], Shixin Wang [1], Gang Qin [1,2], Weijie Zou [1,2] and Jinfeng Zhu [1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Key Laboratory of Earth Observation of Hainan Province, Hainan Aerospace Information Research Institute, Sanya 572029, China
* Correspondence: wangft@aircas.ac.cn; Tel.: +86-010-64879460
† These authors contributed equally to this work.

Abstract: Building extraction based on remote sensing images has been widely used in many industries. However, state-of-the-art methods produce an incomplete segmentation of buildings owing to unstable multi-scale context aggregation and a lack of consideration of semantic boundaries, ultimately resulting in large uncertainties in predictions at building boundaries. In this study, efficient fine building extraction methods were explored, which demonstrated that the rational use of edge features can significantly improve building recognition performance. Herein, a fine building extraction network based on a multi-scale edge constraint (MEC-Net) was proposed, which integrates the multi-scale feature fusion advantages of UNet++ and fuses edge features with other learnable multi-scale features to achieve the effect of prior constraints. Attention was paid to the alleviation of noise interference in the edge features. At the data level, according to the improvement of copy-paste according to the characteristics of remote sensing imaging, a data augmentation method for buildings (build-building) was proposed, which increased the number and diversity of positive samples by simulating the construction of buildings to increase the generalization of MEC-Net. MEC-Net achieved 91.13%, 81.05% and 74.13% IoU on the WHU, Massachusetts and Inria datasets, and it has a good inference efficiency. The experimental results show that MEC-Net outperforms the state-of-the-art methods, demonstrating its superiority. MEC-Net improves the accuracy of building boundaries by rationally using previous edge features.

Keywords: multi-scale edge constraint; building extraction; remote sensing; deep learning; build-building

check for updates

Citation: Wang, Z.; Zhou, Y.; Wang, F.; Wang, S.; Qin, G.; Zou, W.; Zhu, J. A Multi-Scale Edge Constraint Network for the Fine Extraction of Buildings from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 927. https://doi.org/10.3390/rs15040927

Academic Editors: Hongyan Zhang and Xiong Xu

Received: 4 January 2023
Revised: 3 February 2023
Accepted: 5 February 2023
Published: 8 February 2023

1. Introduction

As individuals spend most of their lives in buildings, these structures are an important part of basic geographic information elements. Accordingly, their role in urban planning, real estate management, disaster risk assessment and many other fields cannot be ignored [1–3]. High-spatial-resolution imagery provides richer spatial details, enabling accurate buildings extractions. As the external environment of the building is very complex (e.g., building ancillary facilities and shadow occlusion make feature extraction more difficult), the contrast between the building and some non-buildings (parking lot, bare ground, roads, etc.) is low, which can easily interfere with the recognition results [4]. In addition, the contours, structures and materials of buildings in different regions are significantly different. Therefore, more work is needed to achieve accurate building extraction.

Traditional building extraction methods from optical remote sensing images can be roughly divided into image classification and morphological index-based methods [5–10]. The pixel-oriented image classification method uses a single pixel in the image as the basic unit and independently classifies and extracts the semantic information of each pixel in the image. This type of method uses labeled data for model learning. Further, machine learning algorithms, such as Markov random fields, are widely used [5–7]. Notably, these machine learning methods only consider the spectral characteristics of a single pixel point and do not consider other attribute information of related objects. To increase the neighborhood information, scholars have proposed an object-oriented method that divides the input image into homogeneous polygon objects and uses each object as an analysis unit. Although the classification results are better than those of pixel-oriented methods, their performance often depends on the performance of object segmentation [2,8,9]. The aforementioned building extraction methods based on image classification often require rich samples. Some researchers have proposed a building extraction method that does not require labeled samples (i.e., the building morphological index) [10]. However, they rely on features that are limited, the classification threshold must be selected manually by trial and error and serious misclassification will occur [11]. In addition to optical images, SAR and LiDAR have important application values. SAR data can provide all-day, all-weather observations and have great potential for building extraction [12]. However, compared to optical images, SAR images have large geometric distortions, and high-precision building boundary information is difficult to extract [13]. LiDAR data not only contain high-quality 2D data but also the height of ground objects, which play a role in the task of accurately extracting buildings [14]. However, the acquisition of lidar data is costly and slow, thereby causing difficulty on a large scale [15].

With the accumulation of remote sensing data and the continuous improvement of computer software and hardware computing efficiency, the deep applications in remote sensing display a vigorous development trend. Convolutional neural networks (CNN) have powerful image feature extraction abilities, and fully CNN (FCN) is emphasized in the problem of building extraction. Audebert et al. [16] proposed the transfer of the application of depth FCN from ordinary images to remote sensing images. A multi-core convolution layer was introduced, residual correction was used to fuse the heterogeneous sensor data and a good building extraction result was obtained. Kang et al. [17] proposed EU-Net, which first designed a denser spatial pyramid pool (SPP), which helps extract more buildings of different scales. Thereafter, focal loss is used to reverse the negative effect of false tags on training in order to enable a higher extraction accuracy. Jin et al. [18] proposed BARNet for precisely localizing building boundaries. A gated attention fine fusion unit was developed to better handle cross-layer features in skip connections. Wang et al. [19] designed a new edge-ignoring cross-entropy function for the insufficient context of the edge pixels in a sample image. The morphological building index and image stitching are involved in the training process. Guo et al. [15] adaptively improved building prediction by exploiting the spatial details of low-level features, further improving the effect of building extraction.

At present, the overall accuracy of building extraction is already high; however, the accuracy of the building boundary area is poor [20–23]. In this study, we proposed a multi-scale edge-constrained network (MEC-Net) for the fine extraction of buildings. Image edge features were added to the feature groups that aggregate features of different scales for prior feature constraints, and the attention mechanism was used to suppress the negative effects caused by the edge features of other objects. From the perspective of data, a data augmentation method (denoted as build-building), specifically for buildings, was proposed. Build-building achieves instance-level expansion of the number and style of building instances by simulating the new construction of buildings, thereby improving the generalization of MEC-Net. The main contributions of this study are as follows:

(1) An MEC-Net based on a multi-scale edge constraint (MEC) module was proposed. The network effectively integrated features of different scales and added two attention mechanisms, namely, prior edge constraints and learnable squeeze excitations, to each scale feature group to improve the segmentation effect on the boundary area of buildings.

(2) A data augmentation method named build-building was designed to simulate the construction of new buildings. A building object instance was copied from one image and

pasted onto another image. Before pasting, the styles of the two images must be unified. Increasing the number of building samples can also enrich the background of buildings of the same type.

(3) Our proposed MEC-Net achieved state-of-the-art performance on the WHU, Massachusetts and Inria datasets.

The remainder of this paper is organized as follows: Section 2 describes work related to MEC-Net; Section 3 describes the proposed MEC-Net; Section 4 describes the experimental design, including the datasets, experimental details, accuracy evaluation, comparative experiments and ablation experiments; Section 5 presents the results and analysis of MEC-Net; and Section 6 contains the conclusions.

2. Related Work

2.1. Multi-Scale Feature Fusion

With the introduction of the FCN [24], Ronneberger et al. [25] designed U-Net with a better segmentation performance. U-Net belongs to the encoder-decoder structure, and its name originates from the "U" shape of the entire network. The encoder on the left is used to extract high-dimensional feature information, which is a process from an image to a high-semantic-level feature map. A symmetric decoder for precise localization, which is a process from high-semantic-level feature maps to pixel-level classification score maps, is located on the right. As shown in Figure 1a, the features corresponding to the decoder and encoder implement feature fusion through skip connections. The optimal depth of the U-Net network is unknown a priori; therefore, an architecture search at different depths is required. Notably, skip connections only perform feature fusion on the same scale feature of the encoder and decoder, which has limitations.

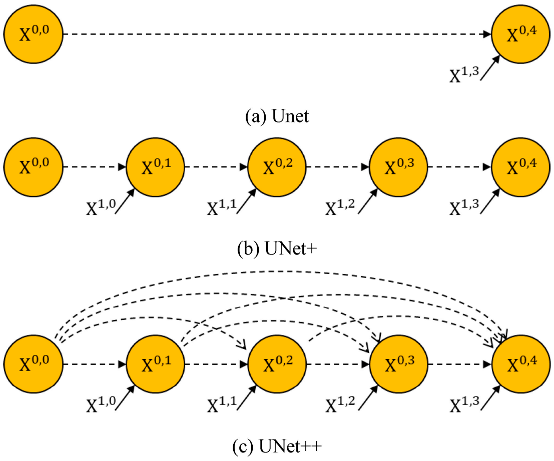


Figure 1. The first skip pathway of UNet, UNet+ and UNet++.

To solve these problems, Zhou et al. [26] proposed UNet+ and UNet++. UNet+ integrates U-Nets at different depths to mitigate the negative impact of the network depth on the results. These U-Net modules share an encoder, learn together and connect two adjacent nodes in the same horizontal layer with short skip connections, enabling the transmission of supervisory signals to the shallower decoder by the deeper decoder, as shown in Figure 1b. Based on UNet+, UNet++ redesigns skip connections to generate densely connected skip connections. As shown in Figure 1c, each horizontal layer is a standard DenseNet structure, enabling a densely propagated feature along skip connections, causing more flexibility in the feature fusion of the decoder.

The multi-scale features of UNet++ are aggregated step-by-step to obtain the final segmentation result, which improves the performance and results in a faster convergence speed, ultimately leading to its wide application in remote sensing image segmentation [27–29].

However, these multi-scale features are all automatically learned by the model, lacking human-controllable prior feature constraints.

2.2. Boundary-Constrained Refined Building Extraction

Accurately identifying building boundaries is critical for many remote sensing applications. To effectively extract objects with clear boundaries, some researchers have added boundary information extraction tasks during the performance of semantic segmentation tasks in the network [20–22]. Building boundary extraction is taken as a separate task to improve the accuracy of boundary regions in the building extraction task. Guo et al. [4] added a boundary refinement module to the network to perceive the orientation of each pixel to the nearest object, thereby refining the building prediction results. Based on these experimental results, by focusing the model on boundary details, the quality of building segmentation can be improved. The boundary of the building is highly correlated with the edge features of the image, and the effective use of this prior knowledge is critical. Shao et al. [23] inputted the edge information of objects into the network, together with the image, which improves the accuracy of building change detection. Liao et al. [30] integrated an additional convolution module to process the edge features and take the structural features that capture the boundary information of buildings to improve the accuracy of the segmentation results. However, building edge information is a feature obtained by operating on the image. The fusion of edge features and features extracted by the convolutional network is more logical than directly splicing the image with the input to the network. In addition, there are also some studies on the boundary constraints of buildings from the perspective of loss functions. You et al. [31] proposed a joint loss function of building boundary weighted cross-entropy and dice loss to strengthen the constraints on building boundaries. Despite the efforts of many scholars, inaccurate boundary identification is still inevitable.

3. Methodology

The existing building extraction network can already achieve a high overall accuracy; however, a large error exists in the building boundary area. Improving the extraction performance of building boundary areas is an urgent problem that must be solved. To solve this problem, a multi-scale edge constraint network (MEC-Net) was proposed to accurately extract building footprints. The overall structure of MEC-Net is shown in Figure 2.

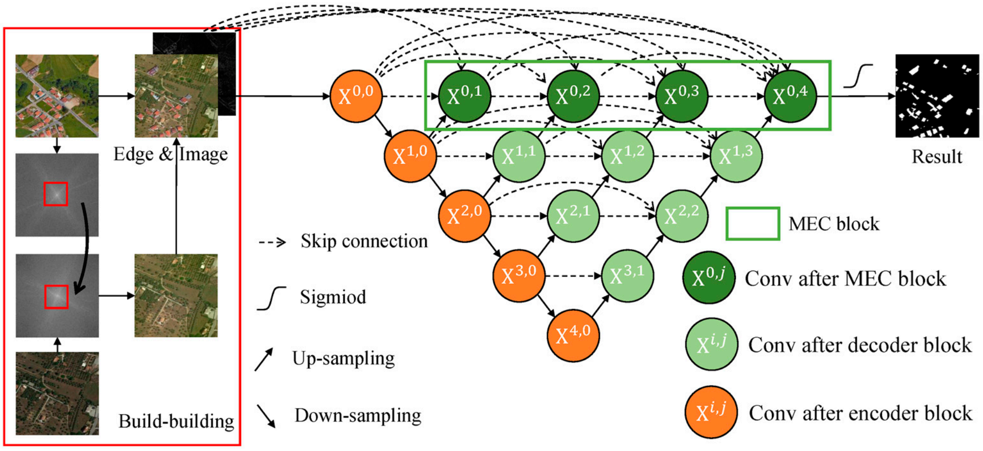


Figure 2. The structure of MEC-Net.

As shown in Figure 2, MEC-Net uses U-Net++ as the basic network and resnet50 as the backbone. The encoder network extracts building features of different scales, and the decoder uses multi-scale feature fusion and edge feature constraints to obtain the same extraction result as the input image size. Section 3.1 discusses the design of an efficient MEC module that utilizes prior edge features to perform attention constraints on feature maps at different scales to refine building boundaries. In addition to prior attention, the self-attention mechanism scSE was added in MEC, as well as other decoders, as described in Section 3.2. Overall, a data augmentation method specifically for building extraction was designed, as shown in Section 3.3. Furthermore, the loss function of MEC-Net was introduced in Section 3.4.

3.1. Multi-Scale Edge Constraint (MEC)

The features in the deep neural network are automatically learned through human supervision, and high-dimensional features are difficult to explain and are uncontrollable. The addition of prior knowledge to a network can improve its learning. In fact, the convolutional layer used in CNN utilizes prior knowledge: the closer the distance between the image pixels, the stronger the correlation. The settings of the anchor size and the aspect ratio in the famous Faster RCNN [32] were also determined a priori.

At present, the accuracy of building boundaries is poor, and building boundaries and edge features are highly correlated. As a result, an MEC block, which adds prior knowledge of edge features to the network to optimize the building boundary, was constructed. As depicted by the green rectangle in Figure 2, the MEC block contains four units, which fuse semantic features and edge features of different scales to obtain $X^{0,j}, j \in \{1,2,3,4\}$. MEC causes the model to be more sensitive to the edges of buildings. The structure of the j-th unit is shown in Figure 3. The edge features were fused as concatenation with the low-dimensional features of the current level and the high-dimensional features of the next level to obtain F_C^j. (bilinear interpolation) with 1 × 1 convolution was used to avoid the checkerboard pseudo-checker [33]. Subsequently, two 3 × 3 convolutions and ReLU activations were performed to obtain the contextual information of a larger receptive field. The edge feature Among these, high-dimensional features must be unified in size before concatenation. Up-sampling input to MEC-Net includes buildings and other objects. To ensure that the model is more focused on the edges of buildings, a self-attention mechanism scSE was added at the end of the unit. The final output $Output^j X^{0,j}$ of the unit can be described using the following formula:

$$F_C^j = [E, [F_L^k]_{k=0}^{j-1}, Conv^{1\times1}(Up(F_H^{j-1}))] \tag{1}$$

$$X^{0,j} = scSE((ConvReLU^{3\times3})^2(F_C^j)) \tag{2}$$

where E represents the edge feature, F_L^k represents the k-th low-dimensional feature, that is,$X^{0,k}$ in Figure 2, and represents the $(j-1)$-th high-dimensional feature, that is, $X^{1,j-1}$ in Figure 2.

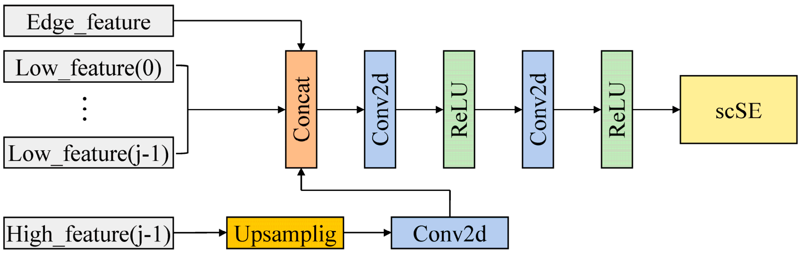


Figure 3. The structure of the *j*-th unit in the MEC block.

The ratio of the feature resolution to the input image was used to quantitatively describe the feature scale. Accordingly, the first unit in the MEC block in the model fuses the multi-scale features of [1/2, 1/4], the second unit fuses the multi-scale features of [1/2, 1/4, 1/8], the third unit fuses the multi-scale features of [1/2, 1/4, 1/8, 1/16] and the fourth unit fuses the multi-scale features of [1/2, 1/4, 1/8, 1/16, 1/32].

The edge features in the MEC block are represented by the gradient values obtained by the Sobel operator and image convolution operation. The Sobel operator [34] is a common first-order derivative edge detection operator that uses two 3 × 3 matrices to convolve the original image and obtain the horizontal gradient E_x and vertical gradient E_y, as shown in Equations (3) and (4). Finally, the edge feature E was obtained by combining the above two results.

$$E_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I \tag{3}$$

$$E_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * I \tag{4}$$

$$E = \sqrt{{E_x}^2 + {E_y}^2} \tag{5}$$

3.2. *Spatial and Channel Sequeeze & Excitation (scSE)*

The attention module is often used for feature screening and optimization to make the model converge better [35–37]. In addition to adding scSE to the MEC block, scSE was added in the last part of each decoder unit. scSE is a type of joint attention that combines spatial and channel attention. Attention allows the model to focus on the information or features of interest, enabling unimportant information to be ignored and the use of more useful information to be enhanced.

Hu et al. [35] proposed a new unit called "sequence and excitation" (SE), which sheds light on image classification by recalibrating channel feature responses. As a result, Roy et al. [36] introduced three SE modules: cSE, sSE and scSE.

The cSE module assigns importance weights to the feature mapping in the channel dimension. For a feature with shape (C, W, H), the shape was first converted to (C, 1, 1) via global average pooling. Thereafter, two 1 × 1 convolution blocks and the Sigmoid function were used to obtain the vector of shape (C, 1, 1), representing the importance of the channel features. Finally, it was multiplied by the original feature map in a channel-wise form to obtain an attention-corrected feature map on the channel.

The sSE module belongs to the spatial attention mechanism, which reassigns the weights to spatial information and obtains features containing different spatial weight information. First, a 1 × 1 convolution was used to directly change the shape of the feature from (C, H, W) to (1, H, W). Thereafter, to obtain a feature representing the importance of each spatial position, the sigmoid function was used for activation. Finally, it was multiplied by the original feature to complete the spatial correction.

The scSE module is a parallel combination of cSE and sSE. In particular, after passing through the sSE and cSE modules, the two results were added to obtain a more accurately corrected feature map, as shown in Figure 4.

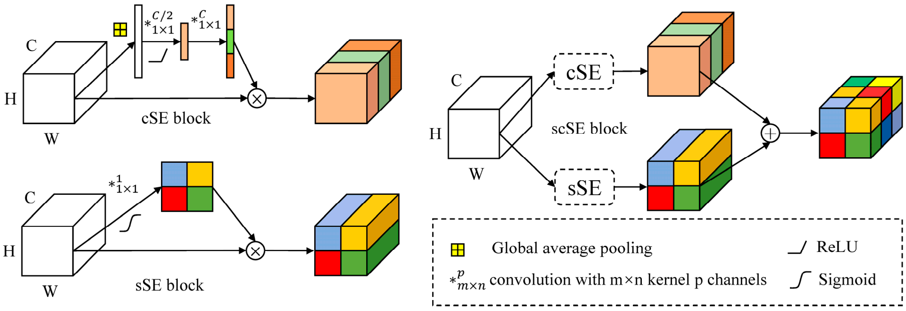


Figure 4. Operation flow chart of cSE, sSE and scSE.

3.3. Build-Building Data Augmentation

Reasonable data augmentation can improve the accuracy and generalization ability. For building extraction, Zhang et al. [38] and Guo et al. [15] used the random rotation and flipping of images, while Jin et al. [18] used random flipping, scaling and Gaussian smoothing. However, these methods are more general and were not designed for building extraction. To enable data augmentation methods to play a significant role in building extraction, a data augmentation method build-building for buildings was proposed, as shown in Figure 5.

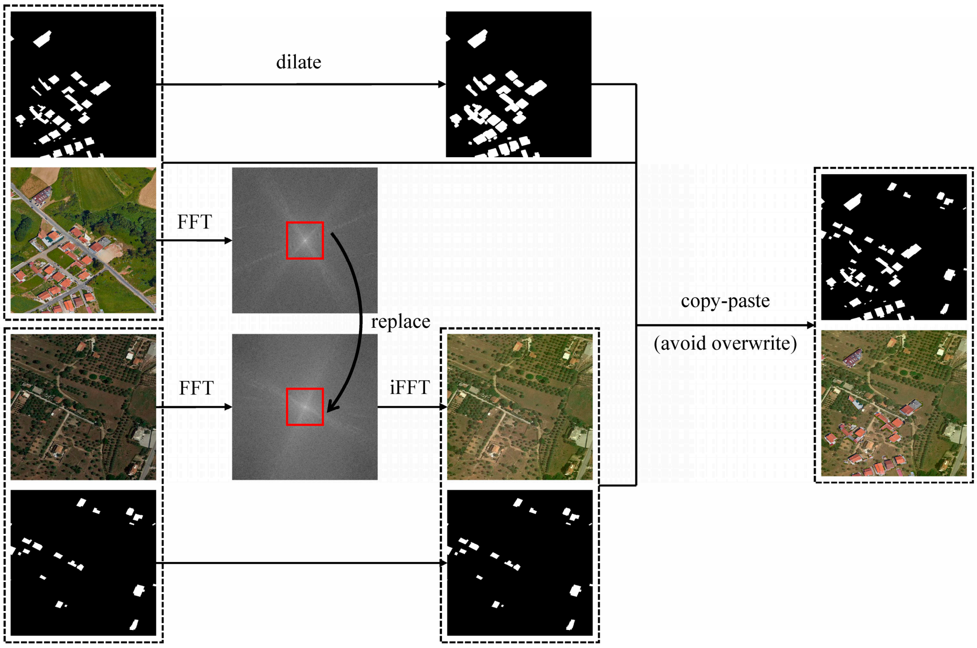


Figure 5. Flow chart of build-building. Among them, the red box area is the low frequency area.

Build-building is inspired by copy-paste [39]. The core idea of copy-paste is to combine the information from different images in an instance-aware manner. Copy-paste is similar to mixup [40] and CutMix [41]; however, only the pixels within the object are copied instead of all pixels in the bounding box (rectangle). Thus, copy-paste can simulate the construction of buildings in open spaces. On one hand, copy-paste can increase the sampling frequency of building samples; on the other hand, it can expand the background style of similar buildings, as shown in Figure 6c.

Owing to different imaging conditions such as illumination and atmosphere, different remote sensing images reflect the brightness and darkness of gray values, which are inconsistent. The direct use of copy-paste can cause the augmented results to appear abrupt, possibly prompting the model to be optimized in the wrong direction. Therefore, a fast Fourier transform (FFT) was employed to unify the style of the two images before copy-paste.

First, FFT was performed on the image $f_{copy}(x,y)$ (width is W and height is H) for the copy operation and the image $f_{paste}(x,y)$ as the paste background:

$$F_{copy}(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f_{copy}(x,y)e^{\frac{-2\pi j}{M}ux+\frac{-2\pi j}{N}vy} \tag{6}$$

$$F_{paste}(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f_{paste}(x,y)e^{\frac{-2\pi j}{M}ux+\frac{-2\pi j}{N}vy} \tag{7}$$

where x and y are image variables, and u and v are frequency variables.

After the spatial domain is converted to the frequency domain, the places where the gray value changes sharply correspond to the high frequency, and vice versa corresponds to the low frequency. Thus, high frequency is mainly a measure of the edge and contour, and low frequency is a comprehensive measure of the intensity. Only the low frequency of $F_{paste}(u,v)$ must be replaced with $F_{copy}(u,v)$ to obtain $F_{copy-paste}(u,v)$; thereafter, by converting it to the spatial domain through the inverse fast Fourier transformation, we can skillfully achieve style unity:

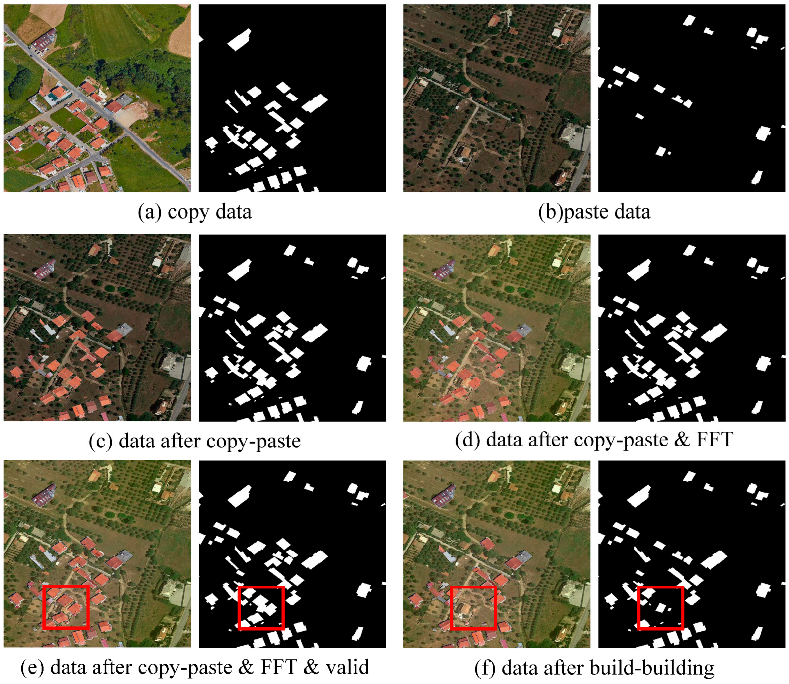
$$F_{copy-paste}(u,v) = \begin{cases} F_{copy}(u,v) & F_{paste}(u,v) \text{ is low frequency} \\ F_{paste}(u,v) & F_{paste}(u,v) \text{ is high frequency} \end{cases} \tag{8}$$

$$f_{copy-paste}(x,y) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F_{copy-paste}(u,v)e^{\frac{2\pi j}{M}ux+\frac{2\pi}{N}vy} \tag{9}$$

The frequency domain of the 2 × 2 area in the center of the FFT result after spectrum centralization is the low-frequency domain that must be converted. The results for the copy-paste after style unification are shown in Figure 6d.

When constructing a building dataset, the top of the building is generally used as the corresponding ground truth value. If only the pixels corresponding to the top of the building are copied, the neighborhood features that are very important for building extraction (such as sloped walls or shadows) are ignored. The expansion operation of the corresponding mask is used during copying, and then copy-paste is applied to solve it. The results are shown in Figure 6e.

During the copy-paste operation, situations in which buildings cover other buildings may arise owing to the overlapping areas of two or more buildings. This occurrence is impossible in reality and may affect the correct direction of model learning. Therefore, the overlapping building objects were deleted, and the copy-paste operation was performed; the results are shown in Figure 6f, where the potential of buildings covering each other is avoided.


(a) copy data

(b)paste data

(c) data after copy-paste

(d) data after copy-paste & FFT

(e) data after copy-paste & FFT & valid

(f) data after build-building

Figure 6. Example results of each stage of build-building. Among them, the red box area is the focus area.

3.4. Loss Function

The categorical cross entropy (CE) loss $Loss_{ce}$ is often used as the loss function for objects extraction, and its expression is as follows:

$$Loss_{ce} = \frac{1}{N}\sum_{i=1}^{N} -\sum_{c=1}^{M} y_c^i \log(p_c^i) \tag{10}$$

where N is the number of pixels, M is the number of categories, y_c^i is the c-th value in the one-hot encoding of the actual class at pixel i and p_c^i is the predicted value of class c at pixel i.

Owing to the CE function, the model can predict the probability of positive classes approaching one and the probability of negative classes approaching zero in the process of network parameter learning and updating; this can lead to over-confidence in the model when the training data are insufficient to cover all forms of building, resulting in overfitting. Soft CE loss $Loss_{sce}$ involves the performance of label smoothing [42] on the ground truth and is followed by cross-entropy calculation with the predicted value, which can improve the generalization to a certain extent. Its expression is as follows:

$$\begin{cases} Loss_{sce} = \frac{1}{N}\sum_{i=1}^{N} -\sum_{c=1}^{M} \overset{\wedge}{y_c^i} \log(p_c^i) \\ \overset{\wedge}{y_c^i} = \begin{cases} 1-\alpha, & c = target \\ \alpha/M, & c \neq target \end{cases} \end{cases} \tag{11}$$

where $\overset{\wedge}{y_c^i}$ is the ground truth after smoothing, and α is the smoothing coefficient.

The positive and negative samples of the building dataset are generally quite different, and the dice loss $Loss_{dice}$ [43] can alleviate the class imbalance phenomenon to a certain extent; however, the training is prone to instability. Therefore, in this study, the joint function $Loss_{seg}$ of $Loss_{sce}$ and $Loss_{dice}$ was adopted as the loss function for this experiment.

$$\begin{cases} Loss_{seg} = Loss_{sce} + Loss_{dice} \\ Loss_{dice} = \frac{2TP}{2TP+FP+FN} \end{cases} \tag{12}$$

where TP is the pixel area correctly classified as the building, FP is the pixel area wrongly classified as the building and FN is the pixel area incorrectly classified as the non-building.

4. Experimental Design

4.1. Datasets

We opted to conduct sufficient experiments on three publicly available and commonly used datasets. The three datasets cover different scopes, scenarios and cities and can measure the pros and cons of the model from different perspectives. The details are as follows.

(1) The WHU dataset includes more than 220,000 individual buildings, covering 450 km^2 [44]. The original spatial resolution is 0.075 m, and the image resolution is down-sampled to 0.3 m for the convenience of task training. Notably, the estimated model accuracy was more convincing with many building examples. The division of the training set, validation set and test set of the WHU dataset is consistent with the original paper.

(2) The Massachusetts dataset consisted of 151 tiles with a spatial resolution of 1 m [45]. The scenarios not only include cities but also suburbs, which are suitable for evaluating the robustness of the model for different scenarios. To facilitate the training, we cropped each tile with a sliding window to obtain a series of 512 × 512 small tiles with a repetition rate of 0.5. The division of the training set, validation set and test set of the Massachusetts dataset is consistent with the original paper.

(3) The Inria dataset covers a total area of 810 km^2 [46]. The training set included different urban settlements. Each city in the training set contained 36 tiles with 5000 × 5000 pixels. To facilitate the comparison of the results, we adopted the same division basis as the other study [4]: tiles with serial numbers 1–5 were selected for testing performance, and the

rest were used for training and validation (tiles with serial numbers 6–10 were used for validation). In addition, owing to the large number of datasets, the sliding window method, with a repetition rate of 0.1, was used to crop all tiles and remove the cropping results that do not contain buildings. Finally, 11,028 training, 1912 verification and 1950 test images were obtained.

4.2. Implementation Details

All experiments in this study were trained on an Nvidia GeForce RTX 3090 GPU, and the models were built using pytorch [47]. The batch size was set to 8, and Adam [48], with weight decay, was used as the optimizer (decay factor = 0.001). The initial learning rate was set as 0.001. The learning rate was adjusted using cosine annealing, and the minimum learning rate was set as 0.0001. The training set was randomly augmented with a probability of 0.5 (including the commonly used image flipping and the build-building proposed in this study; no build-building was performed in the ablation experiment to verify whether the build-building was effective). The maximum number of epochs for model training was 125.

4.3. Evaluation Metrics

In this study, four commonly used accuracy metrics were employed for accuracy evaluation: precision, recall, F1-score and IoU. Precision refers to the proportion of pixels classified as buildings that are actually buildings. Recall is a measure of the number of building pixels that were correctly classified as buildings. F1-score is a composite indicator of the precision and recall. The IoU is the ratio of the intersection over the union of the predicted and ground-truth building areas. We calculated the additional IOU of the buffer area with a radius of two pixels from the building boundary to measure the extraction effect of the building boundary area. The formula for each indicator is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

$$IoU = \frac{TP}{TP + FN + FP} \tag{16}$$

4.4. Benchmark Techniques

We adopted the following four SOTA semantic segmentation networks as our experimental benchmark techniques: PSP-Net [49], Res-U-Net [50], DeeplabV3+ [51] and HR-Net [52].

PSP-Net performs contextual information aggregation in different regions through a pyramid pooling module and embeds difficult scene contextual features into a pixel prediction framework. Res-U-Net combines the advantages of residual learning with U-Net. Deeplabv3+ uses atrous SPP to obtain multi-scale information and introduces a new decoder module to further fuse low-level and high-level features. HRNet changes the link between high and low resolutions from series to parallel, enabling the maintenance of high-resolution representations throughout the network structure.

MEC-Net was compared with recent building extraction methods, including SiU-Net [43], JointNet [53], EU-Net [17], DS-Net [54], MAP-Net [55] and CBR-Net [4].

4.5. Ablation Study Design

Adequate ablation studies have been conducted. As shown in Table 1, UNet++ was the method used in the first ablation study, which was marked as the baseline. In the second ablation study, the method involved the addition of edge features in the first horizontal layer of UNet++, marked as Baseline + E. The method in the third ablation study included the addition of the MEC block described above to UNet++, which is our MEC-Net, marked as Baseline + E + A. The fourth ablation study used build-building when training MEC-Net, labelled Baseline + E + A + B.

Table 1. Composition of the four ablation studies.

Method	UNet++	Edge Feature	Attention (scSE)	Build-Building
Baseline	√			
Baseline + E	√	√		
Baseline + E + A	√	√	√	
Baseline + E + A + B	√	√	√	√

5. Results and Analysis

5.1. Comparison with SOTA Methods

Tables 2–4 show the quantitative evaluation results of MEC-Net. For readability, the highest-scoring values are in bold. MEC-Net had the highest IoU (91.13%, 74.13% and 81.05%) and f1 (95.36%, 85.15% and 89.53%) on all three datasets, outperforming the other methods and demonstrating the effectiveness of our method. Compared with the second-best-performing HR-Net, MEC-Net increased the IoU by 1.27%, 0.68% and 0.57% using the datasets, respectively. The optimization effect of MEC-Net using the Massachusetts and Inria datasets was relatively low, mainly due to the less detail provided by the low-spatial resolution and the smaller discrimination of the building boundary areas. Although the spatial resolution of the Inria dataset is consistent with that of the WHU dataset, the quality of its ground truth is poor. The IoU of the building boundary buffer area (with a radius of two pixels) was calculated and recorded as IoU (boundary) in Tables 2–4. Compared with the second-highest IoU (boundary), MEC-Net improves the IoU of building boundary regions by 1.00%, 0.17% and 0.36% using the three datasets, respectively. The extracted boundary quality of the three datasets improved, indicating that MEC-Net can accurately describe the building boundaries of different sizes in different regions and has strong robustness.

Table 2. Quantitative evaluation of MEC-Net using the WHU dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)	IoU (Boundary) (%)
PSP-Net	92.60	93.83	93.21	87.28	58.88
Res-U-Net	94.00	94.86	94.43	89.45	66.00
DeeplabV3+	94.30	94.50	94.40	89.39	64.93
HRNet	94.67	94.64	94.66	89.86	67.52
MEC-Net(ours)	**94.70**	**96.03**	**95.36**	**91.13**	**68.52**

Table 3. Quantitative evaluation of MEC-Net using the Massachusetts dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)	IoU (Boundary) (%)
PSP-Net	77.28	83.97	80.49	67.35	48.54
Res-U-Net	81.74	86.66	84.13	72.61	55.51
DeeplabV3+	**83.49**	83.77	83.63	71.87	53.49
HRNet	82.82	86.66	84.69	73.45	56.64
MEC-Net(ours)	83.24	**87.14**	**85.15**	**74.13**	**56.81**

Table 4. Quantitative evaluation of MEC-Net using the Inria dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)	IoU (Boundary) (%)
PSP-Net	87.46	88.37	87.91	78.43	47.43
Res-U-Net	88.21	88.64	88.42	79.25	48.52
DeeplabV3+	87.67	88.23	87.95	78.49	47.39
HRNet	**89.31**	89.06	89.18	80.48	50.66
MEC-Net(ours)	89.26	**89.81**	**89.53**	**81.05**	**51.02**

In order to further analyze the advantages of the method proposed in this paper, Figures 7–9 show some inference examples for each dataset. It can be seen that MEC-Net can obtain the most complete building extraction results and the most prominent boundary details. For the small buildings in the first row of Figure 7 and the third row of Figure 8, Deeplabv3+ and HR-Net caused omissions, while MEC-Net can be well extracted. The characteristics of the parking lot in the second row of Figure 7 are very similar to buildings, and other models have been more or less misrepresented, while MEC-Net effectively distinguishes these two objects. For the super-large building in Figure 7, the results extracted by Deeplabv3+ are very disordered, and our MEC-Net successfully avoids this error. For the denser buildings in the second and third rows of Figure 8, MEC-Net achieves more complete and accurate building extraction results than other models. For the building corners and surrounding lawns in Figure 9, MEC-Net is also able to distinguish buildings well. Accurate building localization and boundary extraction will facilitate downstream tasks such as fine building area measurement and fine building vectorization.

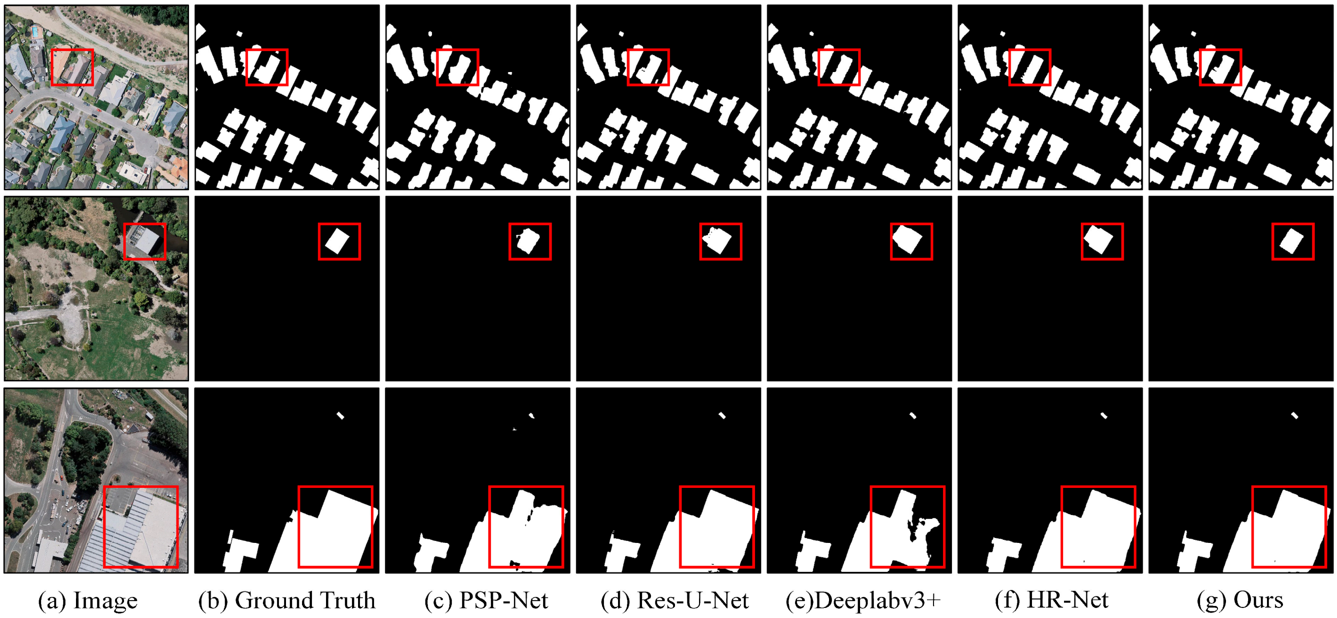


Figure 7. Examples of MEC-Net using the WHU dataset.

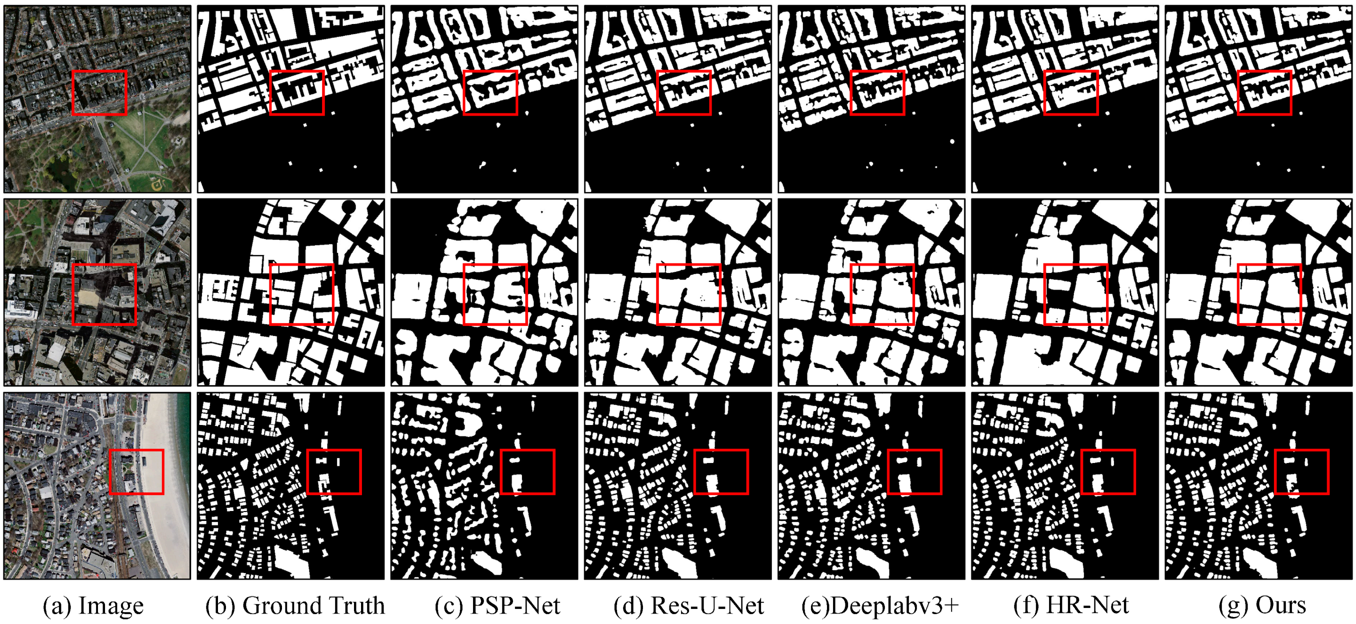


Figure 8. Examples of MEC-Net using the Massachusetts dataset.

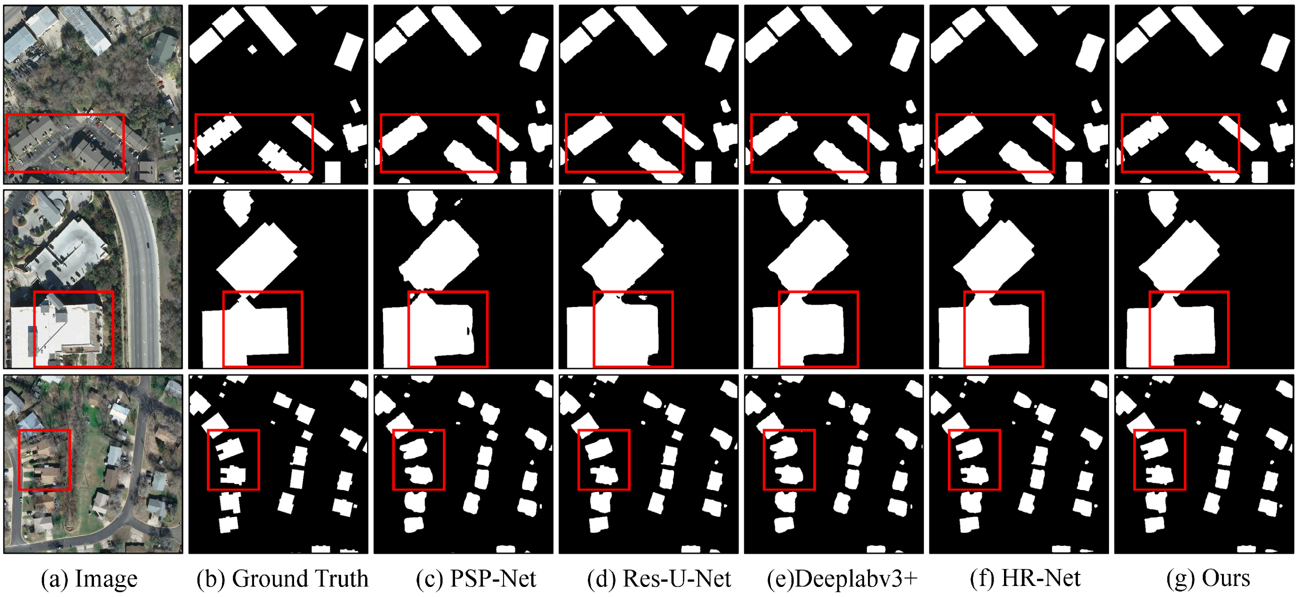


Figure 9. Examples of MEC-Net using the Inria dataset.

5.2. Comparison with Recent Methods

Tables 5–7 show the quantitative comparison results of MEC-Net and the methods described in Section 4.4. Similarly, the highest-scoring values are in bold. Precision, recall and F1 were not evaluated in the study describing DS-Net. As a result, the horizontal lines in Table 7 were employed. MEC-Net was found to achieve the best results. The IoU was 0.04% higher than that of CBR-Net with the WHU dataset, 0.20% higher than that of EU-Net with the Massachusetts dataset and 0.32% higher than that of DS-Net with the Inria dataset.

Table 5. Quantitative comparison of MEC-Net using the WHU dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
SiU-Net	90.30	94.50	92.35	85.80
EU-Net	94.98	95.10	95.04	90.56
MAP-Net	**95.62**	94.81	95.21	90.86
CBR-Net	95.14	95.53	95.34	91.09
Ours	94.70	**96.03**	**95.36**	**91.13**

Table 6. Quantitative comparison of MEC-Net using the Massachusetts dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
SiU-Net	68.10	74.60	71.20	55.20
Joint-Net	86.21	81.29	83.68	71.99
EU-Net	**86.70**	83.40	85.01	73.93
CBR-Net	85.63	83.51	84.56	73.24
Ours	83.24	**87.14**	**85.15**	**74.13**

Table 7. Quantitative comparison of MEC-Net using the Inria dataset.

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
SiU-Net	71.40	84.60	83.30	71.40
DS-Net	-	-	-	80.73
EU-Net	**90.28**	88.14	89.20	80.50
CBR-Net	90.22	88.23	89.21	80.53
Ours	89.26	**89.81**	**89.53**	**81.05**

5.3. Ablation Study

Some examples of the ablation study results are shown in Figures 10–12. As depicted by the red box in the figure, the modules added to the baseline played a positive role to varying degrees. The quantitative performance results are listed in Table 8. Baseline + E improved the IoU of Baseline by 0.63%, 0.39% and 0.69% for the three datasets, respectively, which proves that the multi-scale fusion of edge features is effective. The WHU and Inria datasets had a higher accuracy, as they have a higher resolution and more distinguishing edge features. We visualized the edge features of the sample images (the size of the images was 512 × 512) of the three datasets, as shown in Figure 13. The edge features of the Massachusetts dataset were identified to be coarser than those of other datasets. The addition of attention improved the IoU by 0.46%, 0.34% and 0.61%, respectively, proving the role of the attention strategy of scSE. By assigning weights to the features of different channels and the features of different spatial locations, the added edge features were applied more reasonably. Build-building increased the resulting IoU by 0.50%, 0.68% and 0.49%, respectively. The Massachusetts dataset had the most significant optimization effect, which might be due to this dataset possessing the greatest disparity in the ratio of building and non-building samples (the ratios are 0.23, 0.15 and 0.24 for the WHU, Massachusetts and Inria datasets). Build-building can reasonably (not simply repeat) increase the sampling frequency of building samples during model training; therefore, in theory, the smaller the ratio of building samples, the greater the effect.

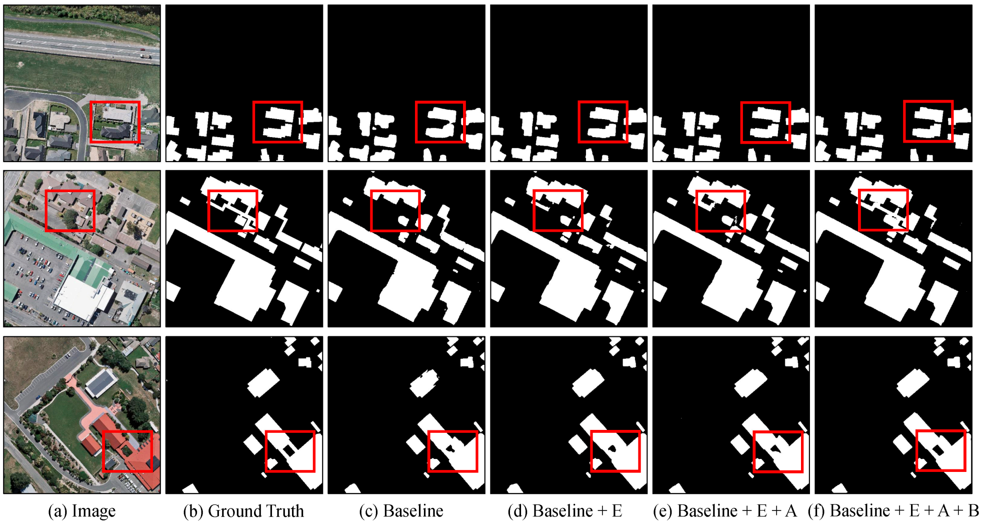


Figure 10. Ablation study results using the WHU dataset.

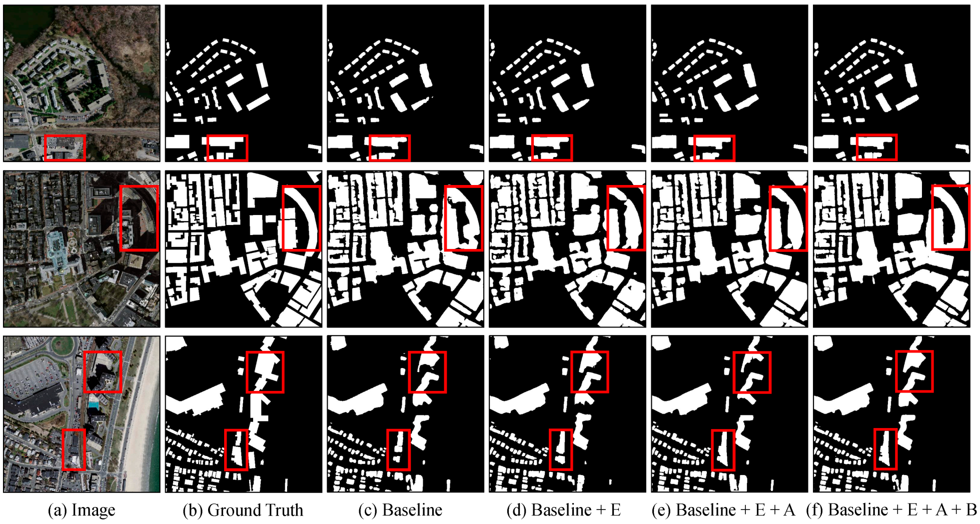


Figure 11. Ablation study results using the Massachusetts dataset.

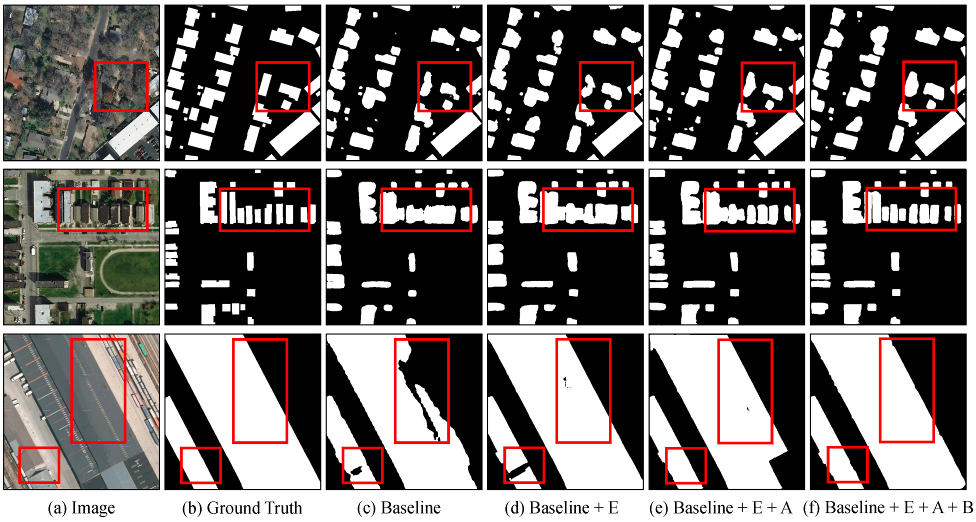
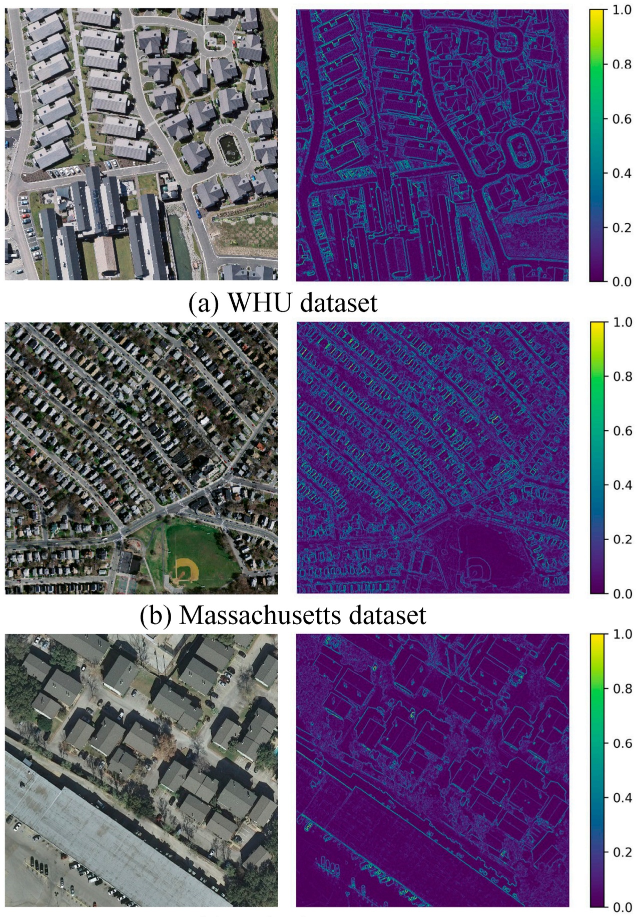


Figure 12. Ablation study results using the Inria dataset.



(a) WHU dataset

(b) Massachusetts dataset

(c) Inria dataset

Figure 13. Edge feature visualization of example images of the three datasets.

Table 8. Quantitative comparison of the four ablation studies.

Method	**WHU Dataset**		**Massachusetts Dataset**		**Inria Dataset**	
	F1 (%)	**IoU (%)**	**F1 (%)**	**IoU (%)**	**F1 (%)**	**IoU (%)**
Baseline	94.03	89.54	83.82	72.72	88.25	79.26
Baseline + E	94.55	90.17	84.23	73.11	88.68	79.95
Baseline + E + A	94.82	90.63	84.74	73.45	88.96	80.56
Baseline + E + A + B	95.36	91.13	85.15	74.13	89.53	81.05

5.4. Model Complexity

Floating point operations (FLOPs) and inference time are often used to measure model complexity. We visualized the FLOPs of the MEC-Net model and IoU using three datasets, as shown in Figure 14. The orange bars represent the IoU with the WHU dataset, the light green bars represent the IoU with the Massachusetts dataset, the light coral bars represent the IoU with the Inria dataset and the blue triangles represent the FLOPs. Although PSP-Net had the lowest FLOPs, its performance was the worst. HR-Net ranked second in terms of performance but had the highest FLOPs. The inference time for a single image slice for the models is presented in Table 9. The inference time of MEC-Net was located in the middle and was slightly more than that of PSP-Net and Res-U-Net. In summary, MEC-Net had the highest IoU, moderate FLOPs and inference speed, and a good overall performance.

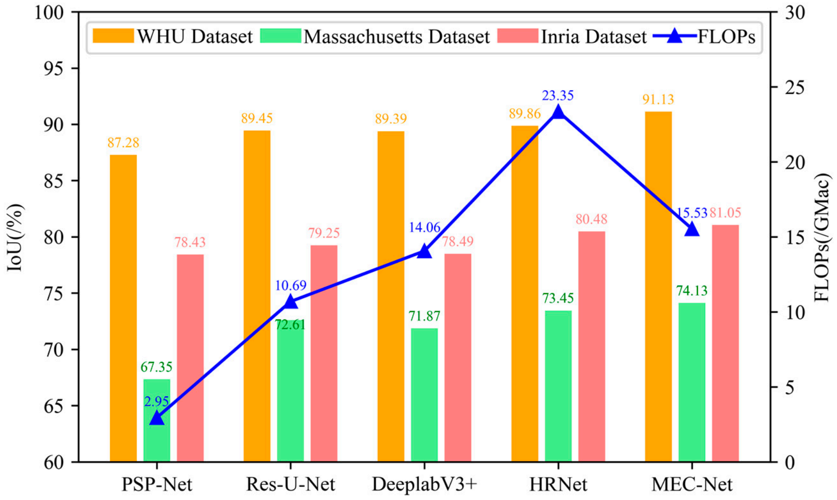


Figure 14. FLOPs for each model and their IoU with the three datasets.

Table 9. Inference time for a single image slice for the models.

Method	PSP-Net	Res-U-Net	DeeplabV3+	HR-Net	MEC-Net
Inference time	0.011 s	0.018 s	0.031 s	0.092 s	0.027 s

6. Conclusions

In this study, a new method for building extraction from remote sensing images, called MEC-Net, was proposed. Using prior knowledge of the high correlation between the boundaries of buildings and edge features, we designed an MEC structure that utilizes edge features to effectively constrain the model at the feature level. The edge features were fused multiple times with features learned by models at different scales to enhance the model's utilization of edge features. To alleviate the interference caused by the edge features of other objects, an attention mechanism was introduced to ensure that the model was more focused on the building area during the learning process. In addition, a building data augmentation method, called build-building, was proposed based on copy-paste combined with remote sensing imaging characteristics. The more unbalanced the positive and negative samples, the greater the improvement in the generalization of the model. Build-building is only used during training. As a result, the inference cost does not increase. MEC-Net was used to perform an experiment with the WHU, Massachusetts and Inria datasets, and it was found to exhibit state-of-the-art performance.

The edge feature calculation tool in MEC-Net is the Sobel operator; however, the CNN-based method updates the edge detection performance of the edge features. In the future, a CNN will be considered to replace the Sobel operator while controlling its additional time cost that comes as much as possible. MEC-Net has a relatively complete building boundary. Thus, converting the raster boundary into a vector polygon composed of building corners will be the focus of future research. The comprehensive utilization of data from other modalities (SAR and infrared data) and optical data is also a future direction. The source codes of MEC-Net are available at https://github.com/WangZhenqing-RS/MEC-Net, accessed on 2 February 2023.

Author Contributions: Conceptualization, Z.W.; Funding acquisition, Y.Z., F.W., S.W. and J.Z.; Methodology, Z.W. and S.W.; Software, Z.W., G.Q. and W.Z.; Supervision, Y.Z. and S.W.; Validation, G.Q. and W.Z.; Writing—original draft, Z.W. and Y.Z.; Writing—review & editing, F.W. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (No. 2022YFC3006400, 2022YFC3006402), the Finance Science and Technology Project of Hainan Province (No. ZDYF2021SHFZ103) and the Fujian Provincial Science and Technology Plan Project (No. 2021T3065).

Data Availability Statement: The data presented in this study are openly available in reference number [44–46].

Acknowledgments: The authors would like to thank the editors and the reviewers for their valuable suggestions.

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

References

1. Alshehhi, R.; Marpu, P.R.; Woon, W.L.; Mura, M.D. Simultaneous Extraction of Roads and Buildings in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 139–149. [CrossRef]
2. Dornaika, F.; Moujahid, A.; Merabet, E.Y.; Ruichek, Y. Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors. *Expert Syst. Appl.* **2016**, *58*, 130–142. [CrossRef]
3. Xiong, C.; Li, Q.; Lu, X. Automated Regional Seismic Damage Assessment of Buildings Using an Unmanned Aerial Vehicle and a Convolutional Neural Network. *Automat. Constr.* **2020**, *109*, 102994. [CrossRef]
4. Guo, H.; Du, B.; Zhang, L.; Su, X. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 240–252. [CrossRef]
5. Turker, M.; Koc-San, D. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *34*, 58–69. [CrossRef]
6. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for Object Segmentation and Fine-grained Localization. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 447–456.
7. Zhang, T.; Huang, X.; Wen, D.; Li, J. Urban Building Density Estimation from High-Resolution Imagery Using Multiple Features and Support Vector Regression. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3265–3280. [CrossRef]
8. Zhao, Y.; Ren, H.; Cao, D. The Research of Building Earthquake Damage Object-Oriented Change Detection Based on Ensemble Classifier with Remote Sensing Image. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, IGARSS, Valencia, Spain, 22–27 July 2018; pp. 4950–4953.
9. Gavankar, N.L.; Ghosh, S.K. Object based building footprint detection from high resolution multispectral satellite image using K-means clustering algorithm and shape parameters. *Geocarto Int.* **2019**, *34*, 626–643. [CrossRef]
10. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *5*, 161–172. [CrossRef]
11. Zhu, Y.; Huang, B.; Gao, J.; Huang, Y.; Chen, H. Adaptive Polygon Generation Algorithm for Automatic Building Extraction. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
12. Zheng, Z.; Ma, A.; Zhang, L.; Zhong, Y. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 254–264. [CrossRef]
13. Sun, Y.; Hua, Y.; Mou, L.; Zhu, X.X. CG-Net: Conditional GIS-aware Network for Individual Building Segmentation in VHR SAR Images. IEEE Trans. *Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
14. Griffiths, D.; Boehm, J. Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours. *ISPRS J. Photogramm. Remote Sens.* **2019**, *154*, 70–83. [CrossRef]
15. Guo, H.; Su, X.; Tang, S.; Du, B.; Zhang, L. Scale-Robust Deep-Supervision Network for Mapping Building Footprints From High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10091–10100. [CrossRef]
16. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *Lect. Notes Comput. Sci.* **2017**, *10111 LNCS*, 180–196.
17. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An Efficient Fully Convolutional Network for Building Extraction from Optical Remote Sensing Images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
18. Jin, Y.; Xu, W.; Zhang, C.; Luo, X.; Jia, H. Boundary-aware refined network for automatic building extraction in very high-resolution urban aerial images. *Remote Sens.* **2021**, *13*, 692. [CrossRef]
19. Wang, Z.; Zhou, Y.; Wang, S.; Wang, F.; Xu, Z. House building extraction from high resolution remote sensing image based on IEU-Net. *J. Remote Sens.* **2021**, *25*, 2245–2254. [CrossRef]
20. Yang, G.; Zhang, Q.; Zhang, G. EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images. *Remote Sens.* **2020**, *12*, 2161. [CrossRef]
21. Zheng, X.; Huan, L.; Xia, G.; Gong, J. Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 15–28. [CrossRef]
22. 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]
23. Shao, R.; Du, C.; Chen, H.; Li, J. SUNet: Change Detection for Heterogeneous Remote Sensing Images from Satellite and UAV Using a Dual-Channel Fully Convolution Network. *Remote Sens.* **2021**, *13*, 3750.

24. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
25. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
26. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef]
27. Peng, D.; Zhang, Y.; Guan, H. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
28. Zhang, Y.; Gong, W.; Sun, J.; Li, W. Web-Net: A novel nest networks with ultra-hierarchical sampling for building extraction from aerial imageries. *Remote Sens.* **2019**, *11*, 1897. [CrossRef]
29. Zhang, X.; Yue, Y.; Gao, W.; Yun, S.; Su, Q.; Yin, H.; Zhang, Y. DifUnet++: A satellite images change detection network based on UNet++ and differential pyramid. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8006605. [CrossRef]
30. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049.
31. 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]
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef]
33. Odena, A.; Dumoulin, V.; Olah, C. Deconvolution and checkerboard artifacts. *Distill* **2016**, *1*, 10. [CrossRef]
34. Kanopoulos, N.; Vasanthavada, N.; Baker, R.L. Design of an image edge detection filter using the Sobel operator. *IEEE J. Solid-State Circuits* **1988**, *23*, 358–367. [CrossRef]
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
36. Roy, A.G.; Navab, N.; Wachinger, C. Concurrent Spatial and Channel 'Squeeze & Excitation' in Fully Convolutional Networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, 16–20 September 2018; Springer International Publishing: Cham, Germany, 2018; pp. 421–429.
37. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 905–909. [CrossRef]
38. Zhang, P.; Du, P.; Lin, C.; Wang, X.; Li, E.; Xue, Z.; Bai, X. A hybrid attention-aware fusion network (Hafnet) for building extraction from high-resolution imagery and lidar data. *Remote Sens.* **2020**, *12*, 3764. [CrossRef]
39. Ghiasi, G.; Cui, Y.; Srinivas, A. Simple copy-paste is a strong data augmentation method for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2918–2928.
40. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
41. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 6023–6032.
42. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:1512.00567.
43. Milletari, F.; Navab, N.; Ahmadi, S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
44. 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.* **2019**, *57*, 574–586. [CrossRef]
45. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
46. 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, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
47. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
48. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
49. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
50. Zhang, Z.; Liu, Q.; Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [CrossRef]
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. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
53. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
54. Zhang, H.; Liao, Y.; Yang, H.; Yang, G.; Zhang, L. A Local-Global Dual-Stream Network for Building Extraction From Very-High-Resolution Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 1269–1283. [CrossRef] [PubMed]
55. 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.* **2021**, *59*, 6169–6181. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.