applied sciences

MDPI

Article

A Multiscale Attention-Guided UNet++ with Edge Constraint for Building Extraction from High Spatial Resolution Imagery

Hua Zhao, Hua Zhang * and Xiangcheng Zheng

School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; huazhao@cumt.edu.cn (H.Z.); ts20160046a31@cumt.edu.cn (X.Z.)
* Correspondence: zhhuacumt@cumt.edu.cn

Abstract: Building extraction from high spatial resolution imagery (HSRI) plays an important role in the remotely sensed imagery application fields. However, automatically extracting buildings from HSRI is still a challenging task due to such factors as large size variations of buildings, background complexity, variations in appearance, etc. Especially, it is difficult to extract both crowded small buildings and large buildings with accurate boundaries. To address these challenges, this paper presents an end-to-end encoder–decoder model to automatically extract buildings from HSRI. The designed network, called AEUNet++, is based on UNet++, attention mechanism and multi-task learning. Specifically, the AEUNet++ introduces the UNet++ as the backbone to extract multiscale features. Then, the attention block is used to effectively fuse different-layer feature maps instead of direct concatenation in the output of traditional UNet++, which can assign adaptive weights to different-layer feature maps as their relative importance to enhance the sensitivity of the mode and suppress the background influence of irrelevant features. To further improve the boundary accuracy of the extracted buildings, the boundary geometric information of buildings is integrated into the proposed model by a multi-task loss using a proposed distance class map during training of the network, which simultaneously learns the extraction of buildings and boundaries and only outputs extracted buildings while testing. Two different data sets are utilized for evaluating the performance of AEUNet++. The experimental results indicate that AEUNet++ produces greater accuracy than U-Net and the original UNet++ architectures and, hence, provides an effective method for building extraction from HSRI.

Keywords: building extraction; high spatial resolution imagery (HSRI); UNet++

Citation: Zhao, H.; Zhang, H.; Zheng, X. A Multiscale Attention-Guided UNet++ with Edge Constraint for Building Extraction from High Spatial Resolution Imagery. *Appl. Sci.* **2022**, *12*, 5960. https://doi.org/10.3390/app12125960

Academic Editor: Antonio Fernández

Received: 19 April 2022
Accepted: 10 June 2022
Published: 11 June 2022

1. Introduction

With the rapid development of high spatial resolution imagery (HSRI), which makes such small objects as buildings identifiable from images, immediate and accurate extraction of buildings from HSRI has become one of the interesting focuses in the remote sensing field [1–4]. Effective extraction of buildings is significant for illegal building detection, urban planning, disaster emergency response, etc. In practice, it is time-consuming to extract buildings from images by manual mode. Automatic and timely building extraction methods are urgent to be developed. Luckily, several methods have been presented to automatically extract buildings from images in recent decades. They can be roughly divided into two categories: traditional image processing and deep learning-based methods. The traditional building extraction method mainly trains an efficient classifier for extraction of buildings from images using a mount of samples containing such handcrafted features as the spectrum, texture, geometry, shadow, etc. [5–11]; these algorithms have made important progress in building extraction. While the handcrafted features vary with different sensor types, building structures, light conditions, etc., the traditional methods are only suitable for specific types of buildings, and their generalization ability is limited.

Recently, deep learning technologies, especially the convolutional neural network (CNN), have shown potential in remote sensing applications [12]. The use of deep learning for building extraction has attracted considerable attention, because it has good feature representation ability [13–15]. However, the typical CNN models, such as AlexNet [16] and VGG networks [17], are usually used to mark the entire image, not to label each pixel to a class. Evolving from CNN, the fully convolutional network (FCN) contributes greatly to semantic segmentation, which can assign a label to each pixel. Inspired by the great success of FCN in semantic segmentation, building extraction can be completed by semantic segmentation. Thus, on the basis of FCN architecture, many FCN-based models have been proposed for extracting buildings [18–30]. However, due to pooling operators in the CNN or the availability of only a portion of features to generate the final feature map, the problems of missing image content details and poor boundary accuracy can arise. Furthermore, due to large intra class variance and small inter class variance in the pixel values and similarities between buildings and their backgrounds, as well as the materials, proportions and illumination of buildings, extensive salt-and-pepper noise and boundary ambiguities exist in the building extraction results produced by the FCN-based methods.

To address the problem of spatial location and content details induced by the downsampling operations in the CNN, many improved methods have been presented for achieving more accurate edges by using more boundary geometric information or post-processing operators. Shrestha et al. [18] designed a new FCN for extraction of buildings from images, in which post-processing conditional random fields (CRFs) are added at the end of the network to refine the coarse pixel level label predictions to produce fine-grained segmentation results. The principle is to transform the pixel classification problem into a probabilistic reasoning problem. Wei et al. [19] designed a deep network for extracting building footprints, in which the polygon regularization is conducted on the initial result to obtain a rectangle building map. Xia et al. [20] presented a CNN model that combined full-scale skip connections and an edge guidance module to improve the location accuracy of buildings. Sun et al. [21] introduced the active contour model to the CNN, combining remote sensing images and LiDAR data for precise building extraction. However, while most of the above methods need accurate edge information, auxiliary data or sophisticated structures, it is difficult to achieve accurate edges due to poor spatial resolution, spectral similarity, and mixed pixels. Furthermore, CRF-based methods do not sufficiently extract features from the images, lack adequate information propagation, and use post-processing that also lowers model performance.

Improving multiscale feature extraction ability is also an important way to enhance building extraction performance. Based on FCN and VGG16, the multiscale decoding network (MSDNet) was designed for the semantic segmentation of images [31]; an inception module, which combines the un-pooling, transposed convolution, and dilated convolution paths, is treated as its decoding part. Ma et al. [32] proposed the GMEDN framework, in which a distilling decoder part is used to extract the multiscale features for precise prediction results. Rastogi et al. [33] provided the UNet-AP model for accurate building footprint extraction from very-high resolution remote sensing satellite imagery. In which atrous convolution is applied to feature extraction for enhancing the representation of objects at different scales in the image. The USPP framework was provided for building segmentation in high-resolution remote sensing [34], which enables extraction of features at multiple spatial scales and at the same time up-samples the feature maps to learn global contextual information by incorporating a spatial pyramid pooling module. These models try to extract and fuse the multiscale features in the network; however, to reduce the complexity of the model or improve its efficiency, the multiscale feature extraction modules are mainly implemented in the decoding part of the network. Thus, extraction of the multiscale features is not enough, for example, lacking refinement of the feature maps extracted by the network to reduce negative feature information, the fusion of deep and shallow features in the encoding stage, the direct multiscale features from the input images, etc.

In order to fully utilize different-level feature maps, residual networks or skip-layer connections are used to fuse the shallow layers and deep layers in some proposed models. For example, the dilated convolutions are used to increase the receptive field instead of using pooling layers [22–24]. However, dilated convolution-based models must perform convolutional operations on high-resolution feature maps through the entire network, which makes such models difficult to train and computationally expensive. Another important approach is to use skip connections, such as U-Net [25], UNet++ [26], SegNet [27], etc. They directly connect a convolutional layer in the encoder and a corresponding layer in the decoder. Thus, the fusion of the feature maps of earlier layers and the discriminative feature maps up-sampled at the end of the encoder step can obviously refine the segmentation map. Therefore, the skip-connections approach is widely used in encoder–decoder models for improving the performance of remote sensing classification [35,36]. Diakogiannis et al. [37] proposed the ResUNet-a model, in which U-Net was taken as the encoder/decoder backbone, and the pyramid scene parsing pooling module, residual connections module and multi-tasking module were combined. The above methods can perform well in most cases, but while these networks are able to improve the overall segmentation results, the boundaries between two different semantic classes often can not be well defined.

Above all, despite the great progress achieved by the above methods, accurate extraction of buildings from remote sensing images is still a challenge. Presently, attention mechanisms [38] and multi-task learning [39] are widely used in image processing; among them, the goal of applying attention mechanisms is to help select effective information used in the network. The aim of multi-task learning is to leverage useful information contained in related tasks to help improve the generalization performance of all tasks. Inspired by attention mechanisms and multi-task learning, in this paper, to address the aforementioned problems based on the traditional UNet++ architecture combined with attention mechanisms and multi-task learning, we designed an improved UNet++ architecture with an attention block and edge preservation (AEUNet++) for accurate extraction of buildings from HRSI. Based on UNet++, we exploited deep structured feature fusion techniques to enhance the feature fusion by giving a trainable weight to the different feature maps using the convolutional block attention module (CBAM) [40]. Further, in order to address the problem of segmentation prediction results with poor boundaries, we incorporated the boundary information of the building mask into the network by introducing a multi-task loss based on the distance class map. The aim was to produce accurate semantic segmentation results in homogeneous regions and preserve image details.

The main advantages of the AEUNet++ are as follows:

(1) An improved UNet++ for the wise fusion of extracted feature maps is proposed, in which the CBAM, including the spatial attention and channel attention gates, is introduced to learn 'where' and 'what' the meaningful representations of the given features are. It significantly suppresses the drawbacks of direct concatenation by averaging the operations in the UNet++ models, thus improving the segmentation accuracy.
(2) To improve the boundary precision of extracted buildings, the boundary geometric information of the building is introduced into the proposed AEUNet++ by using a multi-task loss based on the proposed distance class map.
(3) The proposed AEUNet++ achieved 1.62% and 1.8% F1 and 2.0% and 3.24% intersection over union (IoU) improvements compared with UNet++ on the Massachusetts building data set [41] and the WHU data set [36] and outperforms two other SOTA methods on the two data sets.

2. Methodology

The purpose of this paper was to explore a network to improve the accuracy of building extraction from HSRI, especially for enhancing poor boundaries. By introducing the attention mechanism and multi-task learning into the conventional UNet++, we designed an end-to-end encoder–decoder network for automatically extracting buildings from HSRI (as shown in Figure 1). The proposed model mainly contains three modules: multiscale

features extraction module, attention block, and multi-task learning module. First, remote sensing images are fed into the multiscale feature maps extraction module for extracting multiscale feature maps, then, to enhance the fusion of the multiscale feature maps, the attention block is applied to enhance the fusion of feature maps from different hierarchical layers according to their degrees of importance. Lastly, to further address the problem of poor extracted building boundaries caused by the pooling operations in the AEUNet++, the multi-task learning module is introduced to optimize the segmentation results for producing fine-grained segmentations with accurate boundaries.

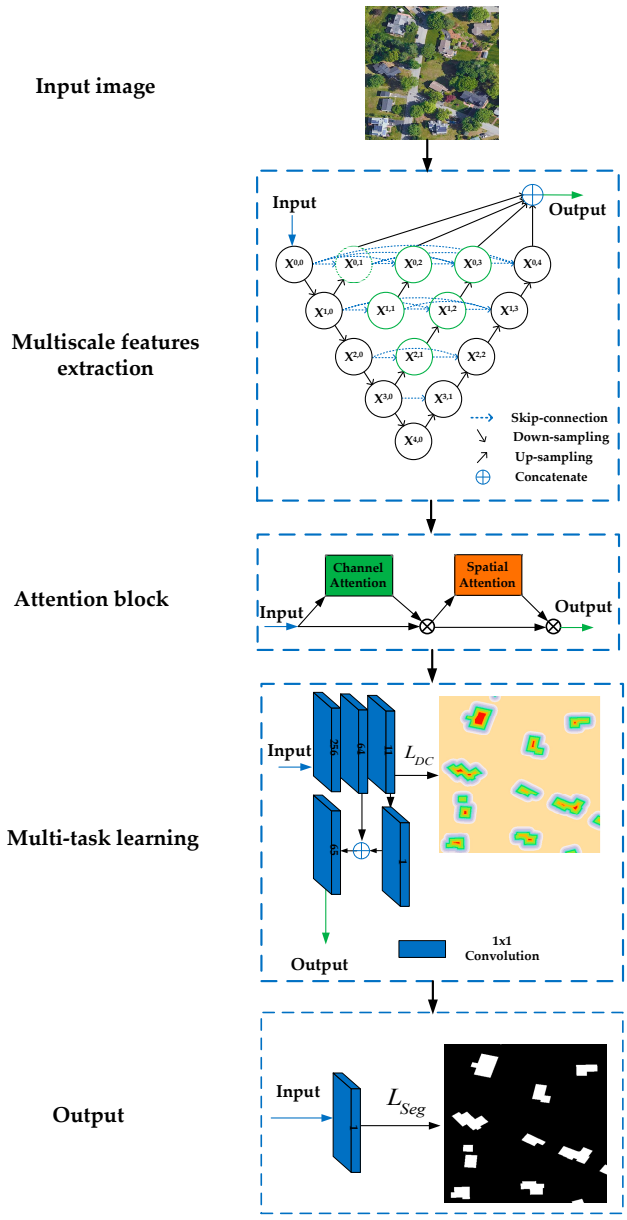

Figure 1. The architecture of the proposed AEUNet++ (the output of previous stage is the input of the next stage).

2.1. Multiscale Feature Extraction Module

Variants of encoder–decoder architectures such as FCN and U-Net had been widely used to extract multiscale features from images due to skip connections, which can combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network. Specially, UNet++ [26]

is a variant of U-Net, which can narrow and fill the information gap between the feature maps of the encoder and decoder prior to fusion. In this study, the multiscale feature extraction module of the proposed AEUNet++ was based on the traditional UNet++, in which the encoder and decoder sub-networks are connected through a series of nested and dense skip pathways. Additionally, long-range connections are also introduced by the encoder and the corresponding decoder parts; thus, different hierarchical feature maps from the encoder can be fully fused in the decoder part, and as a result, the network becomes much more precise and expansible. As shown in the part on multiscale feature extraction in Figure 1, let $x^{i,j}$ denote the output of node $X^{i,j}$, where i denotes ith down-sampling layer along the decoder pathway and j is jth convolution layer along the skip pathway. The fused feature maps $x^{i,j}$ can be represented as:

$$x^{i,j} = \begin{cases} \delta^{cov}(x^{i-1,j}), & j = 0 \\ \delta^{cov}(\delta^{cat}(\delta^{cat}(x^{i,0}, x^{i,1} \cdots, x^{i,j-1}), \delta^{up}(x^{i+1,j-1}))), & j > 0 \end{cases} \quad (1)$$

where δ^{cov} represents a convolution operation including an activation function, δ^{cat} is the concatenation, and δ^{up} represents an up-sampling layer. If $j = 0$, $X^{i,j}$ represents the nodes in the encoder sub-network. If $j > 0$, $X^{i,j}$ represents the concatenation results of all the other nodes in the same level and the up-sampled result of $X^{i+1,j-1}$, which includes the deeper, coarser and sematic information.

2.2. Attention Block

As shown in the part on multiscale feature extraction in Figure 1, $X^{0,1}$, $X^{0,2}$, $X^{0,3}$ and $X^{0,4}$ represent the four predicted feature maps generated by the UNet++, respectively. In the traditional UNet++, the four feature maps are directly concatenated by an averaging operation, which will suppress the better feature maps and raise the negative feature maps on the final output. The reasons may be as follows: for the extracted hierarchical semantic features, the low-level feature maps have poor semantic information but rich spatial location information for the small receptive field, whereas the high-level feature maps have strong semantic information but weak spatial location information because of the large receptive field. Hence, feature maps from different levels should be concatenated discriminately to make networks allocate reasonable attention to the high-level and low-level features. Furthermore, it is necessary to emphasize the important parts and suppress the unimportant parts due to the extracted features that are often spatially affected by similar patterns and noisy backgrounds. Thus, in this paper, in order to select representative features when fusing different level feature maps, we assigned weights to different-layer feature maps as their relative importance in the channel dimension and sub-features in the spatial dimension. Thus, inspired by the great progress in attention mechanisms of neural networks, CBAM was introduced in our proposed network to learn which information to emphasize or suppress, which is a sequential combination of a channel and a spatial attention module (as shown in the attention block in Figure 1).

As shown in Figure 2, given the predicted feature map $F \in R^{C \times H \times W}$ generated from the multiscale features extraction module, $N = H \times W$ denotes the number of spatial pixels, and C is the dimension of the feature map.

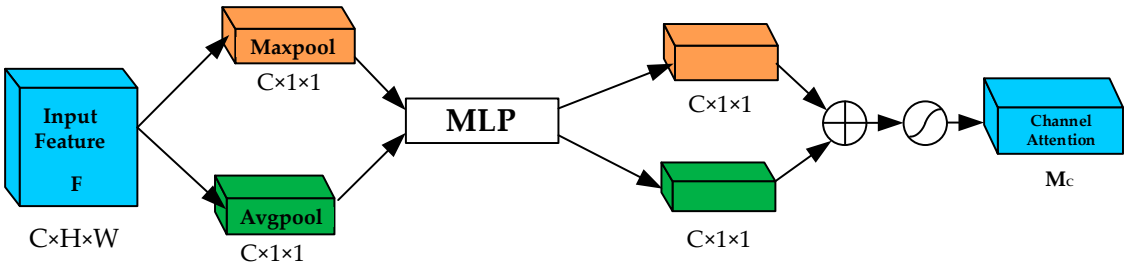

Figure 2. The channel attention mechanism.

The channel attention explores 'what' is useful in an input image by exploiting the inter-channel relationships of features. Firstly, two descriptors including average-pooled features F^C_{AVG} and max-pooled features F^C_{MAX} are generated by using global average-pooling and global max-pooling for aggregating the global information of each channel and clues about distinctive object features, respectively. The two descriptors are then forwarded to a shared multi-layer perceptron (MLP) (as shown in Figure 3) to produce two vectors, which are next merged by an element-wise sum and finally output the channel attention map $M_c(F)$:

$$M_c(F) = \text{Sigmoid}(a_1(a_0(F^C_{AVG})) + a_1(a_0(F^C_{MAX})) \quad (2)$$

where a_0 and a_1 denote the MLP weights.

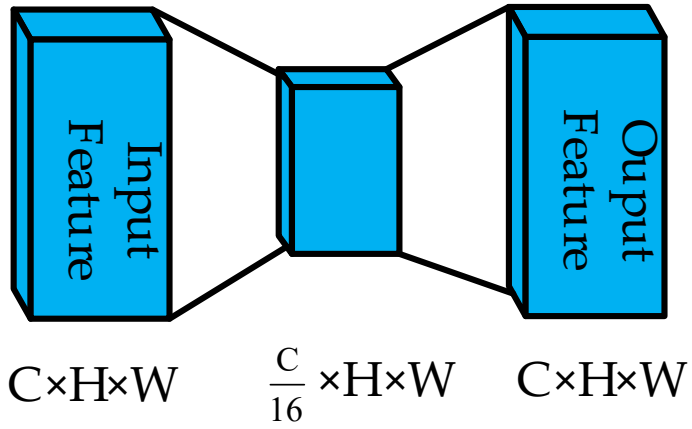

Figure 3. The multi-layer perceptron (MLP) consisting of two convolution layers with a filter size of 1.

The spatial attention explores 'where' the informative regions that should be paid more attention are, just as illustrated in Figure 4. Firstly, based on the channel refined feature F', average-pooled features F^S_{AVG} and max-pooled features F^S_{MAX} across the channel are generated by performing average-pooling and max-pooling for aggregating the channel information, respectively. Then, they are concatenated and convolved by a standard 7×7 convolution layer, producing the spatial attention map:

$$M_s(F) = \text{Sigmoid}(f^{7\times7}([F^S_{AVG}; F^S_{MAX}])) \quad (3)$$

where $f^{7\times7}$ denotes a convolution operation with size of 7×7.

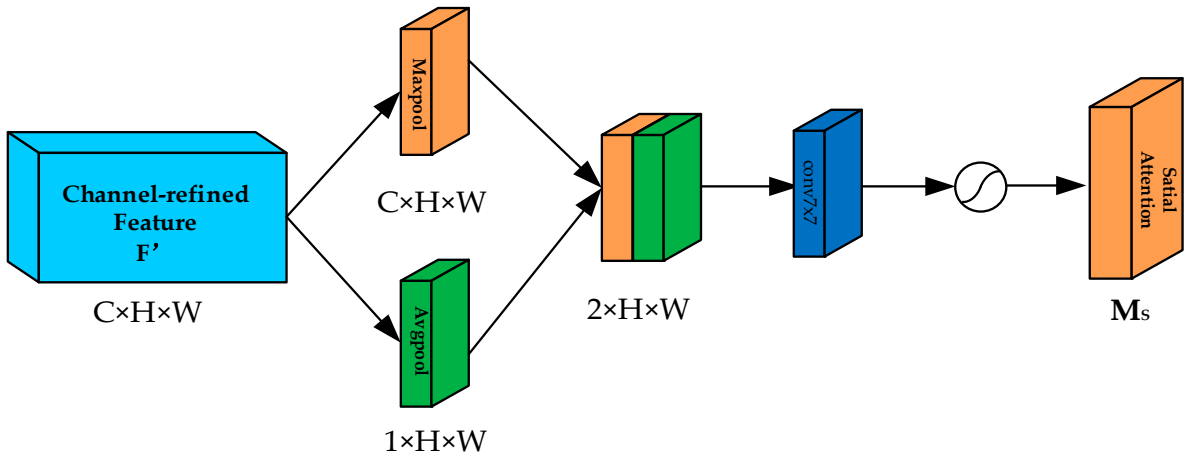

Figure 4. The spatial attention mechanism.

Then, the CBAM sequentially infers the channel attention map $M_c \in R^{C\times1\times1}$ and the spatial attention $M_s \in R^{1\times H\times W}$, and the total attention process can be written as:

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F', \end{aligned} \tag{4}$$

where $\otimes$ represents element-wise multiplication and F'' is the final refined feature map. $M_c(F)$ and $M_s(F')$ are the channel and spatial attention maps as described above, respectively.

At last, the four prediction feature maps generated by the multiscale feature extraction module are fused by assigning adaptive weights to different-layer feature maps as their relative importance, which is refined by the CBAM.

2.3. Multi-Take Learning Module

Due to the downsampling operations in the UNet++ network, image content details and spatial location information often are missed in the result; in particular, buildings are segmented with poor boundaries. To improve the accuracy of boundaries, here, we introduced the building edge geometric information into the proposed network as the constraints to guide the network to produce precise boundaries, which is trained by the multi-task learning based on the distance class map defined in this part. As shown in the multi-task learning part of Figure 1, two convolutional layers L_{DC} and L_{Seg} are added to the AEUNet++ to balance semantic properties and the boundary geometric properties, where L_{DC} is used to predict the distance class to the edges of building, and the segmentation of buildings is predicted by L_{Seg}.

Building masks and boundary maps (as shown in Figure 5) are usually used to incorporate building geometric information into the network, while both of them have their own advantages and shortcomings. If using the boundary maps, due to spectral variation, limited spatial resolution, noise pixels, and the fact that edges only occupy a tiny part of the whole image, it is difficult for the network to produce accurate closed outlines that fit the boundaries of the buildings well. Furthermore, the boundary maps cannot judge whether pixels are inside or outside buildings. Using the building mask maps can yield a better result, but it cannot represent boundaries of adjacent buildings. To solve these problems, we introduced the distance class map to represent the building labels in the proposed AEUNet++. As shown in Figure 5, the signed distance function is firstly introduced to represent the distances of pixels between the boundaries' pixels, and then extracted as output representation constraints. The value of the signed-distance function for a pixel denotes the distance between the pixel and its nearest boundary pixel, and positive and negative values indicate whether the pixel is inside or outside the building, respectively. For convenience, the distance is truncated at a given threshold, and the truncated signed-distance function can be described as:

$$Dist(i) = \delta_d \min(\min_{j\in X} d(i,j), d) \tag{5}$$

where i denotes a pixel in an image, x denotes the set of pixels belonging to the building boundaries, and $\min_{j\in X} d(i,j)$ is the Euclidean distance between the pixel i and its nearest boundary pixel j. d is a threshold, and δ_d denotes a sign function to indicate that the pixel is inside or outside the building; if $\delta_d = 1$, pixel i is inside the building mask, while $\delta_d = -1$ represents that pixel i is outside the building mask.

In this paper, to facilitate training of the network, we defined the distance class map by uniformly quantizing the truncated signed-distance $Dist(i)$ to a limited number of classes at equal intervals. As shown in Figure 5d, different values represent different distance classes. In this case, the boundary prediction is converted into the multi-label segmentation task, because pixels are assigned to finely divided classes based on their distance to boundaries instead of a small number of coarse classes (e.g., building and non-building). The distance

class maps guide the network to differentiate regions with different spatial relations to building masks.

(**a**) (**b**) (**c**) (**d**)

Figure 5. Extracted building representation. (**a**) Original image. (**b**) Building mask. (**c**) Boundary map. (**d**) Distance class map (different numbers stand for different nearest distances from pixels to the boundaries).

2.4. Loss Function of AEUNet++

As described above, our proposed AEUNet++ comprises an edge branch and segmentation branch, and we define the multi-task loss as follows:

$$L = \lambda_1 L_{seg} + \lambda_2 L_{DC} \tag{6}$$

where L_{seg} is the semantic segmentation task loss function and L_{DC} is the distance class map prediction task loss function, and they are weighted by λ_1 and λ_2, respectively. Usually, the weighting terms λ_i are set equal or found through an expensive grid-search. Here, the uncertainty-based multi-task loss [35] is introduced to determine the weighting terms λ_i by using the uncertainty in the model's prediction for each task, in which, depending on the confidence of the individual task prediction, a relative task weight is learned. Equation (6) can be rewritten as follows:

$$L(x; \theta, \sigma_{DC}, \sigma_{seg}) = L_{seg}(x; \theta, \sigma_{seg}) + L_{DC}(x; \theta, \sigma_{DC}) \tag{7}$$

where θ denotes the network parameters, x is the trained images, and $\sigma_{DC}, \sigma_{seg}$ are the corresponding task weights for λ_i, respectively.

If the likelihood of the model for each classification task is represented by the model output $f(x)$ with the uncertainty through a SoftMax function:

$$P(C = 1|x, \theta, \sigma_t) = \frac{\exp[\frac{1}{\sigma_t^2} f_c(x)]}{\sum_{c\prime=1} \exp[\frac{1}{\sigma_t^2} f_{c\prime}(x)]} \tag{8}$$

where P is the multi-task estimation, $f_c(x)$ is the desired output, $f_{c\prime}(x)$ is the original actual input, and σ_t denotes the scaling factor.

Using the negative log likelihood for Equation (4), and expressing the classification loss with uncertainty as follows:

$$\begin{aligned} L(x, \theta, \sigma_t) &= \sum_{c=1}^{C} -C_c \log P(C_c = 1|x, \theta, \sigma_t) \\ &= \sum_{c=1}^{C} -C_c \log\{\exp[\tfrac{1}{\sigma_t^2} f_c(x)]\} + \log \sum_{c\prime=1}^{C} \exp[\tfrac{1}{\sigma_t^2} f_{c\prime}(x)] \end{aligned} \tag{9}$$

Assuming the multi outputs consist of continuous and discrete outputs, they are modeled by Gaussian likelihood and SoftMax, respectively. Then, the loss function can be simplified as follows:

$$\frac{1}{\sigma_t^2}\sum_{c\prime}\exp[\frac{1}{\sigma_t^2}f_{c\prime}(x)] \approx \{\sum_{c\prime}\exp[f_{c\prime}(x)]\}^{\frac{1}{\sigma_t^2}} \tag{10}$$

Combining Equations (9) and (10), the loss function of the network is described as follows:

$$L(x,\theta,\sigma_t) \approx \frac{1}{\sigma_t^2}\sum_{c=1}^{C} -C_c \log P(C_c = 1|x,\theta) + \log(\sigma_t^2) \tag{11}$$

3. Experiments and Results

3.1. Data Sets

To evaluate the performance of AEUNet++, the data sets we used in the experiments were two open building data sets, i.e., the Massachusetts buildings dataset and WHU aerial image data set. In addition, as shown in Figures 6 and 7, we can see that the two data sets cover various building characteristics, such as size, shape, and spatial resolution and distribution, Thus, the two kinds of data sets can be used to evaluate the generalization ability of the proposed model.

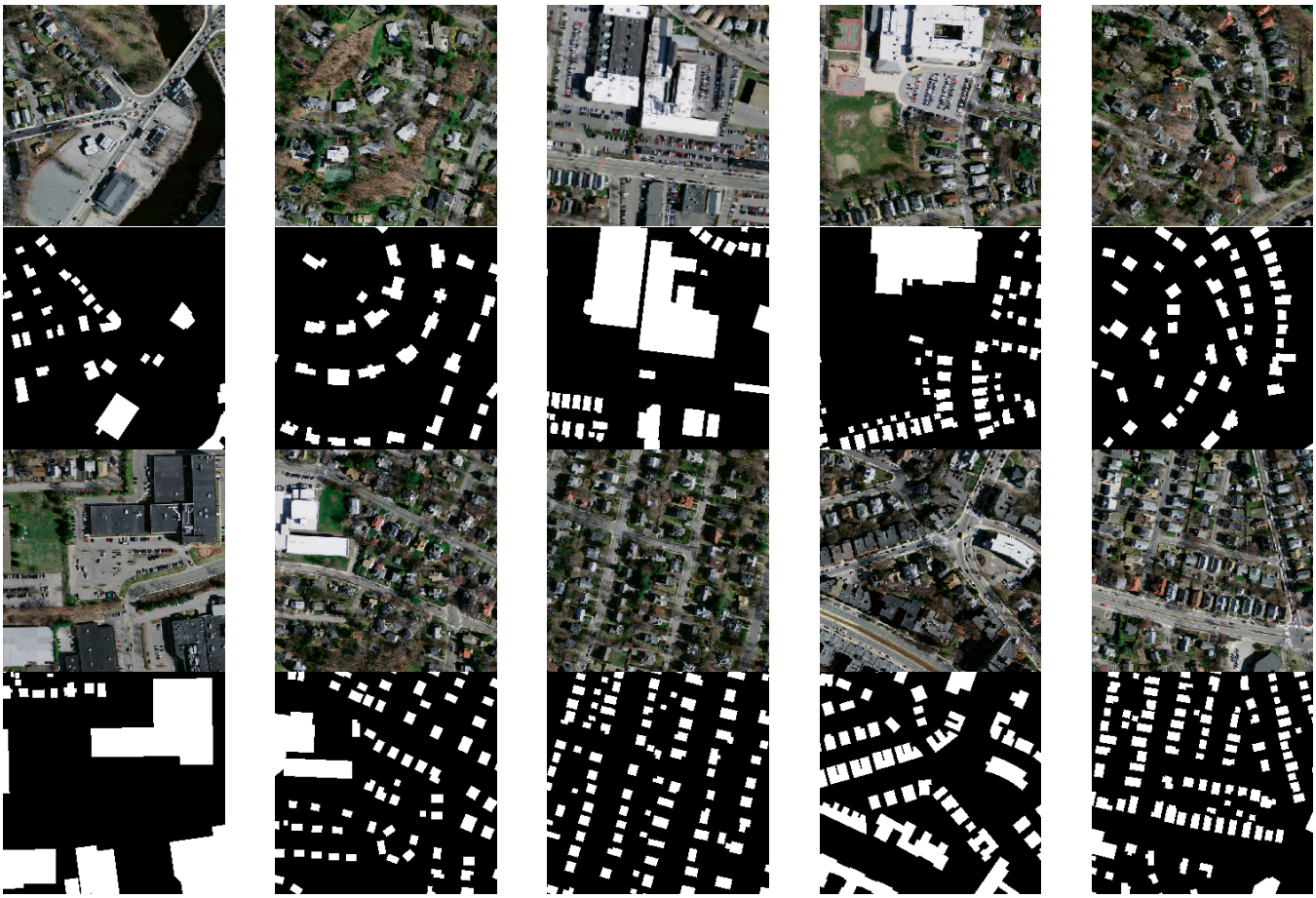

Figure 6. Samples of the Massachusetts buildings data set.

Figure 7. Samples of the WHU aerial image data set.

3.1.1. Massachusetts Buildings Data Set

The Massachusetts building data set [37] consists of many 1 m resolution RGB aerial images with the size of 1500 × 1500 pixels. The original data set contains 137 images for training, 10 images for testing, and 4 images for validating. For computational convenience, in the data preprocessing stage, we cropped all of the images into 512 × 512 pixels and added a flip operation for data augmentation. Through augmentation, there were 4700 images for training and 144 images for validating. Lastly, we transformed labeled images to grayscale images with pixel values of 0 and 1. Some samples are shown in Figure 6.

3.1.2. WHU Aerial Image Data Set

The WHU data set [38] consists of an aerial image sub-dataset and two satellite image sub-datasets. In this experiment, only the aerial image data sets were selected as the test data. The WHU aerial dataset covers 450 km^2 in Christchurch, New Zealand, and about 220,000 independent buildings are contained in it. The spatial resolution of the WHU arterial image data set is 0.3 m. In the data reprocessing stage, images were cropped into 512 × 512 pixels. All of the cropped images were sliced into training, test and validation sets numbering 4736, 2416 and 1036 images, respectively. Some samples of them are shown in Figure 7.

3.2. Network Configurations and Training

In Section 2, we proposed the AEUNet++ model based on the traditional UNet++. It can be regarded as an improved UNet++, so in the two experiments, the performance of the proposed AEUNet++ model was compared with the four SOTA methods, including U-Net, UNet++, SegNet and DeeplabV3+. All networks in the experiments used the same settings: adaptive moment estimation (Adam) optimizer was adopted, the initial learning rate was 0.0001, batch size was 5, and the learning rate was decreased by 0.5 times every 50 epochs. We conducted 100 epochs on the two data sets. All networks were implemented with Pytorch 1.2.0 and python 3.7.9, and were checked using a single Nvidia Tesla P100 GPU with 16 GB.

In addition, before the training stage, it was necessary to generate the defined distance maps used in the multi-task learning module. In all experiments, as shown in Figure 5, based on the building mask in the ground truth samples, the building boundaries were extracted through the canny edge detection. Then, for each of the building boundaries, the truncated signed-distance of each pixel was calculated based on Equation (5). Finally, each building mask described by the truncated signed-distance was categorized into different classes, namely distance class maps (e.g., Figure 5d). Here, we repeated tests using AEUNet++ with different truncation thresholds by step 1 in the interval [15,30] and the different truncated signed-distance class number by step 1 in the interval [5,20] in the WHU data set. Figure 8 shows the intersection over union (IoU) curve obtained with different pairs of thresholds (drawing a point every 10 points); the curve fluctuates and reached the highest IoU at threshold pairs (20, 11); thus, the truncation threshold was set to 20, and the truncated signed-distance class number was set to 11 in the experiment.

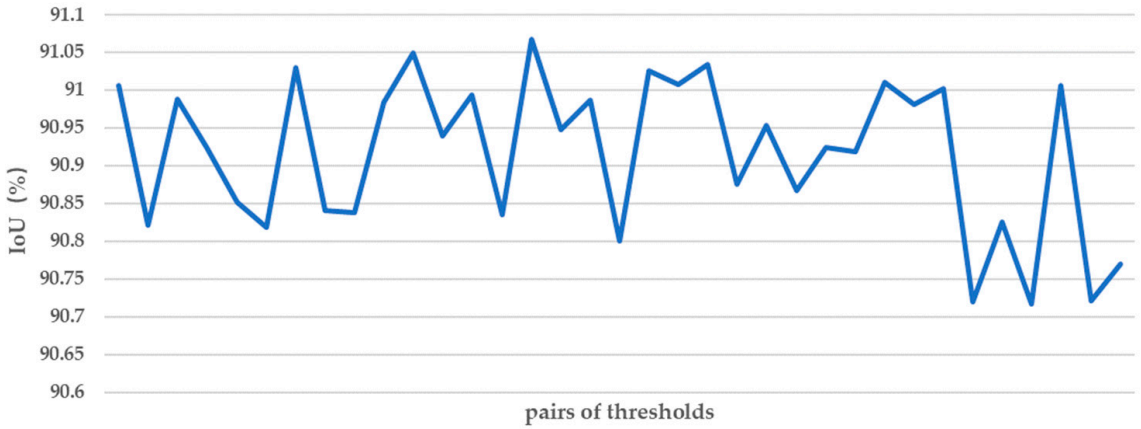

Figure 8. The IoU obtained with different pairs of thresholds for AEUNet++ in the WHU data set.

3.3. Metrics

To quantitatively evaluate the effectiveness of the proposed network, overall accuracy (OA), recall, precision, F1 score, IoU and mIoU were utilized to evaluate its performance. They are defined as follows:

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{12}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{15}$$

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

$$\mathrm{mIoU} = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k} \tag{17}$$

where TP denotes true positives (correctly extracted building pixels), FP denotes false positives (pixels mislabeled as buildings in results), TN denotes true negatives (correctly identified nonbuilding pixels), and FN denotes false negatives (pixels incorrectly labeled as non-buildings or that could be interpreted as missed building pixels). TP_k, FP_k and FN_k denote the TP, FP and FN of class k, respectively, and N is the number of classes.

3.4. Experimental Results

3.4.1. Results with the Massachusetts Buildings Data Set

Figure 9 lists the building extraction results for the Massachusetts buildings data set predicted by the U-Net, UNet ++, SegNet, DeeplabV3+ and AEUNet++, respectively. Seen from the segmentations of U-Net, SegNet and DeeplabV3+, due to mixed pixels and spectral similarity in the image, they produced prediction results including an amount of salt-and-pepper noise and showed weaker performance compared with the other two methods. Additionally, many mislabeled pixels were produced; for example, many road pixels were labeled as building pixels in marked areas A and E, some building pixels were mislabeled as background in marked areas D and F because the plain skip connections are often used in U-Net for fusing multiscale features, and many noise features were extracted by SegNet and DeeplabV3+, which may result in insufficient fusion of feature maps from different layers. Seen from marked areas A, C, D and F, UNet ++ and AEUNet++ presented more homogeneous segmentation maps with accurate boundaries due to fine-grained details and features captured by the nested and dense skip connections introduced in UNet++ and AEUNet++. Compared with UNet++, AEUNet++ produced more homogeneous segmentation maps with accurate boundaries due to the introduce of attention block and multi-task learning to adaptively fuse multiscale features and edge constraints. Although UNet++ nearly eliminated most of the isolated noises and obtained satisfactory results, a number of details in the image content were missing, and many pixels were mislabeled. As we can see from the segmentation results obtained with AEUNet++, most of the isolated noises were eliminated and details of the image content were satisfactorily preserved. Taking the marked area E as an example, many road pixels were mislabeled as building pixels by UNet++. Seen from the marked area F, many shadow pixels were mislabeled as building pixels by UNet++, while in AEUNet++, the pixels were mostly correctly labeled. The main reasons may be as follows: in the traditional UNet++, the total loss was obtained by averaging the final output of extracted feature maps by UNet++; the simple averaged result may suppress the better feature maps and raise the negative feature maps from the final output. However, in AEUNet++, the four prediction feature maps were adaptively refined through the attention block, and as a result, feature maps were learned that should have been emphasized or suppressed. As we can see from Figure 8, AEUNet++ achieved the most accurate boundaries among the three methods, because in the three methods, progressive down-sampling may cause the loss of location information, which will result in poor boundaries. The multi-task learning was introduced into the proposed AEUNet++ network to incorporate building boundary geometric properties, and as a result, more accurate boundaries will be preserved in comparison with U-Net, UNet++, SegNet and DeeplabV3+.

Table 1 gives the quantitative evaluation results. Seen from Table 1, UNet++ and AEUNet++ yielded higher segmentation accuracies than that of U-Net. Amongst the three models, AEUNet++ obtained the greatest accuracy. Compared with UNet++, after adding the attention block and multi-task learning training, the prediction accuracy was significantly improved; AEUNet++ obtained 95.12%, 83.40%, 66.10%, 73.75%, 58.41%, and 76.58% with respect to OA, Precision, Recall, F1, IoU, and mIoU index, respectively. The

accuracy gains with AEUNet++ over UNet++ were 0.37%, 1.78%,2.32%, 2.14%, 2.64% and 1.51%, respectively, yielding improvements of approximately 2.64%, 2.0%, 0.97% and 0.11% IoU compared with the other four methods.

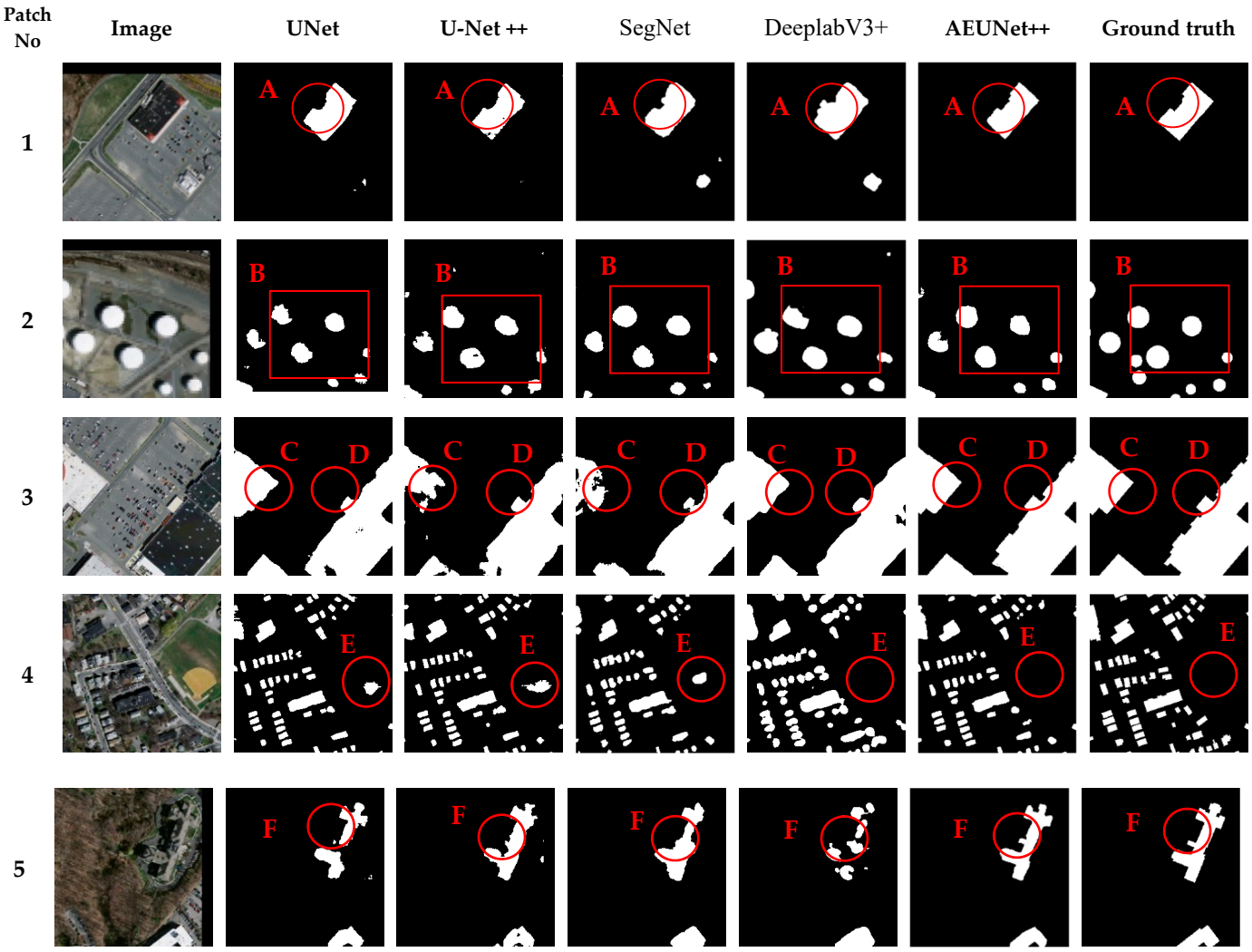

Figure 9. Segmented patches from U-Net, UNet++, SegNet, DeeplabV3+ and AEUNet++ based on the Massachusetts buildings dataset. The white color in the segmented patches represents pixels belonging to buildings (A–F denote the marked areas).

Table 1. Quantitative comparison of OA, Precision, Recall, F1, IoU and mIoU for the Massachusetts buildings data set.

Method	OA	Precision	Recall	F1	IoU	mIoU
U-Net	94.75	81.62	63.78	71.61	55.77	75.07
UNet++	94.82	81.68	64.58	72.13	56.41	75.43
SegNet	94.89	80.99	66.39	72.97	57.44	75.98
DeeplabV3+	93.97	81.11	**67.40**	73.62	58.30	75.64
AEUNet++	**95.12**	**83.40**	66.10	**73.75**	**58.41**	**76.58**

3.4.2. Results with the WHU Aerial Image Data Set

Figure 10 illustrates the segmentation results based on the WHU aerial image data set using U-Net, UNet ++, SegNet, DeeplabV3+ and AEUNet++, respectively. Visually, UNet++ and DeeplabV3+ produced more homogeneous segmentation maps with accurate boundaries than those of U-Net and SegNet, and AEUNet++ had the best performance,

whereas extensive salt-and-pepper noises existed in the U-Net, SegNet and DeeplabV3+ results. UNet++ enhanced the U-Net to some extent, but showed weaker performance than AEUNet++. This was also illustrated in areas A–H. Specially, many small areas of buildings were mislabeled as background by U-Net, UNet++, SegNet and DeeplabV3+ in areas A, B, C and F, but AEUNet++ showed better performance. The main reason may be that insufficient utilization of feature maps from different layers may have occurred due to the use only of plain skip connections by U-Net. Though UNet++ used the dense skip connections to capture more representative feature maps, negative feature maps were produced. Moreover, in order to increase the perceptions of U-Net and UNet++, many max-pooling operators were applied, which may have ignored the details of objects and location information. Compared with U-Net, UNet++, SegNet and DeeplabV3+, AEUNet++ introduced the attention block for adaptively refining feature maps extracted by UNet++ according to their contributions, and the multi-task learning was introduced to incorporate building boundary geometric information for guiding accurate boundaries.

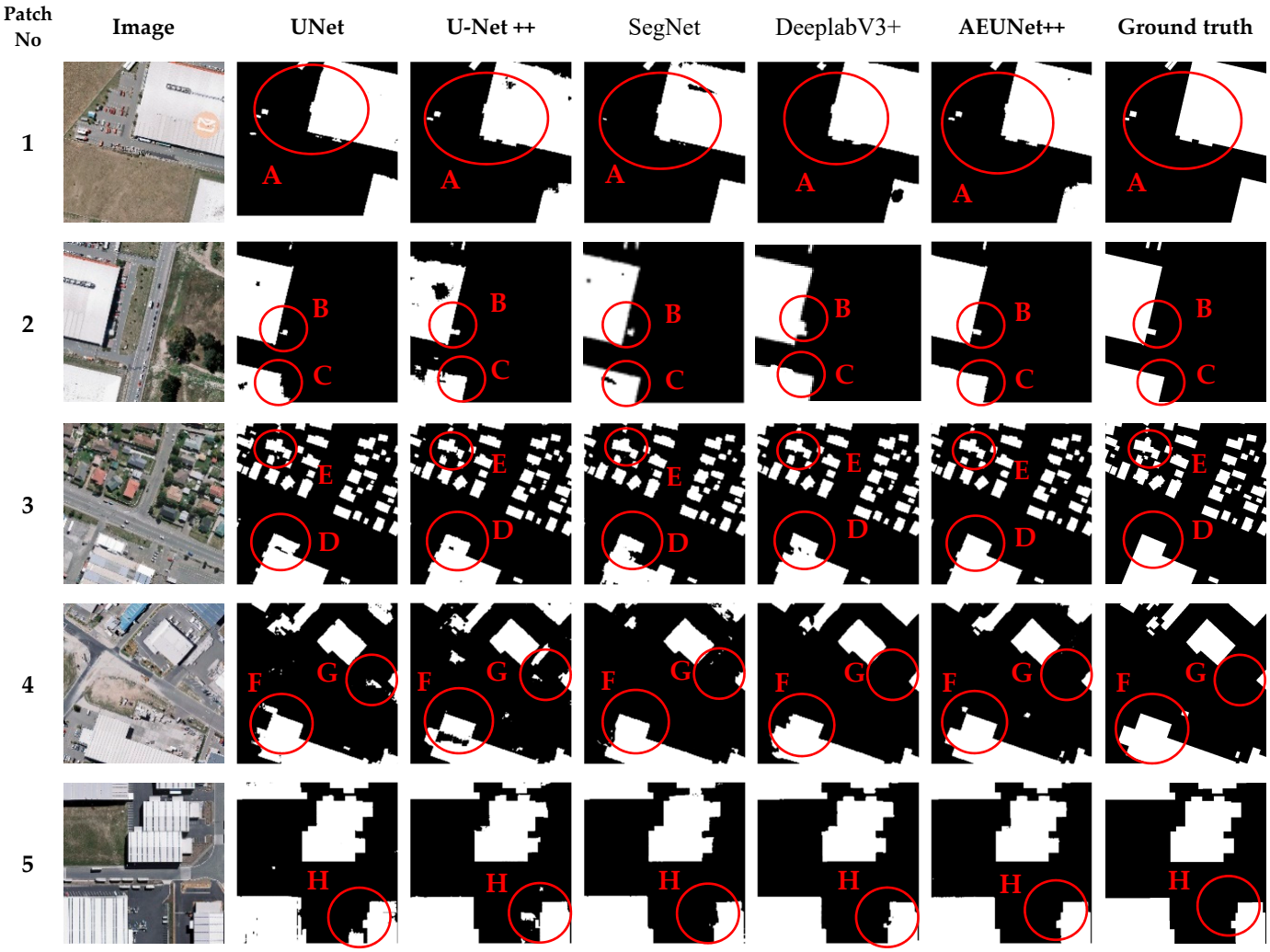

Figure 10. Segmented patches from U-Net, UNet++, SegNet, DeeplabV3+ and AEUNet++ based on the WHU aerial image data set. The white color in the segmented patches represents pixels belonging to buildings (A–H denote the marked areas).

As also illustrated in Table 2, AEUNet++ achieved the greatest accuracy of 98.97%, 95.23%, 95.43%, 95.33%, 91.08%, and 94.97%, respectively, for OA, Precision, Recall, F1, IoU, and mIoU index. In particular, compared with UNet++, Precision and IoU increased by 1.92% and 3.24%, respectively, and AEUNet++ obtained IoU improvements of approximately 6.27%, 3.24%, 2.6% and 0.15% over U-Net, UNet ++, SegNet and DeeplabV3+, respectively.

Table 2. Quantitative comparison of OA, Precision, Recall, F1, IoU and mIoU for the WHU aerial image data set.

Method	OA	Precision	Recall	F1	IoU	mIoU
UNet	98.18	91.21	92.35	91.78	84.81	91.39
UNet++	98.57	93.31	93.74	93.53	87.84	93.12
SegNet	98.67	94.86	92.93	93.88	88.48	93.50
DeeplabV3+	98.96	95.51	94.99	95.25	90.93	94.88
AEUNet++	**98.97**	**95.23**	**95.43**	**95.33**	**91.08**	**94.97**

From the above two experiments, we can draw conclusions from Figures 8 and 9 for the two data sets. Whether for the dense and small scale of building distributions that required smaller and narrower receptive fields, or for the sparse and large scale of building distributions that required larger and wider receptive fields, the AEUNet++ model exhibited very good robustness to building information that could be accurately extracted under different scales and different distributions. Comparing the prediction results from Figures 8 and 9, AEUNet++ produced better performance based on the WHU aerial image data set than that based on the Massachusetts buildings data set, because the resolution of the WHU aerial image was higher than that of the Massachusetts image. Too many mixed pixels existed in the Massachusetts buildings data set, which may have resulted in too many mislabeled pixels in the prediction results. That is to say, the proposed AEUNet++ is more suitable for extracting buildings from the HSRI.

3.4.3. Ablation Study

To assess the advantages of different modules contained in the proposed AEUNet++ model, in this section, we compared the performance of UNet++, AUNet++ (UNet++ with the attention block), and AEUNet++ (UNet++ with attention block and multi-task learning) based on the WHU aerial image data set. As shown in Table 3, the two modules could improve the prediction accuracy; compared with multi-task learning, the combination with the attention block boosted the segmentation performance dramatically. The reasons may be that while the attention module could adaptively refine the four prediction feature maps output by UNet++, feature maps were learned that should have been emphasized or suppressed, which was most important for the final result. Hence, combination with the attention block was able to improve the accuracy further. Although the introduction of multi-task learning using the distance classes map could also improve the accuracy of segmentation by refining the building boundaries, because the proportion of boundary pixels in the total image was relatively small, the accuracy was not improved so much, as seen from Table 3. However, the buildings' boundaries were highly enhanced, as can be drawn from the visualizations in Figures 8 and 9.

Table 3. Ablation comparison of OA, Precision, Recall, F1, IoU and mIoU for the WHU aerial image data set.

Method	OA	Precision	Recall	F1	IoU	mIoU
UNet++	98.57	93.31	93.74	93.53	87.84	93.12
AUNet++	98.92	95.04	95.19	95.12	90.69	94.75
AEUNet++	98.97	95.23	95.43	95.33	91.08	94.97

4. Conclusions

In this paper, based on UNet++, an attention mechanism and multi-task learning, a new AEUNet++ was designed and proposed to automatically extract buildings from HSRI. The proposed method can overcome the drawbacks of segmentation prediction results with poor boundaries in UNet++. It can produce homogeneous building segmentation results while weakening the boundary blurring simultaneously. This can be attributed to the introduced attention block for fusing and refining multiscale features from the different-

layer feature maps in light of their relative importance, and to multi-task learning for integrating the boundary geometric information of buildings into the proposed AEUNet++ network. To test the performance of AEUNet++, experiments on two data sets representing the distribution of buildings at different scales were conducted. Compared with the existing U-Net and UNet++ models, AEUNet++ was more accurate in visual and quantitative evaluations. Therefore, AEUNet++ is an effective method for extraction of buildings from high spatial resolution imagery. In future research, we will explore better ways of expressing multiscale features and boundary constraints and expanding the data sets.

Author Contributions: Conceptualization, H.Z. (Hua Zhao), H.Z. (Hua Zhang) and X.Z.; methodology, H.Z. (Hua Zhao) and H.Z. (Hua Zhang); software, X.Z.; validation, X.Z.; formal analysis, H.Z. (Hua Zhao), H.Z. (Hua Zhang), X.Z.; investigation, H.Z. (Hua Zhao), H.Z. (Hua Zhang), X.Z.; resources, H.Z. (Hua Zhao) and H.Z. (Hua Zhang); data curation; writing—original draft preparation, H.Z. (Hua Zhang) and X.Z.; writing—review and editing, H.Z. (Hua Zhao) and H.Z. (Hua Zhang); visualization, X.Z.; supervision, H.Z. (Hua Zhao) and H.Z. (Hua Zhang); project administration, H.Z. (Hua Zhang); funding acquisition, H.Z. (Hua Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation, China (No. 41971400), and in part by the Fundamental Research Funds for the Central Universities under Grant 2019ZDPY09.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

Conflicts of Interest: No potential conflict of interest was reported by the author.

References

1. Huang, X.; Wen, D.; Li, J.; Qin, R. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote Sens. Environ.* **2017**, *196*, 56–75. [CrossRef]
2. Lin, A.; Sun, X.; Wu, H.; Luo, W.; Wang, D.; Zhong, D.; Wang, Z.; Zhao, L.; Zhu, J. Identifying urban building function by integrating remote sensing imagery and POI data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8864–8875. [CrossRef]
3. Li, L.; Liang, J.; Weng, M.; Zhu, H. A Multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]
4. Zhang, Z.; Guo, W.; Li, M.; Yu, W. GIS-supervised building extraction with label noise-adaptive fully convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 2135–2139. [CrossRef]
5. Li, Z.; Shi, W.; Wang, Q.; Miao, Z. Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 883–899. [CrossRef]
6. Wang, J.; Yang, X.; Qin, X.; Ye, X.; Qin, Q. An efficient approach for automatic rectangular building extraction from very high-resolution optical satellite imagery. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 487–491. [CrossRef]
7. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.* **2012**, *5*, 161–172. [CrossRef]
8. Inglada, J. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS J. Photogramm. Remote Sens.* **2007**, *62*, 236–248. [CrossRef]
9. Du, S.; Zhang, F.; Zhang, X. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 107–119. [CrossRef]
10. Chen, Q.; Wang, L.; Wu, Y.; Wu, G.; Guo, Z.; Waslander, S. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 42–55. [CrossRef]
11. Yuan, J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2793–2798. [CrossRef] [PubMed]
12. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z.L. Building extraction in very high-resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
13. Shi, Y.L.; Li, Q.Y.; 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]
14. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Garcia-Rodriguez, J. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.

15. Zhu, X.; Tuia, D.; Mou, L.; Xia, G.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Mag.* **2017**, *5*, 8–36. [CrossRef]
16. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst. (NIPS)* **2012**, *25*, 1097–1115. [CrossRef]
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Shrestha, S.; Vanneschi, L. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
19. Wei, S.; Ji, S.; Lu, M. Toward automatic building footprint delineation from aerial images using CNN and regularization. *IEEE Trans. Geosci. Remote.* **2020**, *58*, 2178–2189. [CrossRef]
20. Xia, L.; Zhang, J.; Zhang, X.; Yang, H.; Xu, M. Precise extraction of buildings from high-resolution remote sensing images based on semantic edges and segmentation. *Remote Sens.* **2021**, *13*, 3083. [CrossRef]
21. Sun, Y.; Zhang, X.; Zhao, X.; Xin, Q. Extracting building boundaries from high resolution optical images and LiDAR data by integrating the convolutional neural network and the active contour model. *Remote Sens.* **2018**, *10*, 1459. [CrossRef]
22. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2015**, arXiv:1511.07122.
23. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2016**, arXiv:1612.01105.
24. 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. *arXiv* **2016**, arXiv:1606.00915. [CrossRef]
25. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015*; Springer: Cham, Switzerland, 2015; pp. 234–241.
26. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net Architecture for Medical Image Segmentation. *arXiv* **2018**, arXiv:1807.10165.
27. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
28. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
29. Van Noord, N.; Postma, E. Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognit.* **2017**, *61*, 583–592. [CrossRef]
30. Ji, S.P.; Wei, S.Q.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
31. Zhang, X.; Xiao, Z.; Li, D.; Fan, M.; Zhao, L. Semantic segmentation of remote sensing images using multiscale decoding network. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1492–1496. [CrossRef]
32. Ma, J.; Wu, L.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Building extraction of aerial images by a global and multiscale encoder-decoder network. *Remote Sens.* **2020**, *12*, 2350. [CrossRef]
33. Rastogi, K.; Bodani, P.; Sharma, S. Automatic building footprint extraction from very high-resolution imagery using deep learning techniques. *Geocarto Int.* **2022**, *37*, 1501–1513. [CrossRef]
34. Liu, Y.; Gross, L.; Li, Z.; Li, X.; Fan, X.; Qi, W. Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling. *IEEE Access* **2019**, *7*, 128774–128786. [CrossRef]
35. Li, X.; Yao, X.; Fang, Y. Building-A-Nets: Robust Building Extraction from High-Resolution Remote Sensing Images with Adversarial Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3680–3687. [CrossRef]
36. 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]
37. Diakogiannis, F.; Waldner, F.; Caccetta, P.; Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
38. Guo, M.; Liu, H.; Xu, Y.; Huang, Y. Building extraction based on U-Net with an attention block and multiple losses. *Remote Sens.* **2020**, *12*, 1400. [CrossRef]
39. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv* **2017**, arXiv:1705.07115.
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. *arXiv* **2018**, arXiv:1807.06521.
41. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.