remote sensing

MDPI

Article

Building Extraction Based on U-Net with an Attention Block and Multiple Losses

Mingqiang Guo [1], Heng Liu [1], Yongyang Xu [1,*] and Ying Huang [2]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; guomingqiang@mapgis.com (M.G.); cainandangdao@cug.edu.cn (H.L.)
[2] Wuhan Zondy Cyber Technology Co. Ltd., Wuhan 430074, China; huangying@mapgis.com
* Correspondence: yongyangxu@cug.edu.cn

Received: 24 February 2020; Accepted: 27 April 2020; Published: 28 April 2020

Abstract: Semantic segmentation of high-resolution remote sensing images plays an important role in applications for building extraction. However, the current algorithms have some semantic information extraction limitations, and these can lead to poor segmentation results. To extract buildings with high accuracy, we propose a multiloss neural network based on attention. The designed network, based on U-Net, can improve the sensitivity of the model by the attention block and suppress the background influence of irrelevant feature areas. To improve the ability of the model, a multiloss approach is proposed during training the network. The experimental results show that the proposed model offers great improvement over other state-of-the-art methods. For the public Inria Aerial Image Labeling dataset, the F1 score reached 76.96% and showed good performance on the Aerial Imagery for Roof Segmentation dataset.

Keywords: building extraction; attention block; multiple losses; semantic segmentation; remote sensing images

1. Introduction

With the development of remote sensing technology, high-spatial-resolution images provide more and more semantic and detailed information. Automating semantic segmentation is crucial to the development of the industry and has also gained the attention of more and more researchers [1]. Extracting information from remote sensing images efficiently and accurately remains a difficult task. In the past few years, deep learning has enabled remarkable breakthroughs in the field of data and image processing [2]. Convolutional neural networks have been shown to have great advantages in image classification, object detection, and semantic segmentation. Although some image segmentation algorithms have achieved good results, there are still many problems to be solved for remote sensing information extraction [3–5]. Unlike the segmentation of natural scene images, remote sensing image segmentation often requires more professional background knowledge [6].

With the continuous improvement in the spatial resolution of optical remote sensing images, building detection from high-resolution images has attracted increasing attention. There are various methods for building extraction from remote sensing images including feature extraction based on prior knowledge such as strong edge, shape design, roof color, shadow, etc. [7]. Some methods are based on building roof detection techniques such as template matching [8], mathematical morphology [9,10], and active contours [11,12], and graph theory [13,14], random forests [15], and support vector machines [16,17] have also been employed. These methods are based on prior knowledge, taking into account the complexity and diversity of building shape, roof surface, imaging conditions, and spatial environment, etc., and they are easily confined to specific building shape areas [18].

In recent years, because of the rapid development of deep learning in computer vision, compared with traditional building detection or semantic segmentation of remote sensing images, deep learning methods have been seen as advantageous in automatically extracting high-dimensional features. The general convolutional neural network (CNN) model cannot produce accurate building contours. Therefore, in certain circumstances, post-processing of segmented graphs is still necessary [19]. The use of full convolution networks (FCNs) greatly improves the training and prediction efficiency of the model. FCNs provide a learning framework for end-to-end image semantic segmentation by transforming the full connection layer of the CNN into the convolution layer [20]. At present, FCNs have been used in building detection from aviation and satellite images [21,22]. In addition to FCNs, there are some methods based on CNNs for semantic segmentation such as instance segmentation and segmentations based on recurrent neural networks (RNNs).

Instance segmentation is also a pixel-level classification that is specific to individual instances. There are complex structures and overparameterizations [23] in the models, and sometimes, instance segmentation exhibits better performance in terms of accuracy and stability. As the regression of the anchor in the instance segmentation model can better locate the target region, location accuracy is relatively high. However, the effect of boundary processing is not very satisfactory. Moreover, this method does not consider the connection between the pixels to be segmented. Meanwhile, as the network structure is mostly used for the regression of the anchor describing the target location, the area of interest that is not inside the anchor may not be segmented. Moreover, as buildings are relatively dense in urban areas, there are some buildings that are not single entities, which means that instance segmentation is not effective in detecting individual buildings.

In contrast to instance segmentation, with the proposed use of the attention mechanism, the RNN series has been gradually applied in the segmentation field. This mechanism forms a directed cycle by using backward connection to improve the segmentation result. In this way, features are treated as a sequence of time-series structures [24]. Bergado et al. merged recurrent methods into the task of semantic segmentation when ReuseNet learns the content dependency in the label and redefines the segmentation result. ReuseNet applied the semantic segmentation operations in Rcycle, in which each cycle takes the score map of the previous cycle concatenated with the original image as input. The method based on generative adversarial networks (GANs) optimizes the final result by forcing pixels to define the relationship between adjacent pixels. The authors in [25] first proposed the use of GANs to train segmented networks, where the segmentation network is trained with an adversarial network until the adversarial network cannot distinguish between the output of the segmentation network and the label image. However, for image segmentation, GAN-trained models are generally less robust than general learning methods.

In summary, neural networks have made great progress and development in processing the semantic segmentation of remote sensing images. To extract buildings from complex features accurately, there are many semantic problems such as the blocking of buildings by trees, the similarity of roads and some buildings in spectral characteristic curves, and the cover of shadows caused by topography, all of which will lead to the absence of segmentation results [26]. There are two main problems in the segmentation module at the present stage. The first is that the high-dimensional features are insensitive to the response of background information and the target region. Some methods cannot accurately distinguish background information from the target region, which results in the loss of spatial information in the extraction process of high-dimensional features. In the process of upsampling, due to insufficient spatial information, the results of segmentation become grids at the pixel level and salt-and-pepper noise occurs. As the model only classifies at the pixel level and ignores the relationship between pixels, spatial discontinuity of the segmentation results appears. The second aspect is boundary ambiguity. Conventional segmentation methods only consider the intersection ratio (IoU) between the target image and the output result. Although the prediction accuracy of the model is improved, the boundary of the output result will become fuzzy and irregular.

For semantic segmentation, top-down architectures such as UNet [27], SharpMask [28], SegNet [29], and the Feature Pyramid Network (FPN) [30] have been proposed to further improve the efficiency and performance. Generally, the development of these models is driven by the reduction of spatial information input to the image after passing through several convolutional sum aggregation layers, which will lead to blurring of the boundary or poor edge precision of segmentation results. The network structure does not perform well in the semantic segmentation of buildings, mainly because the high-dimensional features are pooled or stride-convolved, and the spatial information of local and regions of interest is lost. Therefore, some scholars compensate for local information [31–33] by designing a highway. However, because of the unexplainability of features, the selection of the number of feature channels is a major problem. In addition, the DeepLab series network [34–36] model was designed by Google, where the feature extraction of images by atrous convolution achieves a larger receptive field and a good generalization effect for multiscale problems in the region of interest.

In this study, a multiloss-based UNet model with an attention block (AMUNet) is proposed. The attention block mainly aims at enhancing the sensitivity of extracting high-dimensional feature information from the model and controls each pixel of the feature map by using a gate to suppress the influence of background information. By considering the lower dimension of information and compensating for less information of the high-dimensional features, we designed the attention block. Therefore, it is necessary to use a large-scale convolution feature map to obtain local information in a larger range of receptive fields, so that it can compensate for the lost spatial information by the high-dimensional features. Concurrently, we also propose a multiloss method to solve the problem that the boundary of the segmentation region is not clear. In this work, we used a loss function based on the pixel level to constrain the amount of segmentation information first, and then an IoU loss function was used to constrain the output result at the pixel level.

2. Methodology

In this work, we explored the structure of the network to overcome the insensitivity problems to the region of interest, adopted a gate control mechanism based on the attention mechanism, and designed an end-to-end feature map processing method (Figure 1). The designed model mainly involved three modules: downsampling, attention, and upsampling. Inspired by multitasking, we designed a structure with multiple losses to optimize the binary results, which made full use of the features as much as possible to ensure segmentation accuracy. However, as there are many types of ground objects in remote sensing images, in the case of three bands, many targets will exhibit certain similarities. This will cause the boundary of the segmentation target area to be blurred and there will be a lot of noise. Based on the multiloss, we proposed an attention mechanism (see Section 2 for details) to solve these problems. Although attention mechanisms have been applied in the field of remote sensing, in this article, inspired by SE-Net, we developed the attention mechanism aimed at the buildings' features in remote sensing imagery. The main improvement was modifying the position of Resampler to enhance the sensitivity of the gate to the feature map. This improvement not only solves the problem that the segmentation model is not sensitive to small buildings, but it also adopts the structure of GAP, which can effectively suppress the background information and optimize the segmentation details.

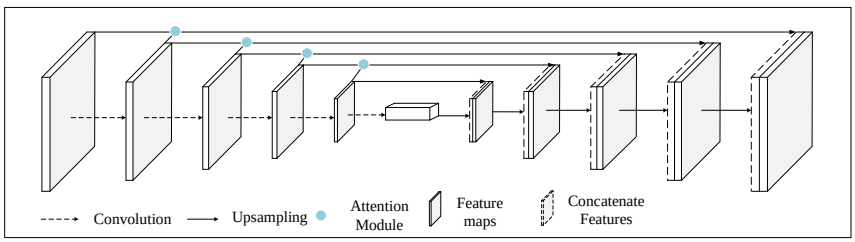


Figure 1. Module structure design based on attention. In the coding encoder part of the network, each scale of the input image is filtered and sampled by two factors, step by step.

2.1. Gate Control Based on the Attention Mechanism

Attention can be viewed, broadly, as a tool [37–41] to bias the allocation of the available processing resources toward locating the target during multitasking and perform well. Research shows that the attention mechanism has obvious advantages in image feature extraction and edge detection [42,43]. It is a light channel selection mechanism that simulates the relationship between channels by calculating the efficiency [44,45], thereby enhancing the feature extraction ability of basic modules in the whole network. While the building extraction processing occurs in the decoder stage, the model's ability to interpret high-dimensional features will be weakened by the need for upsampling. Different scholars have different views on this issue, of which there are two main perspectives. The first is the pyramid scene parsing network (PSPNet), which is composed of feature pyramids by sampling high-order features at different scales, and after spatial information fusion, the upper sampling stage is carried out. For the upsampling stage, there is enough high-dimensional information, so no other information compensation is needed, but multiple loss is needed to constrain the segmentation results. The second is the skip connected approach, which is represented by UNet. In this approach, it is necessary to compensate for the branch information of the feature in the process of upsampling. Based on this, we proposed the attention module. The module is mainly composed of two parts: a gate control process of high-dimensional features and information processing and statistics based on the gate mechanism.

2.1.1. Merging of Features at Different Scales

As the semantic information of features in different stages is a top–down context relationship in space, so far, for semantic information fusion of different scales in space, we can use dense upsampling [46] to fuse high-dimensional information or directly use upsampling to fuse additive features. The former will lead to an exponential increase in the number of parameters, and the model becomes very large. The latter does not make good use of the differences of different dimensional features [47,48], resulting in a hackly boundary. The method proposed maps the features of two stages to the same dimension through the convolution of unsynchronized amplitude for semantic information fusion. For features in different dimensions x_{c1}^{l} and x_{c2}^{l+1}, the final output is as follows:

$$x_i = \sigma_r\left(x_{c1}^{l} * k_{c1,i} + x_{c2}^{l+1} * k_{c2,i}\right) \tag{1}$$

where $*$ represents convolution; $c1$, $c2$, and i represent the dimensions of feature space, if the feature sizes are different, convolution needs to unify the feature sizes in steps greater than one; and σ_r represents a rectified linear unit (ReLU). As shown in Figure 2, the above formula can fuse different scale map features (x_{c1}^{l} and x_{c2}^{l+1}) to a unified dimension (i, i = min ($c1$, $c2$)).

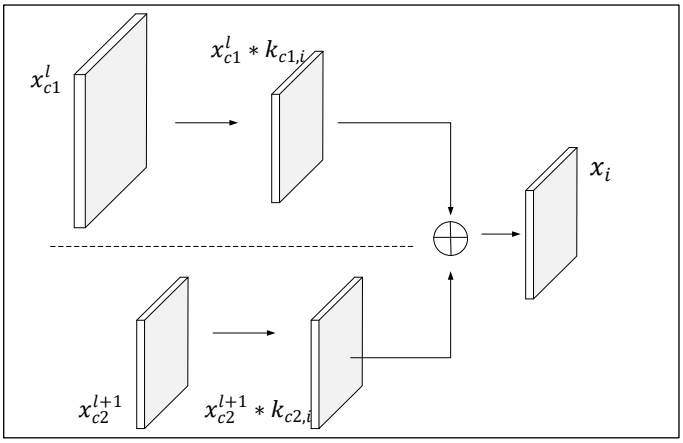


Figure 2. Merging of features at different scales. Mapping features x_{c1}^{l} and x_{c2}^{l+1} to x_i.

2.1.2. Gate Control

Gate control can extract high-dimensional information and suppresses background information by using a gate mechanism. The convolution layer gradually represents the representation of the high-dimensional image (x^l) by local information layer by layer. Finally, pixels are separated according to the differences of target feature semantics in high-dimensional space. Through the continuous change process, the model prediction is based on the information collected in a small-to-large receiving field. Therefore, the feature map x^l is obtained by a linear transformation and a nonlinear activation function at the output of layer l, and the activation function is as follows: $\sigma\left(x_{i,c}^l\right) = \max\left(x_{i,c}^l, 0\right)$, where c represents dimensions. Therefore, the output of the feature can be expressed as the convolution operation represented by $*$ in the formula $x_c^l = \sigma_1(\sum x_{c-1}^{l-1} * k_{c^{i-1},c})$, and k is the convolution kernel size. To capture a large enough receiving domain and obtain context-dependent information, features are gradually downsampled in the standard CNN architecture, taking into account both global information and local information. However, it is still difficult to accurately predict small objects with large differences. To improve the accuracy, most semantic segmentation frameworks [47,48] rely on the following segmentation steps. We prove that the same goal can be achieved by integrating the attention gate (AG) into the standard CNN model, which does not need to train multiple models nor require many additional model parameters. Compared with the multistage CNN localization model, AGs gradually suppress the feature response of the irrelevant background area and do not need to cut regions of interest in the network, and the attention coefficient α_i ensures that the prominent area and the suppressed background information are realized through the activation function. Figure 3 shows a schematic diagram of the attention module. We realized the compensation of low-dimensional information to high-dimensional information through concatenating attention mechanism features and upsampling features. By default, each pixel of the feature map F_l needs to be scalar scaled, so the attention of the multihead needs to be used for processing in the case of multiple categories. Each gate will focus on the corresponding feature map, so the modification of the attention mechanism is

$$q_{att}\left(x_i^l, x_g^{l-1}\right) = S_q\left(H_x^T\left(W_x^T x_i^l + W_g^T x_g^{l-1} + b_g\right) + b_H\right) = S_q\left(H_x^T\left(F_i + b_g\right) + b_H\right) \quad (2)$$

$$\alpha_i^l = \sigma_2\left(q_{att}\left(x_i^l, x_g^{l-1}\right)\right) \quad (3)$$

where S_q represents global average pooling; σ_2 is the sigmoid activation function; and x_g^{l-1} and x^l represent upsampling the feature maps and they are mapped into the same dimension space with convolution coefficient matrix W_x^T, W_g^T, and H_x^T.

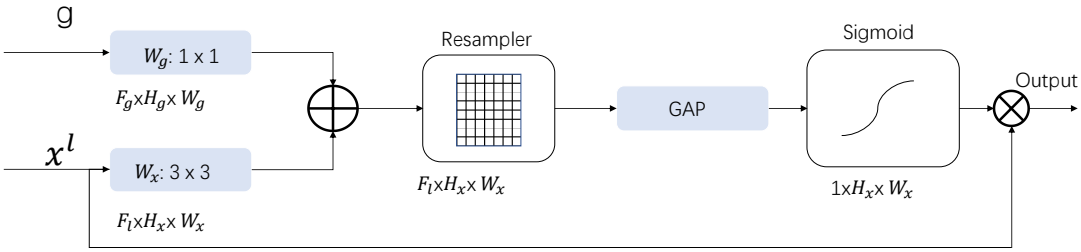


Figure 3. Attention module. The input feature (x) and activation feature maps are calculated. The selection of the spatial region is based on the activation coefficient and context information provided by the sigmoid to collect multiscale information. In the resampling stage, deconvolution is adopted to conduct coefficient resampling of the feature maps.

The purpose is to map pixels that are not in the stage to the same dimension for information processing and exchange. To improve the accuracy of feature response to small target information, a resampler is placed in front of the activation function to provide more accurate and sensitive door control information for the feature map. The output of the AG is processed based on pixel-level change

as follows: $\hat{x}_c^l = \alpha_i \cdot x_c^l$, where $\hat{x}_c^l$ represents the attention result, α_i represents the gate coefficient, and x_c^l is feature maps.

2.2. Multiloss Approach Based on Pixels

To further improve the accuracy of the network, we proposed a multiloss approach to optimize the results of semantic segmentation. Bischke proposed the multitask loss method to enhance the final semantic segmentation results and used different data functions to constrain the data results in different aspects to achieve a good result of constraining the model output. However, before the feature is activated, if the dimension of the feature is large, the network model will not converge, and the accuracy will fluctuate. At the same time, many unnecessary parameters will be introduced, resulting in a higher accuracy of the output of the loss 1 stage than that of the loss 2 stage. To solve this problem, we proposed a multiloss method to optimize the results. Figure 4 shows a structural diagram of the multiloss method. For the last two convolutions, the semantic segmentation results are optimized through different losses, which means that the output of the model is constrained by two loss functions, and the final convolution layer plays the role of fine-tuning the results. In terms of loss function composition, we used the sum of the basic loss and the offset loss to form the final loss. The first is the basic loss. We used pixel-based loss, which mainly measures each pixel value in the upsampling and is divided into two parts. The first part is the total amount of information carried by the pixel and is represented by the binary cross entropy; the second part is based on the pixel value gap measurement, which compares the difference between the segmentation result and the artificially produced label in pairs [49–52]. Pixel-based loss is a function that changes the edge probability index (the probabilistic Rand index, *PRI*) to segmentation statistics. Let $I\left(l_i^S = l_j^S\right)$ be a binary function defined between the label and the segmented pixel pair (x_i, x_j). Then, PRI can be defined as

$$PRI(S, S_k) = \frac{1}{\binom{N}{2}} \sum_{i,j,i \neq j} \left[I\left(l_i^S = l_j^{S_k}\right) p_{ij} + I\left(l_i^S \neq l_j^{S_k}\right) p_{ij} \right] \tag{4}$$

where N represents the number of pixels; S_k represents the set of segmentation labels and S represents the set of ground truth; and p_{ij} represents the probability. According to experience, p_{ij} is calculated from the average of all the divided pixel pairs and the *PRI* is between 0 and 1, where 0 indicates that the segmentation result is opposite to the segmentation label, and 1 indicates that the segmentation result is completely correct and each pixel is accurately classified. This measure appropriately improves the measurement of pixel classification results because it is from the perspective of human observation. A standardized random probability index (NPR) index extends the *PRI* metric, allowing for a comparison of segmentation between different images. Specifically, it uses the expected value of the input image to normalize the NPR, so the average NPR is 0 and the range is greater than the *PRI*. We define the loss as the pixel-based loss (PBL):

$$loss_{pbl} = 1 - PRI \tag{5}$$

In the selection of biased loss for loss 1, the main purpose of $loss_{dice}$ and $loss_{ce}$ is to extract the information step by step and make a preliminary evaluation and impose a constraint on the segmented edge information and total accuracy information. $Loss_2$ mainly uses $loss_{jacaard}$ to refine the output of the previous phase. The formulas are as follows:

$$loss_1 = \alpha_1 loss_{ce} + \beta_1 loss_{dice} \tag{6}$$

$$loss_2 = \alpha_2 loss_{pbl} + \beta_2 losss_{jaccard} \tag{7}$$

where α_1, α_2, β_1, and β_2 are bias parameters; and $\alpha_i + \beta_i = 1$, which is used to determine the sensitivity of the control segmentation boundary. $loss_{ce}$ represents the cross-entropy loss function, $loss_{dice}$ represents the dice coefficient loss function, and $losss_{jaccard}$ represents the Jaccard coefficient loss function. If the

boundary information is more complex, then the value of α needs to be increased; in contrast, if the boundary information is more regular, then the value of β needs to be increased.

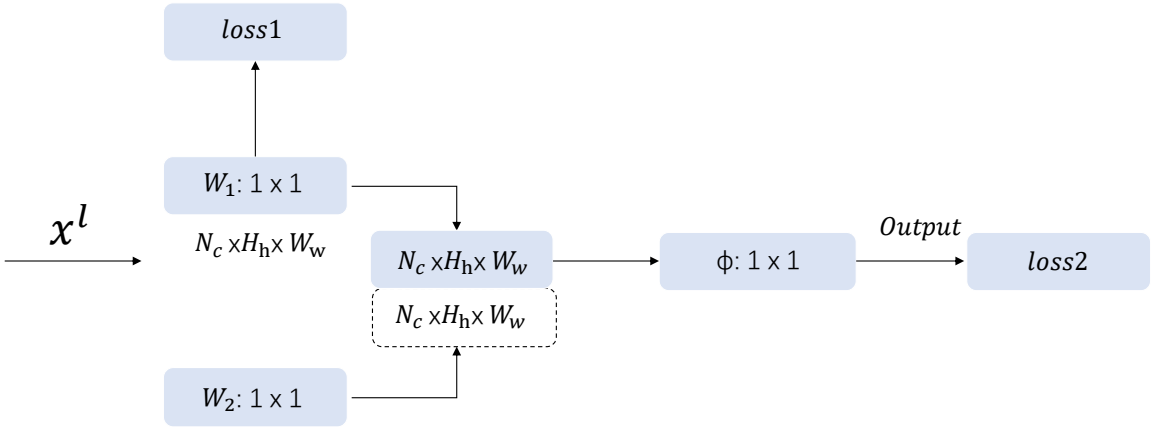


Figure 4. Junction diagram of the multiloss module. Each 1 × 1 convolution in the figure is followed by an activation function.

In terms of the general structure of the network, the attention block was adopted based on Unet. Given the parameter problem, adding random deactivation was chosen to increase the richness of the network parameters in the network. At the same time, the deactivation rate is gradually reduced with the process of the decoder; that is, the deactivation rate is inversely proportional to the level of the decoder. In the initial stage of the decoder, a higher deactivation rate is selected; as the decoder stage progresses, a lower deactivation rate is gradually selected. When the feature is in the high-dimensional stage, the information related to the building needs to be compensated for, so the network needs to randomly filter out some information unrelated to the building. With the progress of the decoder, the number of channels of the feature map is gradually reduced, and the location information carried by the feature in the process of upsampling needs to be compensated for, so the deactivation rate needs to be reduced to yield the real location information.

3. Materials and Results

3.1. Datasets

In this section, two databases were used for the experiments: the Inria aerial image labeling dataset for buildings [21] was used to demonstrate the effect of the model and the Aerial Imagery for Roof Segmentation [53] dataset was used to verify the effect of the model.

3.1.1. Inria Aerial Image Labeling Dataset

The Inria aerial image labeling dataset for buildings is a dataset that automatically labels pixels including a coverage area of ~810 km^2, of which 405 km^2 is used for training and the other 405 km^2 of data is used for testing. The data mainly include 360 high-resolution remote sensing images including several different countries, particularly, Austin, Chicago, Kitsap County, West Tyrol, and Vienna. The terrain and landforms in these areas have obvious differences. The spatial resolution of the images is 0.3 m and the size of each image is 5000 × 5000 pixels. Only 180 images are provided with labels in this dataset, and the other 180 images are only provided for testing the generalization ability of an online test. In this paper, we not only used the first five images of five cities as routine tests, but also used 180 images without labels as the basis of the model generalization test. In fact, on the official website of the Inria Aerial Image Labeling dataset, there is a clear introduction that the first five images of each city are used as test samples, and the remaining images are used as training samples. The data for the five regions are each highly representative. In Austin, there are many trees around buildings, the edges of the building are covered, and the distribution is not very dense. In Chicago, the building scale

is large, but some buildings are small and the distribution is dense. Kitsap County has more green space and the buildings are unevenly distributed. In Vienna, the buildings are very numerous with complex shapes, and the distribution is dense. These remote sensing images cover urban land use in different areas, from densely populated areas (such as San Francisco) to towns in high mountains (such as Lienz in the Austrian Tyrol).

3.1.2. Aerial Imagery for Roof Segmentation Dataset

The Aerial Imagery for Roof Segmentationdataset is a very high-resolution public dataset for building detection. The dataset comprises a data image of the entire city of Christchurch in the southern part of New Zealand. The ortho image covering 457 km^2 includes ~220,000 buildings with a resolution of 0.075 m and is an orthographic image.

Comparing the Aerial Imagery for Roof Segmentationdataset with the Inria aerial image labeling dataset for buildings, although the spatial resolution of the image has been improved, it can be seen that the regional span is not very large. The ability to test the model alone is not strong, so during the experiment, we mainly used the Inria aerial image labeling dataset for buildings to perform the generalization test for large-scale images and the details of subsequent ablation-based experiments. We used the Aerial Imagery for Roof Segmentationdataset to supplement the description of the processing power of the model for large-scale buildings and the processing of boundary details when segmenting ortho-resolution images.

3.2. Results

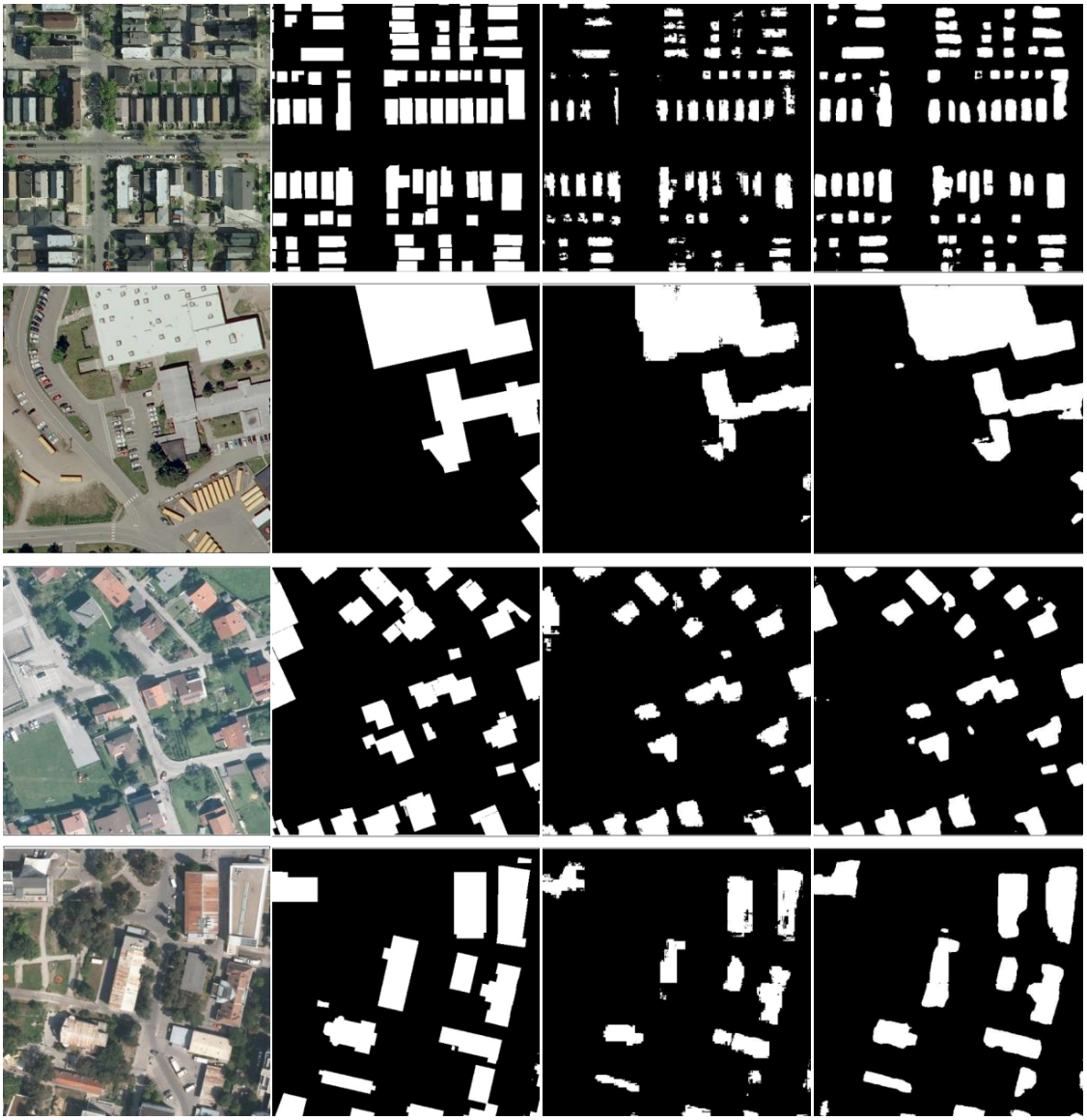
In this paper, the experiments were divided into three stages. The first stage was to make a comparison to the UNet model itself. The second stage entailed improving the accuracy and robustness of the model after adding the improved attention module. The third stage was to verify the proposed model and determine how much of the final accuracy could be achieved in the test set.

A few simple model simplification tests were performed to test how much effect the AMUNet module had on segmentation. The 180 samples with real data from five cities were divided into two parts according to a 31:5 ratio. The first part used the training set (where the training set still needs to be allocated according to a ratio of 4:1) as the model training work. The second part was the test set to take the first five remote sensing images of each city as test images.

Figure 5 shows a schematic diagram of the results made on the evaluation set. Training was performed using 30 images; five images were evaluated to obtain a model for evaluation, and the final output was the result graph. The result maps were from four regions (Chicago, Kitsap, Tyrol, and Vienna), and their distributions and geographic spatial layouts were significantly different. The picture size was 512 × 512. The AMUNet model has shown good generalization effects on different distributions. It can be clearly seen from the results that, after adding the attention module and a multiloss module, its prediction accuracy was significantly improved, the boundary and shape of the region of interest had been qualitatively improved, and the model also well suppressed salt-and-pepper noise and background information. In terms of the overall effect, the AMUNet model also exhibited a very good generalization ability. It can be seen from the figure that, whether for the dense and concentrated environment of Chicago, or in many large areas that require larger and wider receptive fields such as Kitsap, or in suburbs where the buildings are more regularly distributed, the AMUNet model exhibited very good robustness to the building information that could be accurately extracted under different scales and different distributions.

To analyze the ability of the proposed method, a series experiments was conducted done. In the experiments, we evaluated and measured whether the designed model has an enhanced effect on accuracy. The main focus was on how much accuracy could be improved from the basic model to the AMUNet model. It can be seen from Table 1 that once attention was added, accuracy was increased by 1.98% and that of IoU was increased by 17.31%. This was due to the increase in the diversity of location information. The field strengthened the model's generalization ability, increased the recognition

accuracy, and enhanced model performance. The accuracy improvement from UNet to attention UNet was very large, especially the improvement of IoU. It can be seen from the results that the ability of attention to suppress noninterest areas was very strong and enhanced the robustness of the model. After adding the multiloss module, the result performed better. With the addition of the multiloss module, the prediction accuracy of the model further increased by 1.68%, and the segmentation result reached a relatively high accuracy compared with the current model. The overall accuracy of IoU increased by 9.78%, which indicates that the multiloss module improved the results.



(**a**) Input (**b**) Label (**c**) Unet (**d**) AMUNet

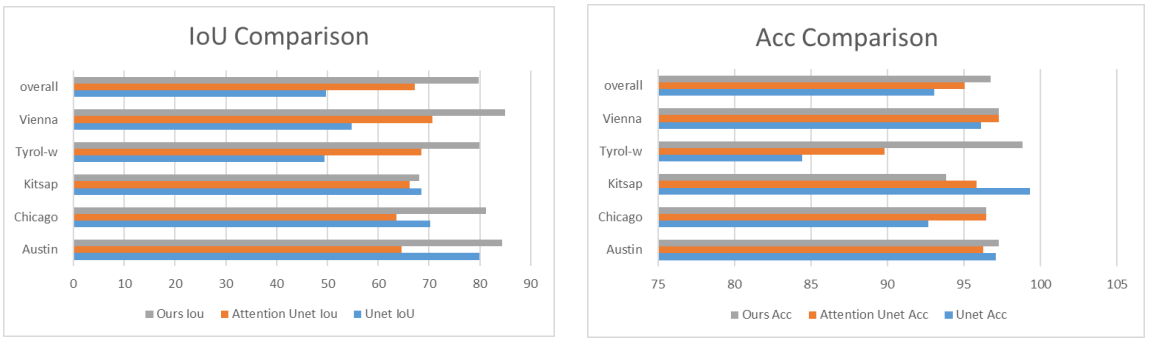
Figure 5. (**a**) Original image selected to process the data into four regions (Chicago, Kitsap, Tyrol, and Vienna) and (**b**) ground truth. (**c**) UNet processing results. (**d**) AMUNet processing results.

From the model accuracy and IoU of UNet, attention UNet, and our method in Figure 6, it can be seen that the accuracy of the three models was not very different: the overall accuracy was not only significant, but the IoU improvement was very significant. While maintaining the accuracy of the model, the model improved signal suppression in the noninterest area and the significantly improved semantic information compensation of the contour. Figure 7 shows a schematic diagram of the test segmentation results of our model. Here, Figure 7a–f are the prediction results on the dataset: blue

indicates true positive pixels, white indicates true negative pixels, magenta indicates, false positive pixels, and cyan indicates negative pixels. It can be seen that the blue area was large and very regular, and had a good segmentation performance on various terrains, while missed and misdetected cases rarely occurred. At the same time, we also validated the generalization ability on the Aerial Imagery for Roof Segmentationdataset and performed fine-tuning based on the original model to verify the generalization effect of the model on ultra-high-resolution remote sensing images. As shown in Figure 8, it can be clearly seen that, when the scale of the image became larger, clearly large blurred areas and unclear borders appeared in the segmentation results, but our model still performed well.

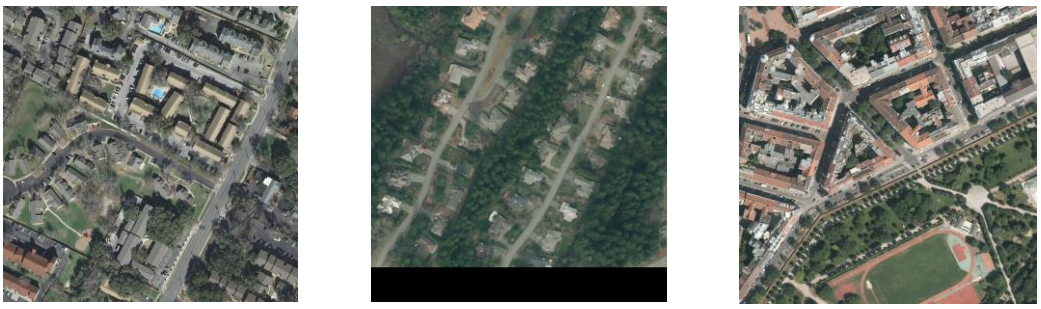
Table 1. Model generalization capability tests on Inria aerial image labeling (Inria Aerial Image Labeling) test datasets.

	Austin	Chicago	Kitsap	Tyrol-w	Vienna	Overall
UNet IoU	79.95	70.18	68.56	49.38	54.82	49.69
UNet accuracy	97.10	92.67	99.31	84.44	96.11	93.07
Attention UNet IoU	64.58	63.54	66.17	68.44	70.69	67.18
Attention UNet accuracy	96.29	96.45	95.83	89.83	97.28	95.05
AMUNet IoU	84.43	81.22	68.13	79.97	85.05	79.76
AMUNet accuracy	97.29	96.45	93.83	98.83	97.28	96.73



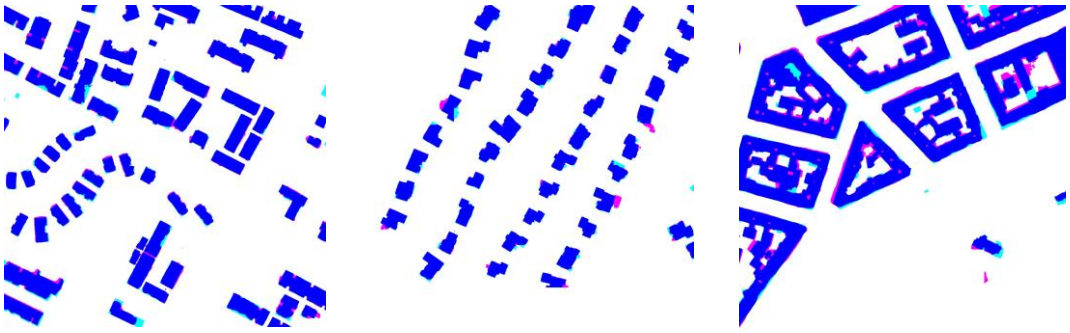
(**a**) IoU (**b**) Acc

Figure 6. Histogram of Intersection over Union (IoU) (**a**) and accuracy (**b**) of UNet, attention UNet, and AMUNet models in the simplified test.



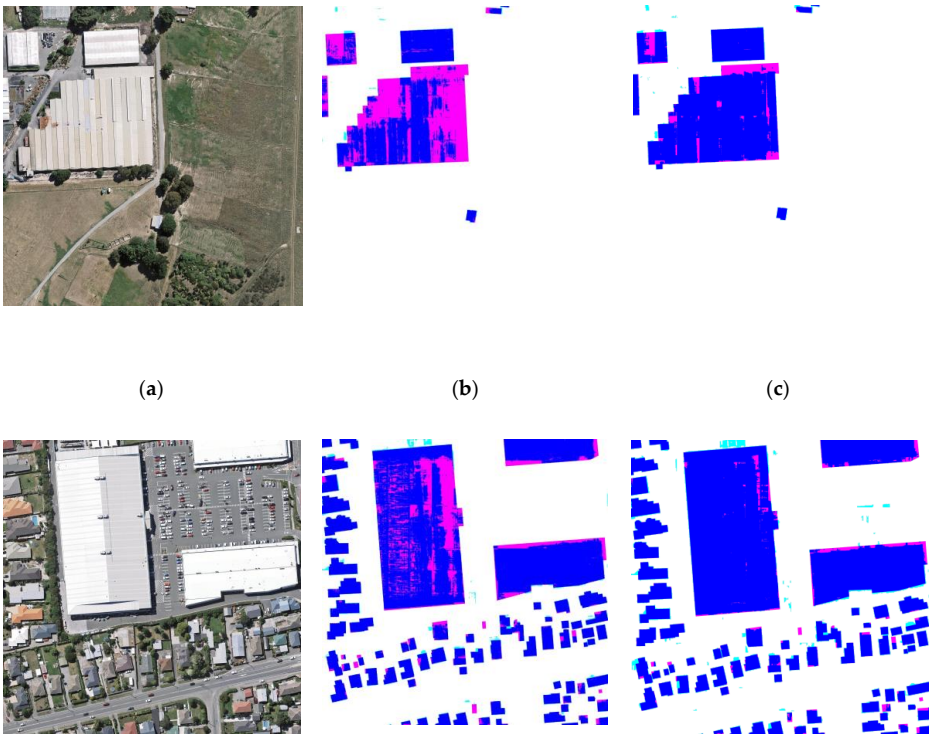
(**a**) (**b**) (**c**)

Figure 7. *Cont.*



(**d**) (**e**) (**f**)

Figure 7. Predicted results of building extraction on an Inria aerial image labeling dataset. (**a**–**c**) are original image and (**d**–**f**) are the results of AMUNet. Blue indicates true positive pixels, white indicates true negative pixels, magenta indicates false positive pixels, and cyan indicates false negative pixels.



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

(**d**) (**e**) (**f**)

Figure 8. Predicted results on Aerial Imagery for Roof Segmentation data of extracted buildings. (**a**,**d**) represent original image, (**b**,**e**) represent the results of Unet and (**c**,**f**) represent the results of AMUnet Blue indicates true positive pixels, white indicates true negative pixels, magenta indicates false positive pixels, and cyan indicates false negative pixels.

4. Discussion

4.1. Comparison to State-of-the-Art Methods

In the current model comparison, we selected the following mainstream remote sensing image segmentation models for horizontal comparison: SegNet multitask loss, 2-level UNet, MSMT, GAN-SCA, and other models, as these networks are very representative. The SM network is

representative of multi-task image segmentation; 2-level Unet and MSMT are representatives of deepened network structure models for image segmentation; and GAN-SCA represents the use of adversarial training segmentation network. Following a common practice, 180 tiles images were divided into a training set and test set. We followed data partitioning in deep learning methods where 80% were used to train the model and 20% were used to valid the model. Therefore, the remote sensing images were divided into a training set (155 tiles) and a test set (25 tiles). The dataset was enhanced to a certain extent, and 125 pieces of 5000 × 5000 clips were expanded to 200,000 pieces of 256 × 256 datasets and 30 pieces of evaluation set data were enhanced to 8000 pieces of 256 × 256 pictures. The remaining 25 pieces were used for the test. The batch size was eight, the optimizer was the stochastic gradient descent, and the input image size was 256 × 256 pixels.

Accuracy and IOU are indicators that can show the accuracy of the semantic segmentation of images. As the IAIL dataset is public and has already been divided into training dataset and testing dataset by the official website, therefore, it can better reflect the performance of the model during a comparison with other state-of-the-art methods. From the data in Table 2, we can see the test results of the generalization ability of the network in the Inria aerial image labeling dataset for buildings. AMUNet performed well not only in Chicago and Vienna, but also in densely populated areas such as Austin.

Table 2. Generalization ability of AMUNet on the Inria Aerial Image Labeling test dataset.

	Austin	**Chicago**	**Kitsap**	**Tyrol-w**	**Vienna**	**Overall**
SegNet + multitask loss [54], IoU	76.76	67.06	73.30	76.68	66.91	73
SegNet + multitask loss accuracy	93.21	99.25	97.84	91.71	96.61	93.07
2-level UNet IoU [55],	77.29	68.52	72.84	75.38	78.72	74.55
2-level UNet accuracy	96.69	92.4	99.25	98.11	93.79	96.05
MSMT [56], IoU	75.39	67.93	66.35	74.07	77.12	73.31
MSMT accuracy	95.99	92.02	99.24	97.78	92.49	96.06
GAN-SCA [53], IoU	81.01	71.73	68.54	78.62	81.62	77.75
GAN-SCA accuracy	97.26	93.32	99.30	98.32	94.84	96.61
AMUNet IoU	84.43	81.22	54.13	79.97	85.05	76.96
AMUNet accuracy	97.29	96.45	93.83	98.83	97.28	96.73

The SegNet + multitask method introduces a multi-task structure based on Segnet to improve performance. Although SegNet + multitask achieved the highest accuracy in Chicago, the overall IOU was not high. The 2-level U-Nets and MSMT had similar accuracy, with the former being superior to the latter in terms of IoU in all regions. This is mainly because the 2-level U-Net is based on U-Net, and its architecture is deeper than that of MSMT. GAN-SCA mainly resorts to the strategy of row training, so most of the indicators were better than the second-level Unet, but the network structure of GAN-SCA became deeper than that of AMUNet. At the same time, as the adversarial training method was used in training, it was more difficult to train the model. Therefore, compared with other models, AMUNet has great advantages. AMUNet has two obvious advantages: first, the attention mechanism and multilosses can segment buildings under different terrains accurately and suppress background information effectively. Second, the designed model has good scalability and can easily be further enhanced by enhancing the encoder structure.

4.2. Ablation Study

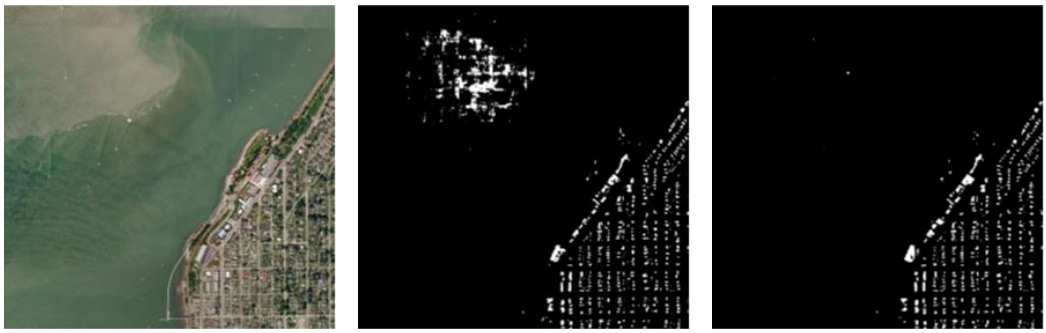
In this section, we conducted a model simplification experiment. In the simplified experiment, we tried to ensure that the experimental conditions were the same, except when comparing the loss, where we used the cross-entropy loss function to compare with the multiloss module proposed.

To test the generalization ability of the model, we used the data of 180 images of five regions with ground truth in the dataset and 180 remote sensing images without ground truth of another five regions (Bellingham, Bloomington, Innsbruck, San Francisco (SFO), and Tyrol) to perform the model

generalization test online, which was used to verify the generalization ability of the model, as given in Table 3. It can be seen in the table that both the accuracy and the IoU for the AMUNet model had been greatly improved. The AMUNet model also had a certain degree of robustness in the case of large noise interference capabilities. As shown in Figure 9, for the test set data from the Bellingham area where strong surface reflection was present, UNet produced a very large negative response, but the AMUNet model did not respond too much, effectively suppressing nontarget areas and improving the generalization ability of the model. Figure 10 shows a detailed view of the segmentation results of UNet and our model. It can be seen that, when the color of the building changed, UNet suffered from very poor segmentation, but the AMUNet model still showed the correct boundary information.

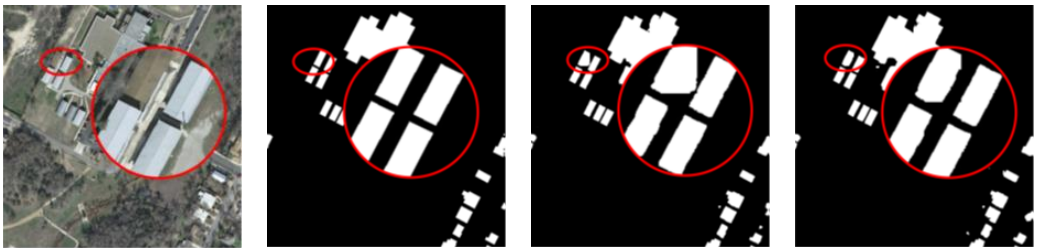
Table 3. Model generalization ability test.

	Bellingham	Bloomington	Innsbruck	SFO	Tyrol-e	Overall
UNet IoU	56.77	49.16	55.90	59.16	59.78	57.10
UNet accuracy	95.71	95.25	94.89	86.67	96.41	93.79
Attention UNet IoU	58.03	53.81	62.86	61.22	64.13	60.47
Attention UNet accuracy	95.56	95.62	95.45	87.29	96.57	94.10
AMUNet IoU	64.46	54.59	68.75	70.78	71.38	67.69
AMUNet accuracy	96.33	95.69	96.20	90.29	97.39	95.18



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

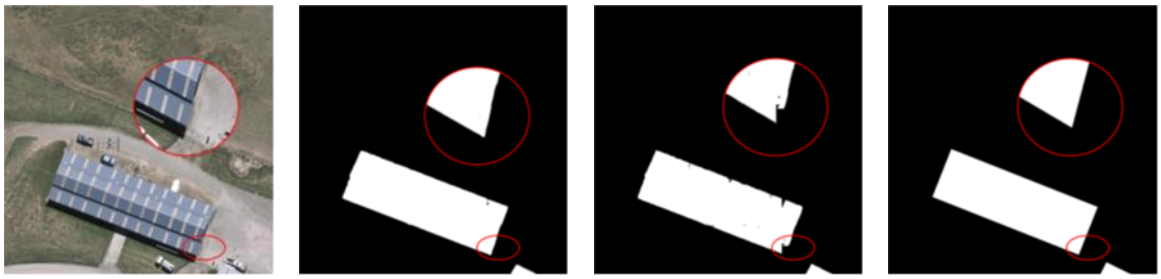
Figure 9. Results of the test generalization effect of the Inria aerial image labeling dataset. (**a**) Original test image. (**b**) Result of UNet processing. (**c**) Result of the AMUNet model test.



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

Figure 10. Segmentation boundary detail diagram. (**a**) Original diagram. (**b**) Ground truth. (**c**) UNet segmentation result. (**d**) AMUNet processing result.

Figure 11 shows the effect on the Aerial Imagery for Roof Segmentation validation set. It can be clearly seen that the accuracy of the AMUNet model for boundary processing and segmentation results had been significantly improved. UNet is not strong in information processing in the field and does not guarantee the correctness of the information control results. In comparison, our model has a stronger and more prominent generalization ability for the refinement of segmentation results.



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

Figure 11. Data results from the Aerial Imagery for Roof Segmentationdatabase. (**a**) Original diagram. (**b**) AMUNet processing result. (**c**) UNet segmentation result. (**d**) Ground truth. It is obvious from the figure that UNet will have serious omissions and suffer from misdetection in the extraction of large-scale buildings, whereas our model has a certain robustness for this situation.

In summary, in the experiment, we compared the method proposed in this paper with the state-of-the-art methods, and verified the effectiveness and technicality of the proposed method. After that, two databases were used for the model simplification experiments, which showed very good results in the evaluation set and test set. As shown in Table 2, compared with the baseline, the overall accuracy of five cities in the Inria Aerial Image Labeling dataset was improved by 3.66%, while the IOU was increased to 30.07%, which shows that AMUNet has a better effect and higher accuracy on shape constraints in extracting buildings.

5. Conclusions

The experimental results demonstrated that the method proposed had a good generalization ability on the Inria Aerial Image Labeling and Aerial Imagery for Roof Segmentation datasets. In addition, the method proposed can also be used in other semantic segmentation applications. The proposed method takes into account feature selection and channel size in space and optimizes the extraction results by learning higher order structural features. First, the space and attention mechanisms help enhance effective information while improving segmentation performance and alleviating oversegmentation. Second, we proposed a multiloss method to use pixel-based loss to constrain the results to avoid post-processing of the segmented images. After designing the attention block and multiloss module in the network structure, the ability to enhance the performance of the network model and optimize the boundary of the segmentation results could be achieved. Compared with existing models, the method proposed performed well in terms of both accuracy and IoU. The experiments were performed on Inria aerial images and on the Aerial Imagery for Roof Segmentationbuilding dataset. The results showed that the spatial and channel attention mechanisms could selectively increase effective information and improve the prediction ability of the model; the multiloss module could further optimize the prediction results in a shorter time.

Furthermore, the attention module did have an information compensation effect on higher order information, and the segmentation effect under the multiloss constraint was also significantly improved. To make the experiment more descriptive, although we chose deeper and more complex networks, the prediction results had higher accuracy. In some current network structures, the decoder structure is single, so the key is discerning how to provide the decoder with richer semantic information and location information in limited network results to improve the progress of semantic segmentation. Due to the need for the model to have some robustness and accuracy guarantee, in the loss function, an auxiliary loss function that quantitatively evaluates the segmentation results at the pixel level needs to be selected for correction.

Although the proposed method performs well as a fully supervised method, it relies on numerous manual label samples. Further research is needed to ease manual annotation. Possible directions that

can be explored include data augmentation techniques and semisupervised semantic segmentation for adversarial learning. Data augmentation techniques can increase training. Adversarial learning of semisupervised semantic segmentation can use unlabeled data to generate self-learning content signals to subdivide the network.

Author Contributions: M.G. and Y.X. proposed the network architecture design and the framework of extracting buildings. M.G. and H.L. performed the experiments and analyzed the data. Y.X. and H.L. wrote the paper. Y.H. revised the paper and provided valuable advice for the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financially supported by the National Natural Science Foundation of China (Nos. 41701446, 41971356).

Conflicts of Interest: We declare that we have no conflict of interest.

References

1. Rees, W.G. *Physical Principles of Remote Sensing*; Cambridge University Press: Cambridge, UK, 2013.
2. Xu, Y.; Chen, Z.; Xie, Z.; Wu, L. Quality assessment of building footprint data using a deep autoencoder network. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1929–1951. [CrossRef]
3. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sens.* **2017**, *9*, 522. [CrossRef]
4. Liu, Y.; Piramanayagam, S.; Monteiro, S.T.; Saber, E. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order CRFs. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017), Honolulu, Hawaii, USA, 21–26 July 2017; pp. 76–85.
5. Pan, X.; Gao, L.; Marinoni, A.; Zhang, B.; Yang, F.; Gamba, P. Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network. *Remote Sens.* **2018**, *10*, 743. [CrossRef]
6. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
7. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3213–3223.
8. Sirmacek, B.; Unsalan, C. Urban-Area and Building Detection Using SIFT Keypoints and Graph Theory. *IEEE Trans. Geosci. Remote* **2009**, *47*, 1156–1167. [CrossRef]
9. Huang, X.; Zhang, L. Morphological Building/Shadow Index for Building Extraction from High-Resolution Imagery over Urban Areas. *IEEE J.-Stars* **2012**, *5*, 161–172. [CrossRef]
10. Zhang, Q.; Huang, X.; Zhang, G. A Morphological Building Detection Framework for High-Resolution Optical Imagery over Urban Areas. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1388–1392. [CrossRef]
11. Ahmadi, S.; Zoej, M.J.V.; Ebadi, H.; Moghaddam, H.A.; Mohammadzadeh, A. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. *Int. J. Appl. Earth Obs.* **2010**, *12*, 150–157. [CrossRef]
12. Liasis, G.; Stavrou, S. Building extraction in satellite images using active contours and colour features. *Int. J. Remote Sens.* **2016**, *37*, 1127–1153. [CrossRef]
13. Li, E.; Xu, S.; Meng, W.; Zhang, X. Building Extraction from Remotely Sensed Images by Integrating Saliency Cue. *IEEE J.-Stars* **2017**, *10*, 906–919. [CrossRef]
14. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [CrossRef]
15. 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]
16. 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.* **2015**, *34*, 58–69. [CrossRef]
17. 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]

18. Han, J.; Zhang, D.; Cheng, G.; Liu, N.; Xu, D. Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. *IEEE Signal Proc. Mag.* **2018**, *35*, 84–100. [CrossRef]
19. 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]
20. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
21. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-Resolution Aerial Image Labeling with Convolutional Neural Networks. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *55*, 7092–7103. [CrossRef]
22. Shrestha, S.; Vanneschi, L. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
23. Allen-Zhu, Z.; Li, Y.; Song, Z. A Convergence Theory for Deep Learning via over-Parameterization. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10 July–15 July 2018; pp. 242–252.
24. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intel.* **1998**, *20*, 1254–1259. [CrossRef]
25. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef]
26. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road extraction from high-resolution remote sensing imagery using deep learning. *Remote Sens.* **2018**, *10*, 1461. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
28. Pinheiro, P.O.; Lin, T.; Collobert, R.; Dollár, P. *Learning to Refine Object Segments*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 75–91.
29. 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]
30. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii Convention Center, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Zhu, S.; Liu, Y. Scene Segmentation and Semantic Representation for High-Level Retrieval. *IEEE Signal Proc. Lett.* **2008**, *15*, 713–716. [CrossRef]
32. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
33. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. *Unet++: A Nested U-Net Architecture for Medical Image Segmentation*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
34. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intel.* **2018**, *40*, 834–848. [CrossRef]
35. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
36. Chen, L.; 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.
37. Cao, C.; Liu, X.; Yang, Y.; Yu, Y.; Wang, J.; Wang, Z.; Huang, Y.; Wang, L.; Huang, C.; Xu, W.; et al. Look and Think Twice: Capturing Top-down Visual Attention with Feedback Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2956–2964.
38. Larochelle, H.; Hinton, G.E. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2010; pp. 1243–1251.
39. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2204–2212.

40. Sountsov, P.; Santucci, D.M.; Lisman, J.E. A biologically plausible transform for visual recognition that is invariant to translation, scale, and rotation. *Front. Comput. Neurosc.* **2011**, *5*, 53. [CrossRef]
41. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 2017–2025.
42. Bluche, T. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 838–846.
43. Miech, A.; Laptev, I.; Sivic, J. Learnable pooling with context gating for video classification. *arXiv* **2017**, arXiv:1706.06905.
44. Stollenga, M.F.; Masci, J.; Gomez, F.; Schmidhuber, J. Deep networks with internal selective attention through feedback connections. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 3545–3553.
45. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
46. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883.
47. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, 21 July–26 July 2017; pp. 2881–2890.
48. Dai, J.; He, K.; Sun, J. Convolutional feature masking for joint object and stuff segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3992–4000.
49. Unnikrishnan, R.; Pantofaru, C.; Hebert, M. Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 929–944. [CrossRef]
50. Fowlkes, E.B.; Mallows, C.L. A method for comparing two hierarchical clusterings. *J. Am. Stat Assoc.* **1983**, *78*, 553–569. [CrossRef]
51. Jiang, X.; Marti, C.; Irniger, C.; Bunke, H. Distance measures for image segmentation evaluation. *EURASIP J. Appl. Signal Proc.* **2006**, *2006*, 209. [CrossRef]
52. Unnikrishnan, R.; Hebert, M. Measures of similarity. In Proceedings of the 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1, Breckenridge, CO, USA, 5–7 January 2005; p. 394.
53. Pan, X.; Yang, F.; Gao, L.; Chen, Z.; Zhang, B.; Fan, H.; Ren, J. Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms. *Remote Sens.* **2019**, *11*, 917. [CrossRef]
54. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1480–1484.
55. Khalel, A.; El-Saban, M. Automatic pixelwise object labeling for aerial imagery using stacked u-nets. *arXiv* **2018**, arXiv:1803.04953.
56. Marcu, A.; Costea, D.; Slusanschi, E.; Leordeanu, M. A multi-stage multi-task neural network for aerial scene interpretation and geolocalization. *arXiv* **2018**, arXiv:1804.01322.