remote sensing

MDPI

Article

High-Precision Segmentation of Buildings with Small Sample Sizes Based on Transfer Learning and Multi-Scale Fusion

Xiaobin Xu [1,2,*,†], Haojie Zhang [1,2,†], Yingying Ran [1,2] and Zhiying Tan [1,2]

[1] College of Mechanical and Electrical Engineering, Hohai University, Changzhou 213022, China; 211619010102@hhu.edu.cn (H.Z.); 200219030004@hhu.edu.cn (Y.R.); zytan@hhu.edu.cn (Z.T.)
[2] Jiangsu Key Laboratory of Special Robot Technology, Hohai University, Changzhou 213022, China
* Correspondence: xxbtc@hhu.edu.cn
† These authors contributed equally to this work.

Abstract: In order to improve the accuracy of the segmentation of buildings with small sample sizes, this paper proposes a building-segmentation network, ResFAUnet, with transfer learning and multi-scale feature fusion. The network is based on AttentionUnet. The backbone of the encoder is replaced by the ResNeXt101 network for feature extraction, and the attention mechanism of the skip connection is preserved to fuse the shallow features of the encoding part and the deep features of the decoding part. In the decoder, the feature-pyramid structure is used to fuse the feature maps of different scales. More features can be extracted from limited image samples. The proposed network is compared with current classical semantic segmentation networks, Unet, SuUnet, FCN, and SegNet. The experimental results show that in the dataset selected in this paper, the precision indicators of ResFAUnet are improved by 4.77%, 2.3%, 2.11%, and 1.57%, respectively, compared with the four comparison networks.

Keywords: semantic segmentation; building extraction; transfer learning; small sample data

Citation: Xu, X.; Zhang, H.; Ran, Y.; Tan, Z. High-Precision Segmentation of Buildings with Small Sample Sizes Based on Transfer Learning and Multi-Scale Fusion. *Remote Sens.* **2023**, *15*, 2436. https://doi.org/10.3390/rs15092436

Academic Editors: Xiangrong Zhang, Yansheng Li, Lichao Mou, Licheng Jiao and Xu Tang

Received: 26 March 2023
Revised: 3 May 2023
Accepted: 3 May 2023
Published: 5 May 2023

1. Introduction

High-resolution remote-sensing images contain rich information on geometry, texture, and spatial distribution, which are widely used in agriculture, geological surveys, disaster detection, environmental protection, transportation, urban planning, and other fields [1–6]. With the development of remote-sensing technology, the extraction of information on target buildings automatically and accurately from high-resolution remote-sensing images has gradually become a key problem. The improvements in remote-sensing-image resolution also create new challenges. For example, when the texture and structure of one building are inconsistent, incomplete building extractions, or internal voids, might occur. At present, researchers are dedicated to the task of automatically extracting information on buildings from remote-sensing images. A variety of extraction algorithms have been proposed, which can be mainly divided into traditional methods and deep-learning-based approaches.

Regarding the traditional methods, the structure and texture, of and prior information on buildings in remote-sensing images are used for extraction. Irvin R. B. et al. proposed a method to extract information on buildings from remote-sensing images by using building shadows [7]. Lee et al. extracted building-height information based on the building-volume shadow analysis (VSA) method [8]. However, this method is not suitable for urban buildings with high densities, whose shadows might overlap with each other. On the other hand, many researchers extract information on buildings by analyzing their structure and texture. Levitt. S proposed a method based on image-texture detection [9]. This method is simple and efficient, but the actual effect depends on the choice of texture. Lin et al. adopted the edge-detection algorithm to extract information on buildings by detecting their roofs, walls, and shadows [10]. They extracted candidate building edges by analyzing line-segment features in remote-sensing images, and determined the candidate

building boundaries by region segmentation, analysis, and region merging [11]. With the development of digital-image-processing technology, methods based on mathematical morphology have emerged. Xu et al. used two-dimensional OTSU segmentation to carry out the rough segmentation of buildings, and used mathematical morphology and other means to detect and eliminate the non-buildings in aerial images [12]. Gavankar et al. proposed an automatic building-extraction method based on mathematical morphology in high-resolution remote-sensing images by using the top-hat filter and K-means algorithm of image morphology [13]. In general, the traditional methods only use the shallow features of buildings. Their extraction accuracy is not high, their generalization ability is poor, limited to specific cases requiring significant building features.

In recent years, the application of deep learning has emerged. Deep learning uses convolutional neural networks to extract and learn features in images, resulting in breakthroughs in classification, detection, and segmentation tasks. Currently, the mainstream convolutional neural networks include AlexNet [14], VGGNet [15], ResNet [16], etc. Compared with traditional feature-extraction algorithms, convolutional neural networks can effectively extract the deep features of images with great levels of generalization. Scholars have designed a series of semantic segmentation networks using convolutional neural networks. Long et al. proposed the full convolutional neural network, FCN [17], in 2015, and applied it in the semantic segmentation of images. The FCN restores the feature map to the image's input size through deconvolution, so as to achieve pixel-by-pixel prediction. Subsequently, many new semantic segmentation networks were derived in order to improve the FCN, including Unet [18], SegNet [19], PSPNet [20], DeepLab [21], etc. In particular, the Unet network has a completely symmetric encoder–decoder structure, which preserves the simplicity of the structure while improving its performance. Many scholars have undertaken efforts based on the application of Unet to the automatic extraction of information on buildings with remote-sensing images. Oktay et al. introduced attention gates into the skip connections between the encoder and decoder in Unet. Attention gates can be used to suppress irrelevant regions and focus on useful salient features during network training [22]. He et al. proposed a first-order and second-order hybrid attention network [23] and applied it to Unet to improve the correlation between the intermediate features. Shi et al. introduced a spatial-channel attention mechanism on the basis of Unet [24] to improve the feature-extraction ability of the neural networks. The structure of Unet can also be improved by using other networks. Ji et al. improved Unet by adapting FPN, and proposed SuUnet [25] to deal with objects of different sizes in remote-sensing images or multi-scale problems in the images of different resolutions. Delibasoglu et al. employed Inception blocks instead of traditional convolutional layers in the coding phase of Unet [26]. This approach enabled the model to capture building features at multiple scales by utilizing parallel convolution kernels of different sizes. Abdollahi et al. employed dense connections, ConvLSTM [27], and the SE mechanism [28] to suppress irrelevant information [29] while incorporating prior information on buildings. Although deep-learning-based methods are superior in the extraction of information on buildings from images, large numbers of building samples are required in the network-training processes. However, in real-world-application scenarios, the sample data sizes are often limited, leading to unsatisfactory learning results when using semantic segmentation networks.

Obtaining strong network performances with limited data-set capacity is a common problem. At present, an intuitive solution is to use transfer learning, which involves the transfer of the pre-training weight of related fields and fine-tuning the network. In medical imaging, label scarcity is a common problem. Therefore, transfer learning has been widely applied in the field of medical imaging, in the form of by domain adaptation (DA). This is a transfer-learning method in which the source task is the same as the target task, but the data distribution of the source domain is different from that of the target domain. In recent years, scholars have proposed a series of DA methods applicable to the medical field. Khan et al. [30] pre-trained the VGG network on the ImageNet dataset and then fine-tuned it using small amounts of labeled magnetic resonance imaging (MRI)

data for the classification of Alzheimer's disease (AD). Gu et al. [31] proposed a two-stage adaptation approach based on an intermediate domain for use in situations in which there are insufficient target samples with which to fine-tune the model directly. All these examples are domain-adaptive methods under supervision. Since data-set annotation is time-consuming, scholars have proposed a series of unsupervised domain-adaptive (UDA) methods. Wollmann et al. [32] proposed an UDA method based on style transfer learning. They first used CycleGAN to transform whole-slide images (WSIs) of lymph nodes from a source domain to the target domain. Next, DenseNet was used to classify breast cancers. However, due to the lack of class-level joint distribution in adversarial learning, the aligned distribution is not discriminative. In order to solve this problem, Liu et al. [33] proposed a novel margin-preserving self-paced contrast learning (MPSCL) model to facilitate the classification of the alignment of perceptual features through cross-domain contrast learning, which effectively improved the segmentation performance of the network. Yao et al. [34] proposed an unsupervised domain-adaptive framework. The framework achieves full image alignment to alleviate the domain-offset problem, and introduces 3D segmentation in domain-adaptive tasks to maintain semantic consistency at deep levels. In some areas of confidential medical imaging, source datasets are often unavailable. To solve this problem, Liu et al. [35] proposed a passive semantic segmentation adaptive framework, which can use knowledge migration to obtain knowledge in the source domain from existing source models, and then achieve domain adaptation. Stan et al. [36] developed an UDA algorithm that does not require access to source-domain data during target adaptation. This method encodes the source-domain information as an internal distribution, which is used to guide the adaptation in the absence of source samples. The successful application of transfer learning in the medical field also provides a new approach to the high-precision extraction of information on buildings with small sample sizes.

To address the challenge of achieving high-precision building segmentation with neural networks with small sampling sizes, this paper proposes a segmentation network, ResFAUnet, based on transfer learning and multi-scale fusion. The contributions of this paper are as follows:

(1) In the encoding part of the network, the ResNeXt101 network with group convolution is used for feature extraction, and pre-trained weights are introduced in the transfer learning, which improves the feature-learning performance of the network with limited samples.
(2) In the decoding part of the network, the original single-chain up-sampling feature-fusion mode is replaced by the multi-chain, while in the decoding part, the features from different up-sampling scales are stacked to realize direct information fusion across scales.

The subsequent sections of this paper are organized as follows. Section 2 introduces the network structure of ResFAUnet. Section 3 introduces the datasets, parameter setting, and evaluation indexes used in the experiment, describes the ablation experiment, compares the performance of ResFAUnet with those of other networks, and analyzes the experimental results. Section 4 discusses the influence of some of the training parameters (data-set capacity, input-image size) on the performance of the ResFAUnet network during the training process. Finally, Section 5 summarizes the conclusions of this study.

2. Materials and Methods

This section describes the construction of ResFAUnet network. Section 2.1 describes the overall architecture of ResFAUnet. Section 2.2 introduces the encoding structure of grouping convolution and transfer learning. Section 2.3 introduces the multi-scale feature-fusion decoding structure.

2.1. ResFAUnet Network Model

In practical applications, sample datasets typically have limited sizes. To achieve greater accuracy in building extraction, high-quality feature extraction is necessary. There-

fore, a semantic segmentation network, ResFAUnet, based on transfer learning and multi-scale fusion, is proposed for limited datasets. The structure diagram is shown in Figure 1. The code and models are available at https://github.com/cjluozi/ResFAUnet-V1.0.

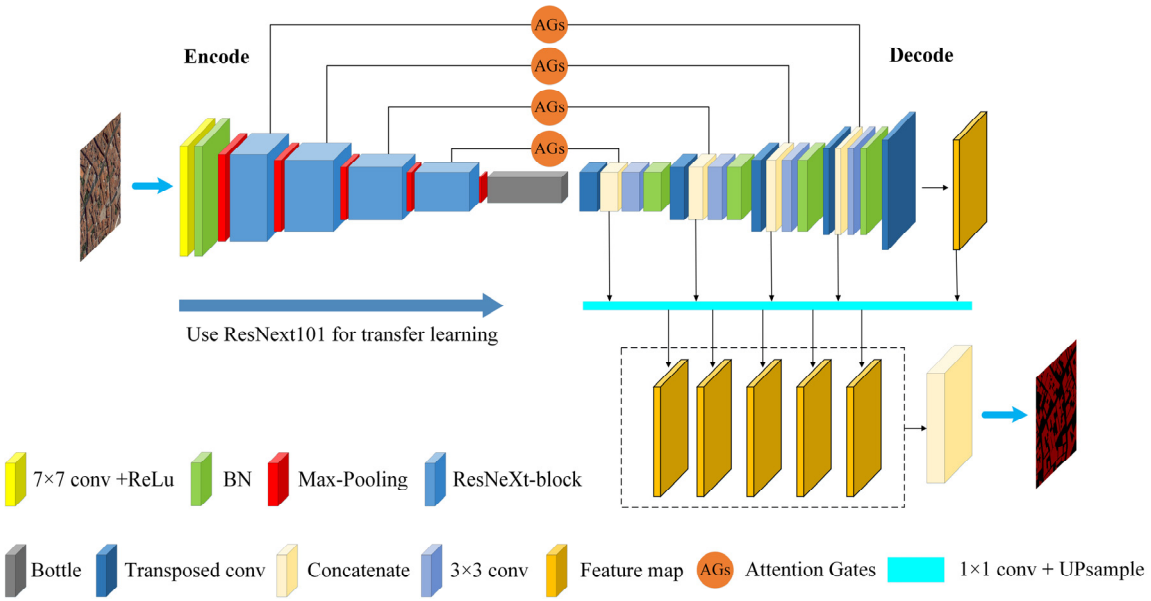


Figure 1. ResFAUnet structure diagram.

The ResFAUnet is built on the basis of AttentionUnet network, which is a variant of Unet network, one of the most commonly used networks in the field of semantic segmentation. It has a completely symmetrical encoding–decoding structure, in which the encoder is responsible for feature extraction, the decoder is used for feature recovery, and the skip connection, which is used to fuse the image features of the encoding part and the decoding part, is introduced between the encoder and the decoder. Unet is widely used in semantic segmentation because of its simple and stable structure. The structure diagram of Unet is shown in Figure 2.

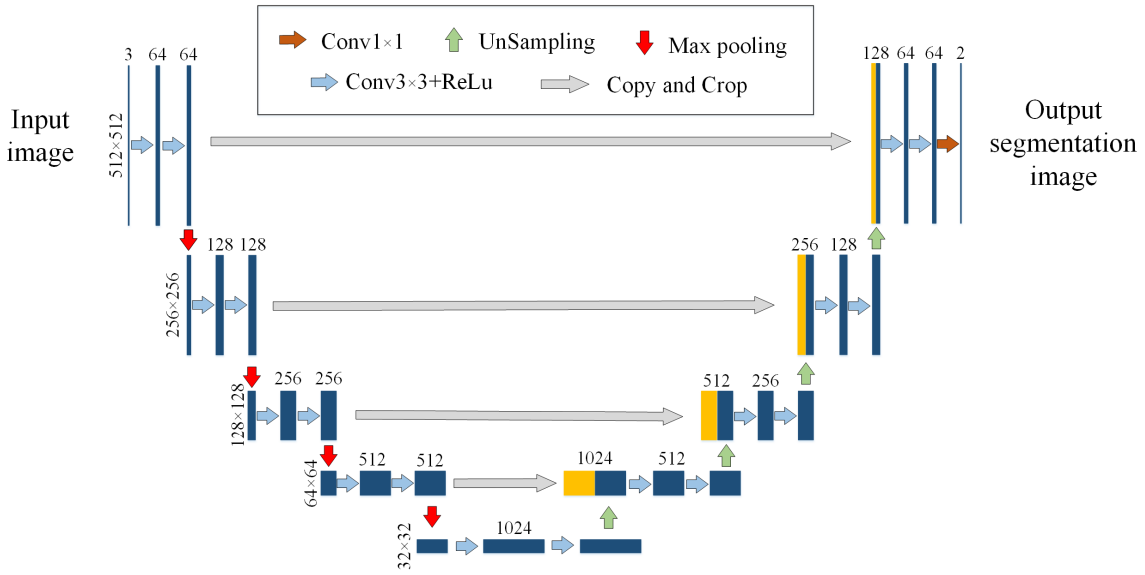


Figure 2. Unet structure diagram.

The AttentionUnet network inherits the "U"-shaped encoding–decoding structure of the classical semantic segmentation network of Unet, and adds attention gates to skip

connection between the encoder and the decoder. In the training process, the self-attention mechanism in attention gate can obtain the weight of attention through forward propagation and backward feedback. Multiplying the weight of attention by corresponding feature images can effectively suppress irrelevant features and focus on useful features. The structure diagram of self-attention mechanism is shown in Figure 3.

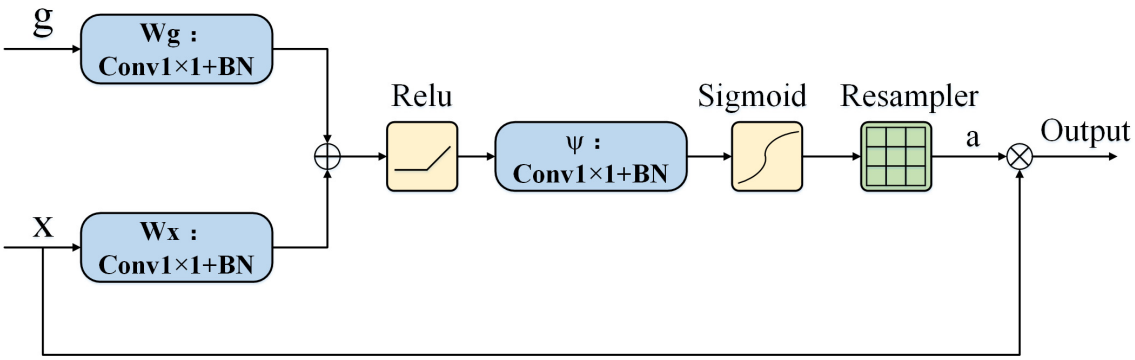


Figure 3. Self-attention-mechanism structure diagram.

Although the addition of attention gates results in the network focusing on the region of interest for learning, and improves the segmentation accuracy compared with Unet network, it does not improve the feature-extraction ability of the network. Therefore, AttentionUnet is improved as follows:

(1) In the encoding stage of AttentionUnet, the backbone of the feature-extraction network is replaced by ResNeXt101 network, which is an improved version of the ResNet101 network. In ResNeXt101 network, group convolution is introduced in the convolution part to improve the feature-extraction ability of the network when the overall framework and computation amount are basically consistent with ResNet101 network. The pre-trained weight of ResNeXt101 on the ImageNet dataset is used for the transfer learning, so that the network converges with reduced training epoch.

(2) In the decoding stage of AttentionUnet, feature maps of different sizes are extracted from the structure of the feature pyramid, and all the feature images of different sizes are stacked at the output to realize the fusion of the features from low to high dimensions to obtain the final predicted image.

2.2. Group Convolution + Transfer Learning

Transfer learning uses the generalization capability of convolutional neural networks to transfer the network structure and weights learned from solving task A to a new task B, resulting in improved performance of task B. Therefore, cases with limited samples, transfer learning can be particularly useful, as it allows the use of general features learned from pre-trained networks, accelerating the training process and providing better results. In this paper, the idea of transfer learning is employed, and ResNeXt101 network is selected to replace the feature-extraction part of AttentionUnet network. With ResNeXt101 and ResNet101 networks with the same widths and depths, group convolution is used to replace the middle layer of Resbolck in ResNet. This enhances the network performance without increasing computational load. The Resblock is the residual block of ResNet network and also the basic module of ResNet network. The ResNet network is a network composed of multiple Resblocks. Similarly, a ResNeXt block is a residual block of the ResNeXt network. Figure 4 shows the ResNeXt block corresponding to ResNeXt. The pre-trained weight of ResNeXt on the ImageNet dataset is used for transfer learning, The output of each ResNeXt_block is connected with the corresponding features of the original AttentionUnet decoding part, so as to realize the feature fusion of the decoding and the encoding parts.

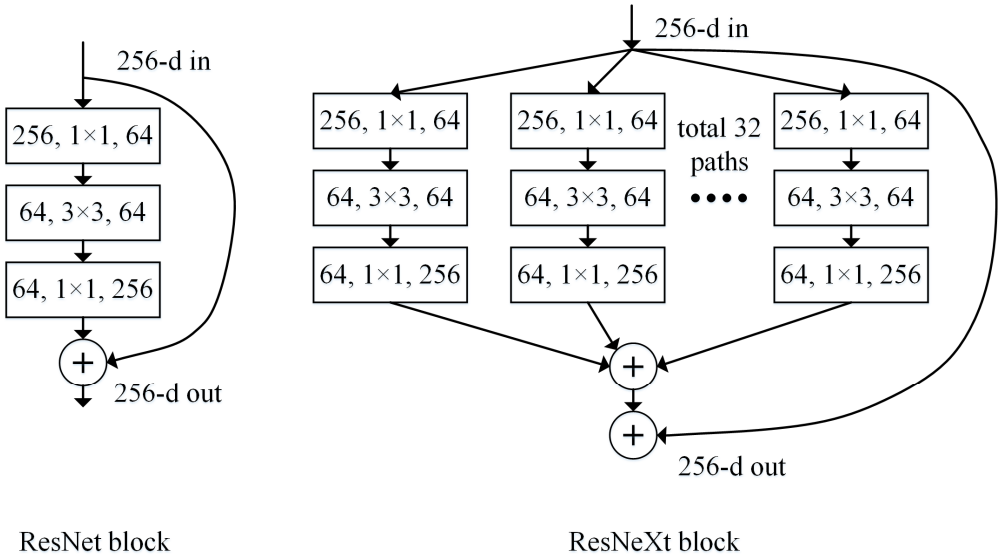


Figure 4. ResNeXt block diagram. Left: residual block of ResNet. Right: residual block of ResNeXt, cardinality = 32, d = 4.

2.3. Multi-Scale Feature Fusion

The Unet network was one of the first algorithms to use full convolutional networks for semantic segmentation. It has a completely symmetrical encoding–decoding structure and introduces skip connections to fuse the image features of the encoding and the decoding parts. The AttentionUnet, a variant of Unet, adds a self-attention mechanism to the skip-connection part of the network, which can improve the segmentation performance by suppressing irrelevant regions. However, both Unet and AttentionUnet only use the last layer of decoding part for image-pixel-category classification. The feature information of the last layer is derived from the up-sampling of the decoder, but the single-chain up-sampling of decoding part can only realize the indirect fusion of information of different scales. It is difficult to realize the complete fusion of information from different scales.

In object-detection field, FPN (feature-pyramid networks) are the most commonly used multi-scale fusion frameworks, and they are used to solve the identification problem of objects with different scales. The structure diagram of the feature-pyramid network is shown in Figure 5. The FPN is a top-down architecture with horizontal connections, which is used to construct high-level semantic feature maps of different scales. By predicting feature maps of different scales and fusing the prediction results of different scales, objects of different scales can be recognized.

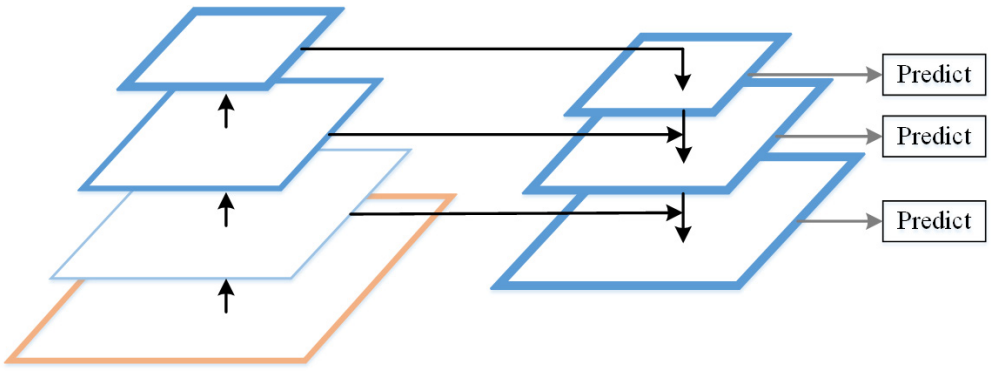


Figure 5. Feature-pyramid-network structure diagram.

In this paper, to improve the feature-extraction ability of AttentionUnet decoding part, a multi-scale fusion framework similar to the golden tower structure is proposed.

As shown in Figure 6, this structure preserves the skip connections in both the encoding and decoding stages of AttentionUnet, and replaces the single-chain up-sampling of the decoding stage with the multi-chain. In the decoding stage, each scale is extracted with the same number of layers as the number of categories, and up-sampled to unify to the output size. Next, the sampled features from different scales are stacked to realize direct fusion of information from different scales in the decoding stage.

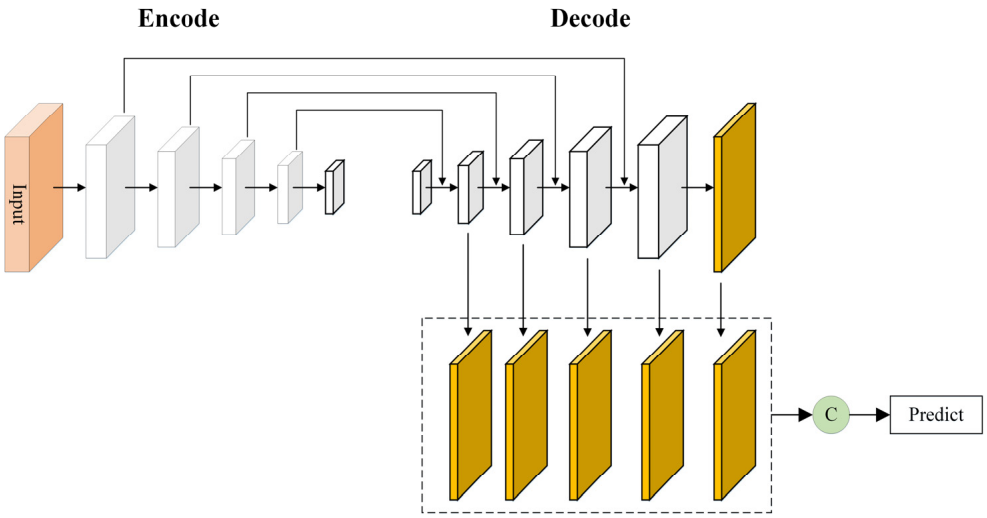


Figure 6. Schematic diagram of multi-scale fusion-frame structure.

3. Experiments and Analysis

In this section, the selection of the public datasets, the setting of the network parameters, and the evaluation index are introduced first. Next, we report the numerous experiments that were conducted to verify the performance of ResFAUnet. Finally, the experimental results are analyzed and discussed.

3.1. Selection of Public Datasets

In this paper, three public datasets were selected to evaluate the performance of the ResFAUnet network. In order to test the performance of ResFAUnet in cases of insufficient data, a total of 500 images were selected as experimental samples from the training set and test set from the three datasets, and the segmentation results were evaluated on the test set after training.

Inria Aerial Image Labeling Dataset: This dataset [37] includes 180 aerial orthographic corrected color images with public labels, covering an area of 810 square kilometers, with a spatial resolution of 0.3 m and an image size of 5000 × 5000 pixels. It covers five different urban settlements (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna), ranging from densely populated areas to alpine towns. Each subset of the dataset contains a different image of each city, which was used to evaluate the network's generalization ability. Figure 7 shows some examples.

Building instances of typical cities in China: This dataset [38] is composed of 7260 images from four Chinese cities (Beijing, Shanghai, Shenzhen, and Wuhan) and 63,886 building examples. The size of each image is 500 × 500 pixels and the spatial resolution is 0.29 m. The dataset was used to test the segmentation performance of the network for localized buildings. Figure 8 shows some examples.

WHU Building Dataset: This dataset [39] is composed of aerial dataset, a global urban-satellite dataset, and an East Asian urban-satellite dataset. The Global Cities Satellite dataset collects remote-sensing images of cities around the world from satellite resources such as QuickBird, Worldview series, IKONOS, and ZY-3. It contains 204 512 × 512-pixel images with spatial resolutions ranging from 0.3 m to 2.5 m. The use of this dataset is

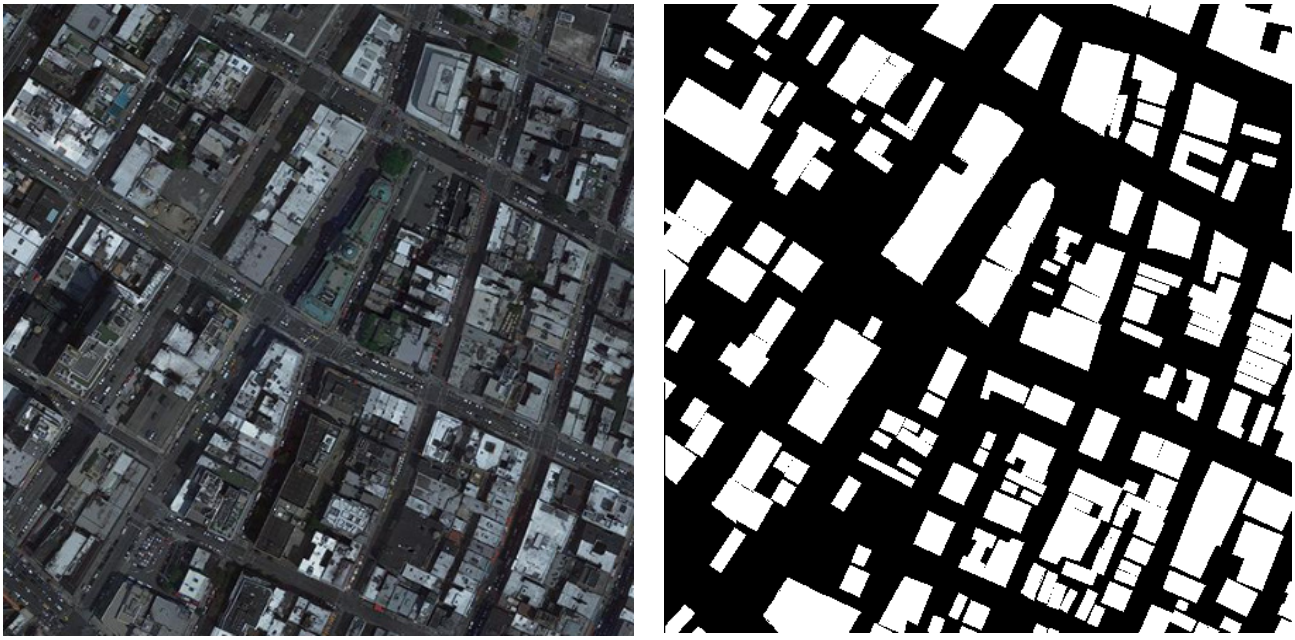
challenging due to the variations between different satellites, sensors, and atmospheric conditions. The dataset was used to test the robustness of the network. Figure 9 shows some examples.

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

Figure 7. An example of the Inria dataset. (**a**) Original image. (**b**) Ground-truth label.

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

Figure 8. An example of the China dataset. (**a**) Original image. (**b**) Ground truth label.



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

Figure 9. An example of the WHU dataset. (**a**) Original image. (**b**) Ground-truth label.

3.2. Parameter Settings

3.2.1. Image Setting

Due to the inconsistency in the image sizes across the different datasets, it was not feasible to use them directly for training. To address this issue, the original images were all preprocessed to obtain 512 × 512-pixel images across the three datasets. For the Inria dataset, the images were expanded from 5000 × 5000 pixels to 5120 × 5120 pixels using bilinear interpolation and then trimmed to 512 × 512 pixels. Similarly, the image sizes in the China dataset were expanded to 512 × 512 pixels by using bilinear interpolation. The WHU datasets were simply retained in their original size without any modification. In order to simulate a limited dataset with a small sample size, for the Inria dataset and the China dataset, 500 images (450 in training set and 50 in test set) were selected. Since the number of images in the WHU dataset was only 204, we also used the 9:1 ratio to divide the training set and the test set.

3.2.2. Training Setting

The Pytorch framework was used to train the ResFAUnet model. During the training, the same parameters were used for all three datasets; the learning rate was set to 0.005, and the batch size was set to 2. This is the maximum size that the graphics card can support at pixel sizes of 512 × 512. The SGD (stochastic gradient descent) optimizer with a momentum item of 0.9 was used to train the network. The SGD optimizer is one of the most commonly used optimization algorithms in deep learning. In each iteration, it randomly selects a small batch of samples to calculate the gradient of the loss function and updates the parameters with the gradient. This randomness makes the algorithm more robust, prevents it from falling into local minima, and improves its training speed. The training graphics card was NVIDIA GeForce GTX 1080Ti (It is made by Nvidia Corporation and sourced from Beijing, China), and the training was performed for 100 epochs. The CE loss (cross-entropy loss) was used to calculate the loss. The CE loss is a loss function based on cross entropy, which can be used to determine how close the actual output is to the desired output. The CE loss uses cross entropy to measure the difference between the output of the network and the label, and uses this difference to update the network parameters through back propagation. The formula for CE loss is as follows:

$$CE(p,q) = -\sum_{i=1}^{C} p_i \log(q_i) \tag{1}$$

where C represents the number of categories, p_i represents the true value, and q_i represents the predicted value.

3.3. Evaluation Indexes

To evaluate the network performances, the *Pre* (precision), *MIoU* (mean intersection over union), *Rec* (recall), and *F*1 (*F*1 score) were employed. The Pre index represents the probability that a certain category of prediction is correct in the forecast results. The *MIoU* index refers to the ratio of the intersection and the union of each type of predicted result and real value, the result of the summation, and then the average. The *Rec* index refers to the probability that a category will be predicted correctly in the true value. The *F*1 index is the harmonic average of *Pre* and *Rec*, and it is used to evaluate the comprehensive performance of the network. The four indicators are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$MIoU = \frac{1}{k}\sum_{i=1}^{k} \frac{TP}{TP + FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 * Rec * Pre}{Rec + Pre} \tag{5}$$

where *TP* (true positive) refers to the part of the input image that is actually positive and predicted to be positive. The *FP* (false positive) represents the part of the input image that is actually negative and predicted to be positive. The *TN* (true negative) represents the part that is actually negative sample and predicted to be negative. The *FN* (false negative) means the part that is actually a positive sample and is predicted to be a negative sample.

3.4. Comparisons and Analysis

In this section, ResFAUnet is compared with Unet, SuUnet, FCN, SegNet, and other mainstream deep-learning networks. The training parameters were kept the same for the different networks to conduct the training using the Inria dataset, China dataset, and WHU dataset. The performance of ResFAUnet is evaluated through qualitative and quantitative analyses of the results.

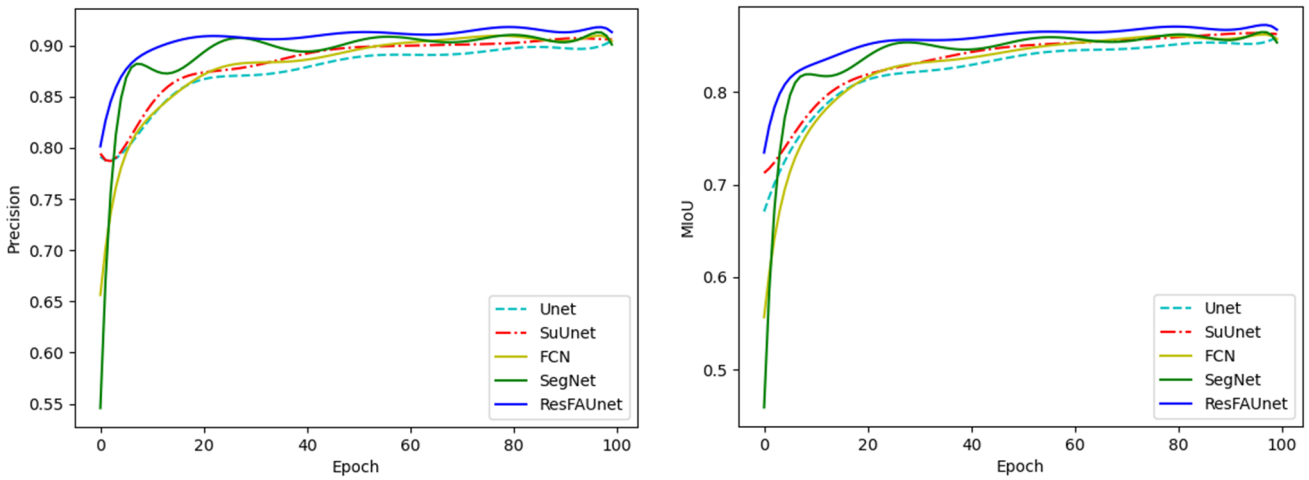
3.4.1. Inria Dataset

Table 1 presents the quantitative evaluation indexes of the different models evaluated on the Inria dataset. As shown in the table, the ResFAUnet network proposed in this paper outperformed the Unet, FCN, SuUnet, and SegNet networks in all four indexes. Due to the high resolution of the Irina dataset, all of the four networks except Unet achieved precision and MIoU values above 90% and 85%, respectively. Compared to the Unet, the precision index of the ResFAUnet network increased by 3.23%, while that of the SuUnet network with the multi-scale fusion structure improved by 0.97%, and those of the FCN network and the SegNet network with the transfer learning achieved improvements of 1.19% and 0.89%, respectively. Among all the comparison networks, ResFAUnet had the highest *F*1 score and the lowest standard deviation, which means that the model was the most stable. Additionally, by taking the MIoU index into account, a better balance was achieved between recall and precision. Figure 10 shows the comparison of the precision and MIoU

indexes of the Inria dataset among the five networks. From the graph, it can be seen that ResFAUnet achieved the highest precision and MIoU indicators by using fewer epochs.

Table 1. Test results of different network models in Inria datasets.

Model	*Precision*	*MIoU*	*Recall*	*F1* Score
Unet	88.81 ± 0.42	84.13 ± 0.31	93.77 ± 0.25	91.22 ± 0.28
SuUnet	91.07 ± 0.29	86.65 ± 0.25	93.69 ± 0.25	92.36 ± 0.27
FCN	90.85 ± 0.28	86.11 ± 0.27	93.92 ± 0.23	92.36 ± 0.24
SegNet	91.15 ± 0.27	86.22 ± 0.27	93.70 ± 0.25	92.41 ± 0.24
ResFAUnet	**92.04 ± 0.25**	**87.26 ± 0.26**	**94.02 ± 0.22**	**93.02 ± 0.23**



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

Figure 10. Test-metrics curves of different models tested on Inria dataset. (**a**) Precision-index graph of different models tested on Inria dataset. (**b**) MIoU-index curves of different models tested on Inria dataset.

Figure 11 shows the test image results under different scenarios. The first imagine is that of a small target building. For the building in the green box in the figure, the Unet, SuUnet, FCN, and SegNet networks did not detect the smallest building, and only the ResFAUnet network completely segmented all the buildings. In the green box in the second image, only ResFAUnet successfully detected the outline of the occluded building. In the third image, the other four networks only segmented part of the outline of the building, while only ResFAUnet segmented the complete outline of the building. In the fourth scenario, the differentiation between the building and the background was not high. In the green box in the image, only ResFAUnet completely divided all the buildings. Therefore, the ResFAUnet network proposed in this paper achieved good results in small-target-building segmentation and building-contour integrity.

3.4.2. China Dataset

Table 2 shows the quantitative evaluation indexes of various models tested on the China dataset. It can be seen that the test scores of the five models using the China dataset were all lower than those using the Inria. This result was attributed to the fact that the China dataset contains image data from Beijing, Shanghai, Shenzhen, and Wuhan, where the building scenes are relatively complex. In particular, the presence of shadows between high-rise city buildings creates significant challenges for segmentation. Additionally, the similarity between city roads and buildings makes it difficult to distinguish between them. Despite these challenges, ResFAUnet still achieved good segmentation results, which were significantly higher than those of Unet, FCN, SuUnet, and SegNet in all four indexes.

Compared to Unet and SuUnet, the precision improved by 5.36% and 3.07%, respectively. Compared to FCN and SegNet, which also adopt transfer learning, the improvements were 1.75% and 1.28%, respectively. Furthermore, the precision and MIoU indexes of the five models using the China dataset were compared, as shown in Figure 12. From the graph, it can be seen that the ResFAUnet curve was the most stable, requiring only a few epochs to achieve convergence, and its performance was superior to those of the four other comparison networks.

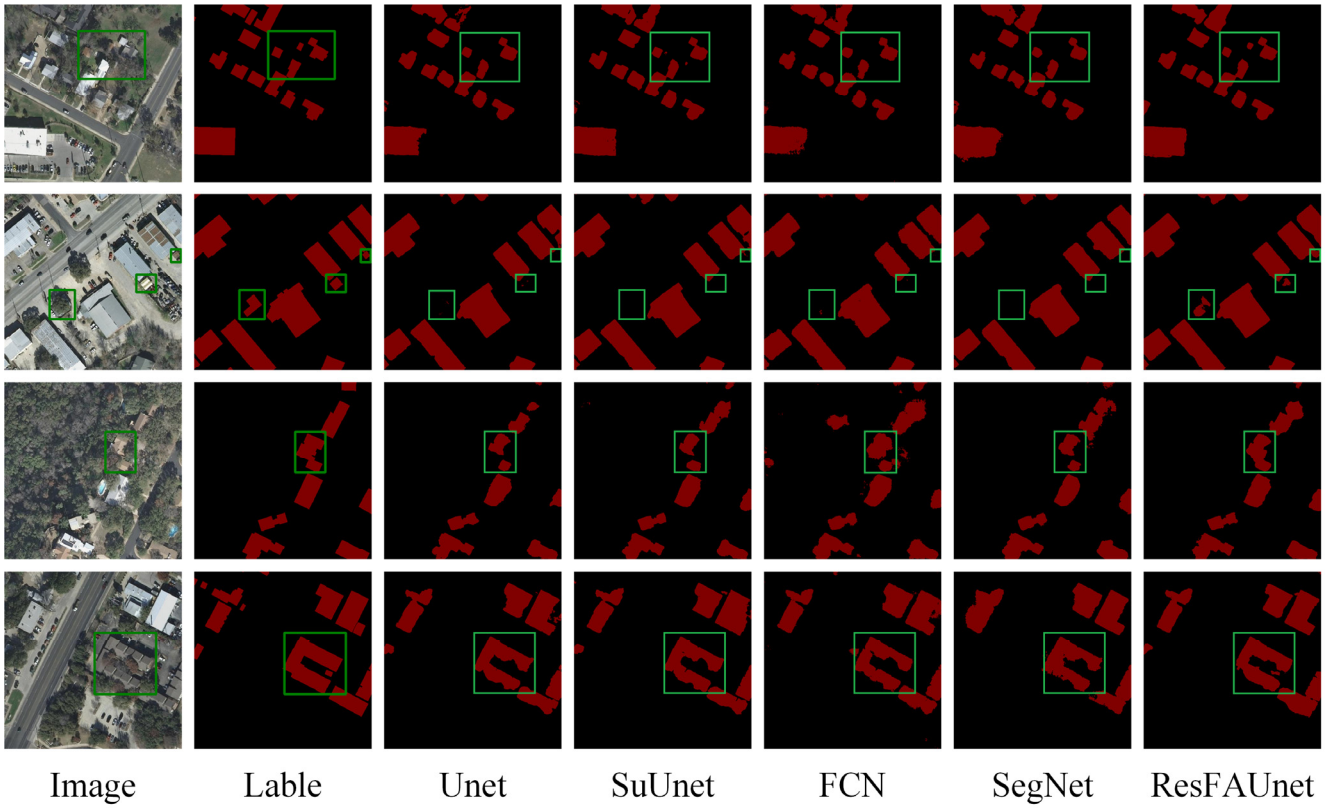


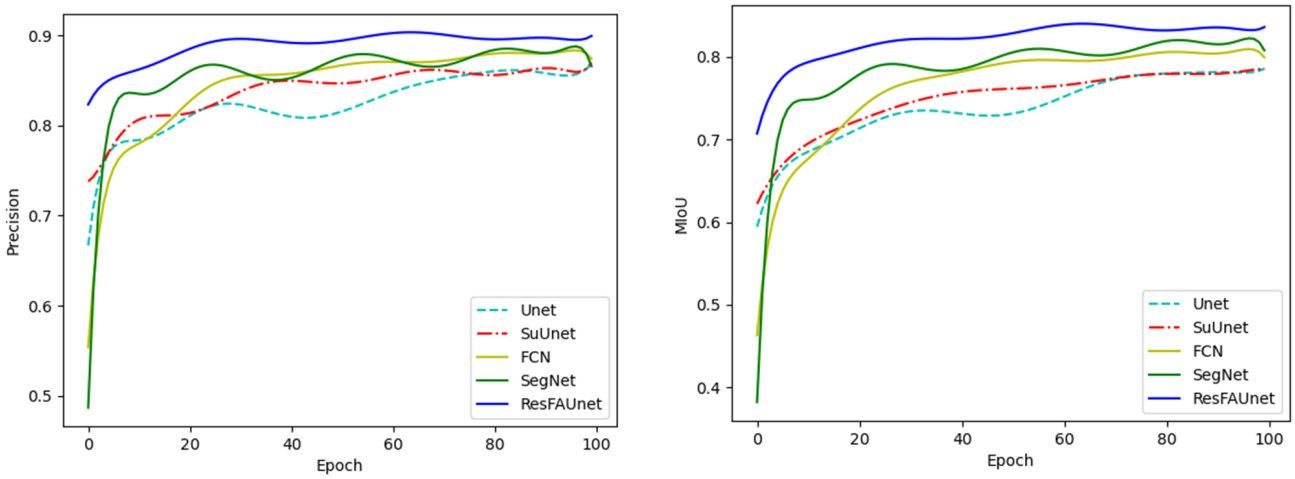
Figure 11. Comparison of test results of different models in Inria dataset.

Table 2. Test results of different network models using China dataset.

Model	*Precision*	*MIoU*	*Recall*	*F*1 Score
Unet	85.86 ± 0.36	78.08 ± 0.35	88.63 ± 0.32	87.22 ± 0.32
SuUnet	87.25 ± 0.34	78.36 ± 0.33	88.54 ± 0.30	87.14 ± 0.30
FCN	89.47 ± 0.30	81.36 ± 0.29	89.25 ± 0.27	89.36 ± 0.25
SegNet	89.94 ± 0.28	81.34 ± 0.28	88.81 ± 0.25	89.37 ± 0.25
ResFAUnet	**91.22 ± 0.27**	**84.88 ± 0.25**	**91.97 ± 0.24**	**91.59 ± 0.23**

Figure 13 shows the segmentation results in various scenarios. In the first scenario, both the building and the road are present. The Unet, SuUnet, and FCN wrongly identified roads as buildings. Although SuUnet performed better in distinguishing between roads and buildings, it failed to identify any buildings in the green-box area; only ResFAUnet accurately segmented all the buildings. In the second and fourth scenarios, large buildings were tested. The Unet, SuUnet, FCN, and SuUnet networks partially segmented the building outlines, and large buildings with cavities inside, while the proposed ResFAUnet network completely segmented the outlines of the buildings. As with the first scenario, the third scenario was used to evaluate the network's ability to distinguish between roads and buildings. Only ResFAUnet was able to segment the entire outlines of the buildings, while Unet and SuUnet did not even recognize the buildings. The fifth scenario featured a ground

area that appeared similar to the roof of a building. Only ResFAUnet successfully identified it, while all the other models falsely identified it as a building. It can be seen that the ResFAUnet network proposed in this paper can distinguish buildings from backgrounds in complex scenarios in Chinese datasets and completely segment the outlines of buildings.



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

Figure 12. Test-index curves of different models tested on China dataset. (**a**) Precision-index curve of different models tested on China dataset. (**b**) MIoU index curves of different models tested on China datasets.

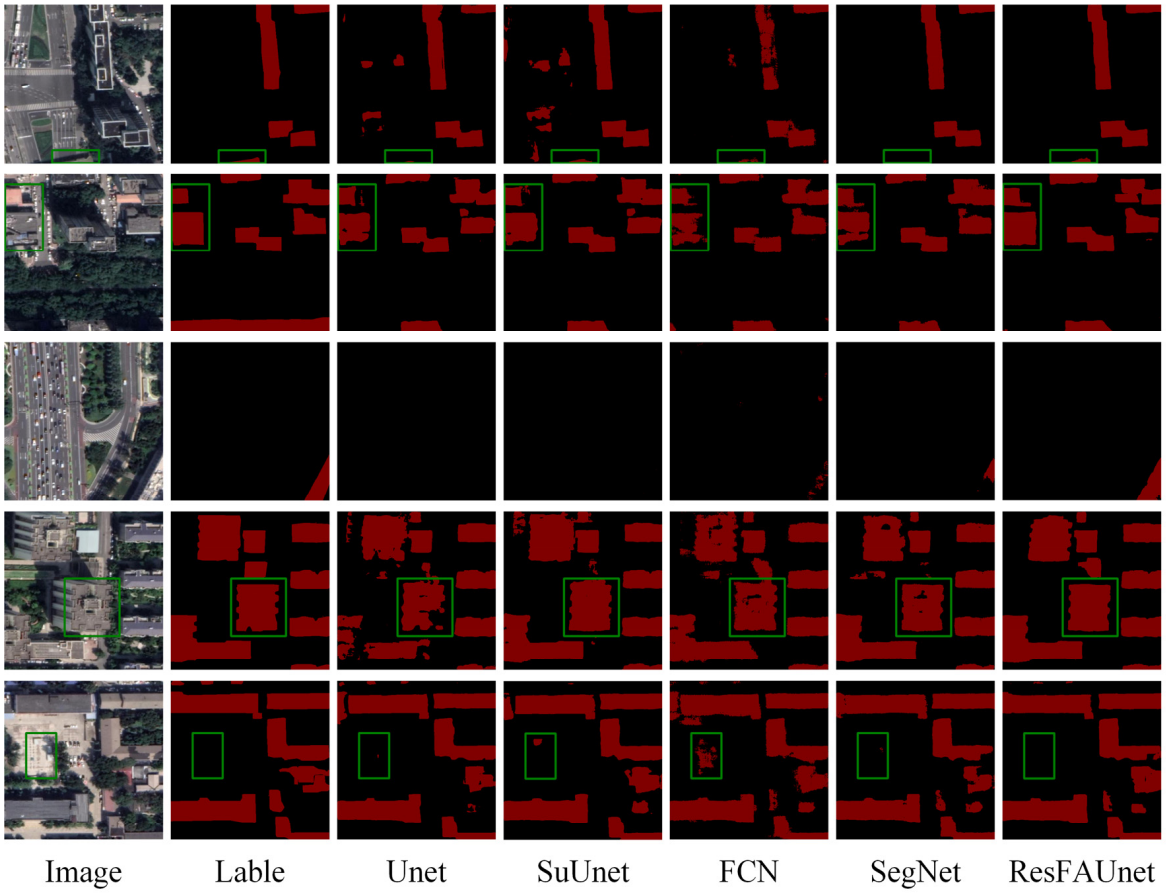


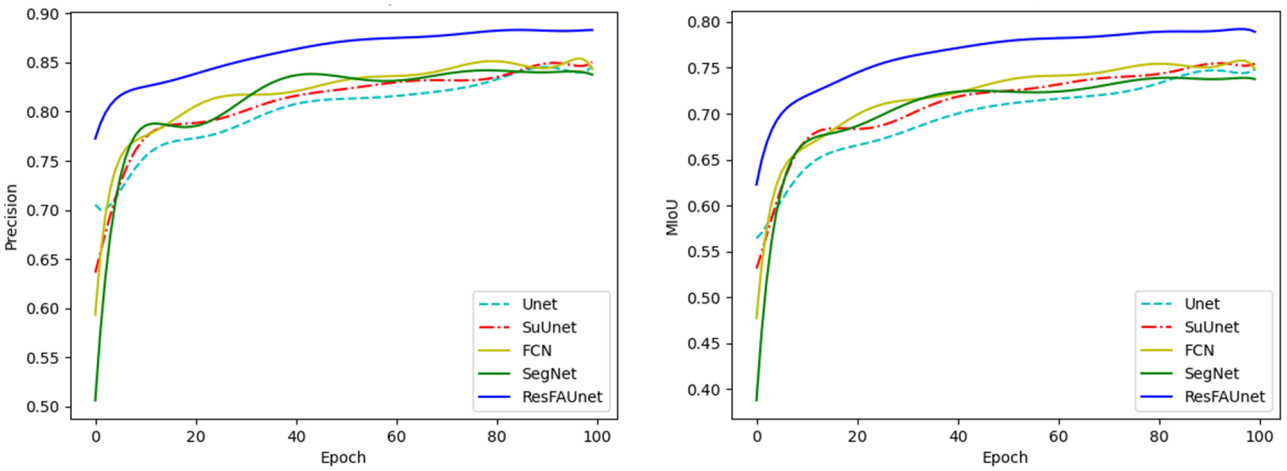
Figure 13. Comparison of test results of different models in China dataset.

3.4.3. WHU Dataset

Table 3 shows the evaluation indexes tested with different models on the WHU dataset. It can be seen from Table 3 that the precisions of all the models were below 90%, and that of the MIoU was less than 85%. This can be attributed to the fact that the number of images in the WHU dataset is small, the image resolution is low, and the environments are different. However, the proposed ResFAUnet still demonstrated acceptable results and outperformed the other networks on all the evaluation indexes. Compared with Unet, SuUnet, FCN, and SegNet, its precision was improved by 5.71%, 2.86%, 3.4%, and 2.54%, respectively. Of the comparison networks, ResFAUnet had the highest *F*1 score and the lowest standard deviation, which means that this model was the most stable. Meanwhile, the comparison of the precision and MIoU indexes of the five networks on the WHU datasets is shown in Figure 14. It can be seen that the precision index and the MIoU index of the ResFAUnet were significantly better, especially in the initial epochs.

Table 3. Test results of different network models using WHU dataset.

Model	*Precision*	*MIoU*	*Recall*	*F*1 Score
Unet	83.10 ± 0.34	73.71 ± 0.38	85.56 ± 0.32	84.31 ± 0.34
SuUnet	85.95 ± 0.32	76.02 ± 0.34	85.85 ± 0.30	85.90 ± 0.30
FCN	85.41 ± 0.28	75.54 ± 0.30	85.71 ± 0.27	85.57 ± 0.28
SegNet	86.27 ± 0.27	74.44 ± 0.29	83.71 ± 0.28	84.97 ± 0.27
ResFAUnet	**88.81 ± 0.25**	**79.11 ± 0.26**	**87.29 ± 0.25**	**88.04 ± 0.24**



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

Figure 14. Test-metrics curves of different models tested on the WHU dataset. (**a**) Curves of precision indicators tested by different models on WHU dataset. (**b**) Curves of MIoU indicators tested by different models on WHU dataset.

Figure 15 displays the test-image results in various scenarios. The first and the second scenarios featured densely populated buildings. The SegNet performed the worst in the first scenario and failed to distinguish the separate buildings accurately. As depicted in the green boxes, only ResFAUnet accurately distinguished the buildings from the ground. All the other four networks misidentified the ground as buildings. In the second scenario, the other four networks did not fully recognize the small buildings in the green boxes; Unet did not recognize any buildings at all. Only ResFAUnet completely recognized the outlines of all the buildings. The third scenario was complex, with buildings and roads both present. In this scenario, the other four networks missed some of the buildings. In the fourth and the fifth scenarios, the contour integrity when the shape of building outline was complex was compared between the models, and only ResFAUnet completely and correctly recognized the contours of the buildings.

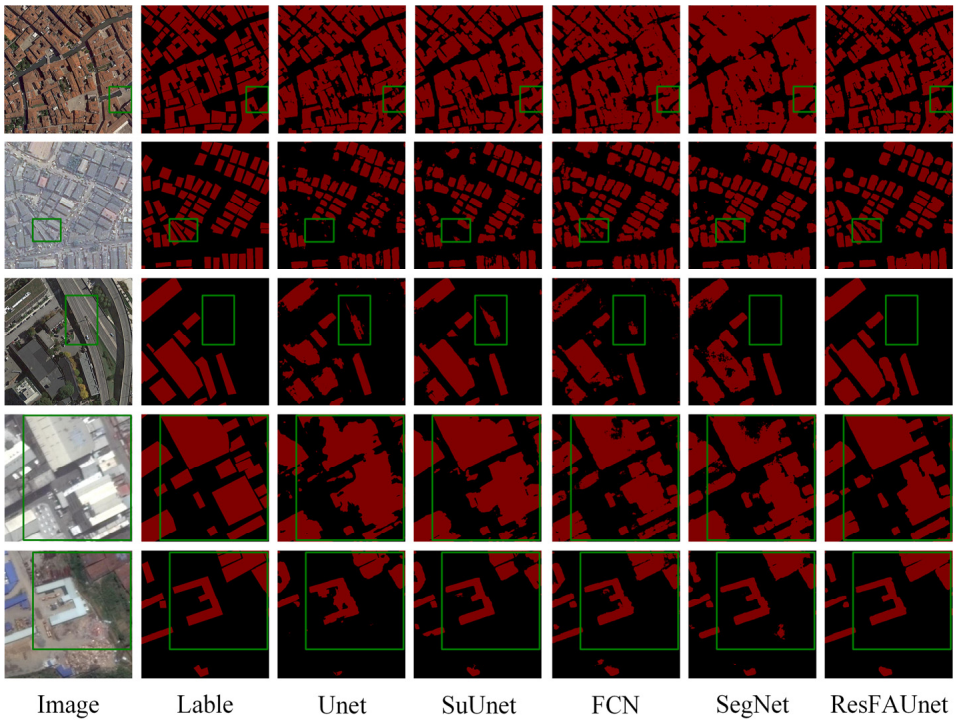


Figure 15. Comparison of test results of different models in WHU dataset.

3.5. Ablation Experiment

This section reports the ablation experiments that were conducted to verify three aspects of the ResFAUnet network. The first aspect was the feature-extraction structure of the ResNeXt101 network in the encoding part (assessed by comparing the performance of AttentionUnet, ResNeXt101 + AttentionUnet, and ResFAUnet networks on different datasets). The second aspect was the multi-scale fusion structure of the decoding part (assessed by comparing the performance of AttentionUnet, FPN + AttentionUnet and ResFAUnet networks on different datasets). The last aspect was the transfer learning (assessed by comparing the performance of AttentionUnet, ResFAUnet_nopred, and ResFAUnet networks on different datasets). The experiments were conducted using the same dataset and parameter settings.

3.5.1. Verifying the Effectiveness of ResNeXt101 Network

In this section, the effectiveness of the ResNeXt101 network as the encoding part for the feature extraction was verified by comparing AttentionUnet, ResNeXt101 + AttentionUnet, and ResFAUnet. Table 4 presents the test results of the three networks on Inria dataset, China dataset, and WHU dataset. The results indicate that after replacing the encoding part of AttentionUnet with ResNeXt101, the performance was significantly improved. The precision indexes of the three datasets increased by 1.93%, 3.88%, and 3.25%, respectively, and the other three indexes also improved. This means that the ResNeXt101 network had a stronger feature-extraction ability.

Replacing AttentionUnet's feature-extraction part with the ResNeXt101 network can improve the feature-extraction ability of the network. For a clearer visualization of this observation, four images are presented from the three datasets, in Figure 16. The first image is from the WHU dataset, and the results reveal that the segmentation by AttentionUnet left holes in the interior of the building, while the outline of the building was completely segmented by ResNeXt101 + AttentionUnet. The second and third images are from the Irina dataset. The AttentionUnet missed buildings that appeared similar to the background, while ResNeXt101 + AttentionUnet accurately recognized them. The fourth image is from the China dataset, in which AttentionUnet did not accurately distinguish the void inside

the building, but ResNeXt101 + AttentionUnet made the correct prediction. In the three datasets, ResFAUnet achieved the best segmentation effect, owing to the introduction of the feature fusion of the decoding part based on ResNeXt101 + AttentionUnet. This demonstrates not only the inheritance of the outstanding feature-extraction ability of ResNeXt101, but also that a more precise segmentation of the building outline was achieved.

Table 4. Test results of verifying the effectiveness of ResNeXt101 network.

Datasets	Models	*Precision*	*MIoU*	*Recall*	*F*1 Score
Inria dataset	AttentionUnet	89.64 ± 0.30	85.31 ± 0.32	93.37 ± 0.29	91.46 ± 0.27
	ResNeXt101 + AttentionUnet	91.57 ± 0.27	86.74 ± 0.27	93.89 ± 0.25	92.72 ± 0.24
	ResFAUnet (ours)	**92.04 ± 0.25**	**87.26 ± 0.26**	**94.02 ± 0.22**	**93.02 ± 0.23**
China dataset	AttentionUnet	86.75 ± 0.34	78.62 ± 0.32	88.36 ± 0.31	87.55 ± 0.30
	ResNeXt101 + AttentionUnet	90.63 ± 0.29	83.92 ± 0.27	91.36 ± 0.25	90.99 ± 0.24
	ResFAUnet (ours)	**91.22 ± 0.27**	**84.88 ± 0.25**	**91.97 ± 0.24**	**91.59 ± 0.23**
WHU dataset	AttentionUnet	84.39 ± 0.32	73.67 ± 0.34	84.50 ± 0.32	84.36 ± 0.30
	ResNeXt101 + AttentionUnet	87.64 ± 0.26	78.81 ± 0.28	86.93 ± 0.25	87.28 ± 0.25
	ResFAUnet (ours)	**88.81 ± 0.25**	**79.11 ± 0.26**	**87.29 ± 0.25**	**88.04 ± 0.24**

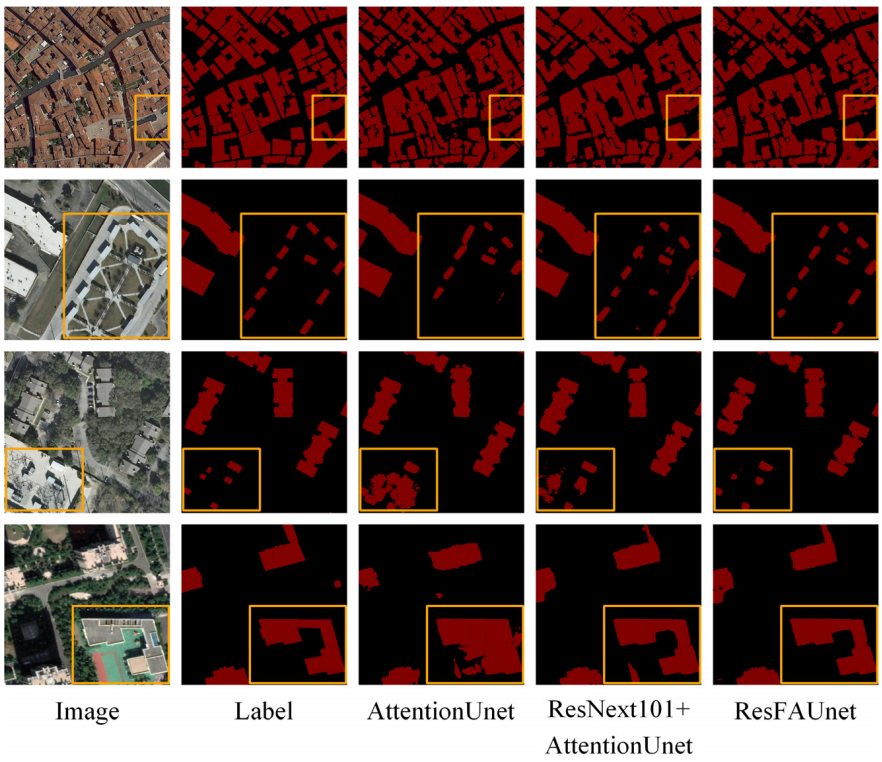


Figure 16. Comparison of samples with and without ResNeXt101 prediction results (WHU dataset, Irina dataset, China dataset (the last two images), respectively).

3.5.2. Verifying the Partial Feature Fusion in Decoding

In this section, the effectiveness of the partial feature fusion in the decoding part is verified by comparing AttentionUnet, feature fusion + AttentionUnet (FPN + AttentionUnet) and ResFAUnet. Table 5 shows the test results for the three networks on the Inria dataset, China dataset, and WHU dataset. It can be seen that the network with feature fusion in decoding (FPN + AttentionUnet) had small improvements in all indexes compared with AttentionUnet network, and that the precision indexes of the three datasets improved by 0.71%, 1.91%, and 0.89%, respectively. This means that the multi-scale feature fusion can effectively improve the feature-extraction ability of the network.

Table 5. Test results of verifying the Partial Feature Fusion.

Datasets	Models	*Precision*	*MIoU*	*Recall*	*F1* Score
Inria dataset	AttentionUnet	89.64 ± 0.30	85.31 ± 0.32	93.37 ± 0.29	91.46 ± 0.27
	FPN + AttentionUnet	90.35 ± 0.32	85.52 ± 0.30	93.72 ± 0.28	92.01 ± 0.27
	ResFAUnet (ours)	**92.04 ± 0.26**	**87.26 ± 0.26**	**94.02 ± 0.22**	**93.02 ± 0.23**
China dataset	AttentionUnet	86.75 ± 0.34	78.62 ± 0.32	88.36 ± 0.31	87.55 ± 0.30
	FPN + AttentionUnet	88.66 ± 0.32	79.64 ± 0.30	87.83 ± 0.29	88.24 ± 0.28
	ResFAUnet (ours)	**91.22 ± 0.27**	**84.88 ± 0.25**	**91.97 ± 0.24**	**91.59 ± 0.23**
WHU dataset	AttentionUnet	84.39 ± 0.32	73.67 ± 0.34	84.50 ± 0.32	84.36 ± 0.30
	FPN + AttentionUnet	85.28 ± 0.30	73.94 ± 0.32	83.45 ± 030	84.44 ± 0.29
	ResFAUnet (ours)	**88.81 ± 0.25**	**79.11 ± 0.26**	**87.29 ± 0.25**	**88.04 ± 0.24**

For a better visualization, one image was selected from each of the three datasets, as shown in Figure 17. In the WHU dataset and China dataset, AttentionUnet missed some of the buildings, while the FPN + AttentionUnet achieved better segmentation results. In the Irina dataset, AttentionUnet incorrectly identified land as buildings, while FPN + AttentionUnet effectively distinguished the ground from buildings. It can be seen that partial feature fusion in decoding can reduce building omissions and false identification to some extent, as well as improving the building-segmentation performance of the network.

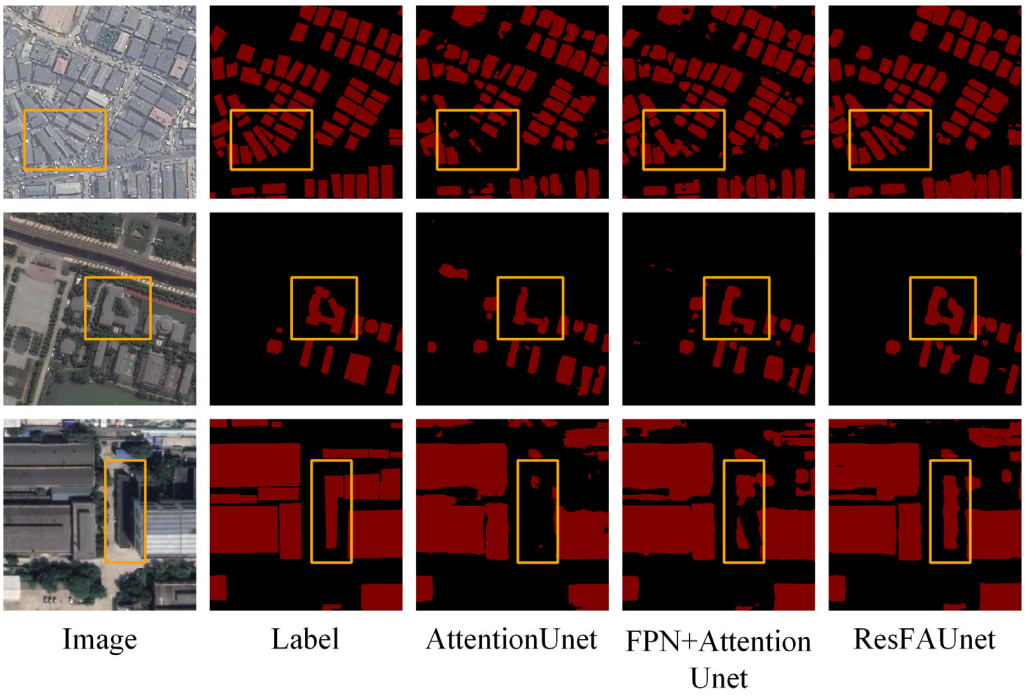


Figure 17. Comparison of samples with or without prediction results of feature-fusion structure (WHU dataset, Irina dataset, China dataset, respectively).

3.5.3. Verifying Transfer Learning

In this section, the effectiveness of transfer learning is verified by comparing AttentionUnet, ResFAUnet_nopred, and ResFAUnet. Specifically, AttentionUnet and ResFAUnet_nopred do not use transfer learning. Table 6 shows the test results for the three networks on the three datasets. The results of ResFAUnet_nopred were all lower than those of AttentionUnet, while the all indexes of ResFAUnet were higher than those of AttentionUnet. Compared to ResFAUnet_nopred, the precision index of ResFAUnet increased by 4.5%, 5.29%, and 8.16% in three datasets, respectively. These results demonstrate that the

introduction of transfer learning can effectively improve network performance in small sample-learning tasks.

Table 6. Test results of verifying transfer learning.

Datasets	Models	*Precision*	*MIoU*	*Recall*	*F*1 Score
Inria dataset	AttentionUnet	89.64 ± 0.30	85.31 ± 0.32	93.37 ± 0.29	91.46 ± 0.27
	ResFAUnet (not pretrained)	87.54 ± 0.32	81.51 ± 0.34	91.49 ± 0.31	89.48 ± 0.29
	ResFAUnet (ours)	**92.04 ± 0.25**	**87.26 ± 0.26**	**94.02 ± 0.22**	**93.02 ± 0.23**
China dataset	AttentionUnet	86.75 ± 0.34	78.62 ± 0.32	88.36 ± 0.31	87.55 ± 0.30
	ResFAUnet (not pretrained)	85.93 ± 0.35	74.88 ± 0.34	84.13 ± 0.32	85.02 ± 0.30
	ResFAUnet (ours)	**91.22 ± 0.27**	**84.88 ± 0.25**	**91.97 ± 0.24**	**91.59 ± 0.23**
WHU dataset	AttentionUnet	84.39 ± 0.32	73.67 ± 0.34	84.50 ± 0.32	84.36 ± 0.30
	ResFAUnet (not pretrained)	80.65 ± 0.34	68.09 ± 0.33	79.7 ± 0.32	80.17 ± 0.32
	ResFAUnet (ours)	**88.81 ± 0.25**	**79.11 ± 0.26**	**87.29 ± 0.25**	**88.04 ± 0.24**

For a clearer visualization, four images were selected from Inria dataset, China dataset, and WHU dataset, and shown in Figure 18. The performance of ResFAUnet_nopred in these four scenarios was the worst, even lower than AttentionUnet's performance, with a large number of incomplete building-contour segmentations. This was mainly because the dataset size was small, making it difficult for ResFAUnet_nopred to learn a good weight. The ResFAUnet achieved the best performance with the same dataset and parameter settings. Thus, transfer learning can help the network gain better training results with small sample sizes.

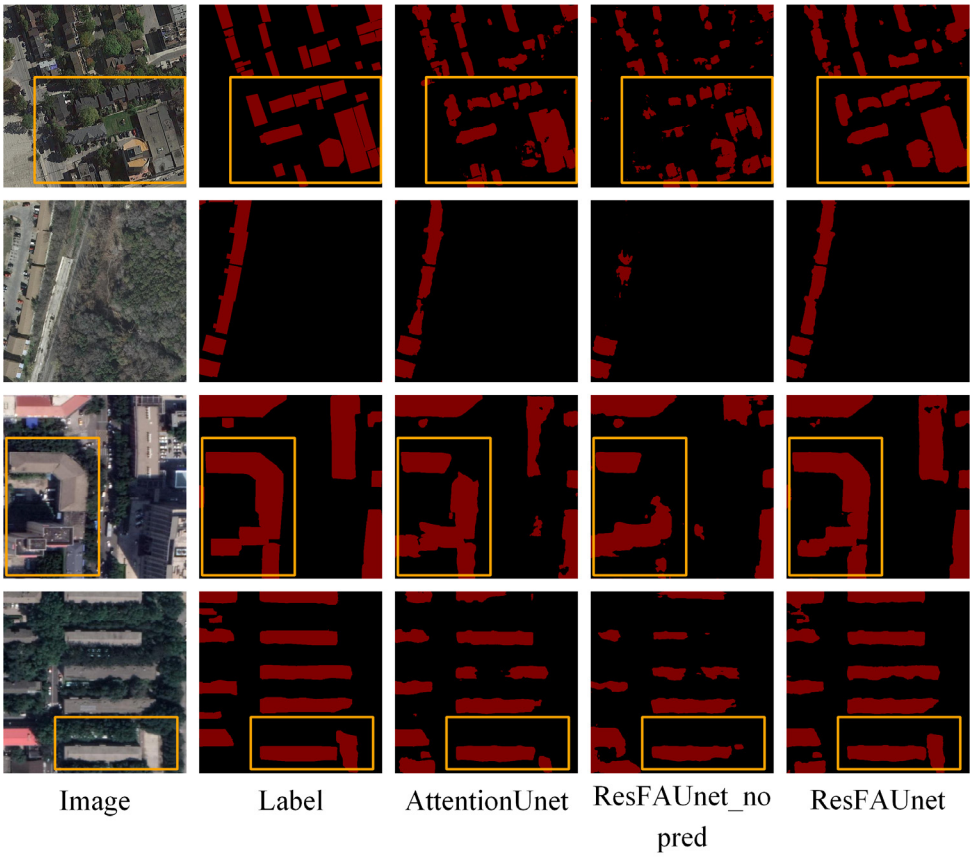


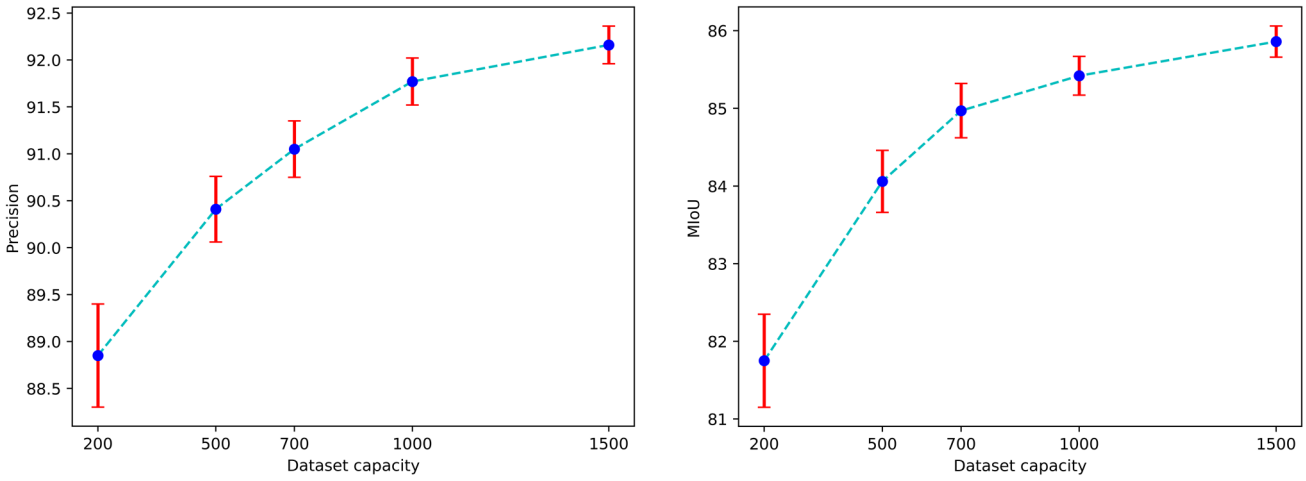
Figure 18. Comparison of samples with or without transfer-learning prediction results (WHU dataset, Irina dataset, China dataset (the last two images), respectively).

4. Discussion

In this section, multiple aspects of the ResFAUnet network are discussed to study the influences of other parameters on the network's performance during the training process. The specific arrangement is as follows. In Section 4.1, the influence of the dataset capacity on ResFAUnet network performance when changing the capacity of the dataset is studied. In Section 4.2, the influence of the image scale on the performance of the ResFAUnet network when changing the size of the input image is studied.

4.1. Dataset Capacity

The dataset is the only source of data in a neural network. The size of the dataset directly affects the feature-learning ability of the neural network. This subsection reports the use of datasets with different capacities to train the ResFAUnet network, and the experimental results obtained using datasets with different capacities are analyzed to study the relationship between the ResFAUnet network's performance and the dataset capacity. In the experiment, the China dataset was selected, and its capacity was set as 200, 500, 700, 1000, and 1500 images; the learning rate was set as 0.01, the image size was set as 512 × 512, and all the other training parameters were kept unchanged to study the effect of the capacity on the network's performance. Figure 19 shows the variation curves between the precision and the dataset capacity, and between the MIoU and the dataset capacity. It can be seen that with the increase in the dataset capacity, the precision index and the MIoU index of the ResFAUnet network both increased as logarithmic functions. When the dataset size was between 200 and 500 images, the precision index and MIoU increased the fastest. When the dataset increased gradually, the two indexes increased slowly, but the overall trend was still upward. Furthermore, with the increase in the dataset capacity, the errors in the precision and MIoU indexes gradually decreased, and the network's performance becomes more stable.



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

Figure 19. Test-metrics curves of ResFAUnet for different capacities of China dataset (the capacities of datasets were 200, 500, 700, 1000, and 1500 images respectively). (**a**) Curves of precision indicators. (**b**) Curves of MIoU indicators.

4.2. Image Size

This subsection reports how the relationship between the image size and the ResFAUnet network performance was investigated by adjusting the size of the input images. In the experiment, the China dataset and the Inria dataset were selected, and three different image sizes were adopted for the experiment. The image sizes were set as 224 × 224, 384 × 384, and 512 × 512, the learning rate was set as 0.01, and the other training parameters were consistent. The experimental results are shown in Table 7. From the results, it

can be observed that the larger the input size of the image, the better the performance of the network. In the China dataset, the result for the image scale of 512 was significantly improved compared with that of 224. The *Precision*, *MIoU*, *Recall*, and *F*1 score were improved by 3.5%, 7.06%, 5.02%, and 4.45%, respectively. In the Inria dataset, compared with the training results with the image scale of 224, the four evaluation indexes with the scale of 512 increased by 4.48%, 6.05%, 3.74%, and 3.92% respectively. Furthermore, with the increase in the image size, the standard deviations of the four evaluation indexes (*Precision*, *MIoU*, *Recall* and *F*1 score) became smaller, and the network became more stable.

Table 7. Test results of different input-image sizes in the China dataset and Inria dataset.

Datasets	Input_Size	*Precision*	*MIoU*	*Recall*	*F*1 Score
China dataset	(224, 224, 3)	87.43 ± 0.53	77.15 ± 0.48	86.09 ± 0.43	86.57 ± 0.46
	(384, 384, 3)	89.95 ± 0.39	83.82 ± 0.33	90.46 ± 0.32	90.41 ± 0.30
	(512, 512, 3)	**90.93 ± 0.32**	**84.21 ± 0.29**	**91.11 ± 0.27**	**91.02 ± 0.25**
Inria dataset	(224, 224, 3)	87.54 ± 0.48	81.16 ± 0.48	90.93 ± 0.43	89.20 ± 0.45
	(384, 384, 3)	91.56 ± 0.35	86.88 ± 0.33	93.92 ± 0.30	92.81 ± 0.35
	(512, 512, 3)	**92.02 ± 0.30**	**87.21 ± 0.28**	**94.67 ± 0.30**	**93.12 ± 0.30**

5. Conclusions

In this paper, a new end-to-end deep-learning network, ResFAUnet, is proposed to solve the problem of the low accuracy of building-information extraction with of small sample sizes. Based on AttentionUnet, the proposed ResFAUnet preserves the attentional mechanism using the skip connection to integrate the shallow features of the encoding part and the deep features of the decoding part. In the encoding part, the grouped convolutional network, ResNeXt101, is used for feature extraction, and the pre-trained weights on the ImageNet dataset are used for transfer learning. This improves the training accuracy of the network with small sample sizes. In the decoding part, the multi-chain up-sampling feature-fusion method is adopted to obtain more abundant feature information and improve the accuracy of the extraction of the building information. In addition, comparative experiments on the Inria dataset, China dataset, and WHU dataset were carried out, and the results showed that the building-information-extraction accuracy of ResFAUnet outperformed the others.

Author Contributions: Conceptualization, H.Z.; methodology, H.Z.; supervision, X.X., Y.R. and Z.T.; visualization, Y.R. and Z.T.; writing—original draft, H.Z.; writing—review and editing, X.X. and Z.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by National Key Research and Development Program of China (grant nos. 2022YFB4201004, 2022YFB4201004-2), Jiangsu Key R&D Program (grant no. BE2020082-1), the Fundamental Research Funds for the Central Universities (grant no. B220202023).

Data Availability Statement: Not applicable.

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

References

1. Sishodia, R.P.; Ray, R.L.; Singh, S.K. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sens.* **2020**, *12*, 3136. [CrossRef]
2. Guo, M.; Li, J.; Sheng, C.; Xu, J.; Wu, L. A Review of Wetland Remote Sensing. *Sensors* **2017**, *17*, 777. [CrossRef] [PubMed]
3. Nichol, J.E.; Shaker, A.; Wong, M.S. Application of high-resolution stereo satellite images to detailed landslide hazard assessment. *Geomorphology* **2006**, *76*, 68–75. [CrossRef]
4. Van der Meer, F.D.; Van der Werff, H.M.; Van Ruitenbeek, F.J.; Hecker, C.A.; Bakker, W.H.; Noomen, M.F.; van der Meijde, M.; Carranza, E.J.M.; de Smeth, J.B.; Woldai, T. Multi-and hyperspectral geologic remote sensing: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *14*, 112–128. [CrossRef]
5. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-of-the-Art Review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]

6. Shi, W.; Zhang, M.; Ke, H.; Fang, X.; Zhan, Z.; Chen, S. Landslide recognition by deep convolutional neural network and change detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4654–4672. [CrossRef]
7. Irvin, R.B.; McKeown, D.M. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1564–1575. [CrossRef]
8. Lee, T.; Kim, T. Automatic building height extraction by volumetric shadow analysis of monoscopic imagery. *Int. J. Remote Sens.* **2013**, *34*, 5834–5850. [CrossRef]
9. Levitt, S.; Aghdasi, F. Texture measures for building recognition in aerial photographs. In Proceedings of the 1997 South African Symposium on Communications and Signal Processing, Grahamstown, South Africa, 9–10 September 1997; pp. 75–80.
10. Lin, C.; Nevatia, R. Building detection and description from a single intensity image. *Comput. Vis. Image Underst.* **1998**, *72*, 101–121. [CrossRef]
11. Li, H.; Xiang, J.; Liu, J. An automatic building extraction method from high resolution satellite image. In Proceedings of the 31st Chinese Control Conference, Hefei, China, 25–27 July 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 4884–4889.
12. Xu, G.; Gao, L.; Yuan, X. Building extraction from aerial imagery based on the principle of confrontation and priori knowledge. In Proceedings of the 2009 Second International Conference on Computer and Electrical Engineering, Dubai, United Arab Emirates, 28–30 December 2009; IEEE: Piscataway, NJ, USA, 2009; Volume 1, pp. 363–366.
13. Gavankar, N.L.; Ghosh, S.K. Automatic building footprint extraction from high-resolution satellite image using mathematical morphology. *Eur. J. Remote Sens.* **2018**, *51*, 182–193. [CrossRef]
14. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
19. 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] [PubMed]
20. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
22. Schlemper, J.; Ozan, O.; Michiel, S.; Mattias, H.; Bernhard, K.; Ben, G.; Daniel, R. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef] [PubMed]
23. He, N.; Fang, L.; Plaza, A. Hybrid first and second order attention Unet for building segmentation in remote sensing images. *Sci. China Inf. Sci.* **2020**, *63*, 140305. [CrossRef]
24. Shi, X.; Huang, H.; Pu, C.; Yang, Y.; Xue, J. CSA-UNet: Channel-Spatial Attention-Based Encoder–Decoder Network for Rural Blue-Roofed Building Extraction from UAV Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3197319. [CrossRef]
25. Shunping, J.I.; Shiqing, W.E.I. Building extraction via convolutional neural networks from an open remote sensing building dataset. *Acta Geod. Cartogr. Sin.* **2019**, *48*, 448.
26. Delibasoglu, I.; Cetin, M. Improved U-Nets with inception blocks for building detection. *J. Appl. Remote Sens.* **2020**, *14*, 044512. [CrossRef]
27. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810. [CrossRef]
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
29. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Multi-Object Segmentation in Complex Urban Scenes from High-Resolution Remote Sensing Data. *Remote Sens.* **2021**, *13*, 3710. [CrossRef]
30. Khan, N.M.; Abraham, N.; Hon, M. Transfer learning with intelligent training data selection for prediction of Alzheimer's disease. *IEEE Access* **2019**, *7*, 72726–72735. [CrossRef]
31. Gu, Y.; Ge, Z.; Bonnington, C.P.; Zhou, J. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 1379–1393. [CrossRef] [PubMed]
32. Wollmann, T.; Eijkman, C.S.; Rohr, K. Adversarial domain adaptation to improve automatic breast cancer grading in lymph nodes. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 582–585.
33. Liu, Z.; Zhu, Z.; Zheng, S.; Liu, Y.; Zhou, J.; Zhao, Y. Margin preserving self-paced contrastive learning towards domain adaptation for medical image segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 638–647. [CrossRef]

34. Yao, K.; Su, Z.; Huang, K.; Sun, J.; Coenen, F. A novel 3D unsupervised domain adaptation framework for cross-modality medical image segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4976–4986. [CrossRef]
35. Liu, Y.; Zhang, W.; Wang, J. Source-free domain adaptation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1215–1224.
36. Stan, S.; Rostami, M. Domain Adaptation for the Segmentation of Confidential Medical Images. *arXiv* **2021**, arXiv:2101.00522.
37. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3226–3229.
38. Fang, F.; Wu, K.; Liu, Y.; Li, S.; Wan, B.; Chen, Y.; Zheng, D. A Coarse-to-Fine Contour Optimization Network for Extracting Building Instances from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2021**, *13*, 3814. [CrossRef]
39. Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]