remote sensing

MDPI

Article

AFL-Net: Attentional Feature Learning Network for Building Extraction from Remote Sensing Images

Yue Qiu [1], Fang Wu *, Haizhong Qian, Renjian Zhai, Xianyong Gong, Jichong Yin, Chengyi Liu and Andong Wang

Institute of Geospatial Information, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
* Correspondence: wufang_630@126.com

Abstract: Convolutional neural networks (CNNs) perform well in tasks of segmenting buildings from remote sensing images. However, the intraclass heterogeneity of buildings is high in images, while the interclass homogeneity between buildings and other nonbuilding objects is low. This leads to an inaccurate distinction between buildings and complex backgrounds. To overcome this challenge, we propose an Attentional Feature Learning Network (AFL-Net) that can accurately extract buildings from remote sensing images. We designed an attentional multiscale feature fusion (AMFF) module and a shape feature refinement (SFR) module to improve building recognition accuracy in complex environments. The AMFF module adaptively adjusts the weights of multi-scale features through the attention mechanism, which enhances the global perception and ensures the integrity of building segmentation results. The SFR module captures the shape features of the buildings, which enhances the network capability for identifying the area between building edges and surrounding nonbuilding objects and reduces the over-segmentation of buildings. An ablation study was conducted with both qualitative and quantitative analyses, verifying the effectiveness of the AMFF and SFR modules. The proposed AFL-Net achieved 91.37, 82.10, 73.27, and 79.81% intersection over union (IoU) values on the WHU Building Aerial Imagery, Inria Aerial Image Labeling, Massachusetts Buildings, and Building Instances of Typical Cities in China datasets, respectively. Thus, the AFL-Net offers the prospect of application for successful extraction of buildings from remote sensing images.

Keywords: building extraction; remote sensing; image segmentation; feature fusion; feature refinement

Citation: Qiu, Y.; Wu, F.; Qian, H.; Zhai, R.; Gong, X.; Yin, J.; Liu, C.; Wang, A. AFL-Net: Attentional Feature Learning Network for Building Extraction from Remote Sensing Images. *Remote Sens.* **2023**, *15*, 95. https://doi.org/10.3390/rs15010095

Academic Editors: Khaled Rabie, Pascal Lorenz, Muhammad Asghar Khan, Syed Agha Hassnain Mohsan and Muhammad Shafiq

Received: 26 November 2022
Revised: 17 December 2022
Accepted: 21 December 2022
Published: 24 December 2022

1. Introduction

Buildings are the main spaces where humans gather. The extraction of buildings from remote sensing images has been applied widely in fields such as digital city construction [1,2], urban planning [3,4], change detection [5–7], damage assessment [8,9], and digital mapping [10].

Before the development of building extraction methods based on the building features in images, buildings were manually and individually labelled, making such extraction a time-consuming and labor-intensive task. Conventional building extraction methods include those based on feature detection [11,12], regional segmentation [13,14], and combined auxiliary information [4,15]. These methods mainly rely on experts to design an algorithm with appropriate parameter thresholds by studying basic features of buildings, such as the light spectra, shapes, texture, and shades. However, such methods cannot utilize the deep-level information contained in images, implying limited accuracy in the building extraction results and a low degree of automation.

The rapid advance of computer vision has led to the development of convolutional neural networks (CNNs) that show excellent performance in semantic segmentation tasks such as car navigation [16], scene analysis [17], and medical image segmentation [18]. In

the training process, a CNN generates feature maps of various resolutions and levels by stacking the convolutional layers. When a training dataset contains abundant data, the CNN can stably obtain rich amounts of feature information. Accordingly, ample study has been conducted on utilizing CNNs to extract buildings from remote sensing images. Despite such abundant research, two unresolved problems remain.

(1) With feature maps, the features obtained from images are neither rich enough nor adequately utilized, resulting in building over- or under-segmentation. In generating feature maps, conventional CNNs usually reduce and subsequently increase their resolution [19]. As the downsampling process reduces the resolution of the feature map, a certain amount of detailed information is lost [20]. This can cause small buildings and buildings with rich feature information on their roofs to be neglected in the extraction process. Several studies [21,22] have attempted to recover some of the lost detailed information through fusion with low-level feature maps during the upsampling process. The results showed that, despite abundant redundant information being available for fusion with the low-level feature maps, a limited amount of such information could be utilized [23].

A popular method [24,25] for feature utilization is to directly concatenate the generated feature maps after unifying their sizes. This newly generated feature map is a simple fusion of all feature information. As the weights of each piece of feature information are not adjusted adaptively according to the segmentation target, it is difficult to selectively retain effective information and filter out noise. Consequently, the feature information cannot be fully utilized. In building extraction tasks, this aspect causes difficulty in distinguishing buildings from their surrounding background pixels, leading to the false detection of buildings.

(2) With the target, the shapes and features of buildings are not maintained adequately, making it difficult to distinguish buildings from their surrounding ground-based objects. In conventional CNNs, the size of the receptive field of the generated feature map is limited [26]. Moreover, the range of the receptive field is regular. When the receptive field is too small or the feature map contains irregularly shaped buildings, it is difficult to capture the complete shape feature of an entire building for identifying a round-shaped shed or irregularly shaped buildings, such as art museums. To expand the receptive field, various feature pyramid modules have been designed based on dilated convolution or large-scale convolutional kernels, such as the receptive field block [27] and atrous spatial pyramid pooling module [28]. Using these modules for building extraction tasks [29,30] has improved the ability to extract buildings of various sizes. Nevertheless, it remains difficult to accurately identify irregularly shaped buildings. Buildings in urban areas are arranged in a certain manner with regular orientation, but buildings in suburban areas are arranged differently, distributed randomly, and may even overlap each other. Moreover, diverse types of ground-based objects are located around buildings in suburban areas, including trees, shrubs, bare ground, and concrete pavements. Existing algorithms are usually unable to consider the spatial relationship between buildings and the surrounding ground-based objects, leading to the false or missed detection of buildings.

To solve these problems, we propose an attentional feature learning CNN for building extraction from remote sensing images. In particular, our proposed network was able to solve unfocused feature fusion and incomplete building shape retention. The main contributions are as follows.

1. We designed an attentional multiscale feature fusion (AMFF) module to adjust the weight of each piece of feature information during the feature fusion process. Therefore, more effective information conducive to the separation of buildings from backgrounds was retained and irrelevant noise information was filtered out. Employing our module ensured highly efficient use of the feature information and enhanced the integrity of building segmentation.

2. We designed a shape feature refinement (SFR) module that featured a receptive field that is not limited to a regular area and an expanded receptive field range. Therefore, the network adaptively learned the shape features of buildings and reduced interference

from complex environments at the interface between buildings and backgrounds, thereby maintaining the shape patterns of the extracted buildings.

3. We conducted a comparison of the results from the application of our model on four benchmark datasets. This indicated that the performance of the proposed AFL-Net was the most advanced of the compared models. In addition, we conducted an ablation study to verify the effectiveness of the proposed AMFF and SFR modules.

2. Methodology

2.1. AFL-Net Architecture

The proposed AFL-Net followed the 'encoder-decoder' structure shown in Figure 1.

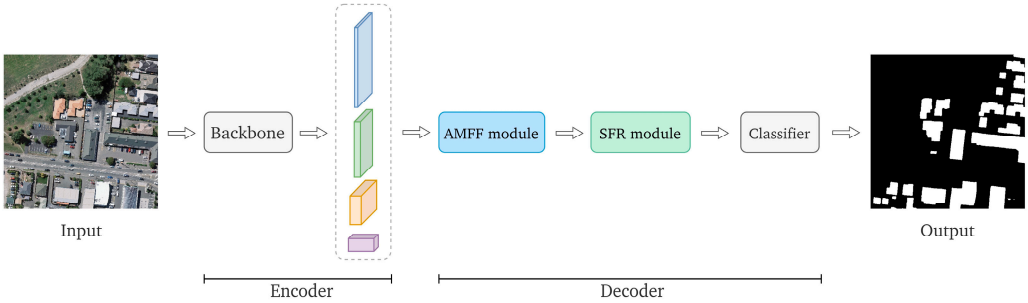

Figure 1. Schematic diagram of the AFL-Net framework. The encoder extracts the features through the backbone, outputting four feature maps at different scales. The decoder fuses the feature maps via the AMFF module, optimizes the building shape features via the SFR module, and outputs the building segmentation mask after the classifier.

The encoder captured the features from the input images. The decoder separated the buildings from the background according to the features acquired, outputting the segmentation masks of the buildings, and resizing the feature maps to those of the input images. Using HRNetV1 [31] as the feature extraction backbone during the encoding process had two major advantages. The first is that high-resolution features were maintained throughout the network structure, which was conducive to accurately identifying small buildings in the images. The second was that feature maps with different resolutions were fused together, which was conducive to information exchange, and substantiated the feature information contained in the feature map at each resolution. For each input image, the backbone can output four feature maps. The feature maps underwent multiscale feature fusion via the AMFF module during the decoding process. The fused feature map was subsequently input into the SFR module for refined learning of the shape features. Finally, the 1×1 convolution in the classifier was used to adjust the number of channels of the input images to match the number of classes, and to upsample the results bilinearly to the size of the input images to obtain the semantic segmentation results of the buildings.

2.2. AMFF Module

A conventionally used feature fusion method stacks the feature maps via a concatenating operation or adds the feature maps directly. However, concatenating and adding procedures are both simple operations to fuse features, making it difficult to selectively utilize the features at each scale. Inspired by the polarized self-attention mechanism [32], a self-attention module was designed and embedded in the proposed AMFF module to learn the importance of spatial features at various scales and correlate channel features, adaptively and selectively retaining effective features and eliminating useless features. The AMFF structure is shown in Figure 2.

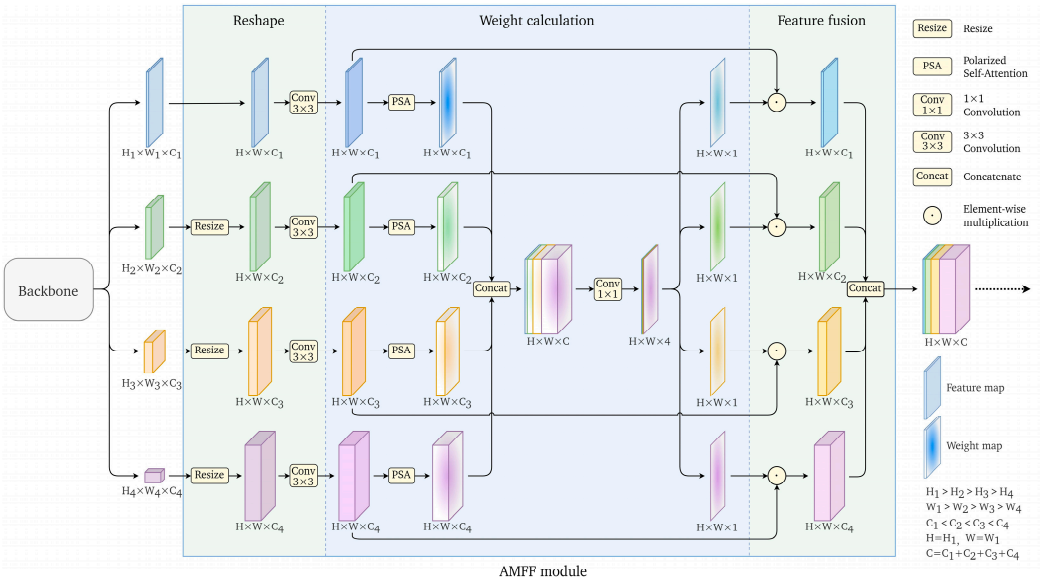

Figure 2. Schematic diagram of the AMFF module structure. The attention mechanism facilitates adaptive adjustment of the weights of the distinctive features.

The AMFF module mainly comprises three steps of reshaping, weight calculation, and feature fusion. In the first step, the feature maps x_i (i∈{1,2,3,4}) were resized to the same size using nearest neighbor interpolation to obtain x_i'. The 3 × 3 convolution (C_3) was subsequently used on x_i' to enhance the feature information and obtain $C_3(x_i')$. In the second step, the attention mechanism was used to calculate the weights of $C_3(x_i')$. The four weighted feature maps were concatenated, and 1 × 1 convolution was used to adjust the number of channels, obtaining an output weighted feature map w_i (i∈{1,2,3,4}). In the third step, $C_3(x_i')$ and w_i were multiplied pixel by pixel to obtain a weight-adjusted feature map. Finally, the weight-adjusted feature maps were concatenated to obtain the final feature map after fusion.

In the second step above, the polarized self-attention mechanism was inspired by the characteristic of polarized lenses that filter light in random directions, only allowing light orthogonal to the transverse direction to pass through. A polarization filtering mechanism was established in the attention calculation. In other words, the spatial dimensional features along the orthogonal direction were folded when calculating channel-only attention, whereas the channel dimensional features along the orthogonal direction were folded when calculating spatial-only attention. In this way, a high resolution was maintained at the attention calculation dimension, reducing information loss. The attention mechanism used in the AMFF module is shown in Figure 3.

As calculated using Function (1), $X \in R^{H \times W \times C}$ is defined as the feature tensor of a sample, where H, W, and C are the height, width, and number of channels of X, respectively. The attention weights output by the attention mechanism $W_A(X)$ are the sum of the channel-only attention weight $W_c(X)$ and spatial-only attention weight $W_s(X)$.

$$W_A(X) = W_c(X) + W_s(X), \tag{1}$$

where the channel-only attention weight $W_c(X)$ is calculated by Function (2):

$$W_c(X) = F_{Si}(L(C_3(R_1(C_1(X)) \times F_{So}(R_2(C_2(X)))))), \tag{2}$$

and the spatial-only attention weight $W_s(X)$ is calculated by Function (3):

$$W_s(X) = F_{Si}(R_3(F_{So}(R_1(G(C_2(X)))) \times R_2(C_1(X)))). \tag{3}$$

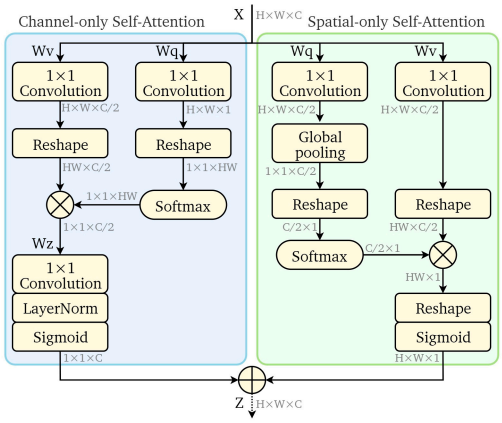

Figure 3. Structure diagram of attention mechanism in the AMFF module. Attentional weights are calculated at the channel and spatial dimensions.

In the above functions, F_{So} represents the softmax operation, which increases the dynamic range of the attention through normalization; F_{Si} represents the sigmoid operation, and the softmax–sigmoid combination is a probability distribution function that results in nonlinear mapping that fully utilizes the high-resolution information stored in the attention branch; L represents the layer normalization operation; C_i represents the i × i convolution operation; R_i represents the tensor reshape operation; G represents the global average pooling operation; and × represents the cross product.

The weights acquired from the various feature maps were spliced and fused to retain the weights of specific features in the same feature map and the importance of the weights of the various feature maps. This factor allowed weight distribution of the distinct pieces of feature information from the same or different feature maps in the third step (feature fusion). In this way, features conducive to building segmentation were retained adaptively and selectively, while unimportant features were filtered out.

2.3. SFR Module

To refine building extraction from high-resolution images, a shape feature refinement module was designed and added before the classifier to allow the extraction of irregularly shaped buildings, and to effectively separate the buildings from the background. The structure of the SFR module is shown in Figure 4.

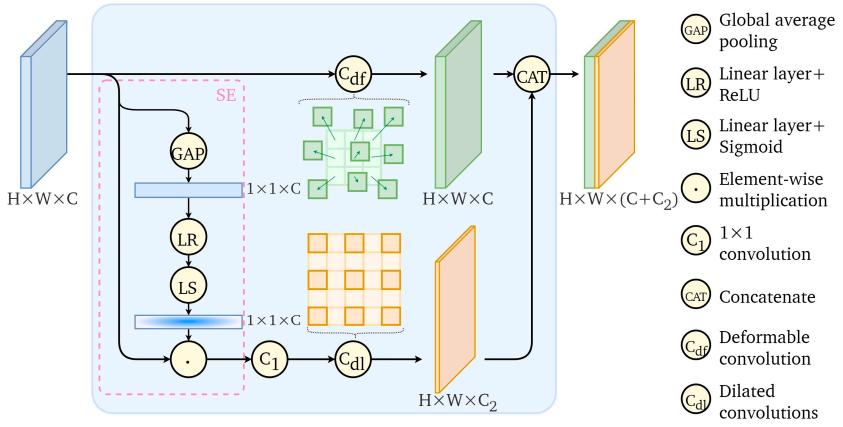

Figure 4. Structure diagram of the SFR module. Deformable convolution and dilated convolution expand the receptive field.

The SFR module contains two branches. Branch one is the deformable convolution [33] and Branch two contains the squeeze-and-excitation channel attention mechanism [34]

and the dilated convolution [35] in series. The two branches were spliced together to output a new feature map. The sampling position of the deformable convolution in Branch one was adjusted adaptively according to the building shape, enabling acquisition of the shape features of the buildings, and enhancing the capability to identify buildings with irregular shapes. Branch two connected the rectified linear unit (ReLU) activation function through a global average pooling and fully connected layers. It subsequently performed the sigmoid operation after another fully connected layer to obtain the channel attention weights. The channel attention weights were multiplied by the original input feature map to allow adaptive feature correction for obtaining the feature map after adaptive learning of the fused features. The dilated convolution expanded the receptive field. In the proposed structure, two dilated convolutions with dilation rates of three and five were used to expand the receptive field and capture contextual relationships. Accordingly, the SFR module learned the building shape features through Branch one and the relationships between the buildings and the surrounding backgrounds through Branch two. This method improved the separation of buildings from the surrounding backgrounds and the retention of building shapes during building segmentation.

3. Experiments

3.1. Dataset Details

Four publicly available datasets were used in the experiments, including WHU Building Aerial Imagery dataset (WHU dataset) [36], Inria Aerial Image Labeling dataset (Inria dataset) [37], Massachusetts Buildings dataset (Massachusetts dataset) [38], and Building Instances of Typical Cities in China (BITCC) dataset [39]. The details of each dataset are listed in Table 1.

Table 1. Details of each dataset.

Dataset	Resolution	Pixels	Coverage Area	Source Area
WHU dataset	0.3 m	512 × 512	450 km^2	Christchurch, New Zealand
Inria dataset	0.3 m	5000 × 5000	810 km^2	San Francisco, Chicago, the Alps, and others
Massachusetts dataset	1.0 m	1500 × 1500	240 km^2	Boston area, USA
BITCC dataset	0.29 m	500 × 500	120 km^2	Beijing, Shanghai, Shenzhen, and Wuhan, China

Before the experiment, we cropped the original images in the Inria and Massachusetts datasets to 512 × 512 pixels. The images in the BITCC dataset were also adjusted from 500 × 500 pixels to 512 × 512 pixels before inputting to the model. According to the dataset division rules, images in the Inria and BITCC datasets were randomly separated in an 8:1:1 ratio for training, validation, and testing, respectively. The WHU and Massachusetts datasets provided the divided training, validation, and testing sets. We directly used their default ratios for training, validation, and testing, respectively. The final training, validation, and testing set divisions of each dataset are listed in Table 2.

Table 2. Settings used for each dataset in the experiments.

Dataset	Training Set (Tiles)	Validation Set (Tiles)	Test Set (Tiles)
WHU dataset	4737	1036	2416
Inria dataset	14,418	1782	1800
Massachusetts dataset	1233	36	90
BITCC dataset	5790	716	723

3.2. Experimental Settings

Prior to the experiments, the pixels with labels indicating buildings in the four datasets were set to a value of one, whereas the pixels belonging to the background were set to a value of zero. The models employed for comparison included U-Net, PSPNet, DeepLab v3+, HRNetV2 [40], and CFENet [41]. The U-Net, PSPNet, DeepLab v3+, and HRNetV2 models are popular semantic segmentation models with proven effectiveness for semantic segmentation of remote sensing images. CFENet is an excellent building extraction model proposed recently. To ensure fairness in the experiments, the proposed AFL-Net was trained in the same experimental environment as the comparison models and with the same training parameters. The operating system of the experimental platform was Windows 10 x 64 (Microsoft Corporation, Redmond, Washington, USA) and the graphics processing unit (GPU) was GeForce RTX 3090 (24 GB, Nvidia Corporation, Santa Clara, California, USA). The parameter settings were as follows: the batch size was set to 12, number of epochs to 120, an Adam Optimizer Algorithm was used, the loss function was the sum of dice loss and focal loss, the benchmark learning rate was 0.0005, the learning rate was updated by cosine annealing [42], and the number of warmups was 10. The data enhancement techniques used included flipping, rotation, scaling, color enhancement, and Gaussian blur.

3.3. Evaluation Metrics

Four widely used metrics were employed for evaluating the reliability and accuracy of the corresponding building extraction models. The metrics employed to quantitatively evaluate the semantic segmentation results of buildings were Intersection over Union (IoU), F1 score, Recall, and Precision. The ratio between the intersection and union of the pixels predicted to contain buildings and the pixels with labels indicating buildings was represented by IoU. Recall represented the ratio of the correctly predicted pixels containing buildings to the pixels with labels indicating buildings. Precision represented the ratio of the correctly predicted pixels containing buildings to the pixels predicted to contain buildings. The F1 score was the harmonic mean of Recall and Precision.

The building segmentation results were compared pixel by pixel with the corresponding labels. Pixels with correctly predicted results were True, and pixels with falsely predicted results were False. Pixels belonging to buildings were Positive, and pixels belonging to the background were Negative, with TP, TN, FP, and FN used to represent True Positive, True Negative, False Positive, and False Negative, respectively.

The formulae for the calculation of IoU, F1 score, Recall, and Precision are as follows:

$$\text{IoU} = \text{TP}/(\text{FN} + \text{FP} + \text{TP}), \tag{4}$$

$$\text{F1 score} = 2\text{TP}/(2\text{TP} + \text{FN} + \text{FP}), \tag{5}$$

$$\text{Recall} = \text{TP}/(\text{FN} + \text{TP}), \tag{6}$$

$$\text{Precision} = \text{TP}/(\text{FP} + \text{TP}), \tag{7}$$

4. Results and Discussion

4.1. Comparative Experiments

4.1.1. Quantitative Results

Figures 5–8 show the evaluation metrics (IoU, F1 score, Recall, and Precision) of the comparison models on the four datasets.

The figures show that the performance of the proposed AFL-Net was superior for the IoU and F1 scores. The IoU scores of the AFL-Net were 0.73, 1.01, 1.49, and 0.71% higher, on the WHU, Massachusetts, Inria, and BITCC datasets, respectively, than those of the second-best performing model, the HRNetV2. The F1 scores of the AFL-Net were 0.40, 0.61, 1.00, and 0.44% higher on WHU, Massachusetts, Inria, and BITCC datasets, respectively, compared with those of HRNetV2.

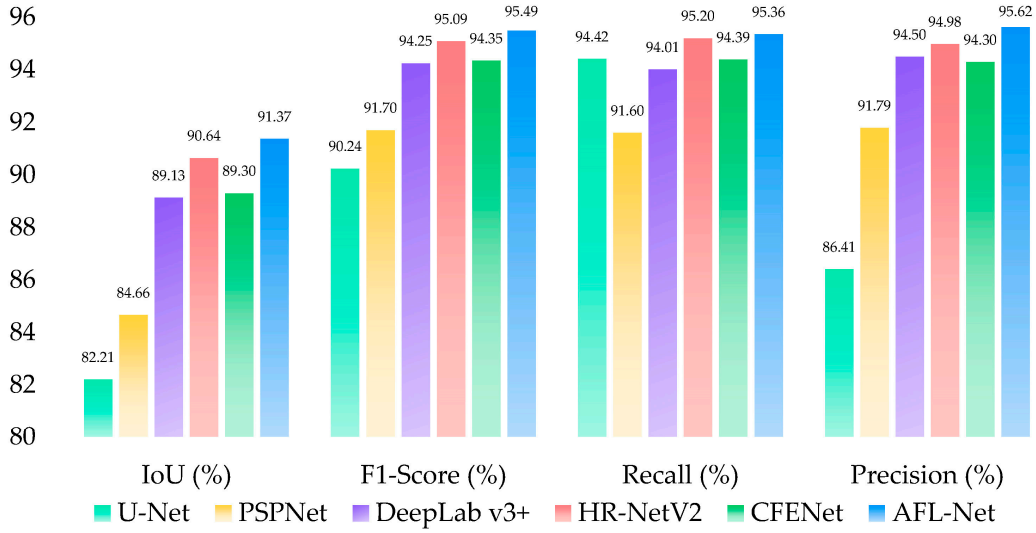

Figure 5. Comparative results from the selected models on the WHU dataset.

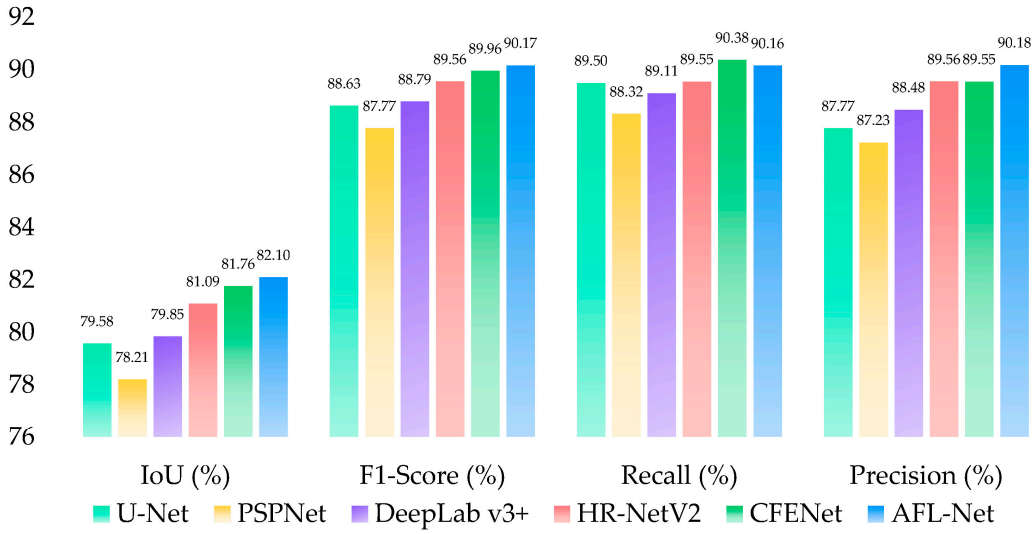

Figure 6. Comparative results from the selected models on the Inria dataset.

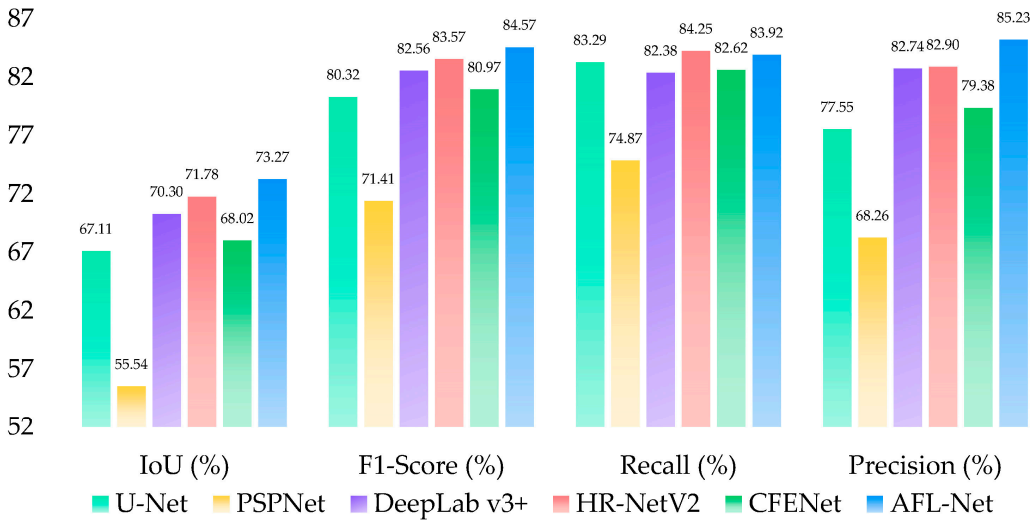

Figure 7. Comparative results from the selected models on the Massachusetts dataset.

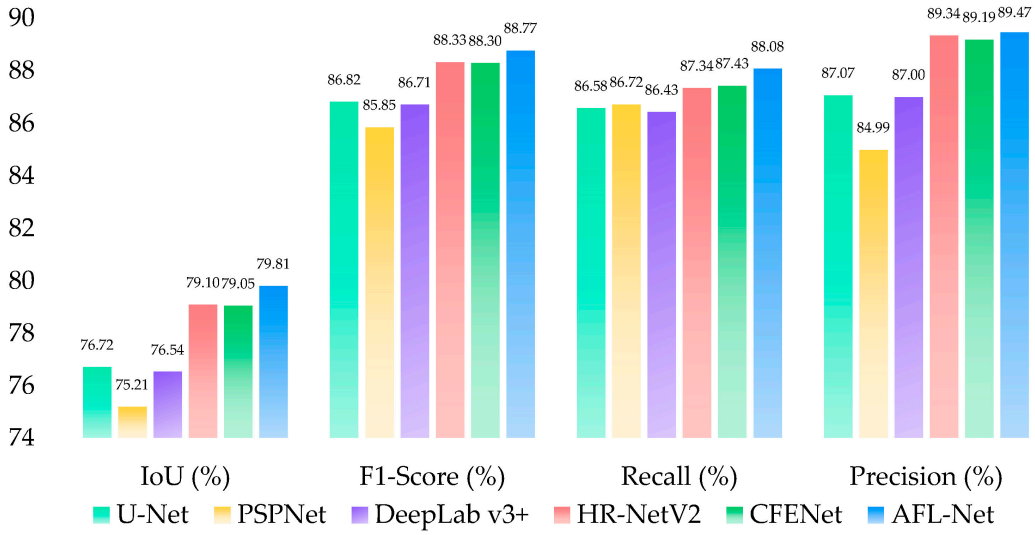

Figure 8. Comparative results from the selected models on the BITCC dataset.

On the WHU dataset, the IoU achieved by AFL-Net was 9.16, 6.71, and 2.24% higher than that of U-Net, PSPNet, and DeepLab v3+, respectively. U-Net achieved higher Recall than PSPNet and DeepLab v3+, while the other metrics were lower than PSPNet and DeepLab v3+. On the Inria dataset, U-Net, PSPNet, and DeepLab v3+ achieved a comparable IoU. On the Massachusetts dataset, AFL-Net improved substantially in IoU and F1 scores compared to the HRNetV2. The Recall of the AFL-Net was 0.33% lower than that of the HRNetV2, but its Precision was 2.33% higher. PSPNet performed poorly and achieved the lowest results in all four metrics. The reason may be that PSPNet tends to overfit on datasets with smaller data volumes. On the BITCC dataset, the performances of U-Net and DeepLab v3+ were comparable, and the results of all four metrics were very close.

Among the four models used for comparison, the performance of HRNetV2 was superior to that of U-Net, PSPNet, and DeepLab v3+. This result was ascribed to both U-Net and DeepLab v3+ fusing low-level feature maps with insufficient semantic information in the upsampling process, and PSPNet fusing only low-resolution feature maps through skip connections. In contrast, HRNetV2 used HRNetV1 as its backbone to extract features at four specific scales in parallel. Therefore, the output features of each scale contained rich semantic information, and the final fused feature map contained more feature information.

4.1.2. Qualitative Results

Figure 9 shows the test images, corresponding labels, and building extraction results from applying the selected models to four sample areas from the WHU dataset. In the extraction results, black represents the background area, white represents the correctly detected building area, red represents the falsely detected building area (i.e., background falsely identified as buildings), and blue represents building areas that were not detected (i.e., building areas falsely identified as background).

In these sample areas, all the models employed for comparison returned some false and missed detection results. For test image (a), UNet, DeepLab v3+, and HRNetV2 could not adequately distinguish the orange building from the surrounding ground of a similar color, leading to the surrounding ground being falsely identified as a building. The PSPNet and DeepLab v3+ models failed to extract a small building. For test image (b), all the comparison models except CFENet failed to identify the buildings with rooftops having special texture features. For test image (c), UNet, PSPNet, HRNetV2, and CFENet falsely identified the container box as a building, and DeepLab v3+ could not detect a conjoined building blocked by trees. For test image (d), UNet, PSPNet, and DeepLab v3+ could not adequately extract the relatively large cross-shaped building, HRNetV2 returned minor cases of missed detection, and CFENet mistakenly identified other objects as buildings.

In contrast, the building extraction results from AFL-Net showed completely segmented buildings with smooth edges and sharp corners, with extremely rare cases of false or missed detection.

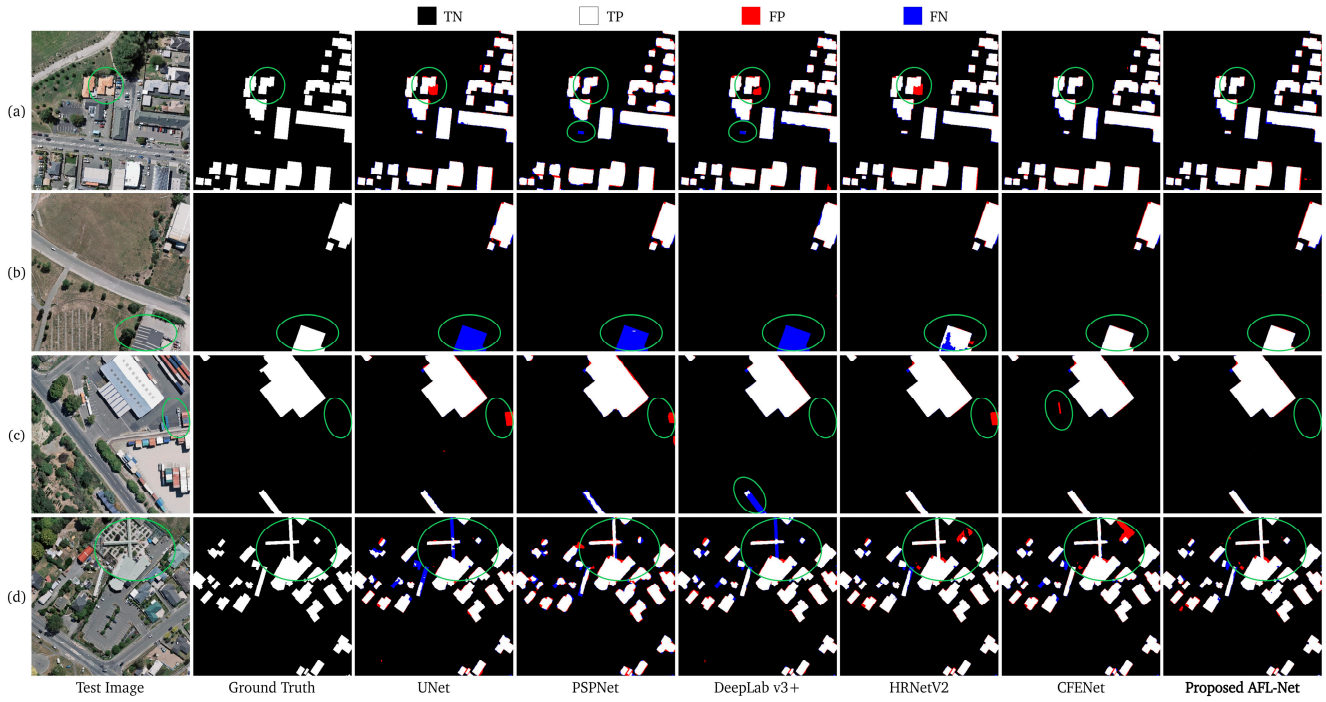

Figure 9. (**a**–**d**) Sample building segmentation results from the selected models with the WHU dataset (comparative experiment).

Figure 10 shows the test images, corresponding labels, and building extraction results after applying the models to four sample areas from the Inria dataset.

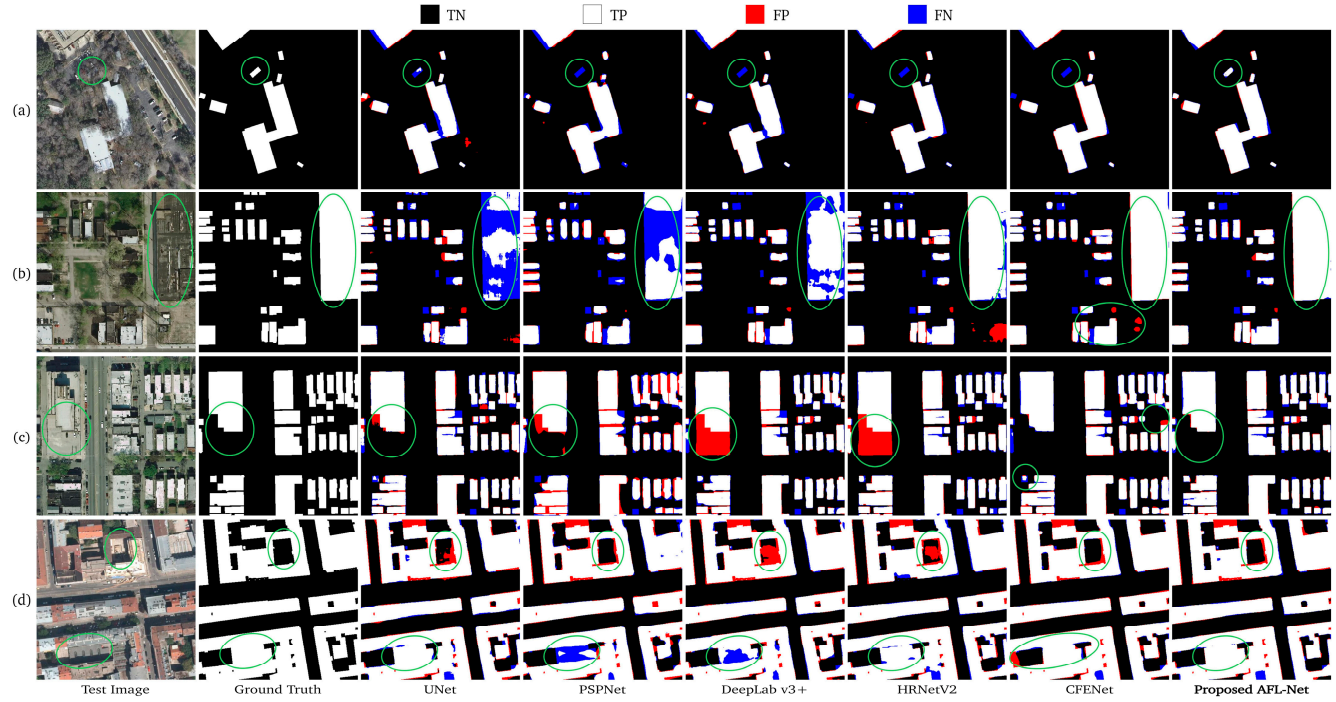

Figure 10. (**a**–**d**) Sample building segmentation results from the selected models with the Inria dataset (comparative experiment).

For test image (a), a portion of the buildings was blocked by trees and none of the comparative models could extract the blocked area. For test image (b), the building circled in green has complex rooftop texture features of diverse colors and shades, and all comparison models except for CFENet failed to completely extract this rooftop, while the CFENet falsely detected three areas. For test image (c), the color of the building circled in green was similar to that of the ground, making it extremely difficult for the models to accurately detect the building. For test image (d), the building circled in green had a square open space in the center, with complex rooftop structures and various colors. U-Net, PSPNet, DeepLab v3+, and HRNetV2 falsely detected the central open space as a building, and CFENet failed to adequately distinguish the complex roof. In contrast, AFL-Net not only detected the blocked area and distinguished the rooftop and ground with similar colors, but it also adequately detected the rooftops with complex structures and colors.

Figure 11 shows the test images, corresponding labels, and building extraction results after applying the selected models to four sample areas from the Massachusetts dataset.

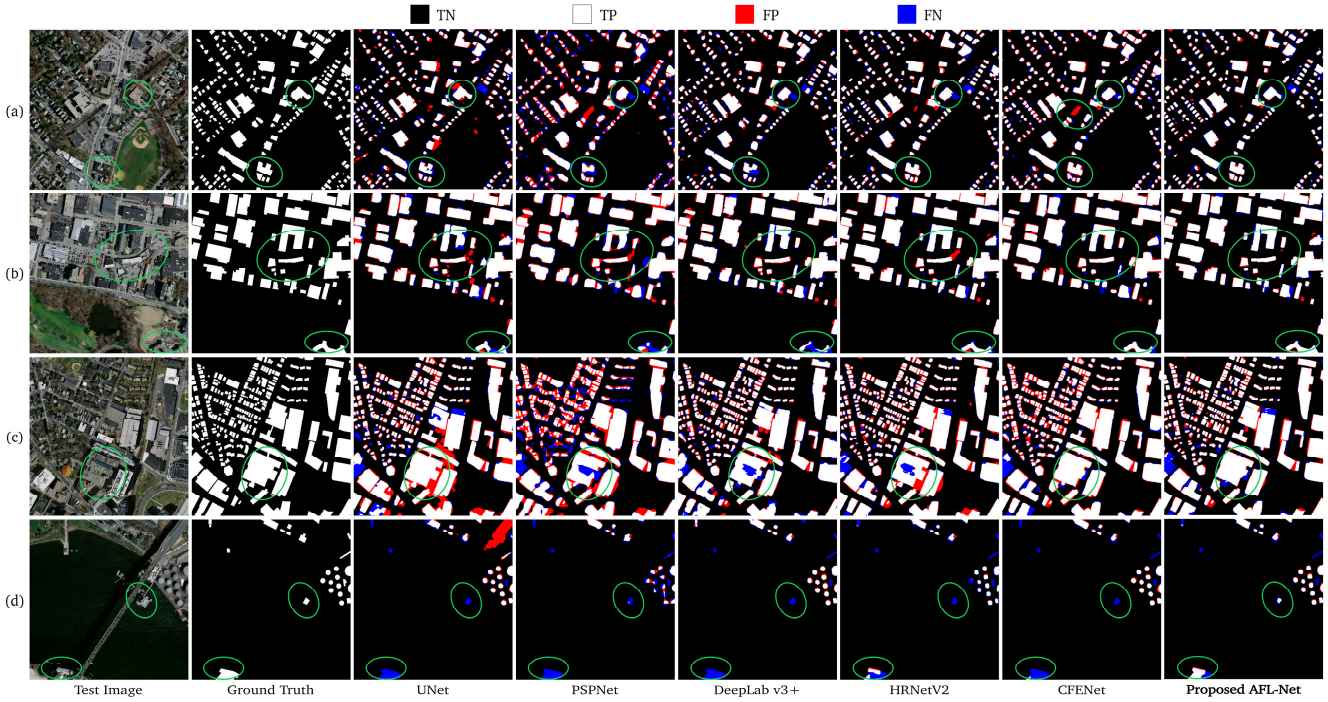

Figure 11. (**a**–**d**) Sample building segmentation results from the selected models with the Massachusetts dataset (comparative experiment).

The Massachusetts dataset has low resolution, and the buildings are mostly shown as scattered patches, posing challenges when extracting small buildings. For test images (a) and (c), all the comparison models failed to adequately detect the buildings with composite structures. The UNet, PSPNet, and HRNetV2 models failed to distinguish the building marked in green and the background in test image (b), with a relatively substantial number of false and missed detections. All the comparative models failed to detect the two buildings on the water in test image (d). As mentioned, the Massachusetts dataset has low resolution, and the lower the resolution of the images, the more insufficient the features that could be extracted would be. In contrast, because of the high resolution of the output feature map owing to the feature extraction backbone, AFL-Net maintained an adequate building extraction performance for images with relatively low resolution, with substantially reduced occurrences of missed detection.

Figure 12 shows the test images, corresponding labels, and building extraction results after applying the selected models to four sample areas from the BITCC dataset. For test image (a), all the comparative models falsely identified the yard next to the building as a

building. For test image (b), the rooftop with composite structure was difficult to extract, and all the comparison models missed the detection to varying degrees. Just as with test image (d) in the previous dataset, test image (c) also contains a building with an open space in the center, and all the comparison models failed to accurately detect this building with an irregular shape. For test image (d), AFL-Net and CFENet accurately separated the building and the ground, with smooth edges and clear boundaries in the extraction results. For the satellite images, AFL-Net was able to completely remove the complex background and accurately retain only the building area.

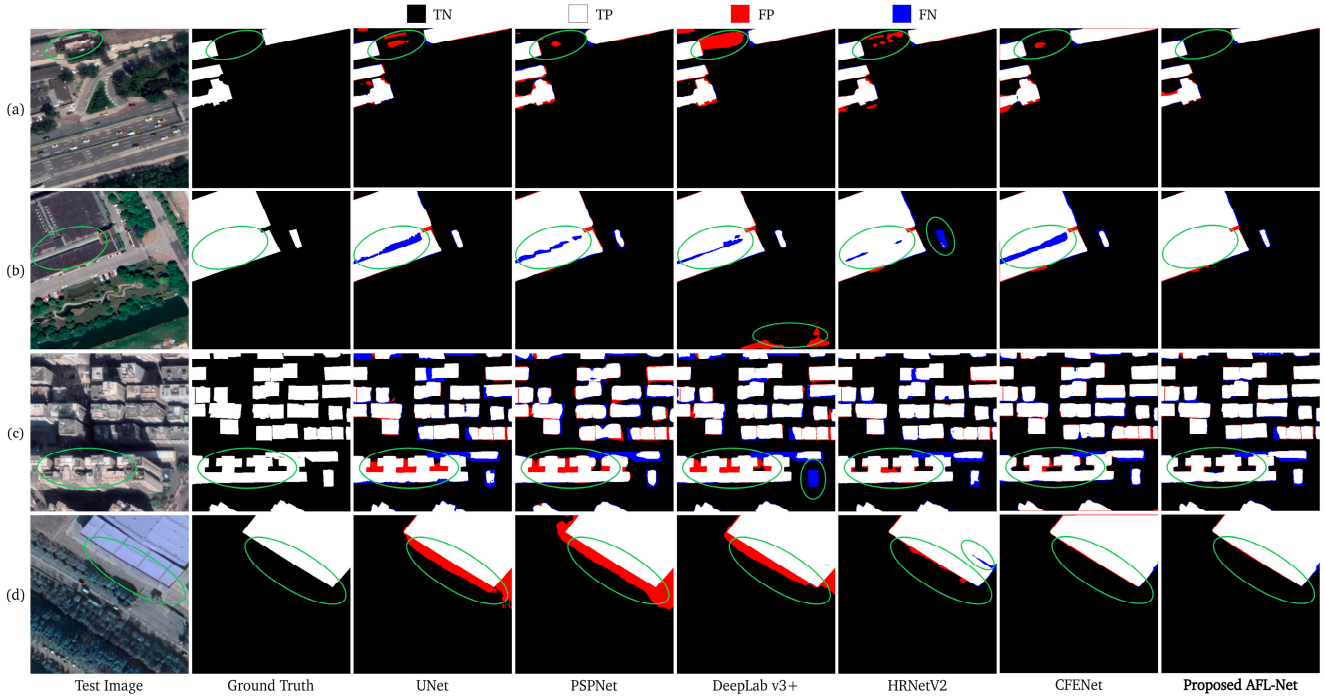

Figure 12. (**a**–**d**) Sample building segmentation results from the selected models with the BITCC dataset (comparative experiment).

In summary, the four datasets contain buildings of a range of styles, scales, and shapes, with AFL-Net achieving superior performance on all four datasets, proving the robustness of our proposed method. The high-resolution feature map provided by the backbone network of AFL-Net reduced the occurrences of missed detection of small buildings. For scenarios where the buildings have features similar to the background, or the building rooftops have complex structures, the AMFF module drove the AFL-Net to adaptively learn the relationship between the buildings and the background. Consequently, AFL-Net could distinguish clearly between the buildings and the background and retain the complete building area. The SFR module drove the AFL-Net to optimize the detection of irregularly shaped buildings in a targeted manner, ensuring the accuracy and smoothness of the edges of the irregularly shaped buildings detected.

4.1.3. Comparison with Recent Methods

Due to the label accuracy and the wide application of the WHU dataset, it was used to demonstrate the accuracy of the proposed AFL-Net in comparison with other building extraction models proposed over the last 2 years, as shown in Figures 13 and 14. The selected building extraction models included SST [43], MHA-Net [44], MAP-Net [45], B-FGC-Net [46], and AGs-Unet [47]. The evaluation metrics for the comparison were IoU, the number of parameters, and the floating-point operations (FLOPs). The higher the IoU, the higher was the accuracy of the model. A model with lower number of parameters and FLOPs usually has fewer complex algorithms and is more convenient for practical

applications. The values of the metrics of these building extraction models were taken directly from the respective publications.

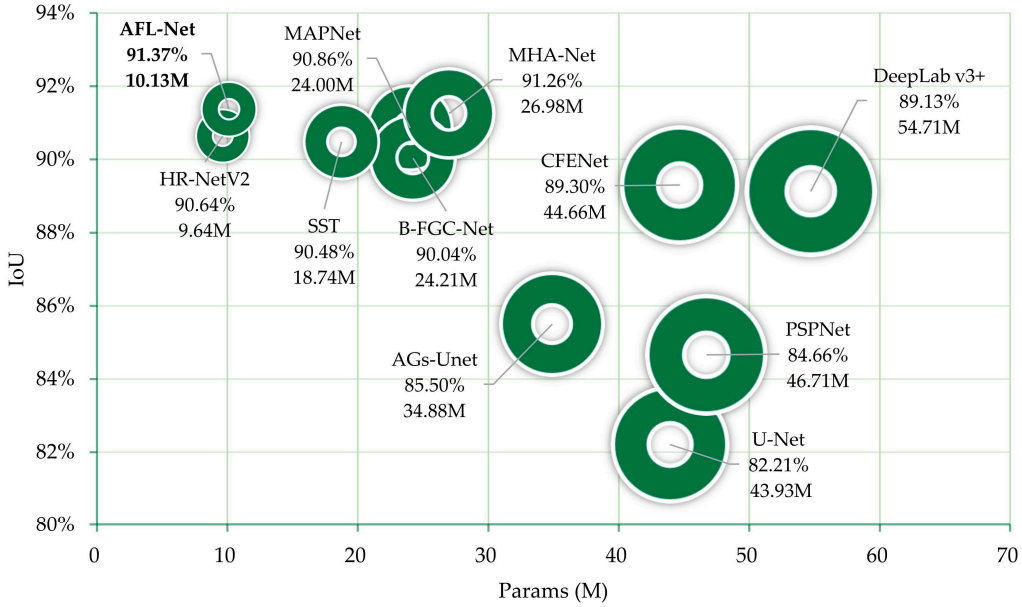

Figure 13. Comparison of the accuracy and complexity of selected models, annotated with IoU and parameters in each model.

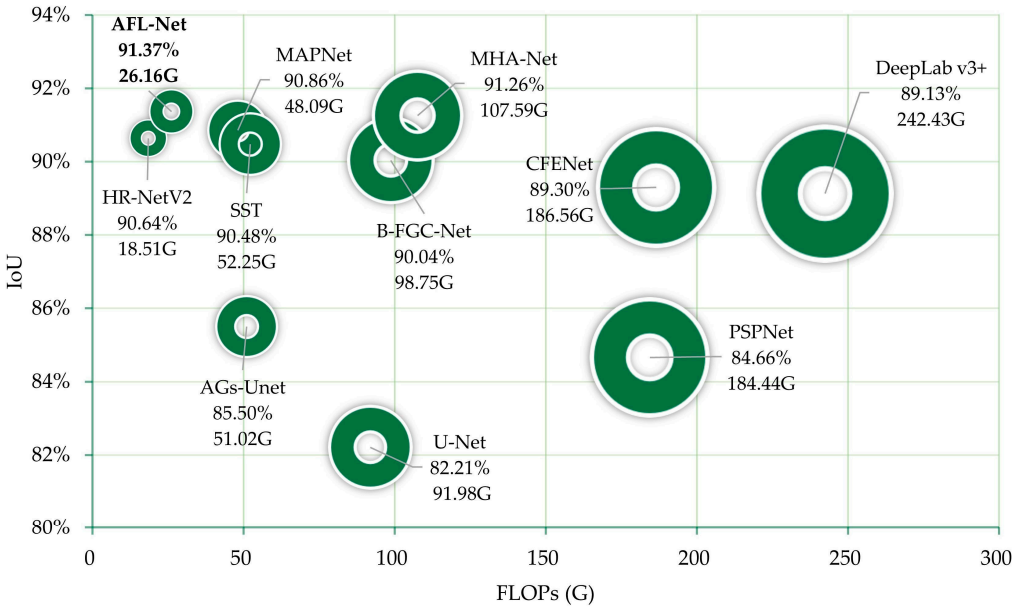

Figure 14. Comparison of the accuracy and complexity of selected models, annotated with IoU and computational cost in each model.

As shown in Figure 13, the IoU of the proposed AFL-Net was higher than those of the other recently proposed building extraction models. In particular, the IoU of AFL-Net was 0.89, 0.11, 0.51, 1.33, and 5.87% higher, respectively, than that of SST, MHA-Net, MAP-Net, B-FGC-Net, and AGs-Unet. Further, the number of parameters of AFL-Net (10.13 M) was approximately 37.55% of that of MHA-Net (26.98 M), which achieved the second highest IoU score. As shown in Figure 14, the FLOPs of AFL-Net (26.16 G) is lower than that of most models, accounting for only 24.31% of that of MHA-Net (107.59 G), proving the excellent balance achieved by AFL-Net between accuracy and computational cost.

4.2. Ablation Study

4.2.1. Quantitative Analysis

HRNetV2 was selected as the Baseline, and an ablation study was conducted on the four datasets to quantitatively analyze the contribution of the proposed modules in AFL-Net. Additionally, we conducted ablation experiments to investigate the effects of the AMFF module and the SFR module on the training speed and inference speed of the Baseline. Speed was measured by the number of frames per second (FPS). The speed tests were conducted on an NVIDIA RTX 3090 GPU with one (for inference speed test) or two (for training speed test) input images of three channels and a size of 512 × 512 pixels. The evaluation results are shown in Table 3.

Table 3. Ablation study of various module combinations with four datasets.

Method	Parameters	Training Speed	Inference Speed	WHU		Inria		Massachusetts		BITCC	
				IoU (%)	F1 Score (%)	IoU (%)	F1 Score (%)	IoU (%)	F1 Score (%)	IoU (%)	F1 Score (%)
Baseline	9.64 M	26.02 FPS	34.86 FPS	90.64	95.09	81.09	89.56	71.78	83.57	79.10	88.33
Baseline+ SFR	9.83 M	22.32 FPS	32.68 FPS	91.10	95.34	81.65	89.90	72.85	84.29	79.51	88.59
Baseline+ AMFF	9.94 M	23.73 FPS	34.05 FPS	91.21	95.41	81.82	90.00	72.94	84.35	79.46	88.56
Baseline+ SFR + AMFF	10.13 M	20.83 FPS	30.67 FPS	91.37	95.49	82.10	90.17	73.27	84.57	79.81	88.77

Table 3 shows that the SFR module improved the IoU of the model by 0.46, 0.56, 1.07, and 0.41%, respectively, for the WHU, Inria, Massachusetts, and BITCC datasets. After adding the AMFF module, IoU was improved by 0.57, 0.73, 1.16, and 0.36%, respectively, for the four datasets compared with the Baseline. The combination of the SFR and AMFF modules improved the IoU by 0.73, 1.01, 1.49, and 0.71%, respectively, over the Baseline. The SFR and AMFF modules increased the parameters by approximately 0.19M and 0.3M, respectively. Both the SFR and the AMFF modules improved the model performance and, in combination, the improvement was greater than either of the two alone. This result indicated that these two modules only conflicted with each other to a minor degree and could both independently improve the model performance.

The SFR module decreased the training and inference speeds by 3.70 and 2.18 FPS, respectively, compared to the Baseline. The AMFF module decreased the training and inference speeds by 2.29 and 0.81 FPS, respectively, compared to the Baseline. The results show that, although the AMFF and SFR modules improve the accuracy of the model, they also decrease the training and inference speeds of the model.

4.2.2. Qualitative Analysis

To qualitatively analyze the influence of the proposed modules on the model performance, the extraction results of Baseline, Baseline + SFR, Baseline + AMFF, and Baseline + SFR + AMFF with the four datasets were compared qualitatively, as shown in Figure 15. The last layer of each of the four networks was visualized to compare and analyze the influence of each module on the feature map, as shown in Figure 16.

Figure 15 shows that the SFR module could reduce false detection. For example, the feature map Baseline + SFR shows that the SFR module adequately distinguished between the building edges and the surrounding ground-based objects (circled in orange in the WHU and BITCC samples), while ensuring the smoothness of the building edges. Additionally, the SFR module successfully identified the background areas between two buildings and in the center of one building (circled in orange in the Massachusetts and Inria samples), preventing adjacent buildings from being identified as one building in the results, while retaining relatively complete building shape features. The area circled in orange corresponds to missed detection in the samples by Baseline, which was improved substantially by Baseline + AMFF. This result indicated that the AMFF module is superior in complex rooftop detection and retained the complete building segmentation results. The

Baseline + SFR + AMFF model could take full advantage of the two modules, avoiding both false and missed detections, achieving superior extraction results.

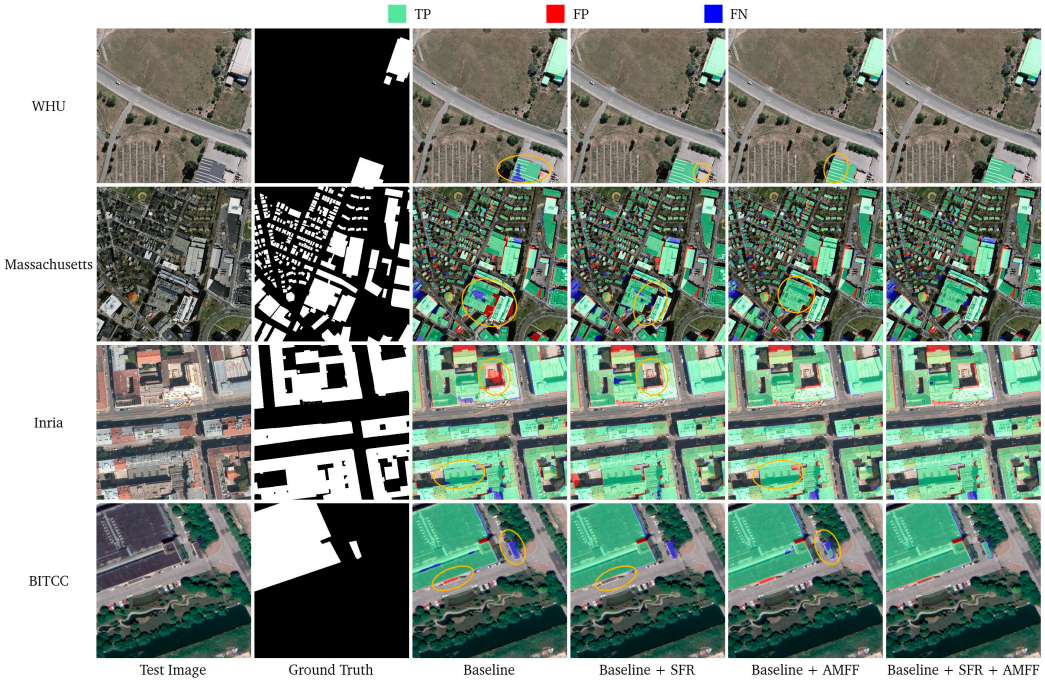

Figure 15. Samples of building extraction results by different models with four datasets (ablation study).

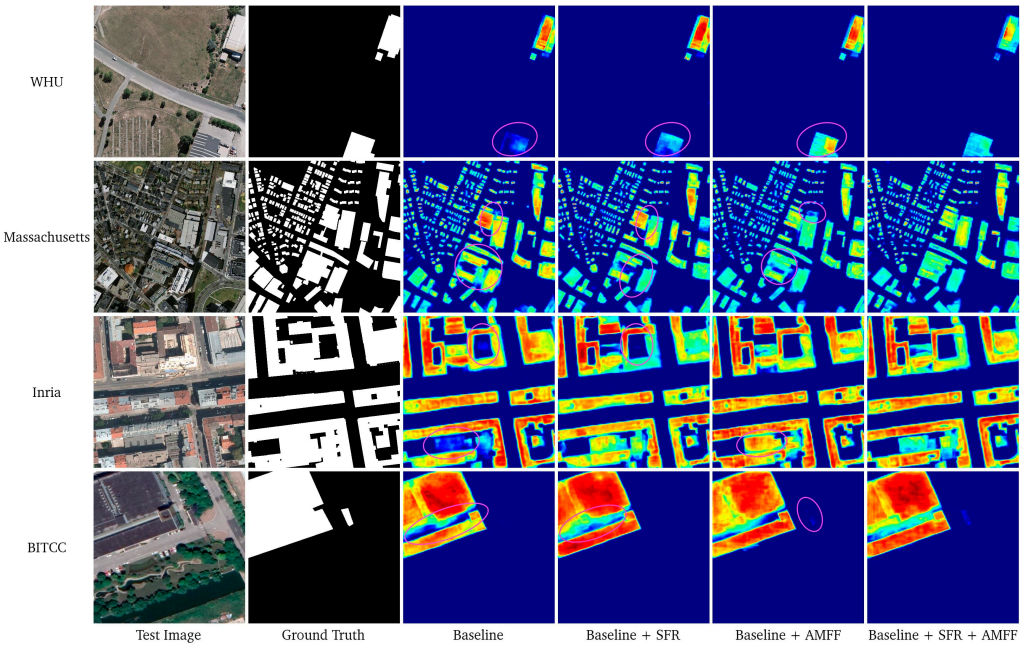

Figure 16. Visualization of feature maps by the selected models with four datasets (ablation study).

The deeper red the area in the feature map, the higher the attention of the model. The more attention the model pays to the background area, the more likely it is to return false detections, whereas the less attention the model pays to the building area, the more likely it is to return missed detections. Figure 16 shows that the SFR module reduced attention to the background area (circled in magenta) in the samples of the Massachusetts and Inria datasets, reducing the interference from background noise on the building extraction results.

Moreover, it increased attention to the building area (circled in magenta) in the samples of the WHU and BITCC datasets, reducing the missed detections caused by insufficient attention to the building area. The AMFF module increased the attention to the building area (circled in magenta) in the samples of all four datasets, leading to more complete extraction results.

Generally, both the SFR and the AMFF modules reduced the occurrence of over- or under-segmentation. The SFR module learned the shape features to overcome over-segmentation caused by the background areas between buildings, also ensuring the smoothness and completeness of the building edges. The AMFF module adaptively learned the building features through the self-attention mechanism, enhancing the capability of the model to detect complex rooftops, ensuring the completeness of the detection results, and reducing the occurrence of under-segmentation. The performance of the combined SFR and AMFF modules achieved performance that was superior to that of either one of the two modules. This finding was attributed to the advantages of combining the two modules (i.e., not interfering with each other but, rather, leading to more complete building extraction results, with the original shape being retained).

4.3. Limitations and Future Work

Although the proposed AFL-Net has achieved excellent performance with all four datasets, there is room for improvement.

In the ablation study in Section 4.2, we found that the improvement of IoU compared with Baseline on the BITCC dataset was less than that on the other datasets. This could be ascribed to the side views of buildings demonstrating features similar to those of the rooftops in the satellite images, interfering with the extraction of the rooftops. In contrast, aerial images are mostly orthophotos that do not contain the side views of buildings, which facilitates the model extracting the roof features of buildings.

We conducted comparative experiments on training and inference speeds of each model, as shown in Figure 17.

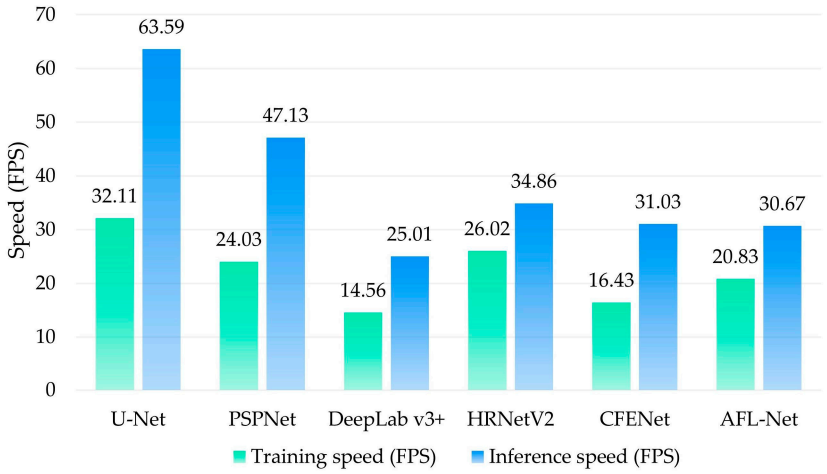

Figure 17. Comparison of the training and inference speeds of the selected models.

The inference speed of AFL-Net was lower than that of U-Net, PSPNet, and HRNetV2, and was comparable to that of CFENet. The training speed of AFL-Net was higher than that of CFENet and DeepLabv3+, but lower than that of other models. Although the number of parameters and computational cost of AFL-Net was less than that of most models, the speed of the model was insufficient. A lower training speed costs more time for model training, and a lower inference speed limits the application of the model in real-time processing tasks.

In future work, the capability of the model to identify the side views of buildings will be improved, and the model will be applied further to detect other types of ground-based

objects. Additionally, we will attempt to improve the training and inference speeds of the model so that it can be used in more building extraction tasks.

5. Conclusions

In this study, we proposed an attentional feature learning neural network (AFL-Net) for building extraction. The backbone of AFL-Net allowed the extracted feature map to retain a high resolution, reducing the loss of detailed information in the downsampling process and ensuring that the features acquired were abundant. The multiscale feature fusion module based on the self-attention mechanism adaptively learned the relationships between various features, ensuring sufficient utilization of building features, while reducing the occurrence of under-segmentation. The shape feature refinement module adaptively learned the shape features of the buildings and improved the smoothness of building edges, while reducing the occurrence of over-segmentation. We conducted experiments with four publicly available datasets, which showed that the accuracy of AFL-Net was superior to that of all the other tested models, proving the robustness and effectiveness of AFL-Net. An ablation study was conducted with the WHU dataset, with the results indicating that, in comparison with other recently proposed models, AFL-Net had significantly fewer parameters and computational cost, while achieving the highest accuracy, indicating an excellent balance between model complexity and accuracy. Both the quantitative and qualitative analyses of the ablation study proved the effectiveness of the proposed AMFF and SFR modules. In future work, we will attempt to improve the efficiency of our model and apply it to the extraction of other types of ground-based objects.

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

Funding: This research was funded by the Natural Science Foundation for Distinguished Young Scholars of Henan Province under grant number 212300410014; the National Natural Science Foundation of China under grant number 42201491.

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

Acknowledgments: We thank the editors and reviewers for their constructive and helpful comments that led to the substantial improvement of this paper.

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

Abbreviations

AFL-Net	Attention feature learning network
AMFF	Attentional multiscale feature fusion
CNN	Convolutional neural network
FLOPs	Floating-point operations
FN	False negative
FP	False positive
FPS	Frames per second
IoU	Intersection over union
PSA	Polarized self-attention
ReLU	Rectified linear unit
SFR	Shape feature refinement
TN	True negative
TP	True positive

References

1. Li, W.; Fu, H.; Yu, L.; Cracknell, A. Deep Learning Based Oil Palm Tree Detection and Counting for High-Resolution Remote Sensing Images. *Remote Sens.* **2017**, *9*, 22. [CrossRef]
2. Zhang, B.; Wang, C.; Shen, Y.; Liu, Y. Fully Connected Conditional Random Fields for High-Resolution Remote Sensing Land Use/Land Cover Classification with Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1889. [CrossRef]
3. 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]
4. Gao, X.; Wang, M.; Yang, Y.; Li, G. Building Extraction from RGB VHR Images Using Shifted Shadow Algorithm. *IEEE Access* **2018**, *6*, 22034–22045. [CrossRef]
5. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
6. Gao, Y.; Gao, F.; Dong, J.; Wang, S. Change Detection from Synthetic Aperture Radar Images Based on Channel Weighting-Based Deep Cascade Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 4517–4529. [CrossRef]
7. Kang, M.; Baek, J. SAR Image Change Detection via Multiple-Window Processing with Structural Similarity. *Sensors* **2021**, *21*, 6645. [CrossRef]
8. Cooner, A.J.; Shao, Y.; Campbell, J.B. Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. *Remote Sens.* **2016**, *8*, 868. [CrossRef]
9. Xiong, C.; Li, Q.; Lu, X. Automated Regional Seismic Damage Assessment of Buildings Using an Unmanned Aerial Vehicle and a Convolutional Neural Network. *Automat. Constr.* **2020**, *109*, 102994. [CrossRef]
10. Chen, Q.; Wang, L.; Waslander, S.L.; Liu, X. An End-to-End Shape Modeling Framework for Vectorized Building Outline Generation from Aerial Images. *ISPRS J. Photogramm. Remote Sens* **2020**, *170*, 114–126. [CrossRef]
11. Jung, C.R.; Schramm, R. Rectangle Detection Based on a Windowed Hough Transform. In Proceedings of the 17th Brazilian Symposium on Computer Graphics and Image Processing, Curitiba, Brazil, 20–20 October 2004; pp. 113–120.
12. Simonetto, E.; Oriot, H.; Garello, R. Rectangular Building Extraction from Stereoscopic Airborne Radar Images. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 2386–2395. [CrossRef]
13. Wei, D. Research on Buildings Extraction Technology on High Resolution Remote Sensing Images. Master's Thesis, Information Engineering University, Zhengzhou, China, 2013.
14. Zhao, Z.; Zhang, Y. Building Extraction from Airborne Laser Point Cloud Using NDVI Constrained Watershed Algorithm. *Acta Optica Sin.* **2016**, *36*, 503–511.
15. Maruyama, Y.; Tashiro, A.; Yamazaki, F. Use of Digital Surface Model Constructed from Digital Aerial Images to Detect Collapsed Buildings During Earthquake. *Procedia Eng.* **2011**, *14*, 552–558. [CrossRef]
16. 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]
17. 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, 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 Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241.
19. Li, Q.; Mou, L.; Hua, Y.; Shi, Y.; Zhu, X.X. Building Footprint Generation Through Convolutional Neural Networks with Attraction Field Representation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [CrossRef]
20. Luo, L.; Li, P.; Yan, X. Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review. *Energies* **2021**, *14*, 7982. [CrossRef]
21. Qiu, Y.; Wu, F.; Yin, J.; Liu, C.; Gong, X.; Wang, A. MSL-Net: An Efficient Network for Building Extraction from Aerial Imagery. *Remote Sens.* **2022**, *14*, 3914. [CrossRef]
22. Yin, J.; Wu, F.; Qiu, Y.; Li, A.; Liu, C.; Gong, X. A Multiscale and Multitask Deep Learning Framework for Automatic Building Extraction. *Remote Sens.* **2022**, *14*, 4744. [CrossRef]
23. Zhu, Q.; Zhang, Y.; Wang, L.; Zhong, Y.; Guan, Q.; Lu, X.; Zhang, L.; Li, D. A Global Context-Aware and Batch-Independent Network for Road Extraction from VHR Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 353–365. [CrossRef]
24. Hosseinpour, H.; Samadzadegan, F.; Javan, F.D. A Novel Boundary Loss Function in Deep Convolutional Networks to Improve the Buildings Extraction from High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4437–4454. [CrossRef]
25. Wang, Z.; Xu, N.; Wang, B.; Liu, Y.; Zhang, S. Urban Building Extraction from High-Resolution Remote Sensing Imagery Based on Multi-Scale Recurrent Conditional Generative Adversarial Network. *GISci. Remote Sens.* **2022**, *59*, 861–884. [CrossRef]
26. Sun, Z.; Zhou, W.; Ding, C.; Xia, M. Multi-Resolution Transformer Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo Inf.* **2022**, *11*, 165. [CrossRef]
27. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 404–419.

28. 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.* **2018**, *40*, 834–848. [CrossRef]
29. Deng, W.; Shi, Q.; Li, J. Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2611–2620. [CrossRef]
30. Wen, Q.; Jiang, K.; Wang, W.; Liu, Q.; Guo, Q.; Li, L.; Wang, P. Automatic Building Extraction from Google Earth Images Under Complex Backgrounds Based on Deep Instance Segmentation Network. *Sensors* **2019**, *19*, 333. [CrossRef] [PubMed]
31. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.
32. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized Self-Attention: Towards High-Quality Pixel-Wise Regression. *arXiv* **2021**, arXiv:2107.00782.
33. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable Convnets V2: More Deformable, Better Results. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308.
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
35. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv* **2016**, arXiv:1511.07122.
36. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
37. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
38. Mnih, V. Machine Learning for Aerial Image Labeling. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 2013.
39. Wu, K.; Zheng, D.; Chen, Y.; Zeng, L.; Zhang, J.; Chai, S.; Xu, W.; Yang, Y.; Li, S.; Liu, Y.; et al. A Dataset of Building Instances of Typical Cities in China. *China Sci.* **2021**, *6*, 182–190.
40. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
41. Chen, J.; Zhang, D.; Wu, Y.; Chen, Y.; Yan, X. A Context Feature Enhancement Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 2276. [CrossRef]
42. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2017**, arXiv:1608.03983.
43. Chen, K.; Zou, Z.; Shi, Z. Building Extraction from Remote Sensing Images with Sparse Token Transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
44. Cai, J.; Chen, Y. MHA-Net: Multipath Hybrid Attention Network for Building Footprint Extraction from High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5807–5817. [CrossRef]
45. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction from Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]
46. Wang, Y.; Zeng, X.; Liao, X.; Zhuang, D. B-FGC-Net: A Building Extraction Network from High Resolution Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 269. [CrossRef]
47. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGS-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. *Sensors* **2022**, *22*, 2932. [CrossRef]