BUILDING EXTRACTION FROM HIGH-RESOLUTION REMOTE SENSING IMAGERY BASED ON MULTI-SCALE FEATURE FUSION AND ENHANCEMENT

Y. Chen [1] *, H. Cheng [1], S. Yao [1], Z. Hu [2]

[1] Nanjing University of Posts and Telecommunications, 210003 Nanjing, China - chenyixiang@njupt.edu.cn
[2] MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area, Shenzhen 518060, China - zwhoo@szu.edu.cn

Commission III, WG III/I

KEY WORDS: High-Resolution Remote Sensing, Building Extraction, Encoder and Decoder, Multi-Scale Features, Dual Attention.

ABSTRACT:
The accurate detection and mapping of buildings from high-resolution remote sensing (HRRS) images have attracted extensive attention. However, as an artificial target, buildings not only have various types, but also have multi-scale characteristics and complex context, which brings great challenges to the accurate identification of buildings. To deal with this problem, a semantic segmentation model based on multi-scale feature fusion and enhancement (MSFFE) is proposed for building extraction from HRRS images. Specifically, the proposed model uses the network structure of encoder and decoder. In the encoding stage, densely connected convolutional neural network is used as an encoder to extract multi-level spatial and semantic features. To effectively use the multi-scale features of buildings, a multi-scale feature fusion (MSFF) module between encoder and decoder is designed to distinguish buildings of different scales in complex scenes. In the decoding stage, an attention weighted semantic enhancement (AWSE) module is introduced into the decoder to assist the up-sampling process. It not only makes full use of the multi-level features output by the encoder, but also highlights the key local semantic information of the building. To verify the effectiveness of the proposed model, experiments were conducted on two building segmentation data sets, WHU and INRIA. The preliminary results show that the proposed model can effectively identify buildings with different scales in complex scenes, and has better performance than the current representative networks including FCN, U-net, DeeplabV3+ and MA-FCN.

1. INTRODUCTION

At present, the wide availability of high-resolution remote sensing images provides the possibility for the acquisition of individual targets. Automatic extraction of buildings from high-resolution remote sensing (HRRS) images has become one of the important research topics in remote sensing and related application fields. This is because the geospatial information related to buildings plays an important role in urban construction, national defense, earthquake relief and people's production and life.

Buildings are man-made objects with typical geometric features such as shape, structure and height. They are the main basis of the traditional building extraction method based on handcrafted features (Wang et al., 2015; Turker et al., 2015; Hermosilla et al., 2011). Although these methods have achieved good results in some targeted applications or specific tasks, they still have great limitations for complex and changeable image scenes due to the diversity of building types and sizes and the complexity of scenes.

In recent years, the rise and development of deep learning has changed the traditional machine learning paradigm based on handcrafted features. The research based on deep learning has made a major breakthrough in visual tasks such as image classification, target detection and semantic segmentation. Building extraction is actually a semantic segmentation problem, which aims to label buildings at the pixel level. This is a very challenging task due to the complexity of remote sensing image scenes. Using the labelled sample data, the deep semantic segmentation network can establish the mapping between the multi-scale feature representation of the image and the pixel category label through end-to-end learning. The proposal of fully convolutional network (FCN) (Long et al., 2015) promotes the rapid development of semantic segmentation network. A series of network models such as U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), DeepLab series (Chen et al., 2014, 2017, 2018a, 2018b) have been proposed successively to solve the visual segmentation task of ordinary images.

In the field of remote sensing target recognition, a variety of network models based on deep learning have been developed for building extraction. For example, Ji et al. (2019) created the widely used WHU aerial and satellite image dataset, and proposed a Siamese U-Net for building segmentation, which significantly improved the accuracy of building extraction. Liu et al. (2019) used the FCN-based method to extract buildings. The proposed spatial residual inception (SRI) module can obtain multi-scale context information, which is used to accurately detect large buildings. Focusing on the accuracy and integrity of building extraction, Shao et al. (2020) proposed a building residual refine network (BRRNet). This model includes a prediction module based on encoder-decoder structure and a residual refinement module, which are used to obtain global features and refined detection results respectively.

Due to the diversity of building sizes and the complexity of scenes, how to extract multi-scale buildings and accurately locate their boundaries has attracted extensive attention. Liu et al. (2020) proposed a multi-scale U-shaped convolutional neural network with edge constraints (EMU-CNN) for the extraction of building instances from HRRS images. This model has strong robustness to the variation of building scale. Deng et al. (2021) introduced the attention gate and atrous spatial pyramid pooling module into the encoder-decoder network to overcome the high intraclass variance and complexity of building scenes. In addition, multi-path attention networks, such as MAP-Net (2020) and MHA-Net (2021), are proposed to further improve the extraction accuracy

* Corresponding author

of multi-scale buildings (especially small buildings) and enhance the robustness to complex scenes.

In short, the diversity of building scale and the complexity of scene are the main challenges of building extraction in HRRS images. To deal with this problem effectively, this paper proposes a semantic segmentation model based on multi-scale feature fusion and enhancement (MSFFE) for building extraction in HRRS images. Specifically, the model adopts the network structure of encoder and decoder. In the encoding stage, densely connected convolutional neural network (DenseNet) (Huang et al., 2017) is used as the encoder to extract multi-level spatial and semantic features. Through dense connection, this network realizes the reuse of features, which can greatly reduce the number of network parameters and improve the training speed. Further, to make effective use of the multi-scale features of buildings, a multi-scale feature fusion (MSFF) module between encoder and decoder is designed to distinguish buildings of different scales in complex scenes. In the decoding stage, the decoder introduces an attention weighted semantic enhancement (AWSE) module to assist the up-sampling process. It not only makes full use of the multi-level features of the encoder output, but also highlights the key local semantic information of the building.

2. METHOD

The proposed model adopts the encoder-decoder structure, which is a relatively stable structure in the field of image semantic segmentation. The encoder is used to learn the feature maps of different scales of the input image through a series of operations such as convolution and pooling. The role of the decoder is to restore the feature maps output by the encoder to the size of the original image through up-sampling or deconvolution.

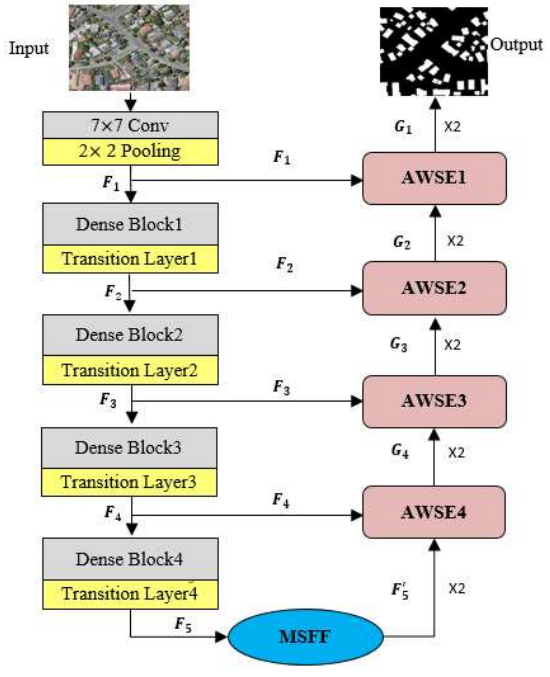

Figure 1. Overview of the proposed model architecture.

Specifically, the overall structure of the model is shown in Figure 1. In the encoding part of the network, we use the densely connected network (i.e. DenseNet) for multi-scale representation of features. It includes four dense blocks. Between the adjacent dense blocks is the transition layer, which is composed of a 1 × 1 convolution layer and a 2 × 2 average pooling layer. At the end of the encoder is a multi-scale feature fusion (MSFF) module, which is used to integrate building features of different scales. In the decoder, in order to make full use of the guiding role of the shallow features output by the encoder in each stage, the attention weighted semantic enhancement (AWSE) module is designed to realize the semantic enhancement of the key information related to buildings in the process of recovering the size of the feature maps through step-by-step up-sampling, so as to improve the accuracy of building segmentation.

2.1 Dense block

The core part of DenseNet is the dense block. By establishing the connection relationship between different layers (the input of each layer comes from the output of all previous layers), the dense block strengthens the transmission of gradient, realizes the reuse of features, and reduces the number of parameters of the model to a certain extent. In this paper, we use a simpler dense block, which contains only four composite functions. The growth rate of feature maps is set to 24, that is, 24 feature maps will be generated after each composite function. The simplified dense block can not only effectively extract the multi-scale features of buildings, but also greatly reduce the amount of parameters in the encoding stage.

2.2 Multi-scale feature fusion module

Our proposed multi-scale feature fusion module adopts the structure of multi-branch parallel and then fusion. The feature maps finally output by the encoder is used as the input of the module. The first branch contains three atrous convolution layers with dilation rates of 1, 2 and 3 respectively. The second branch also uses three groups of such convolution layers, but their dilation rates are 1, 3 and 9 respectively. The third branch first passes through an average pooling layer, and then restores the size of the feature map through an upper sampling layer. Similarly, the fourth branch first passes through a maximum pooling layer, and then the upper sampling layer. Further, after the convolution or pooling layer, all the above four branches pass through a BN layer and a ReLU activation layer to adjust the distribution of the previously output feature maps and accelerate the convergence speed of the network. Finally, the feature maps obtained from the four branches are concatenated, and then a 1× 1 convolution layer is used to realize the fusion of multi-scale features

2.3 Attention weighted semantic enhancement module

In the decoding stage, four attention weighted semantic enhancement（AWSE） modules are used to restore the feature maps output by MSFF module to the size of the original image step by step. In this process, attention mechanism is introduced to guide the up-sampling process of feature maps, which can effectively fuse the shallow features output by each stage of the encoder. The AWSE module we constructed includes two branches: channel attention and spatial attention, which are respectively used to highlight the feature channels or feature locations useful for building discrimination.

The overall structure of the module is shown in Figure 3. The upper branch is a channel attention structure based on SE-Net (Jie et al., 2017). This branch first passes through a global average pooling layer, and then learns the importance of the characteristics of each channel through two full connection layers and nonlinear activation layers, that is, obtains the weight vector representing the importance of the features of each channel; The spatial attention branch first passes through a parallel structure composed of average pooling and maximum pooling layers.

After the output results are concatenated, they pass through a 1 × 1 convolution layer and a sigmoid function to learn the importance of each position in the feature map, that is, to obtain the weight matrix representing the feature importance of each pixel position. Further, the generated weight vector and weight matrix are used to perform channel weighting and spatial weighting operations on the up-sampled input feature map respectively. The output of the former first passes through a 1 × 1 convolution layer to adjust the number of channels, and then is concatenated with the output of the latter. Finally, a 1 × 1 convolution layer is used to realize their fusion. In this way, the feature channels or pixel positions related to building discrimination can be selectively highlighted, and the semantics of buildings in key channels or positions in the feature map can be enhanced.

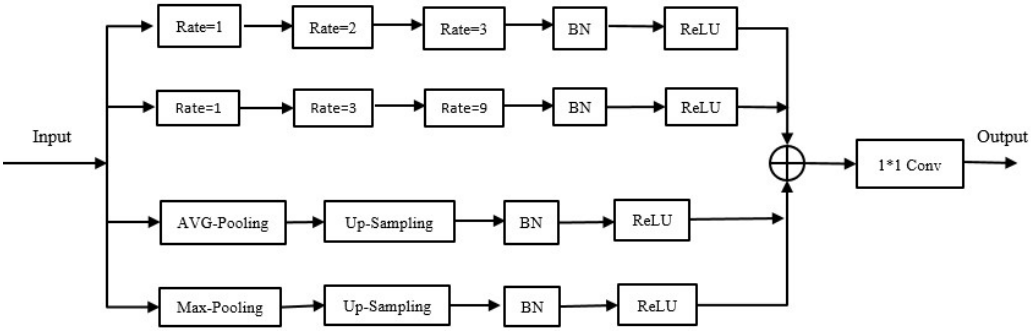

Figure 2. Structure of the MSFF module.

.

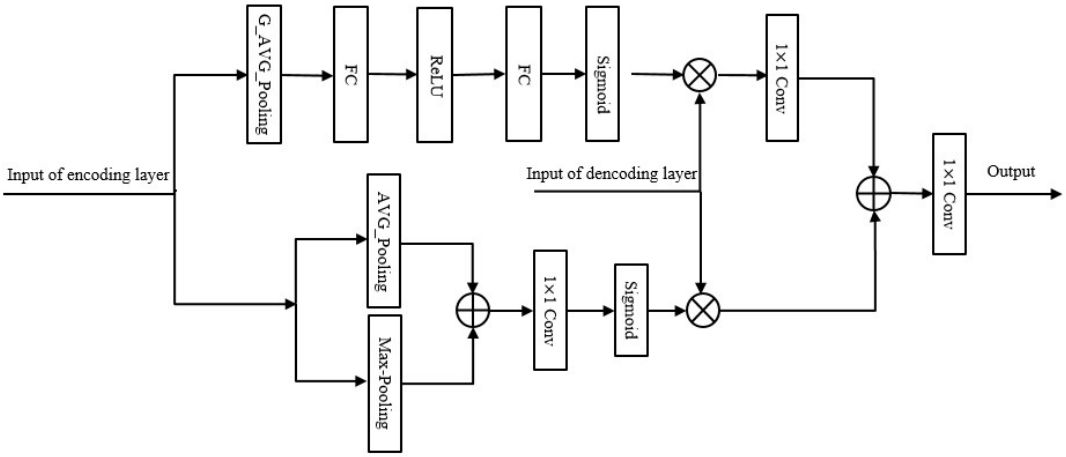

Figure 3. Structure of the AWSE module.

3. EXPERIMENT AND ANALYSIS

To evaluate the effectiveness of the proposed model, two public aerial image datasets, WHU and Inria (Maggiori et al., 2017), were used to perform building extraction experiments. On each dataset, we first train and test the proposed model, and then compare it with the current representative network models including FCN, U-Net, SegNet, DeeplabV3+ (Chen et al., 2018b) and MA-FCN (Wei et al., 2020) in terms of accuracy to evaluate its overall performance. Further, ablation experiments were performed on the WHU dataset to prove the effectiveness of each module in our network model.

3.1 Dataset description and preprocessing

The WHU dataset used contains 8189 sample images with a spatial resolution of 0.3 m (obtained by downsampling). The size of each sample image is 512 × 512. The sample set is divided into three parts, including 4736 training samples, 1036 verification samples and 2416 test samples, with a ratio of approximately 4:1:2.

The Inria data set consists of two parts, a training set and a test set, both of which cover five different regions. The five regions we selected include Austin, Chicago, Kitsap, Tyrol-W and Vienna. There are 36 images in each region, with a size of 5000 ×5000 pixels and a spatial resolution of 0.3m. Due to the large amount of data, we only use 36 images in the training set and cut them into 512 × 512 sample images to form our training set, verification set and test set, with a ratio of 3:1:1. For each region, this provides us with 1750 training set images, 583 verification set images and 583 test set images.

Before training the network, several data preprocessing strategies are implemented to avoid over fitting and improve the generalization ability of the model. The data preprocessing used

includes random rotation at different angles (90°, 180◦, 270◦, and 360◦), random vertical or horizontal mirror flipping, and spectral enhancement of low brightness and low contrast images.

3.2 Experimental setup

The proposed model is constructed and implemented using tensorflow framework with CUDA10.0 and cuDNN7.6. The version of tensorflow is 1.14. The graphics card we use is GeForceRTX1080Ti, and the GPU memory size is 11G. In the training process of the model, the batchsize is set to 4 to fit the video memory size. The binary cross entropy loss is selected as the loss function, Adam is used as the optimizer, and the learning rate is fixed at 0.0001.

3.3 Evaluation metrics

In the evaluation part, the evaluation metrics used include Precision, Recall, F1-Score (F1) and Intersection over Union (IoU). These metrics are widely used as evaluation criteria for semantic segmentation and building extraction. They can be calculated as follows:

$$\text{Precision}=\text{TP}/(\text{TP}+\text{FP}) \quad (1)$$

$$\text{Recall}=\text{TP}/(\text{TP}+\text{FN}) \quad (2)$$

$$\text{IoU}=\text{TP}/(\text{TP}+\text{FP}+\text{FN}) \quad (3)$$

$$\text{F1-Score}=2\times\text{Precision}\times\text{Recall}/(\text{Precision}\times\text{Recall}) \quad (4)$$

In the above formula, TP and FP respectively represent the number of pixels correctly and incorrectly classified as buildings in all test images, while FN and TN refer to the number of pixels incorrectly and correctly classified as background, respectively.

3.4 Experimental results and comparison

The experiment was first conducted on the WHU data set, and the accuracy evaluation statistics of the results are shown in Table 1. The highest value of each metric is highlighted in bold. Compared with other models, our proposed model obtains the highest IoU value and F1-Score, which are 0.8824 and 0.9462 respectively, indicating that our model shows better performance on the whole. In addition, the multi-scale aggregation FCN (MA-FCN) also achieved good performance, indicating that multi-scale features play a very important role in building discrimination. The good performance of the proposed model implies that it has good multi-scale feature representation ability. Figure 4 shows several representative examples of building extraction results. The results obtained by using the proposed model are closer to the ground truth, and the extracted buildings have better shape integrity and boundary accuracy.

Model	IoU	Recall	Precision	F_1
FCN-8s	0.8421	0.9250	0.9135	0.9192
U-net	0.8663	0.9387	0.9298	0.9342
SegNet	0.8572	0.9308	0.9196	0.9252
DeeplabV3+	0.8613	0.9388	0.9291	0.9339
MA-FCN	0.8749	**0.9451**	0.9464	0.9457
Proposed	**0.8824**	0.9432	**0.9492**	**0.9462**

Table 1. Accuracy evaluation results of different models on the WUH dataset.

Compared with the WHU aerial building dataset, the Inria dataset covers five different areas, and its building types and sizes are more diverse. In addition, the contrast between buildings and their background in the image is lower, which adds difficulties to the discrimination of buildings. We conducted experiments on each data set in these five different regions, and calculated the average value of these metrics. The results are shown in Table 2 It can be found that for each metric, the proposed method obtains the highest value, its Precision, Recall and F1-Scores are 0.9465, 0.9480 and 0.9472 respectively, and its IoU value is 0.7103. This shows that our model is still the best. The building extraction results of several example images are shown in Figure 5 By comparing the buildings extracted by different models, it can be found that the proposed model not only preserves a complete shape boundary for large buildings, but also has stronger robustness for small and dark buildings.

Method	IOU	Recall	Precision	F_1
FCN-8s	0.4837	0.8829	0.9015	0.8920
U-net	0.6417	0.9139	0.9304	0.9219
SegNet	0.5920	0.9070	0.9200	0.9133
DeeplabV3+	0.6756	0.9211	0.9346	0.9277
MA-FCN	0.6895	0.9286	0.9446	0.9364
Proposed	**0.7103**	**0.9465**	**0.9480**	**0.9472**

Table 2. Accuracy evaluation results of different models on the Inria dataset.

3.5 Ablation experiments

To verify the effectiveness of our proposed multi-scale feature fusion (MSFF) module and attention weighted semantic enhancement (AWSE) module, the ablation experiments of the two modules were performed on the WHU dataset.

3.5.1 Ablation experiments on the MSFF module

For the ablation experiment of MSFF module, we compared three models: the model obtained by removing the MSFF module from the proposed model (called baseline), the model we proposed (called baseline + MSFF), and the model obtained by replacing MSFF module with atrous spatial pyramid pooling (ASPP) in the proposed model (called baseline + ASPP).

Table 3 shows the performance statistics of these models on the WHU dataset. It can be seen that the model with MSFF module has better segmentation performance. The addition of MSFF module can significantly improve the extraction accuracy of the baseline model, and this module has better performance than ASPP module.

Model	IoU	Recall	Precision	F_1
Baseline	0.8512	0.9302	0.9376	0.9339
Baseline+MSFF	**0.8824**	**0.9512**	**0.9478**	**0.9495**
Baseline+ASPP	0.8735	0.9469	0.9428	0.9448

Table 3. Accuracy statistics of ablation experiments of MSFF module

3.5.2 Ablation experiments on the AWSE module

To evaluate the effectiveness of the proposed ASWE module, we compared the performance of the model with and without (/) AWSE module. Moreover, to verify the generality of this module,

it was also added to the U-Net model and compared with the original U-Net.

As can be seen from Table 4, the ASWE module greatly improves the extraction accuracy of the model. The IoU value of the proposed model with ASWE module is 4.61% higher than that after removing this module, and the Precision, Recall and F1 Score are also greatly improved. Further, for U-Net, after adding the AWSE module, its performance is also improved, and the IoU value and F1-Score increased by 0.76% and 0.96% respectively. These facts show the effectiveness and generalization of ASWE module

Model	IoU	Recall	Precision	F_1
Proposed	**0.8824**	**0.9512**	**0.9478**	**0.9495**
Proposed / ASWE	0.8363	0.9236	0.9342	0.9289
U-Net	0.8663	0.9387	0.9298	0.9342
U-Net + AWSE	0.8739	0.9421	0.9456	0.9438

Table 4. Accuracy statistics of ablation experiments of the AWSE module.

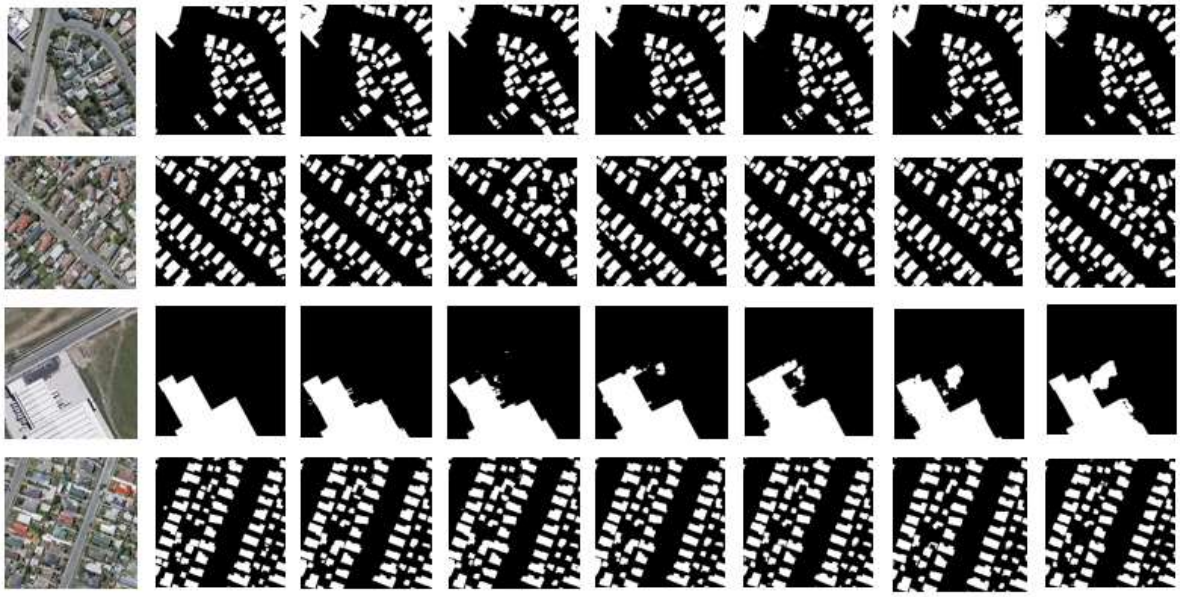

Image Ground Truth Proposed MA-FCN DeeplabV3+ U-Net SegNet FCN

Figure 4. Visualization of building extraction results by different models on the WHU dataset.

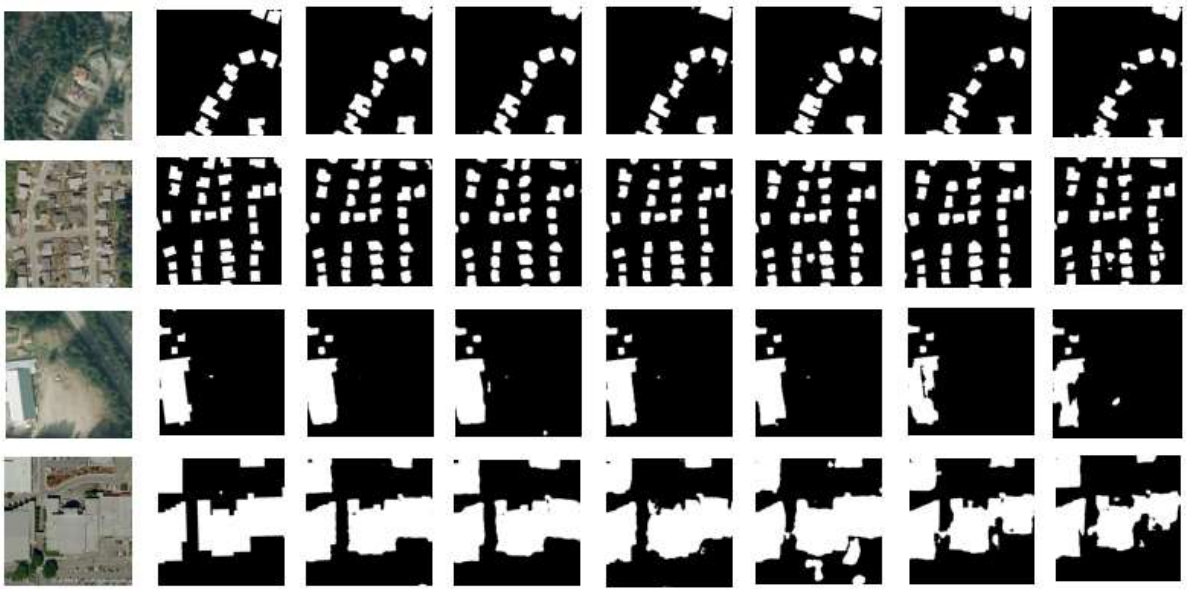

Image Ground Truth Proposed MA-FCN DeeplabV3+ U-Net SegNet FCN

Figure 5. Visualization of building extraction results by different models on the Inria dataset.

4. CONCLUSION

Due to the diversity of the size and type of buildings and the uncertainty of their boundaries, building extraction from high-resolution remote sensing images is still facing great challenges. To deal with these challenges, a semantic segmentation model based on multi-scale feature fusion and enhancement is proposed for the accurate extraction of buildings in complex scenes. Experiments based on WHU and Inria aerial building datasets show that the proposed model can obtain higher extraction accuracy than the current representative building extraction network (including FCN, U-Net, SegNet, DeeplabV3+ and MA-FCN) and has stronger robustness to buildings of different scales and image scenes with low contrast. The extracted buildings retain better shape integrity and boundary accuracy. Further ablation experiments verify the effectiveness of the proposed two modules (including MSFF and AWSE). In the future, we will further optimize our model structure to better improve its building extraction performance in complex scenes.

ACKNOWLEDGEMENTS

This work was partially supported by the MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area (2019004).

REFERENCES

Cai, J., Chen Y., 2021. MHA-Net: Multipath Hybrid Attention Network for building footprint extraction from high-resolution remote sensing imagery. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 14, 5807-5817.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully Connected CRFs. [Online] Available: https://arxiv.org/abs/1412.7062

Chen, L., C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A., L., 2018a. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell*. 40(4), 834–848.

Chen, L., C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation.[Online].Available:https://arxiv.org/abs/1706.05587.

Chen, L., C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. Eur. Conf. Comput. Vision, pp. 801–818.

Deng, W., Shi, Q., Li, J., 2021. Attention Gate Based Encoder-Decoder Network for Automatical Building Extraction. IEEE *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens.*, 14, pp. 2611-2620.

Hermosilla, T., Ruiz, L. A., Recio, J. A., Estornell, J., 2011. Evaluation of Automatic Building Detection Approaches Combining High Resolution Images and LiDAR Data. *Remote Sens*. 3, 1188-1210.

Huang, G., Liu, Z., Laurens, V., Weinberger, K., Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700-4708.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens*., 57(1), 574–586.

Jie, H., Li, S., Gang, S., Albanie, S., 2017. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell*. 42(8), 2011-2023.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 39(4), 640-651.

Liu, P., Liu, X., Liu, M., Shi, Q., Yang, J, 2019. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.*, 11(7): 830.

Liu, Y., Chen, D., Ma, A., Zhong, Y., Xu, K., 2020. Multiscale U-Shaped CNN Building Instance Extraction Framework with Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery. IEEE *Trans. Geosci. Remote Sens*, 59(7), 6106-6120.

Maggiori, E., Tarabalka, Y., 2017. Charpiat, G., Alliez, P., Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp.3226-3229.

Pandey, A., Mishra, A., 2016. Building detection and extraction techniques: A review. In: Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 3816-3821.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. vol. 9351, pp. 234–241.

Shao, Z., Tang, P., Wang, Z., Saleem, N., Sommai, C., 2020. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.*, 12(6): 1–18.

Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. *Int. J. Appl. Earth Observ. Geoinf.*, 24, 58–69.

Wang, J., Yang, X., Qin, X., Ye, X., Qin, Q., 2015. An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery. *IEEE Geosci. Remote Sens. Lett*., 12(3): 487–491.

Wei, S., Ji, S., Lu, M., 2020. From aerial images using CNN and regularization. *IEEE Trans. Geosci. Remote Sens*. 58(3): 2178–2189.

Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2020. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction from Remote Sensed Imagery. IEEE *Trans. Geosci. Remote Sens.* 59(7), 6169-6181.