Building Extraction From Very High-Resolution Remote Sensing Images Using Refine-UNet

Weiyan Qiu, Lingjia Gu, *Member, IEEE*, Fang Gao, and Tao Jiang

***Abstract*— Accurate building extraction from very high-resolution (VHR) remote sensing images plays an important role in urban dynamic monitoring, planning, and management. However, it is still a challenging task to achieve building extraction with high accuracy and integrity due to diverse building appearances and more complex ground background in VHR remote sensing images. Recently, unity networking (UNet) has been proven to be capable of feature extraction and semantic segmentation of remote sensing images. However, UNet cannot achieve sufficient multiscale and multilevel features with larger receptive fields. To address these problems, an improved network based on UNet structure (Refine-UNet) is proposed for extracting buildings from the VHR images. The proposed Refine-UNet mainly consists of an encoder module, a decoder module, and a refine skip connection scheme. The refine skip connection scheme is composed of an atrous spatial convolutional pyramid pooling (ASPP) module and several improved depthwise separable convolution (IDSC) modules. Experimental results on the Jilin-1 VHR datasets with a spatial resolution of 0.75 m demonstrate that compared with UNet, pyramid scene parsing network (PSPNet), DeepLabV3+, and a deep convolutional encoder-decoder architecture for image segmentation (SegNet), the proposed Refine-UNet can obtain more accurate building extraction results and achieve the best precision of 95.1% and intersection over union (IoU) of 87.0%, indicating the great practical potential.**

***Index Terms*— Atrous convolution, building extraction, unity networking (UNet), very high-resolution (VHR) remote sensing imagery.**

I. Introduction

IN RECENT years, with the advancements of remote sensing technology, more detailed land covers information is contained in very high-resolution (VHR) remote sensing images. Consequently, extracting buildings automatically and accurately from VHR data is of great significance in urban dynamic monitoring, land use management, and other geospatial applications [1]. However, with the increase of spatial resolution, more obvious variations of size, shape, and texture of diverse buildings have affected the building extraction accuracy and integrity. Thus, exploiting automatic and robust methods to further improve building extraction accuracy from VHR images is still meaningful and challenging.

Many methods of building extraction have been focused to extract discriminative features to distinguish buildings from nonbuilding. At present, the existing building extraction methods can be sorted as the traditional building extraction methods and deep learning (DL)-based methods [2]. The traditional building extraction methods mainly rely on the low-to-mid-level building features extracted from spectrum, shape, texture, and object-oriented information of buildings [3]. Such building indices include the normalized difference building index (NDBI) [4] based on spectral features and the morphological building index (MBI) [5] based on morphology. The basic idea of MBI is to build a relationship between the implicit characteristics of buildings and the properties of morphological operators. Zhang et al. [6] proposed to use morphologically spatial pattern analysis as postprocessing to further optimize the MBI result and removed the commission errors. To solve the multidirection morphological operations in the MBI, Bi et al. [7] proposed a method to compute building indices using multiscale filtered contours for building extraction. Gu et al. [8] proposed two spectral indices for building extraction, including the normalized spectral building index (NSBI) and the difference spectral building index (DSBI). Chaudhuri et al. [9] combined directional morphological enhancement and internal gray variance clustering techniques to segment buildings. These traditional building extraction methods can detect buildings in a specific scene, intuitively utilizing the characteristics of buildings, including geometry, spectrum, texture, and other information. However, they can hardly capture high-level building features, especially in VHR remote sensing images with a more complex background. Meanwhile, manually designed traditional building extraction methods also have limited capabilities for generalization. The traditional methods cannot address the requirements of VHR datasets and the pursuit of higher performance.

Benefited from the wide application of deep convolutional neural networks (CNN) in remote sensing fields, a series of DL-based methods have been applied to improve the accuracy and efficiency of building extraction. Based on CNN architecture, Long et al. [10] converted a classical CNN to fully convolutional neural networks (FCNs) to achieve pixel-level segmentation. Furthermore, FCN models have become the mainstream in the DL semantic segmentation of remote sensing images. However, the designed limitation of FCNs impacts the accuracy of semantic segmentation. Therefore,

Manuscript received 24 October 2022; revised 28 December 2022 and 31 January 2023; accepted 5 February 2023. Date of publication 9 February 2023; date of current version 20 February 2023. This work was supported in part by the Project of Jilin Province Development and Reform Commission under Grant 2021C044-7 and in part by the Technological Research and Development Projects of Jilin under Grant 20220201017GX. *(Corresponding author: Lingjia Gu.)*

Weiyan Qiu and Lingjia Gu are with the College of Electronic Science and Engineering, Jilin University, Changchun 130012, China (e-mail: qiuwy20@mails.jlu.edu.cn; gulingjia@jlu.edu.cn).

Fang Gao is with Chang Guang Satellite Technology Company Ltd., Changchun 130000, China, and also with the College of Computer Science and Technology, Jilin University, Changchun 130012, China (e-mail: gaofang@163.com).

Tao Jiang is with the Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130012, China (e-mail: jiangtao@iga.ac.cn).

Digital Object Identifier 10.1109/LGRS.2023.3243609

many FCN-based semantic segmentation networks have been developed, such as a deep convolutional encoder-decoder architecture for image segmentation (SegNet) [11], deconvolution network (DeconvNet) [12], and unity networking (UNet) [13]. To further improve the accuracy of image segmentation, DeepLab series models [14], [15], [16] were continuously optimized DL structure and attach context modules, including dense conditional random fields (DCRFs) and atrous spatially parallel pyramids (ASPPs). Moreover, other representative networks were also investigated for building extraction, such as pyramid scene parsing network (PSPNet) [17] that utilizes the pyramid pooling module to capture and fuse multiscale features for improving segmentation accuracy. However, there are still problems despite the promising building extraction results achieved by these state-of-the-art DL-based methods from VHR images due to the complex diversity of building structures.

Following the abovementioned works, a refined network is proposed, which is inspired by the ideas of the UNet architecture and called Refine-UNet. The method performs promisingly in extracting building with high accuracy and integrity. The main contributions of this study are given as follows.

1) A novel Refine-UNet architecture is proposed for accurate building extraction from VHR images, which retains an encoder–decoder structure and incorporates a refine skip connection scheme.
2) In the refine skip connection scheme, the incorporation of an atrous spatial convolutional pyramid pooling (ASPP) module and several improved depthwise separable convolution (IDSC) modules can extract multiscale and multilevel building features from VHR images, which promotes building extraction accuracy.
3) The effectiveness of the proposed method is validated and compared with other DL-based methods based on Jilin-1 VHR remote sensing datasets.

II. Method

A. Architecture of Refine-UNet

UNet [13] is a well-known semantic segmentation method with a popular encoder–decoder structure as shown in Fig. 1, and it is mainly composed of the encoder, decoder, and skip connections. However, the features of UNet are concatenated in the channel layers on the equal level of the encoder and the decoder, which is insufficient for more complete and accurate building extraction. Based on the UNet encoder–decoder structure, the architecture of the proposed Refine-UNet for VHR building extraction is shown in Fig. 2.

It can be seen from Fig. 2 that Refine-UNet is comprised of three parts: the encoder module, the decoder module, and the refine skip connection scheme. The refine skip connection scheme can be further divided into ASPP and IDSC module, as shown in Fig. 2. The building VHR image is first put into the encoder module. The encoder module is composed of four units, including a 3×3 kernel convolution layer and a max-pooling layer. The main function of encoder is to extract multilevel discriminative feature maps and generate low-resolution feature maps with improved discrimination from the input VHR image. Since the number of designed convolutional

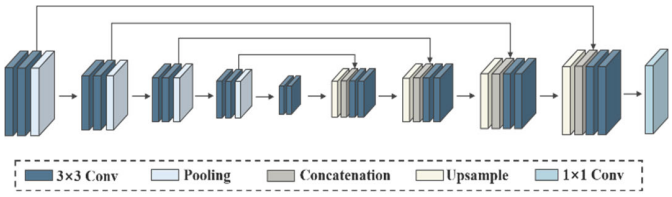

Fig. 1. Architecture of UNet.

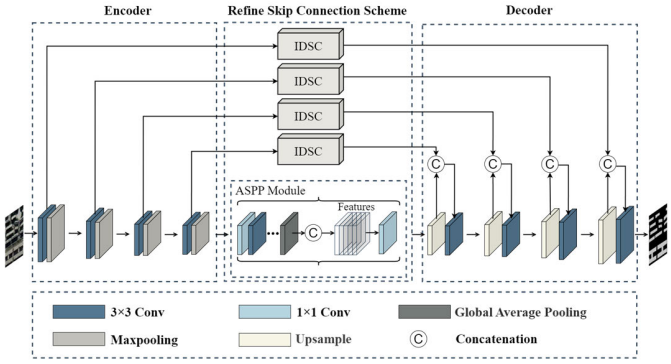

Fig. 2. Architecture of Refine-UNet.

layers in the encoder of Refine-UNet is simplified compared with that of UNet, only low-level features can be extracted from the encoder module. The refine skip connection scheme enables the decoder to flexibly fuse multiscale features of different semantics into the decoder subnetwork and achieve the effective capture of features at multilevels. There are two modules (IDSC and ASPP) contained in the refine skip connection scheme for accurately and efficiently extracting building features. IDSC module is an improvement for DSC composed of depthwise convolution operation and pointwise convolution operation and can extract multilevel features of buildings. The ASPP module samples a given input image in parallel with an atrous convolution at different sampling rates, which can capture rich contextual information by performing pooling operations at different resolutions and can extract multiscale information and obtain global features of buildings. The decoder module is symmetrical with respect to the encoder module. In the decoder module, the refined features from the ASPP module are upsampled and further concatenated with the refined features from the IDSC module. After the concatenation, the features are passed through a 3×3 convolution. The decoder module gradually recovers the details and spatial information of highly discriminant feature maps and obtains the final building extraction result.

B. Design of ASPP Module

The ASPP module is proven to be an effective component to capture more sufficient context information in DeepLabV3+ [16]. In this study, an ASPP module is employed in the refine skip connection scheme between the encoder and decoder modules to further extract multiscale features of images.

Fig. 3 shows the difference between the normal convolution and the atrous convolution. As shown in Fig. 3(a), the receptive field using the normal convolution with the dilation rate of one is only three. By contrast, the input image is independently convolved with different dilation rate settings in the atrous convolution. When the dilation rate is two and four, the

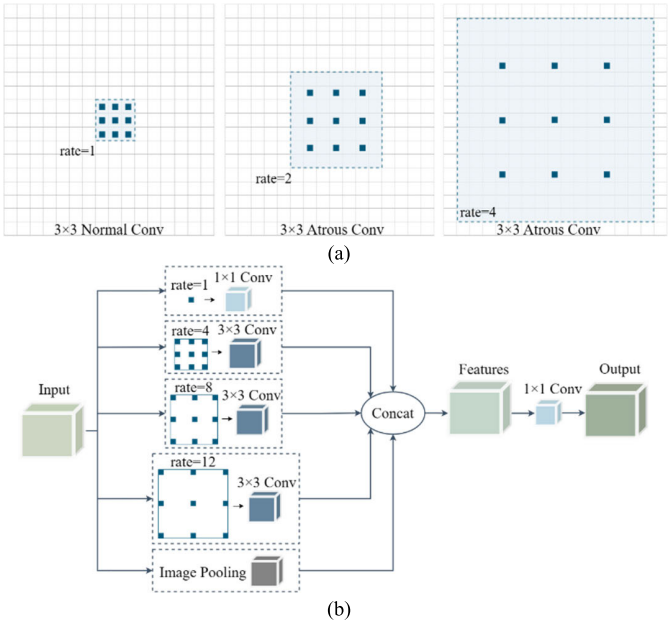

Fig. 3. Atrous convolution and ASPP module. (a) Normal convolution and atrous convolution. (b) Architecture of ASPP module.

large receptive fields with five and nine can be obtained by using the atrous convolution, respectively. Compared with the normal convolution, atrous convolution expands the area of the receptive field due to the difference of dilated rate. Fig. 3(b) shows the convolutional layers in the designed ASPP module, including a 1 × 1 dilation convolution, three 3 × 3 dilation convolutions with a dilation rate of 4, 8, and 12, and a global average-pooling layer. Since the resolution of the feature map would be decreased with the continuous extraction of building features in the encoder module, the selected dilation rate is lower than the default value (6, 12, and 18) in the original ASPP module, which is more conducive to improve extraction features from the feature map with lower resolution. The features extracted at each given dilation rate are processed in a separate branch. At the same time, the global average-pooling operation is performed on the feature maps. Finally, these parallel branches are concatenated together and passed through a 1 × 1 convolution to extract more multiscale features.

C. Design of IDSC Module

Fig. 4(a) shows the architecture of DSC module [18], which is mainly divided into two parts, including depthwise convolution and pointwise convolution. Depthwise convolution applies a single 3 × 3 kernel convolution for each input channel and extracts the spatial information channel-by-channel. The number of feature maps after depthwise convolution is the same as the number of channels of the input layer. However, depthwise convolution only filters input channels, and it does not combine them to create new features. Therefore, a pointwise convolution with a simple 1 × 1 convolution is further employed to create a linear combination of the output feature maps of the depthwise convolution for the cross-channel information integration. Compared with the standard convolutions, the DSC module drastically reduces computation complexity through applying a depthwise convolution followed by a pointwise convolution. Based on DSC, an IDSC module is

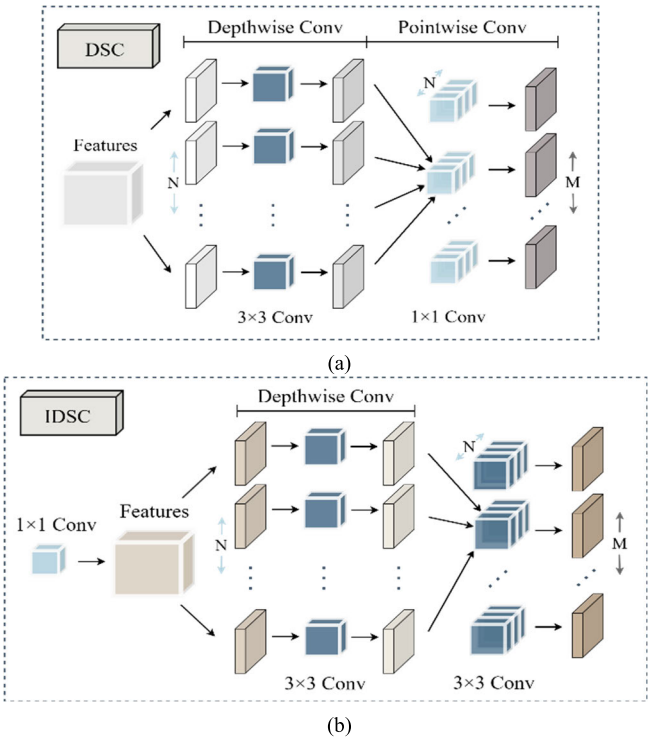

Fig. 4. Architecture of (a) DSC module and (b) IDSC module.

designed, as shown in Fig. 4(b). For the input feature maps from the encoder module, they are first convoluted through a 1 × 1 convolution. Then, the depthwise convolution operation is performed for each input channel. Finally, the output feature maps from the depthwise convolution are convoluted through a 3 × 3 convolution to generate the final feature maps at multilevels.

III. Experiments

A. Study Area and Dataset

The experiments were conducted on Jilin-1 VHR datasets. The parameters of Jilin-1 VHR image are listed in Table I. The Jilin-1 VHR datasets with a spatial resolution of 0.75 m are prepared after preprocessing operations, including radiometric correction, atmospheric correction, orthorectification, and image fusion of panchromatic and multispectral images, which include 1296 RGB images as the training dataset and 216 RGB images as the test dataset. Each image size is 512 × 512 pixels. The ground-truth (GT) data are provided using two semantic classes in the pixel level, including building (white) and nonbuilding (black), which are disclosed only for the training dataset. One group of the original image and its corresponding label is shown in Fig. 5. It clearly shows the typical buildings with regular- and irregular-shaped roofs.

B. Evaluation Criteria

In order to comprehensively evaluate the performance of the proposed network, precision, recall, $F1$ score, and intersection over union (IoU) are used to evaluate the building extraction accuracy [18]. By comparison with GT, true positive (TP), false positive (FP), and false negative (FN), respectively, represent the number of correctly extracted buildings, incorrectly

TABLE I
PARAMETERS OF JINLIN-1 SATELLITE

Satellite	Spatial Resolution/m	Band	Spectral Setting/nm
Jilin-1	0.75	PAN	450-800
	3	Blue	450-510
		Green	510-580
		Red	630-690
		NIR	770-895

(a) (b)

Fig. 5. Buildings training data from Jilin-1 VHR datasets. (a) Image. (b) Label.

extracted buildings, and missed buildings. Using these counts, precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{2}$$

$F1$ score is a representation of the harmonic mean of precision and recall, and is calculated by

$$F1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FN} + \text{FP}}. \tag{3}$$

The value of IoU is applied to characterize the accuracy at a segment level and is calculated by

$$\text{IoU} = \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}}. \tag{4}$$

C. Comparison of Building Extraction

To further demonstrate the effectiveness and feasibility of the Refine-UNet in building extraction, some state-of-the-art semantic segmentation models were selected as comparisons, including PSPNet, SegNet, UNet, and DeepLabV3+. To achieve fair comparisons, the same training datasets were applied to optimize the models and the same test datasets were used to evaluate their performances. Several examples of visual comparison results are given in Fig. 6, where there are five pairs of VHR images, including various buildings with different shapes, sizes, distribution patterns, and texture characteristics. In the whole, all the DL-based methods can perform well for regular building extraction. However, for the irregular building extraction, the proposed Refine-UNet achieves better building extraction accuracy and visual effects compared with other DL-based methods, such as the region marked by the red square in Fig. 6. Due to the effect of shadow and similarity in background features at the edges of buildings, the areas detected by all methods have inaccurate results, especially around the edges. The building extraction results of UNet are more accurate than those of PSPNet, SegNet, and DeepLabV3+. Based on the improvement of UNet, Refine-UNet can extract complex buildings more completely than the

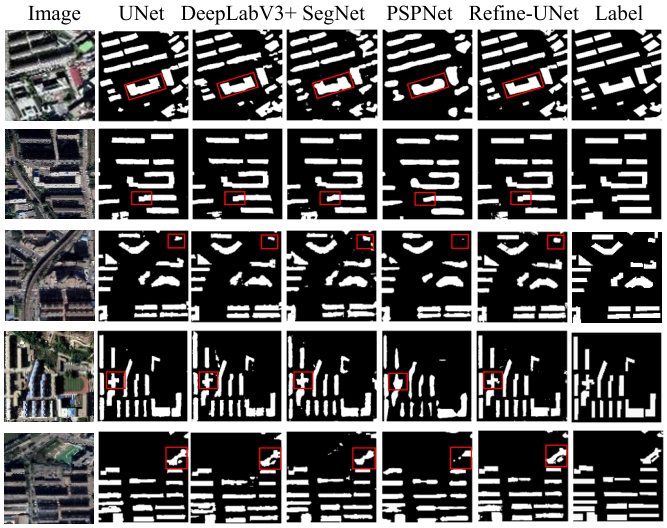

Fig. 6. Comparison of different network results.

TABLE II
AVERAGE ACCURACY OF COMPARISON EXPERIMENT

	Precision	Recall	F1	IoU
UNet [13]	0.887	0.866	0.876	0.785
DeepLabV3+ [16]	0.864	0.838	0.850	0.745
SegNet [11]	0.860	0.875	0.867	0.773
PSPNet [17]	0.773	0.720	0.745	0.596
Refine-UNet	**0.951**	**0.910**	**0.930**	**0.870**

comparative methods, which can extract more complete shape and location of regular and irregular buildings.

The comparison of quantitative analysis results is presented in Table II. Among the compared methods, PSPNet gets the most unsatisfied results, several buildings are not detected by PSPNet, and the edges of the buildings are blurred. The performance of SegNet is better than that of DeepLabV3+, and the performance of UNet is better than that of SegNet. The precision, recall, $F1$, and IoU of Refine-UNet are further improved by 6.4%, 4.4%, 5.4%, and 8.5%, respectively, compared with those of UNet. Refine-UNet can accurately extract multiscale and multilevel building features from VHR images, which shows more robust performance under complex scenarios.

D. Ablation Experiment

To verify the effectiveness of Refine-UNet, a series of ablation experiments were carried out through the module changes in the refine skip connection scheme. Fig. 7 shows the sample results of the ablation experiment. DSC represents that the ASPP module was removed and the IDSC module was replaced by the DSC module in the refine skip connection scheme. IDSC represents that the ASPP module was removed from the refine skip connection scheme, and ASPP represents that the IDSC module was removed from the refine skip connection scheme. From Fig. 7, it can be seen that there is more misdiscrimination of buildings only using the DSC module, IDSC module, or ASPP module than that using Refine-UNet. Although there are inaccurate results especially around the edges using Refine-UNet, they are relatively close to the actual situation.

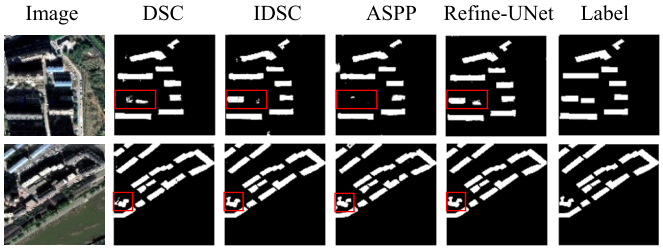

Fig. 7. Ablation experimental results.

TABLE III
AVERAGE ACCURACY OF ABLATION EXPERIMENT

	Precision	Recall	F1	IoU
DSC	0.897	0.868	0.881	0.789
IDSC	0.913	0.891	0.902	0.822
ASPP	0.896	0.824	0.855	0.750
Refine-UNet	**0.951**	**0.910**	**0.930**	**0.870**

The average quantitative evaluation results of the ablation experiment for test datasets are shown in Table III. Due to the removal of modules in the refine skip connection scheme, the feature information extraction is insufficient and the accuracy is obviously degraded. Compared with the DSC module, the IDSC module of the refine skip connection scheme helps to improve precision and IoU by approximately 1.6% and 3.3%, respectively. This shows that the proposed IDSC module has a higher effectiveness. From Table III, the results indicate that the ASPP module helps to improve precision and IoU by approximately 3.8% and 4.8%, respectively, and the IDSC module helps to improve precision and IoU by approximately 5.5% and 12%, respectively. The contribution of the IDSC module to the improvement of building detection accuracy is higher than that of the ASPP module. The Refine-UNet can extract sufficient multilevel and multiscale features to accurately extract buildings, which can be contributed to the refine skip connection scheme.

IV. CONCLUSION

In this letter, a novel network called Refine-UNet is proposed based on the combination of encoder–decoder structure and refine skip connection scheme composed of IDSC and ASPP module to accurately extract buildings from VHR remote sensing images. Through adding IDSC and ASPP modules, the proposed network enhances the multiscale and multilevel feature extraction capability and improves the accuracy of buildings extraction.

Experimental results demonstrate that the proposed method achieves better performance on Jilin-1 VHR datasets, in terms of both numerical metrics and visual results. However, it still needs further improvement in the quality and accuracy of buildings extraction, especially in the boundary areas and shapes of irregular buildings. Future work attempts to combine the digital elevation model (DEM) and shadows of building as supplementary information to further improve the segmentation accuracy.

ACKNOWLEDGMENT

The Jilin-1 very high-resolution (VHR) remote sensing imagery is provided by Chang Guang Satellite Technology Company Ltd. (CGSTL), Changchun, China.

REFERENCES

[1] S. Chen, W. Shi, M. Zhou, M. Zhang, and Z. Xuan, "CGSANet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1526–1542, 2022.

[2] Y. Wang, L. Gu, X. Li, and R. Ren, "Building extraction in multitemporal high-resolution remote sensing imagery using a multifeature LSTM network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1645–1649, Sep. 2021.

[3] J. Wang, X. Yang, X. Qin, X. Ye, and Q. Qin, "An efficient approach for automatic rectangular building extraction from very high resolution optical satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 487–491, Mar. 2015.

[4] C. Li et al., "Attention enhanced U-Net for building extraction from farmland based on Google and WorldView-2 remote sensing images," *Remote Sens.*, vol. 13, no. 21, p. 4411, Nov. 2021.

[5] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," *Photogramm. Eng. Remote Sens.*, vol. 77, no. 7, pp. 721–732, 2011.

[6] Q. Zhang, X. Huang, and G. Zhang, "A morphological building detection framework for high-resolution optical imagery over urban areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 9, pp. 1388–1392, Sep. 2016.

[7] Q. Bi, K. Qin, H. Zhang, Y. Zhang, and K. Xu, "A multi-scale filtering building index for building extraction in very high-resolution satellite imagery," *Remote Sens.*, vol. 11, no. 5, p. 482, 2019.

[8] L. Gu, Q. Cao, and R. Ren, "Building extraction method based on the spectral index for high-resolution remote sensing images over urban areas," *J. Appl. Remote Sens.*, vol. 12, no. 4, p. 5501, Nov. 2018.

[9] D. Chaudhuri, N. K. Kushwaha, A. Samal, and R. C. Agarwal, "Automatic building detection from high-resolution satellite images based on morphology and internal gray variance," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1767–1779, May 2016.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[12] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Aug. 2015, pp. 1520–1528.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[15] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[16] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[18] Y. Qiu, F. Wu, J. Yin, C. Liu, X. Gong, and A. Wang, "MSL-Net: An efficient network for building extraction from aerial imagery," *Remote Sens.*, vol. 14, no. 16, p. 3914, Aug. 2022.