Scale-Robust Deep-Supervision Network for Mapping Building Footprints From High-Resolution Remote Sensing Images

Haonan Guo, Xin Su, Shengkun Tang, Bo Du, and Liangpei Zhang

Abstract—**Building footprint information is one of the key factors for sustainable urban planning and environmental monitoring. Mapping building footprints from remote sensing images is an important and challenging task in the earth observation field. Over the years, convolutional neural networks have shown outstanding improvements in the building extraction field due to their ability to automatically extract hierarchical features and make building predictions. However, as buildings are various in different sizes, scenes, and roofing materials, it is hard to precisely depict buildings of varied sizes, especially in large areas (e.g., nationwide). To tackle these limitations, we propose a novel deep-supervision convolutional neural network (denoted as DS-Net) for extracting building footprints from high-resolution remote sensing images. In the proposed network, we applied deep supervision with an extra lightweight encoder, which enables the network to learn representative building features of different scales. Furthermore, a scale attention module is designed to aggregate multiscale features and generate the final building prediction. Experiments on two publicly available building datasets, including the WHU Building Dataset and the Massachusetts Building Dataset, show the effectiveness of the proposed method. With only a 0.22-M increment of parameters compared with U-Net, the proposed DS-Net achieved an IoU of 90.4% on the WHU Building Dataset and 73.8% on the Massachusetts Dataset. DS-Net also outperforms the state-of-the-art building extraction methods on the two datasets, indicating the effectiveness of the proposed deep supervision and scale attention.**

Index Terms—**Building footprint extraction, convolutional neural network, deep learning, remote sensing image.**

I. Introduction

Building footprint extraction is one of the research hotspots in the remote sensing field due to the broad application of building information [1], [2]. In recent years, the demand for generating precise and up-to-date building footprint information is increasing, due to its potential for sustainable urban planning, smart city construction, and automatic driving [3]. Traditionally, the building footprint information is generated by manual surveying and vectorization, which is expensive and time-consuming, especially for large-scale practice [4]. To facilitate building footprint generation, extensive researches have focused on automatically extracting building footprints from remote sensing images [5].

Conventional building extraction methods mainly include two steps: handcrafted feature extraction and classification. Features that can well represent buildings are firstly extracted from the input image, such as color [6], geometry [7], texture [8], context [9], and shadow [10], followed by a classifier that classifies the features into the building and nonbuilding categories. For example, Zha *et al.* [6] designed a normalized difference built-up index for mapping urban areas using spectral information provided by the multispectral remote sensing images. Huang *et al.* [9] proposed a morphological building index for building extraction by modeling the implicit features using building morphological operators. Ok [11] built the spatial relationship between buildings and shadows and adopted the graph cut algorithm to extract building regions. Although these methods work well to some extent, they rely heavily on the selection of handcrafted features. However, the hand-crafted feature selection process is usually subjective and empirical, hampering the robustness of these traditional methods [12], [13].

Over the years, the rapid development of deep learning methods in the computer vision field provides an alternative technique for extracting building footprints in a time-saving and inexpensive way [14], [15]. Compared with the traditional methods, the convolutional neural network (CNN) can automatically extract hierarchical features from the raw images [16]. Moreover, the feature extraction and classification processes of CNNs are integrated into a single model and can be trained in an end-to-end manner [17]. Moreover, the fully convolutional networks (FCNs) derived from CNNs are capable of making pixel-wise classifications [18]. However, due to the large intra-class variance and low inter-class variance of remote sensing images, it has been proven that directly applying deep methods designed for natural images to remote sensing images leads to accuracy drop [2], [19], [20].

To tackle this problem, various studies have tried to improve the traditional FCNs for building footprint extraction from very

Manuscript received June 27, 2021; revised July 27, 2021; accepted August 21, 2021. Date of publication September 10, 2021; date of current version October 15, 2021. This work was supported in part by Chang'an University, Xi'an, China, through the National Key Research and Development Program of China under Grant 2020YFC1512000 and in part by the National Natural Science Foundation of China under Grant 61801332. *(Corresponding author: Xin Su.)*

Haonan Guo and Liangpei Zhang are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: guohnwhu@163.com; zlp62@whu.edu.cn).

Xin Su and Shengkun Tang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xinsu.rs@whu.edu.cn; shengkuntang@whu.edu.cn).

Bo Du is with the National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430079, China (e-mail: gunspace@163.com).

Digital Object Identifier 10.1109/JSTARS.2021.3109237

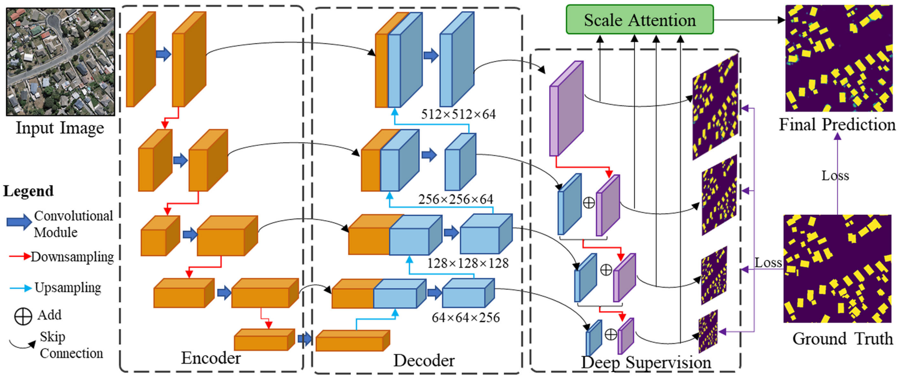


Fig. 1. Structure of the proposed deep-supervision network.

high-resolution remote sensing images. For example, Bittner *et al.* [21] introduced fused-FCN4s to fuse the information from the early layers to reduce the information loss in the building extraction process. To balance the accuracy and network complexity, Shrestha and Vanneschi [22] designed an FCN architecture with the exponential linear unit as the activation function for building extraction. Moreover, the conditional random fields were adopted to suppress the false predictions and sharpen the building boundaries. With the development of semantic segmentation algorithms, the encoder-decoder framework of FCNs has shown outstanding performance on the localization of the building footprints due to the use of skip connection to fuse local-detailed features in the decoder stage [23]. Based on the encoder-decoder architecture, Liu *et al.* [24] designed an Inception-style Res-Net and a densely upsampling module to aggregate the spatial information in the building extraction process. Taking advantage of the properties of capsule networks, Yu *et al.* [25] combined the feature pyramid network with capsule network to fuse different levels of capsule features for building extraction.

As buildings vary in different colors, shapes, and scales, how to fully utilize the multiscale features extracted from the encoder network is one of the key factors for maintaining both building localization accuracy and building edge accuracy [26]. Low-level features are adequate with detailed information, while high-level features are rich in global semantic information [27]. In order to fuse multiscale features, Sun *et al.* [28] designed an SVM-based Multiscale CNN for fusing multiscale building prediction. Li *et al.* [29] designed a multiple-feature reuse network that enables the direct use of hierarchical features in each layer. Inspired by the state-of-the-art (SOTA) HRNet [30], Zhu *et al.* [31] proposed a parallel multipath network to extract high-level semantic features while retaining spatial-detailed building information. Liu *et al.* [32] modified ResNet-101 by a spatial residual inception module to aggregate multiscale contexts. Moreover, depthwise separable convolutions and convolution factorization were introduced to reduce the GPU memory use and increase the computation speed. For better hierarchical building feature representation, Zhang *et al.* [33] designed a local-global dual-stream network that combines local-global branches with deep feature sharing. To tackle the building instance segmentation challenge, Liu *et al.* [34] introduced a multiscale U-shape network for multiscale building instance segmentation. In addition, an edge-constrain network was designed to further refine the building edges. Although much effort has been made to fuse the multiscale building features, there is still a tradeoff between the benefits of low-level details and high-level semantics; how to adaptively integrate information of different scales remains a challenge.

To tackle these problems, in this article, we designed a novel deep-supervision FCN (denoted as DS-Net) for building footprint extraction. The proposed DS-Net is based on the encoder-decoder architecture. In DS-Net, we designed a lightweight and effective deep supervision sub network to boost the model's robustness to buildings of different scales, considering the large scale-variant among buildings in the real-world datasets. This deep supervision subnetwork is capable of generating multiscale building predictions, which enables the model to learn more representative deep features of buildings with varying scales. Moreover, the gradient generated by our proposed deep supervision strategy can flow through the whole network. Furthermore, a scale aggregation module is proposed to calculate the contribution of each scale and generate the final building predictions by aggregating the multiscale predictions. The overall architecture of our proposed DS-Net is shown in Fig. 1. Compared with the traditional methods that directly apply supervision to the decoder network [35], the main contributions of our proposed method are as follows.

1) We proposed a DS-Net for accurate and effective building footprint extraction through the deep-supervision convolutional neural network. Different from the traditional methods that directly apply supervision on the decoder, we designed a lightweight deep supervision subnetwork for generating multiscale building predictions combining deep features and high-resolution features. In this way, the gradients generated by multiscale predictions can flow through the whole network by backpropagation and will strengthen building feature representation.
2) We designed a scale attention module (SAM) to compute global-local attention of different scales. The scale attention vectors represent the contribution of each scale. The final building extraction results can be generated by combing the attention vectors and the multiscale building predictions.
3) The proposed DS-Net outperforms other SOTA methods on two publicly available building datasets, including the WHU Building Dataset and the Massachusetts Building

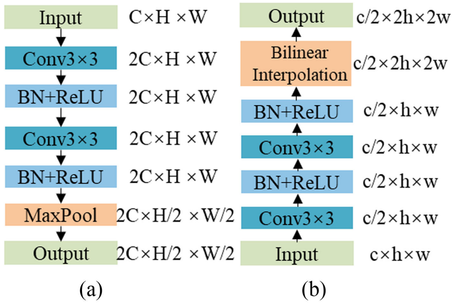


Fig. 2. Structure of the convolutional modules in the (a) encoder subnetwork and (b) decoder subnetwork.

Dataset, which indicates the effectiveness of our proposed method.

The rest of this article is organized as follows. Section II introduces the structure of our proposed DS-Net in detail. Section III describes the dataset and the experimental results, followed by the analysis of the experimental results, and the conclusions are presented in Section IV.

II. Methodology

In semantic segmentation, FCNs have shown great potential in building footprint extraction by making pixel-wise predictions. Inspired by these deep learning methods, we proposed an end-to-end FCN for building footprint extraction namely DS-Net. Fig. 1 shows the architecture of the proposed DS-Net, which includes three components: an encoder–decoder architecture, a deep supervision subnetwork, and an SAM. The input image is first fed into the encoder subnetwork for hierarchical building feature extraction. Then, the decoder subnetwork is used to refine the feature resolution and generate multiscale refined features. After that, the features are fed into the deep supervision subnetwork and generate multiscale building predictions combing low-level features and high-level refined features. Finally, the SAM computes the contributions of each scale and generates the final building prediction. In this section, these three parts are introduced in succession.

A. Encoder-Decoder Architecture

The encoder–decoder architecture in DS-Net follows the basic architecture of U-Net [36], which is an effective network that follows the encoder–decoder design for binary semantic segmentation. Various U-Net-based building extraction methods such as [37] and [35], have shown their effectiveness in building extraction. Here we will introduce the encoder–decoder architecture applied in our method.

Given an input remote sensing image, a convolution layer of kernel size (3,3) is first applied to increase the number of features to 64 channels. After each convolutional operation, a batch norm layer and a nonlinear activation layer are applied to increase the features' nonlinear representation. The output features are then fed to another convolutional layer of kernel size (3,3), followed by a max-pooling operation that halves the input feature size. Then follows four convolutional modules. The architecture of the convolutional modules is shown in Fig. 2(a). The number of the channels in the output features is twice the number of those in the input features; while spatial resolution is half that of the input features. As a result, the output features of the encoder network contain 1024 channels, and the spatial resolution is one-sixteenth of the input image.

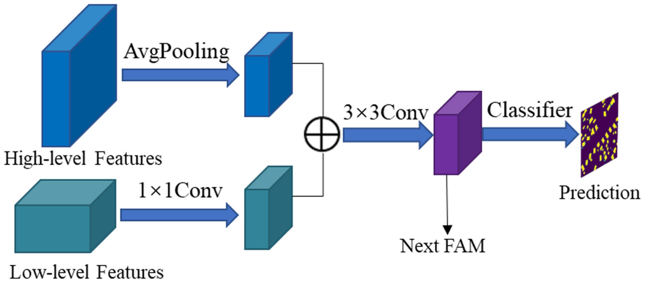


Fig. 3. Architecture of the proposed feature aggregation module (FAM).

Then follows the decoder network, which contains several convolutional modules. The convolutional modules in the decoder take the concatenation of deep features and features transmitted from the encoder subnetwork as input. As shown in Fig. 2(b), the architecture of the convolutional module in the decoder subnetwork is similar to the module in the encoder. For parameter efficiency, we adopted bilinear interpolation instead of deconvolution for up-sampling. The output features are half of the input features in the number of channels, and twice of the input features in spatial resolution. It should be noted that the output features of the last convolutional layer have 64 channels rather than 32 to contain more representative building features.

B. Deep Supervision Subnetwork

The encoder–decoder architecture is capable of extracting hierarchical representative building features, which are beneficial to the final prediction. Considering the large intraclass scale variance of buildings in the real-world datasets, applying the multiscale supervision strategy can enhance models' generalization abilities to buildings of different scales. However, if we directly apply multiscale supervision to the decoder network like [35], the shortcomings lie in two aspects. First, the gradients of generated by multiscale building supervision can only update parameters of the shallower layers, hampering deep layers to benefit from multiscale predictions. Moreover, with the limited inception field of the shallow layers, the extracted features could not well represent buildings, and the predictions directly made by shallow features may be unconvincing. To tackle these problems, we designed an extra deep supervision subnetwork to generate multiscale building predictions and make deep supervision. The proposed deep supervision subnetwork makes up-bottom predictions by refining low-level predictions with high-level features. In this way, the gradients generated by deep predictions can flow through the whole network, which further enhances the model's robustness to buildings of different scales.

The proposed deep supervision subnetwork includes three successive FAMs, as shown in Fig. 3. The proposed FAM module takes the high-level features and the low-level features transmitted from the decoder subnetwork as the input. For the high-level features, an average pooling operation is adopted to halve spatial

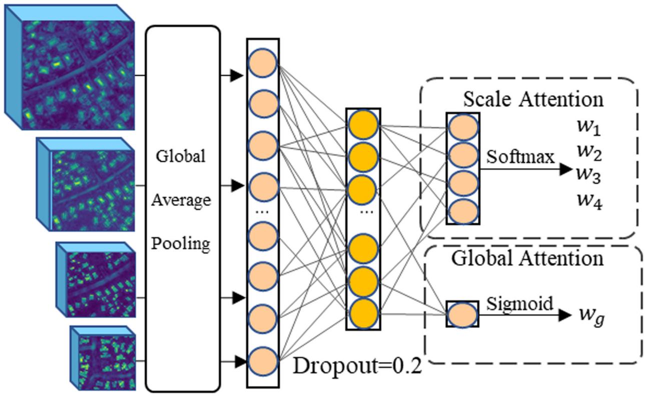


Fig. 4. Structure of SAM.

resolution. For the low-level features, a convolutional layer of kernel size (1,1) is adopted to reduce the feature channel to 64. Then, the low-level features and high-level features are added, followed by a convolutional layer of kernel size (3,3) to aggregate multilevel features. The output features of FAM are the high-level features of the next FAM. A classifier is applied to the output features and makes multiscale building predictions. FAM aggregates multilevel features as follows:

$$F^t = C_3^{64}\left(C_1^{64}\left(f^t\right) + Pool\left(F^{t-1}\right)\right),\ \mathrm{t} \in \{2, 3, 4\} \tag{1}$$

where C_n^i denoted convolutional layers of kernel size n and output channel $i.f^t$ represents features of scale t that are transmitted from the decoder subnetwork. F^t represents the building features of scale t after feature aggregation. It should be noted that FAM is designed lightweight for parameter efficiency. The channel of features in each FAM remains 64, and the whole deep supervision subnetwork only increases 0.2M in parameters.

C. Scale Attention Module

The deep supervision module is capable of generating multiscale building predictions in an up-bottom manner. However, how to integrate multiscale predictions remains a challenge. High-level predictions are high-resolution with detailed edge information. On the contrary, low-level predictions may contain better localization but are low-resolution. If we simply add the multiscale predictions equally, it may fail to fully utilize the benefits of the multiscale predictions and introduce false predictions. To tackle this problem, we designed an SAM to automatically decide the contribution of each scale.

More specifically, the multiscale building features, each of which contains 64 channels, are fed into a global average pooling module to capture global representations of different scales, as shown in Fig. 4. Then the representative vectors are concatenated and fed into a fully connected layer with a weight parameter of 256 × 64. To avoid overfitting, we also adopted a dropout strategy with a probability of 0.2. Then we fed the output vector to two separate fully connected layers, and generate two attention vectors of size 1 × 4 and 1 × 1, respectively. The scale attention vector of size 1 × 4 determines the significance of each scale, and the refined building prediction can be generated as follows:

$$\begin{aligned} P^r &= w_1 \cdot P^1 + w_2 \cdot P^2 + w_3 \cdot P^3 + w_4 \cdot P^4 \\ \mathrm{w} &= \mathrm{softmax}\left(v_{1\times 4}\right) \end{aligned} \tag{2}$$

where P^i denotes the building prediction of scale i. w is generated by applying softmax operation to the scale attention vector, and w_i represents the weight of scale i. Based on our observation, high-resolution predictions contain more detailed structural information and should contribute more to the final prediction. Thus, we designed a global attention module to determine the contribution between the refined building prediction and the highest-resolution prediction. The final prediction can be generated as follows:

$$P^f = \mathrm{sigmoid}\left(v_{1\times 1}\right) \cdot P^1 + \left(1 - \mathrm{sigmoid}\left(v_{1\times 1}\right)\right) \cdot P^r. \tag{3}$$

III. Experimental Results

A. Dataset Descriptions

To evaluate the performance of the proposed method, we validated our proposed method on two publicly available building datasets, including the WHU Building Dataset [37] and the Massachusetts Building Dataset [38]. Both datasets contain high-resolution remote sensing images and their corresponding building labels, as shown in Fig. 5.

1) WHU Building Dataset: Ji *et al.* [37] proposed the WHU Building Dataset for building footprint extraction. We selected the aerial subset of the WHU Building Dataset, which covers various buildings of different appearances and scales. The dataset covers over 450 km 2 areas, with more than 187 000 buildings. The spatial resolution of the aerial images is 0.3 m, with each image size of 512 × 512. The dataset includes 8188 tiles, including 4736 for training, 1036 for validation, and 2416 for testing. Our experiments are conducted on the original dataset partition.
2) Massachusetts Building Dataset: The Massachusetts Building Dataset is proposed by Mnih *et al.* [38]. The dataset contains 151 R-G-B remote sensing images and their corresponding building masks covering approximately 340 km^2 in Boston, Massachusetts. Each of the images contains 1500 × 1500 pixels, and the spatial resolution is 1 m. The whole dataset was split into the training set (137 tiles), validation set (4 tiles), and test set (10 tiles). Following the data preprocessing method in [20], we cropped the dataset into 256 × 256 patches with an overlap rate of 0.5. Then the tiles without buildings were removed. As a result, we obtained 14 705 patches for training, 454 patches for validation, and 1116 patches for testing. Some examples of the Massachusetts Building Dataset are shown in Fig. 5(b).

B. Experimental Details

The proposed DS-Net was implemented using Pytorch [39] on two NVIDIA RTX 2080Ti GPUs. The parameters of the network

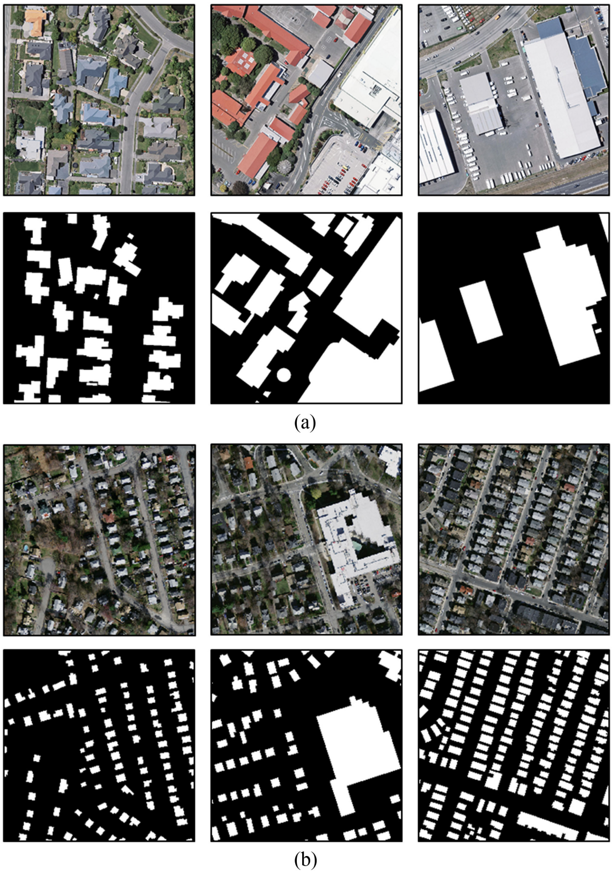


Fig. 5 Examples of the WHU Building Dataset and the Massachusetts Building Dataset.

were optimized with Adam optimizer [40]. The initial learning rate was set to 0.01 and was reduced to a quarter when the accuracy on the validation set failed to increase. The weight parameters in the DS-Net were initialized using kaiming uniform [41], and the bias is initialized with zeros. For the fully connected layers in the SAM, the weights are initialized using a normal distribution with 0.01 as the standard deviation and 0 as the mean value.

The proposed DS-Net is capable of generating multiscale building predictions, and the final prediction can be generated by combing scale attention vectors and multiscale predictions. We adopted binary cross-entropy loss for training DS-Net, which can be described as

$$\begin{aligned}\mathrm{L} = \mathrm{c}_1 \cdot \mathrm{CE}\left(y_{512}, P^f\right) + \mathrm{c}_2 \cdot \mathrm{CE}\left(y_{512}, P^1\right)\\ + \mathrm{c}_3 \cdot \mathrm{CE}\left(y_{256}, P^2\right) + \mathrm{c}_4 \cdot \mathrm{CE}\left(y_{128}, P^3\right)\\ + \mathrm{c}_5 \cdot \mathrm{CE}\left(y_{64}, P^4\right)\end{aligned} \tag{4}$$

where P^i denotes the building prediction of scale i, and P^f denotes the final prediction generated by (3). y_j represents the building ground truths that are resampled to j × j using bilinear interpolation. c denotes the weight of the loss function among different scales. High-resolution predictions contain more structural details, so c_1, c_2 are set to 1. $\mathrm{c}_3, \mathrm{c}_4, \mathrm{c}_5$ are set to 0.3 for deep supervision. After the model converged on the training set, the performance of DS-Net is assessed on the test set using several evaluation metrics.

C. Evaluation Metrics

In this study, four evaluation metrics, including intersection over union (IoU), precision, recall, and F1-score, were selected to evaluate the performance of the proposed DS-Net and other comparative methods from different aspects.

First, the confusion matrix between the building predictions and the building ground truths were calculated, including true-positive (TP), false-positive (FP), and false-negative (FN). Then the IoU, precision, recall, and F1-score can be calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{5}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

$$\mathrm{F1} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{8}$$

D. Comparison Methods

Five SOTA semantic segmentation methods, including SegNet [42], Deeplab v3+[43], Res-U-Net [44], SRI-Net [2], and MA-FCN [35] are selected as the comparison methods to evaluate the performance of the proposed DS-Net. These methods have been proven effective in semantic segmentation and/or building footprint extraction. Here we will give a brief description of these methods.

1) SegNet: SegNet [42] is a fundamental FCN for semantic segmentation designed by Badrinarayanan in 2015. SegNet is designed based on the encoder–decoder architecture. In the encoder part of SegNet, the pooling indices are recoded and are transferred to the unpooling layers. In this way, the upsampling operation can be done without model learning, and the integrity of high-frequency details can be preserved in the segmentation process.
2) Deeplab v3+: Deeplab v3+[33] is the masterpiece of the Deeplab series. By retaining the advantages (e.g., atrous spatial pyramid pooling) of the previous version of Deeplab, the network used a modified Xception model as the backbone and improved the decoder module. Deeplab v3+ reached SOTA performance on several semantic segmentation benchmarks.
3) Res-U-Net: Xu *et al.* [44] introduced Res-U-Net model for building footprint extraction in 2018. Res-U-Net takes the deep residual network (ResNet) as the encoder, which is effective in avoiding gradient vanishing and gradient explosion phenomena. Moreover, handcrafted features and guider filters are designed to further improve the accuracy of building extraction. Experimental results showed that Res-U-Net is more effective in building extraction.
4) SRI-Net: The spatial residual network [32], termed as SRI-Net, is a building extraction FCN designed by Liu *et al.* SRI-Net is capable of retaining global semantic

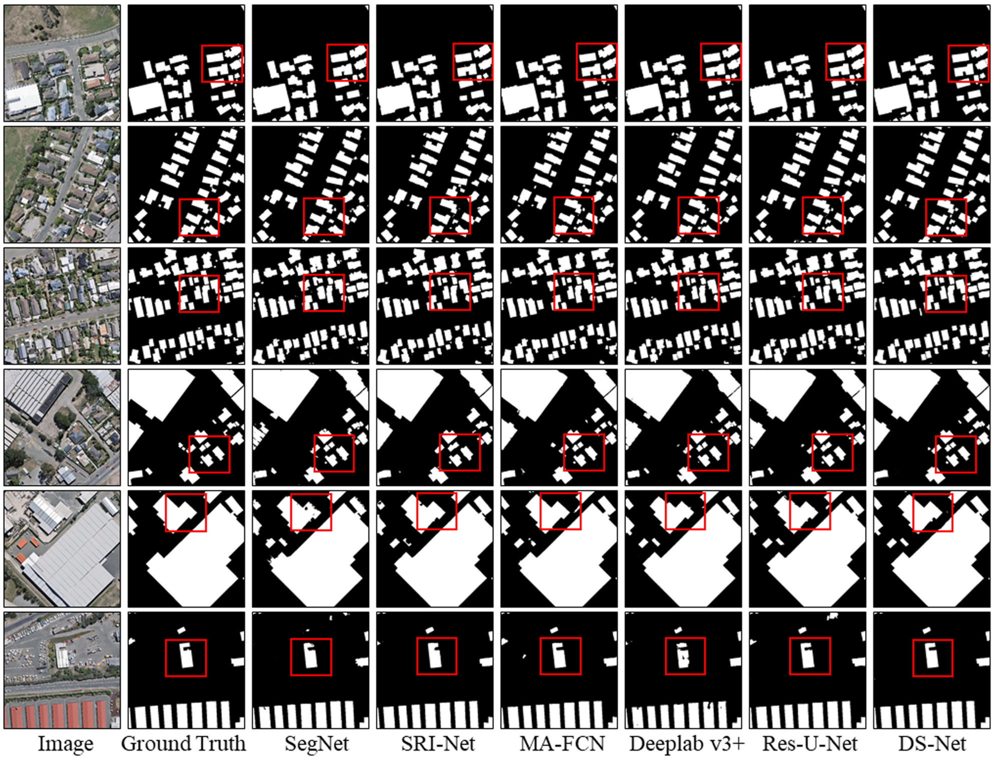


Fig. 6. Visualizations of building extraction results by our proposed DS-Net and other comparison methods on the WHU dataset.

information and local details by the proposed spatial residual inception module. Moreover, depthwise separable convolution was introduced to improve computational efficiency. Experimental results show promising performance for building extraction on a large scale.

5) MA-FCN: Wei *et al.* [35] designed MA-FCN for automatic building footprint extraction and delineation from aerial images in 2020, which is one of the most recent building extraction networks. MA-FCN adopted VGG-16 as the encoder, and multiscale feature aggregation was adopted in the decoder part to generate scale-robust building prediction. Meanwhile, postprocessing strategies were introduced to refine and vectorize the segmentation maps. MA-FCN reached SOTA performance on the WHU dataset.

E. Results and Analysis

1) Results on the WHU Building Dataset: Fig. 6 displays some visualization results of our proposed DS-Net and the SOTA comparison methods on the WHU test dataset. As shown in Fig. 6, DS-Net obtains the most precise building boundary with the least false prediction pixels visually. It can be seen from the highlighted red blocks in Fig. 6 that DS-Net is capable of segmenting buildings of different scales precisely. From rows 1–4 of Fig. 6, we can see that the boundaries of the small buildings can be depicted completely by the DS-Net. By aggregating multiscale building features, the proposed DS-Net is capable of detecting tiny building objects from the remote sensing images, as proven in the first and fourth row of Fig. 6. On the contrary, Deeplab v3+ and SegNet perform worse on the small buildings, which can be explained by the insufficient use of low-level features. Among the SOTA models, SegNet performs the worst. It is because the low-level features extracted by SegNet cannot be transferred to the decoder part, and the local details are lost with continuous downsampling operations. For buildings with complex shapes, as shown in Fig. 6 row 5–6, DS-Net, MA-FCN, and SRI-Net perform better than other comparison methods, which demonstrates that the methods designed for building extraction generally perform better than the traditional semantic segmentation methods. In general, DS-Net performs the best on the WHU Building dataset with good edge accuracy and localization accuracy on buildings of different scales, which benefits from the proposed deep supervision strategy and SAM. Although MA-FCN also integrates multiscale features based on the architecture of U-Net, the multiscale supervision is directly applied to the decoder. In this way, the high-level features are forced to make building predictions, which may affect the representative ability of the features. On the contrary, in DS-Net, we designed an extra lightweight encoder for deep supervision, and the advantage is twofold. First, the high-resolution building predictions generated by the decoder of DS-Net can be adaptively refined by deep features. Second, the gradient of deep supervision can flow through the whole network, which enables the model to learn more compact features.

The quantitative results of DS-Net and the comparison methods are shown in Table I. Values in bolded represent the highest value of the evaluation metrics among the methods. From Table I, we can see that the proposed DS-Net outperforms the SOTA comparison methods. DS-Net generates the highest-quality building extraction results with an IoU of 90.4% and an F1-score of 94.96%. Compared with MA-FCN, the DS-Net

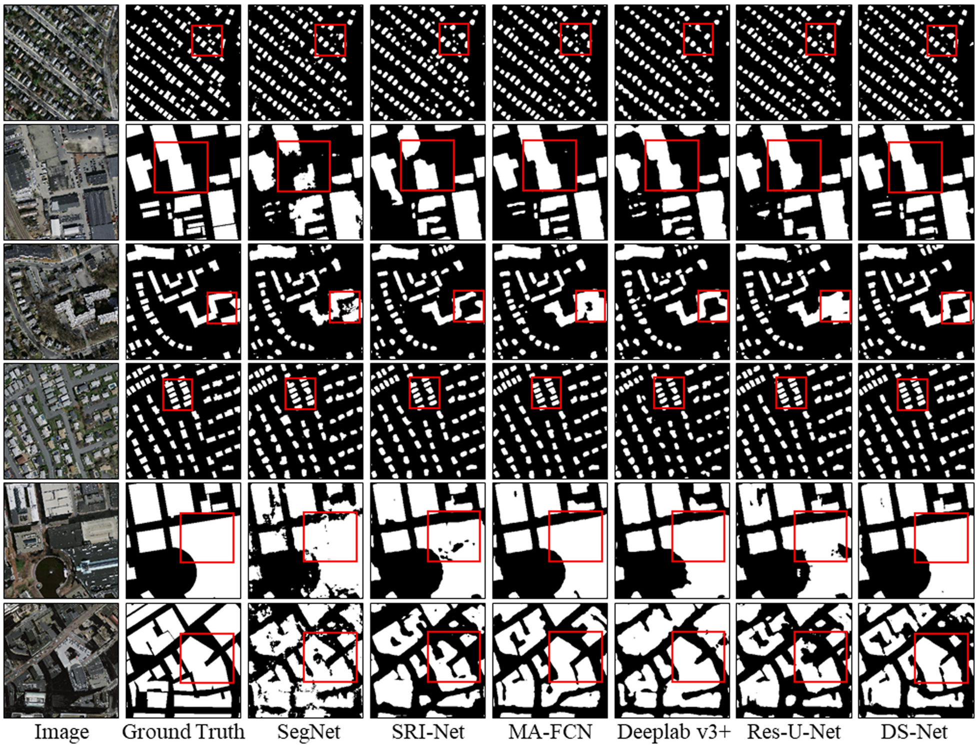


Fig. 7. Visualizations of building extraction results by our proposed DS-Net and other comparison methods on the Massachusetts Building Dataset.

TABLE I
QUANTITATIVE RESULTS FOR DS-NET AND THE COMPARISON METHODS ON THE WHU BUILDING DATASET

Method	IOU	Precision	Recall	F1-Score
DS-Net	**90.4**	**94.85**	**95.06**	**94.96**
Res-U-Net	89.46	94.29	94.53	94.43
SRI-Net	89.88	96.17	93.21	94.67
MA-FCN	89.5	94.58	94.23	94.41
DeeplabV3+	89.61	94.68	92.36	94.52
SegNet	86.45	92.82	92.64	92.73

The entities marked in bold indicate the highest score of the evaluation metric.

TABLE II
QUANTITATIVE RESULTS FOR DS-NET AND THE COMPARISON METHODS ON THE MASSACHUSETTS BUILDING DATASET

Method	IOU	Precision	Recall	F1-Score
DS-Net	**73.79**	**87.77**	82.24	**84.91**
Res-U-Net	70.57	85.59	80.08	82.75
SRINet	71.8	85.43	81.82	83.58
MAFCN	73.14	86.30	**82.75**	84.49
DeeplabV3+	71.63	86.8	80.39	83.47
SegNet	66.57	82.4	77.6	79.93

The entities marked in bold indicate the highest score of the evaluation metric.

exhibits approximately 1%, 0.3%, 0.8%, and 0.5% increment in IoU, precision, recall, and F1-score, respectively. It indicates that the deep supervision strategy in DS-Net is more effective than directly applying auxiliary classifiers on the decoder. It should be noted that deep supervision only increases the parameter by 0.22M, but the improvement in building extraction results is outstanding.

2) Results on the Massachusetts Building Dataset: Based on the above-mentioned analysis, the proposed DS-Net model achieved high building extraction accuracy on the WHU Building dataset, and can precisely depict buildings of different scales. To verify the generalization ability of the DS-Net model, we also conducted experiments on the Massachusetts Building Dataset. The comparison results are shown in Fig. 7, from which it can be seen that DS-Net outperforms the comparison methods visually by the least false predictions and omissions. Especially for the buildings in complex scenes, our proposed DS-Net can successfully distinguish buildings from the complex background. In the results of SegNet, MA-FCN, and Res-U-Net, some nonbuilding areas are misclassified as buildings, resulting in the over-segmentation phenomenon. Especially for SegNet,

TABLE III
SETTINGS OF THE ABLATION EXPERIMENTS OF DS-NET ON THE WHU BUILDING DATASET

Method	**Baseline**	**M**	**D**	**S**	**Description**
Description	U-Net	Multiscale Supervision	Deep Supervision	Scale Attention	
Ablation 1	√				Directly training U-Net
Ablation 2	√	√			Generating multiscale predictions from U-Net
Ablation 3	√		√		Adding deep supervision subnetwork to U-Net
Ablation 4	√		√	√	**DS-Net**

The entities marked in bold indicate the highest score of the evaluation metric.

there are many false predictions and omissions in the irregular largescale buildings. The lack of local detailed information may lead to low edge accuracy, while the lack of semantic information may lead to holes and false predictions. The proposed DS-Net is capable of detecting both largescale buildings and tiny buildings by retaining global semantic characteristics and local details, as shown in the area highlighted in the red boxes in Fig. 7. In general, the model can extract buildings of different scales and ensure the integrity of the extracted buildings.

The quantitative results of the proposed DS-Net and the comparison methods are summarized in Table II, from which we can see the outstanding performance of the proposed DS-Net over other SOTA FCNs. DS-Net generated the best building extraction results with an IoU of 73.79% and F1-Score of 84.91%. Compared with SRI-Net, the DS-Net reached approximately 2% and 1.4% improvement in IoU and F1-Score, respectively. Among the comparison methods, MA-FCN performed the best with an IoU of over 73%, which indicates the importance of utilizing multiscale features in the process of building extraction. DS-Net performs better than MA-FCN with fewer FP predictions. The improvements in the result confirm the effectiveness of applying deep supervision in the process of building extraction.

TABLE IV
ABLATION EXPERIMENTAL RESULTS OF DS-NET ON THE WHU BUILDING DATASET

Method	**IOU**	**Precision**	**Recall**	**F1-Score**
Baseline	89.62	94.7	94.34	94.52
Baseline+M	89.96	95.14	94.29	94.71
Baseline+D	90.19	95.95	93.76	94.84
Baseline+D+S	**90.4**	**98.88**	**94.85**	**94.96**

The entities marked in bold indicate the highest score of the evaluation metric.

3) Ablation Study: To further validate the contribution of the proposed modules in improving the performance of the proposed DS-Net, ablation experiments were conducted on the WHU Building dataset. The baseline was constructed base on U-Net. Then, the multiscale prediction was added to U-Net to quantitatively testify the contribution of utilizing multiscale predictions. It should be noted that, unlike DS-Net that applied deep supervision on an extra subnetwork, the multiscale supervision strategy (denoted as M) direct added the auxiliary classifiers on the decoder to generate multiscale predictions. Then follows our proposed deep supervision, which is denoted as D. In terms of the deep supervision, losses are calculated between the multiscale predictions generated by the deep supervision module and the ground truths. In the test phase, only the highest-level prediction with the highest resolution was outputted as the building extraction results. We further testified the effectiveness of the proposed SAM, which is denoted as S. The SAM is in charge of calculating the contributions of each scale and generating the building extraction results with the weight vector, and the global attention was to determine the weight of fusing multiscale predictions and the highest-resolution prediction. Baseline+D+S denotes the complete DS-Net. More detailed settings of the ablation experiments can be found in Table III.

The ablation study results are shown in Table IV, in which the bolded values represent the highest value of the evaluation metrics. From Table IV, we can see that multiscale supervision can improve the accuracy of building extraction results to some extent, but the improvement is rather small. If we applied deep supervision to U-Net by designing a lightweight encoder to aggregate multiscale features, the performance was improved by 0.57% on IoU. If we further combine deep supervision and scale attention, the model achieved the highest accuracy, with an IoU of 90.4% and F1-Score of 94.96%, which indicates the effectiveness of our proposed method.

IV. Conclusion

With the rapid development of deep learning techniques, FCNs have shown great performance in building footprint extraction. However, because buildings are various among different scales, scenes, and roofing materials, there appear false predictions and omissions in the building extraction results due to the inadequate use of multiscale features. In this article, a novel deep-supervision network with an SAM is proposed. In the proposed network, an extra lightweight subnetwork is proposed to aggregate multiscale features in an up-bottom manner. In this way, the gradients generated by deep supervision can flow through the network. In other words, high-resolution building predictions can be refined by deep features with rich semantic information, and the deep supervision can also guide the learning of shallow convolutional layers. Moreover, an SAM is designed to predict the contribution of each scale and generate the final building prediction by combining building extraction results of different scales. Based on these modules, our proposed DS-Net is scale-robust to buildings by retaining both high-level semantic information and local details. Experiments on the openly available datasets show that DS-Net can effectively extract building of different scales with fewer omissions and false predictions. In our future works, we plan to focus on the vectorization of the building footprint extraction results, as semantic labeling is only a part of building extraction; how to convert building segmentation results into vectors remains a challenge to be solved. Moreover, with more SOTA classification networks appear, the building extraction accuracy can be enhanced by using different classification networks as the backbone. We will consider further improve the accuracy of building extraction with the SOTA backbones.

References

[1] Y. Shi, Q. Li, and X. X. Zhu, "Building segmentation through a gated graph convolutional neural network with deep structured feature embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020, doi: 10.1016/j.isprsjprs.2019.11.004.

[2] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, Jan. 2019, Art. no. 7, doi: 10.3390/rs11070830.

[3] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, early access, pp. 1–16, Jun. 29, 2021, doi: 10.1109/TGRS.2021.3085870.

[4] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589, doi: 10.1016/j.rse.2021.112589.

[5] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Deep multisensor learning for missing-modality all-weather mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 174, pp. 254–264, Apr. 2021, doi: 10.1016/j.isprsjprs.2020.12.009.

[6] Y. Zha, J. Gao, and S. Ni, "Use of normalized difference built-up index in automatically mapping urban areas from TM imagery," *Int. J. Remote Sens.*, vol. 24, no. 3, pp. 583–594, Jan. 2003, doi: 10.1080/01431160304987.

[7] Z. J. Liu, J. Wang, and W. P. Liu, "Building extraction from high resolution imagery based on multi-scale object oriented classification and probabilistic Hough transform," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2005, vol. 4, pp. 2250–2253. doi: 10.1109/IGARSS.2005.1525421.

[8] T. Zhang, X. Huang, D. Wen, and J. Li, "Urban building density estimation from high-resolution imagery using multiple features and support vector regression," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3265–3280, Jul. 2017, doi: 10.1109/JSTARS.2017.2669217.

[9] 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, Jul. 2011, doi: 10.14358/PERS.77.7.721.

[10] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[11] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogramm. Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013, doi: 10.1016/j.isprsjprs.2013.09.004.

[12] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[13] Y. Dong, W. Shi, B. Du, X. Hu, and L. Zhang, "Asymmetric weighted logistic metric learning for hyperspectral target detection," *IEEE Trans. Cybern.*, early access, pp. 1–14, May 26, 2021, doi: 10.1109/TCYB.2021.3070909.

[14] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, early access, pp. 1–18, Jul. 02, 2021, doi: 10.1109/TGRS.2021.3091758.

[15] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5085–5102, Jun. 2021.

[16] Q. Shi, X. Tang, T. Yang, R. Liu, and L. Zhang, "Hyperspectral image denoising using a 3-D attention denoising network," *IEEE Trans. Geosci. Remote Sens.*, early access, pp. 1–16, Jan. 08, 2021, doi: 10.1109/TGRS.2020.3045273.

[17] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

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

[19] Y. Xie *et al.*, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, 2020, doi: 10.1109/JSTARS.2020.2991391.

[20] X. Jiang, X. Zhang, Q. Xin, X. Xi, and P. Zhang, "Arbitrary-shaped building boundary-aware detection with pixel aggregation network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, early access, Aug. 19, 2020, p. 1, doi: 10.1109/JSTARS.2020.3017934.

[21] K. Bittner, F. Adam, S. Cui, M. Körner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[22] S. Shrestha and L. Vanneschi, "Improved fully convolutional network with conditional random fields for building extraction," *Remote Sens.*, vol. 10, no. 7, Jul. 2018, Art. no. 7, doi: 10.3390/rs10071135.

[23] Q. Shi *et al.*, "Domain adaption for fine-grained urban village extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020.

[24] H. Liu *et al.*, "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sens.*, vol. 11, no. 20, 2019, Art. no. 2380.

[25] Y. Yu *et al.*, "Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 895–899, May 2020.

[26] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[27] Y. Dong, T. Liang, Y. Zhang, and B. Du, "Spectral–spatial weighted kernel manifold embedded distribution alignment for remote sensing image classification," *IEEE Trans. Cybern.*, vol. 51, no. 6, pp. 3185–3197, Jun. 2021.

[28] G. Sun, H. Huang, A. Zhang, F. Li, H. Zhao, and H. Fu, "Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images," *Remote Sens.*, vol. 11, no. 3, Jan. 2019, Art. no. 3, doi: 10.3390/rs11030227.

[29] L. Li, J. Liang, M. Weng, and H. Zhu, "A multiple-feature reuse network to extract buildings from remote sensing imagery," *Remote Sens.*, vol. 10, no. 9, Sep. 2018, Art. no. 9, doi: 10.3390/rs10091350.

[30] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," Feb. 2019, *arXiv:1902.09212*, Accessed: Nov. 5, 2020. [Online]. Available: https://arxiv.org/abs/1902.09212

[31] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Oct. 2020.

[32] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 830.

[33] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local-global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–15, Dec. 16, 2020, doi: 10.1109/TNNLS.2020.3041646.

[34] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6109–6120, Jul. 2021.

[35] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using CNN and regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2178–2189, Mar. 2020.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," May 2015, *arXiv:1505.04597 [cs]*, Accessed: Sep. 12, 2020. [Online]. Available: https://arxiv.org/abs/1505.04597

[37] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[38] M. Volodymyr, "Machine learning for aerial image labeling," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2013.

[39] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," Dec. 2019, *arXiv:1912.01703 [cs, stat.]*, Accessed: Jan. 19, 2021. [Online]. Available: https://arxiv.org/abs/1912.01703

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Jan. 2017, *arXiv:1412.6980 [cs]*, Accessed: Jun. 04, 2021. [Online]. Available: https://arxiv.org/abs/1412.6980

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," Feb. 2015, *arXiv:1502.01852 [cs]*, Accessed: Jan. 19, 2021. [Online]. Available: https://arxiv.org/abs/1502.01852

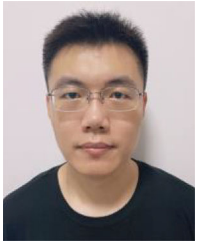
[42] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," Oct. 2016, *arXiv:1511.00561 [cs]*, Accessed: Jan. 19, 2021. [Online]. Available: https://arxiv.org/abs/1511.00561

[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," Aug. 2018, *arXiv:1802.02611 [cs]*, Accessed: Sep. 9, 2020. [Online]. Available: https://arxiv.org/abs/1802.02611

[44] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, Jan. 2018, Art. no. 1, doi: 10.3390/rs10010144.

Haonan Guo received the B.S. degree in Sun Yat-sen University, Guangzhou, China, in 2020. He is currently working toward the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

His research interests include deep learning, building footprint extraction, urban remote sensing, and multisensor image processing.



Xin Su received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in image and signal processing from Telecom ParisTech, Paris, France, in 2015.

He was a Postdoctoral Researcher with the Team SIROCCO, Institut National de Recherche en Informatique et en Automatique (INRIA), Rennes, France. He is currently an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University, China. His research interests include multitemporal remote sensing image processing, multiview image processing, and 3-D video communication.

Shengkun Tang is currently working toward the B.S. degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include object detection, remote sensing image processing, deep learning, and 3D reconstruction.

Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010.

He is currently a Professor with the School of Computer Science, Wuhan University, and the Institute of Artificial Intelligence, Wuhan University. He is also the Director of the National Engineering Research Center for Multimedia and Software, Wuhan University. He has authored or coauthored more than 80 research articles in the IEEE Transactions on Image Processing (TIP), IEEE Transactions on Cybernetics (TCYB), IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Geoscience and Remote Sensing (TGRS), IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), and IEEE Geoscience and Remote Sensing Letters (GRSL).Thirteen of them are Essential Science Indicators (ESI) hot articles or highly cited articles. His research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du regularly serves as a Senior Program Committee (PC) Member of the International Joint Conferences on Artificial Intelligence (IJCAI) and the Association for the Advancement of Artificial Intelligence (AAAI). He is also a Reviewer for 20 Science Citation Index (SCI) magazines, including IEEE TPAMI, IEEE TCYB, IEEE TGRS, IEEE TIP, IEEE JSTARS, and IEEE 2 GRSL. He was the recipient of the Highly Cited Researcher by the Web of Science Group in 2019 and 2020, the IEEE Geoscience and Remote Sensing Society (GRSS) 2020 Transactions Prize Paper Award, the IJCAI Distinguished Paper Prize, and received the IEEE Workshop on Hyperspectral Image and Signal Processing Best Paper Award in 2018. He was the IEEE Data Fusion Contest Champion. He was an Area Chair for the International Conference on Pattern Recognition (ICPR). He is an Associate Editor for *Neural Networks, Pattern Recognition, and Neurocomputing*.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from Xi.an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi.an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He is a Chair Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has authored or coauthored more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution (HR) remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a Fellow of the Institution of Engineering and Technology (IET), London, U.K. He was the recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes of the IEEE Geoscience and Remote Sensing Society (GRSS) 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of IEEE GRSS Wuhan Chapter. He is also an Associate Editor or Editor for more than ten international journals. He is currently as an Associate Editor for the IEEE Geoscience and Remote Sensing Letters.