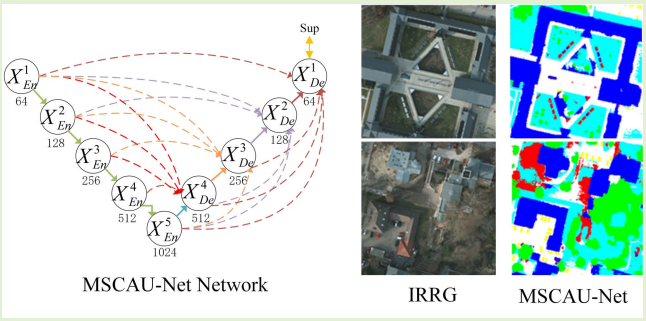
Semantic Segmentation of High-Resolution Remote Sensing Images Using Multiscale Skip Connection Network

Bifang Ma and Chih-Yung Chang, *Member, IEEE*

***Abstract*—Semantic segmentation of remote sensing images plays a vital role in land resource management, yield estimation, and economic evaluation. Therefore, this paper proposes a multi-scale skip connection network with the Atrous convolution to deal with the segmentation problems of the multi-modal and multi-scale high-resolution remote sensing images. Firstly, we applied the Atrous convolution in the encoder to enlarge the convolution kernel's receptive field. Secondly, based on the U-Net network, we merged the light and deep features of different scales by redesigning the skip connection and combining multi-scale features in each U-Net layer. Finally, we applied a pixel-by-pixel classification method and obtained the semantic segmentation results of remote sensing images. The effectiveness of the proposed algorithm is verified. The experimental results show that the mF1 scores are 89.4% and 90.3% on the open dataset of ISPRS Vaihingen and ISPRS Potsdam, respectively, which are better than the state-of-the-art algorithms.**



***Index Terms*— Deep convolutional neural network, high-resolution remote sensing image, multi-scale skip connection, semantic segmentation.**

I. Introduction

With the advancement of science technology, and the development of remote sensing technology, the ability of humanity to probe aerospace has dramatically improved, manifested by the many satellites already orbiting our planet in recent years. The obtained resolution of remote sensing images has been continuously improved. The information it contains is also more affluent and provides convenient conditions for better serving human beings. Object recognition and semantic segmentation of high-resolution remote sensing images have gradually become a focus of research contents in remote sensing image analysis and interpretation. Semantic segmentation of remote sensing images is an essential task that divides each pixel into specific categories. Image segmentation can be applied to disaster assessment, urban planning, and many other directions. Traditional methods used for remote sensing image semantic segmentation relied mainly on the color, texture, shape, and spatial position relationship of ground objects to extract features. And they accomplish the semantic segmentation task using refinements algorithms such as clustering and threshold algorithms. The feature extraction method mainly relies on manual interpretation, but it's impractical to rely on it entirely.

On the one hand, this method is efficient and low-cost, which is hard to guarantee high quality of segmentation results. On the other hand, it's easy to cause visual fatigue and damage to the interpreter. Therefore, to reduce human and material resources and improve the level of intelligence of remote sensing image segmentation, the automation of remote sensing image object segmentation has gradually become an essential topic in remote sensing. Effectively segmenting different object types on remote sensing images is the research direction of intelligent remote sensing applications, which can accelerate the construction of digital China and intelligent cities.

To reduce manual participation and make the feature extraction of remote sensing images more automated, we explore using deep learning to extract high-resolution remote sensing

Manuscript received November 20, 2021; accepted December 28, 2021. Date of publication December 30, 2021; date of current version February 11, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62071123 and in part by the Guiding Project Fund of Science and Technology Department of Fujian Province under Grant 2020H0027. The associate editor coordinating the review of this article and approving it for publication was Prof. Houbing Song. *(Corresponding author: Chih-Yung Chang.)*

Bifang Ma is with the School of Electronic and Mechanical Engineering, Fujian Polytechnic Normal University, Fuqing 350300, China (e-mail: mabf555@126.com).

Chih-Yung Chang is with the Department of Computer Science and Information Engineering, Tamkang University, New Taipei 25137, Taiwan (e-mail: cychang@mail.tku.edu.tw).

Digital Object Identifier 10.1109/JSEN.2021.3139629

image features. And then, we realize the pixel-by-pixel classification of images, which can improve the accuracy of image segmentation and does not require manual feature design. Automatic image segmentation is the fundamental motivation of this paper.

In U-Net++ [1], nested dense skip connections replace ordinary skip connections in U-Net, which significantly enhances the ability of skip connections and narrows the semantic gap between encoder and decoder. Layer features and deep features are effectively merged, which can improve the ability to segment features. A larger receptive field can better merge features from a more extensive context range in the high-resolution images rather than segmentation in a smaller local context. One of the most effective ways to expand the receptive field is to increase the number of convolutional layers or the size of the convolutional filter. Le *et al.* [2] introduced Atrous convolution to achieve high-efficiency image segmentation. In the process of network feature preservation, they used a 3D Atrous residual network and an encode-decode network to segment both large and small objects effectively. Zheng *et al.* [3] introduced the concept of the effective receptive field and used Atrous convolution to expand the effective receptive field. Atrous convolution can be combined with a simple decoder to provide a powerful segmentation model, and the model performed well.

In summary, ordinary skip connections in the U-Net network cannot effectively connect the shallow and deep features in the feature map. It is usually confronted when segmenting high-resolution remote sensing images containing objects of the surface category, which can easily cause high-resolution remote sensing segmentation accuracy to lower and miss segmentation. This paper proposed a Multi-Scale Skip Connection combined with Atrous Convolution Network (MSCA-Net) to perform semantic segmentation on high-resolution remote sensing images. The contributions of this paper are itemized as follows.

1) **Improving the object segmentation effect.** This paper uses Atrous convolution in the encoder to expand the receptive field of the convolution kernel and capture the multi-scale feature information of remote sensing images to improve the object segmentation effect.

2) **Redesigning the multi-scale skip connection.** This paper redesigns the multi-scale skip connection method based on U-Net. We merge the shallow and deep features of different scales to effectively improve the segmentation accuracy of various object categories in high-resolution remote sensing images.

3) **Using the element-wise multiplication strategy.** This paper classifies the images and obtains the final semantic segmentation results through an element-wise multiplication strategy.

The remainder of this paper is organized as follows. Section II reviews the related works. And then, we formulate the problem definition in Section III. Section IV illustrates the semantic segmentation model of high-resolution remote sensing images. Section V introduces the multi-scale skip connection network of the proposed approach combined with the Atrous convolution. The experiments and results are demonstrated in Section VI. Section VII offers a conclusion and future works.

II. Related Work

Traditional semantic segmentation on remote sensing images was mainly performed by extracting the shallow features of the image, and most of the segmentation results were not ideal.

According to the usege of prior knowledge, traditional remote sensing images semantic segmentation methods were usually divided into unsupervised and supervised classifications. The unsupervised classification methods, including ISODATA classification and K-means classification, etc., did not require any prior knowledge and mainly relied on the spectral or spatial characteristics of the ground objects for classification, which saves time and energy of manual labeling. However, this type of method required a lot of analysis and subsequent processing. It was not easy to achieve a one-to-one correspondence with the category of ground objects. To avoid the reclassification of unsupervised classification, we used prior knowledge to determine the category of objects in the supervised classification method. It was often used in the task of segmentation and classification of objects. The commonly used methods of supervised classification included support vector machines, decision trees, and ensemble models. These methods often need to create feature descriptors in advance and then analyze the relationship between features to find the best classification thresholds and parameters. In this type of supervised classification method, features often require a lot of professional knowledge. If the designed features are unreasonable, the classification accuracy will decline, and there are significant drawbacks in the semantic segmentation of remote sensing images.

Deep learning technology in machine learning has been gradually applied to various fields [4]–[6]. Convolutional Neural Network (CNN) created new records in the area of natural image classification continuously. Deep learning doesn't need to manually design features but extracts features by learning from training samples. At the same time, the CNN method can learn more detailed and deeper features from a large number of sample images and avoid the disadvantages of manual design of extracting shallow features in traditional algorithms. The methods were widely used in remote sensing images semantic segmentation. Several popular CNNs developed in recent years, such as AlexNet [7], VGGNet [8], GoogLeNet [9], have been used in scene classification. Zheng *et al.* [10] proposed a Foreground-Aware Relation Network (FarSeg), which enhanced the ability to distinguish foreground features by learning the association between foreground and background and using foreground-related context. To achieve the purpose of false segmentation. Furthermore, A state-of-the-art method [11] was proposed based on the semi-supervised to address the semantic segmentation of natural imagery effectively. It introduced the two variants of the self-correction module using either linear or convolutional functions. FCN was the first fully convolutional neural network that was designed for semantic segmentation. FCN used skip

connections to refine feature maps and upsample the output feature maps to the size of the input origin data. The Fully Convolutional Networks (FCN) has achieved a breakthrough and overcome many of the limitations of CNNs in semantic segmentation. However, the abstraction ability of FCN for the high-level features was not enough to consider useful global context information. The precise boundary cannot be accurately restored by eight upsampling times at the end of the FCN.

Immediately afterward, two types of structures appeared in the semantic segmentation task. One is the backbone style, such as DeepLabV3 [12]. This type of network removed some pooling layers and reduced the degree of downsampling in the feature map. These networks used the pyramid pooling module [13] or atrous spatial pyramid pooling(ASPP) [12] module to extract and fuse feature information of different scales and receptive fields. However, the precise boundary cannot be accurately restored by eight upsampling times at the decoder in the methods. The other type was the encoder-decoder style. Most of the encoder-decoder type networks used the idea of transfer learning [14], [15]. and typical networks included U-Net [16] and SegNet [17], etc. The basic network structure of these networks was an encoder-decoder structure. The advantage of this structure was that the encoder-decoder can generate feature maps with different semantic information. The skip connections can connect different semantic information, which can effectively improve the performance of the encoder-decoder framework. The U-Net network was a distinct semantic segmentation network with a skip-connected encoder-decoder structure. Therefore, the U-Net network has shown excellent performance in the semantic segmentation task. However, the common skip connection method in the U-Net network cannot connect the shallow information and the deep information contained in the feature map effectively, resulting in lower accuracy of image segmentation.

III. Problem Definition

In the problem of remote sensing image semantic segmentation, the traditional semantic segmentation method based on graph partitioning is to abstract the image into a graph form G=(V, E). Among them, V is the graph node, and E is the boundary of the graph, thus completing. Later, the remote sensing image semantic segmentation method based on deep learning has appeared one after another, the semantic information of various types of objects has been added based on it.

We assume that the category of each pixel in the remote sensing image is regarded as a variable $x_i \in \{y_1, y_2, \cdots, y_n\}$, where x_i represents a pixel and y_i represents a category, then we treat one or more target images as:

$$P(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_v(x_i, x_j) \tag{1}$$

where $\psi_u(x_i)$ represents the semantic category corresponding to the pixel, which will be obtained from the prediction result of the semantic segmentation model, and j represents the semantic category adjacent to i. For example, the probability that pixels such as "sky" and "bird" are adjacent in physical space exceeds the adjacent probability of "sky" and "fish". Therefore, the main objective of the proposed work is to predict the probability of the target and its neighbors by extracting the features of the target feature, thereby developing a high-resolution remote sensing image semantic segmentation model. The motivation behind this study is that in the segmentation task, high-resolution remote sensing image target features usually have problems such as inaccurate feature segmentation, over-segmentation, and mis-segmentation. However, we improve the segmentation accuracy of target features by designing a semantic segmentation model suitable for high-resolution remote sensing images.

IV. Semantic Segmentation Model of High-Resolution Remote Sensing Images

A. Multi-Scale Skip Connection

The U-Net network proposed by Ronneberger *et al.* [16] has a simple skip connection, connecting the shallow semantic information and the deep semantic information in the feature map. The U-Net++ network proposed by Zhou *et al.* [1] adds a nested tight skip connection based on the U-Net network. The skip connection dramatically reduces the semantic gap between the encoder and the decoder and improves the segmentation ability of the feature category. However, the skip connection of U-Net and U-Net++ can't fully use multi-scale information and effectively extract semantic features of different scales. Therefore, this paper designs a multi-scale skip connection, as shown in Figure 1. The figure includes simplified diagrams of U-Net, U-Net++, and MSCA-Net. Figure 1(c) shows the specific connection method of the skip connection detailed in the MSCA-Net network proposed in this paper, which is the normal skip connection of U-Net in Figure 1(a) and the nested of U-Net++ in Figure 1(b). Compared with the skip connection, MSCA-Net redesigns the skip connection combined with multi-scale information, converts the interplay between the encoder and decoder, and captures the interaction between the encoder and decoder. It better extracts the fine-grained semantic information and coarse-grained semantic information in the feature maps, improving the accuracy of high-resolution remote sensing images segmentation.

To further explain the multi-scale skip connection, we described the structure of the encoder layer. Figure 2 is an example of the construction. The extracted features can be directly inputted into the decoder because the third layer in the encoder has the same-level encoder layer. However, the skip connection method of the U-Net network cannot effectively connect the shallow and deep features. In contrast, the multi-scale skip connection of the MSCA-Net network encoder-decoder can transfer more low-level semantic information from the small-scale encoder layer by using the maximum pooling operation. For low-level semantic information, a series of skip connections within the decoder connect the large-scale decoder layer X_{De}^4 and then X_{De}^5 transmits high-level semantic information using bilinear interpolation. To further unify the number of channels and reduce the redundant information, it occurred to us that the convolution with 64 filters of size 3×3 can be a satisfying choice.Through this operation,

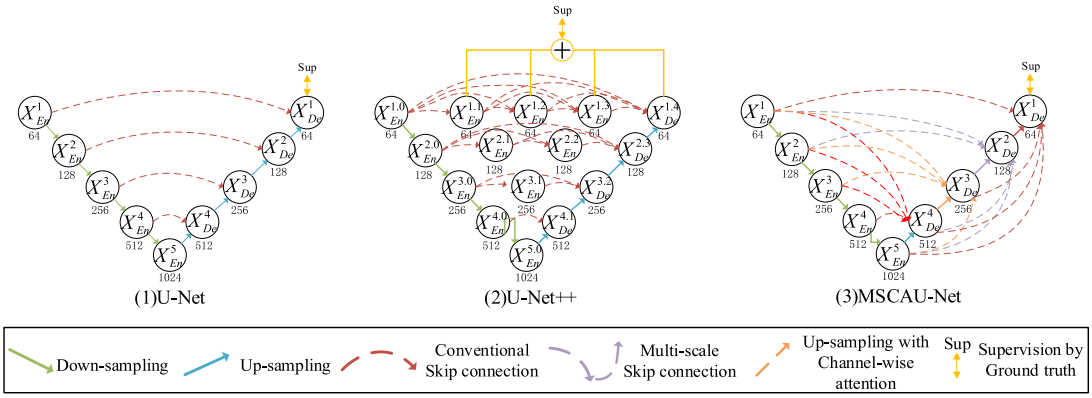


Fig. 1. Comparison of skip connection of U-Net, U-Net++, and MSCA-Net network.

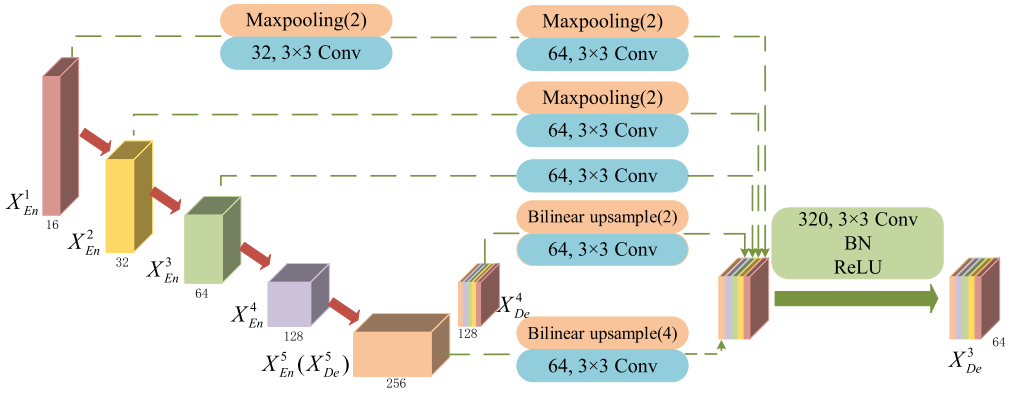


Fig. 2. Illustration of how to construct the multi-scale feature map of the third decoder layer X_{De}^{3}.

we obtain feature maps of the same resolution. Each feature map contains a wealth of remote sensing images information. We seamlessly merge the shallow exquisite information with deep semantic information.This paper further performs a feature aggregation mechanism, the concatenated feature map from five scales, consisting of 320 filters of size 3×3, a batch normalization, and a ReLU activation function. The above operation process is defined as follows:

$$X_{De}^{i} = \begin{cases} X_{En}^{i}, & i = N \\ \mathrm{Y}\left(\underbrace{\mathrm{G}(\mathrm{D}(X_{En}^{k}))_{k=1}^{i-1}}_{Scales:1^{th}\sim i^{th}}, \underbrace{\mathrm{G}(\mathrm{U}(X_{De}^{k}))_{k=i+1}^{N}}_{Scales:(i+1)^{th}\sim N^{th}} \right), & i = 1, \ldots, N-1 \end{cases} \tag{2}$$

where N refers to the total number of the encoder, $\mathrm{Y}(\cdot)$ realizes the feature aggregation mechanism with a convolution followed by a batch normalization and a ReLU activation function, $\mathrm{U}(\cdot)$ and $\mathrm{D}(\cdot)$ indicates up-sampling and down-sampling operation, function $\mathrm{G}(\cdot)$ denotes a convolution operation and $[\cdot]$ represents the concatenation.

B. Atrous Convolutional

We can obtain a larger receptive field without changing the image size through Atrous convolution. Its main advantage is that it allows flexible adjustment of the receptive field size

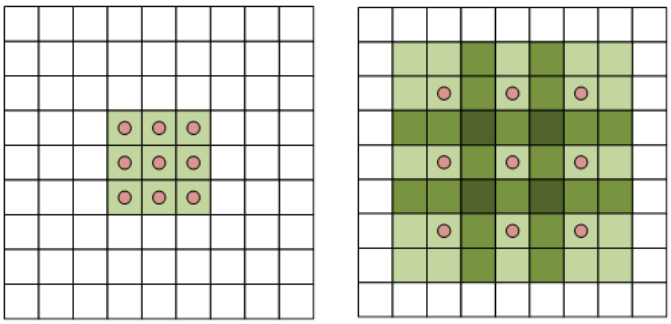


Fig. 3. Atrous convolution diagram.

to capture multi-scale information as much as possible and improve the multi-object performance of segmentation tasks. The two-dimensional Atrous operator is defined as follows:

$$f_{a,b}(x_{\ell}) = \sum_{g=0}^{G_{\ell}} \mu_{m,n}^{a,b} \bullet x_{\ell}^{g} \tag{3}$$

$f_{a,b}$is the convolution operation $\mathbb{R}^{H_{\ell} \times W_{\ell} \times G_{\ell}} \rightarrow \mathbb{R}^{H_{\ell+1} \times W_{\ell+1}}$on the input feature map, $\bullet$ represents the convolution operator, $x_{\ell} \in \mathbb{R}^{H_{\ell} \times W_{\ell} \times G_{\ell}}$ is the feature map belonging to the channel $g \in \{0, 1, \ldots, G_{\ell}\}$ in the a row and b column, $\mu_{m,n}$ is the Atrous convolution with convolution kernel size m and expansion rate $n \in \mathbb{Z}^{+}$. The convolution

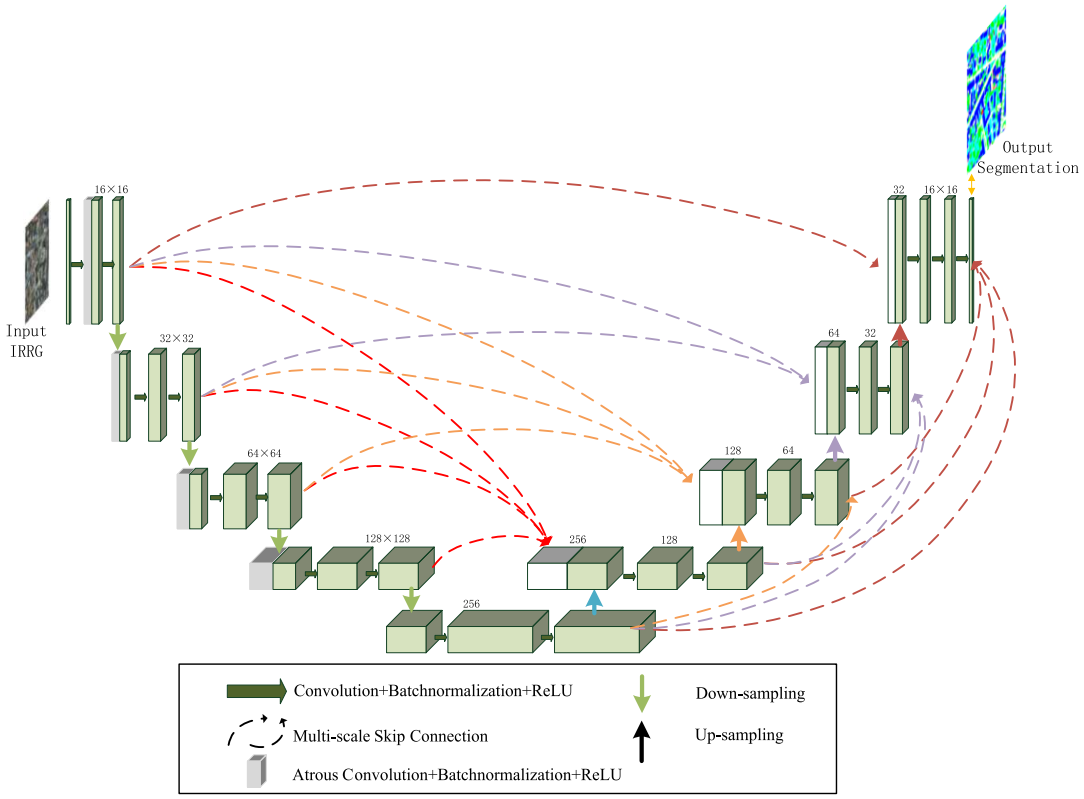


Fig. 4. The structure of Atrous convolution and Multi-scale skip connection U-Net.

kernel size m increases to $m + (m-1)(n-1)$ Atrous convolution, and Atrous convolution is equivalent to standard convolution $n = 1$. Figure 3 shows a schematic diagram of Atrous convolution, and Figure 3(a) is a 3×3 Atrous convolution kernel with an expansion rate n=1, which is the same as the common convolution operation. After the convolution operation, the receptive field is 3×3. Figure 3(b) is a 7×7 Atrous convolution kernel with an expansion rate n=2, and its receptive field is 7×7. The receptive field of the convolutional layer of the standard convolution is related to the size and step length of all the previous convolution kernels. The receptive field grows linearly, while the Atrous convolution can realize the exponential growth of the receptive field.

V. Multi-Scale Skip Connection Network Model Combined With Atrous Convolution

The MSCA-Net structure of the high-resolution remote sensing image semantic segmentation network proposed in this paper is shown in Figure 4. The network consists of the encoder with Atrous convolution, the decoder, and a series of multi-scale skip connections.This paper uses Atrous convolution in the encoder part to increase the receptive field and uses the U-Net network in the backbone network. The multi-scale skip connection method connects semantic features of different scales. For the input high-resolution remote sensing image, firstly, we increase the receptive field of the convolution kernel by introducing Atrous convolution. We perform the feature encoding on the encoder in the backbone network. Secondly, we input the encoded feature map into the decoder for up-sampling operation through the multi-scale skip connection. Finally, we obtain the semantic segmentation result using convolution operation and SoftMax operation.

U-Net's structure is shaped like the letter U, consisting of a convolutional compression layer on the left end and a transposed convolutional amplification layer on the right end. In the operation of transposed convolution on the right end, the results of the previous convolution on the left end will be stitched together to ensure more information. The encoder-decoder structure is used as the basic structure of U-Net. The encoder performs the high dimensional feature extraction and down-sampling operations on the image; this part is the contraction path. The compression path consists of 4 blocks, each using 3 efficient convolution kernels and 1 maximum pooling for the down-sampling operation. After each down-sampling operation, the number of the feature maps is multiplied by 2, so the size of the feature maps becomes larger. The decoder performs an up-sampling operation on the extracted feature maps, whose expansive path comprises 4 blocks. Before each up-sampling operation, the feature maps size is multiplied by 2 through a deconvolution operation, and their amount is halved. They are merged with the feature maps of the symmetric compression path on the left. Due to the different sizes of the feature maps obtained from the compressed and expanded paths, the encoder-decoder structure in U-Net plays a prominent role in retaining the detailed features of the ground objects in large amounts. Therefore, U-Net as the backbone network for semantic segmentation

of high-resolution remote sensing images can achieve good semantic segmentation results. Unlike the U-Net network, the MSCA-Net adds multi-scale skip connections in the encoder-decoder and inside the decoder, which is conducive to learning better feature expression improving the accuracy of remote sensing image segmentation.

At the same time, the MSCA-Net merges the smaller and the same-scale feature maps from the encoder and the larger-scale feature maps from the decoder at each decoder layer to capture the full-scale fine-grained semantics and coarse-grained semantics. We pool and downsample a group of encoder-decoder skip connections from the smaller-scale encoder layer through a non-overlapping maximum pooling operation to transmit the bottom layer's low-level semantic information. Bilinear interpolation is used to transmit high-level semantic information from the large-scale decoder layer through the series mentioned above of internal decoder skip connections. At the same time, we can obtain a larger receptive field under the same feature map by the Atrous convolution, thereby obtaining more dense data. The combination of Atrous convolution and multi-scale skip connection can achieve a larger receptive field. Under this condition, more low-level information of the feature category is retained, and high-level semantic information is efficiently extracted while obtaining dense data. This process is conducive to learning better feature expression and improving the segmentation accuracy of remote sensing images.

VI. Experiments and Results

A. Datasets

This paper carried out our experiments on the two image sets of the ISPRS [19] 2D Semantic Labeling Challenge to validate the proposed method. These datasets comprise aerial images of a very high resolution taken from over two cities in Germany: Vaihingen and Potsdam. Our goal is to do semantic labeling on images belonging to six classes: buildings, impervious surfaces (e.g., roads), low vegetation, trees, cars, and clutter. Two leaderboards (one for each city) and report test metrics, obtained on held-out test images, are available online.

ISPRS Vaihingen: The Vaihingen dataset has a 9 cm/pixel resolution with tiles of approximately 2494 × 2064 pixels Collected in an area of 1.38 square kilometers within the city. Altogether this dataset has 33 images, of which 16 have public ground truth. The Infrared-Red-Green (IRRG) images and the DSM data are used as supplementary data.

ISPRS Potsdam: The Potsdam dataset has a resolution of 5 cm/pixel with tiles of 6000 × 6000 pixels and 38 images, of which 24 have public ground truth. The Infrared-Red-Green-Blue (IRRGB) multispectral images and the DSM data are supplementary data in the tiles. The nDSM data is also included with the dataset, but none of our experiments use this data.

B. Multi-Scale Data: Enhancement and Standardization

The purpose of data enhancement is to generate new sample instances. Data enhancement plays a crucial role in improving the network's generalization ability when training samples are fewer than the needed number. The ISPRS data set is divided into two parts: training and testing. 12 images were selected from the Vaihingen dataset for training, and 4 images were selected for testing. The number of datasets is small. Our method first cropped the images in the training part into a size of 256 × 256 randomly and then expanded the data through rotation and translation operations to prevent overfitting. Finally, this paper used the obtained 12,000 patches for training. This paper selected 18 images from the Potsdam dataset for training and 6 images for testing. After cropping each of the images into a size of 256×256 randomly, this paper obtained 27,000 patches for training. The expanded dataset is used for training MSCA-Net. All the bands (IRRG) of the high-resolution remote sensing images this paper used are standardized within the [0, 1] interval.

The random numbers of neural networks' parameters and activation functions are usually initialized to be between [0, 1]; standardized methods are needed to avoid gradient explosion and dispersion. The z-score standardization method approximates the value of each pixel in the input image to a normal distribution, which helps improve the network convergence speed. The standardized formula is:

$$X_{out} = \frac{\left(X \middle/ \mathrm{Max}(X)\right) - \lambda}{\sigma} \tag{4}$$

where X is the input value, Max(X) is the maximum input value, λ and σ are the mean and standard deviation of X / Max(X), respectively.

C. Network Training

The remote sensing images in the data set used to carry out the experiments are high-resolution. They cannot be directly processed within the deep network, so the sliding window method is used as a pre-step to extract small 256 × 256 size blocks. The size of the overlapping area between two consecutive small blocks thus extracted (using the sliding window method) is also defined by this pre-step. During training, as extracting training samples in smaller sizes contributes greatly to data expansion, we focus on obtaining more smaller sizes training samples. We used a 64px stride for the Potsdam dataset and a 32px stride for the Vaihingen dataset. We adjusted the step size and used a 32px stride to test the Potsdam dataset and a 16px stride to test the Vaihingen dataset. A smaller step size allows the average prediction of overlapping areas to improve overall accuracy. In this paper, the initial learning rate is set to 0.01, and the learning rate is divided by 10 every 5 epochs until 0.00001. The experimental hardware environment is Intel Core i9-9900k + NVIDIA RTX2080Ti 64GB memory + 11GB video memory, and the software environment is Linux Ubuntu 18.04 + CUDA10.0 + cudnn7.6 + Pytorch16.0.

D. Results and Analysis

1) Evaluation Standard: For an evaluation of the performance of the deep learning network, the following formula

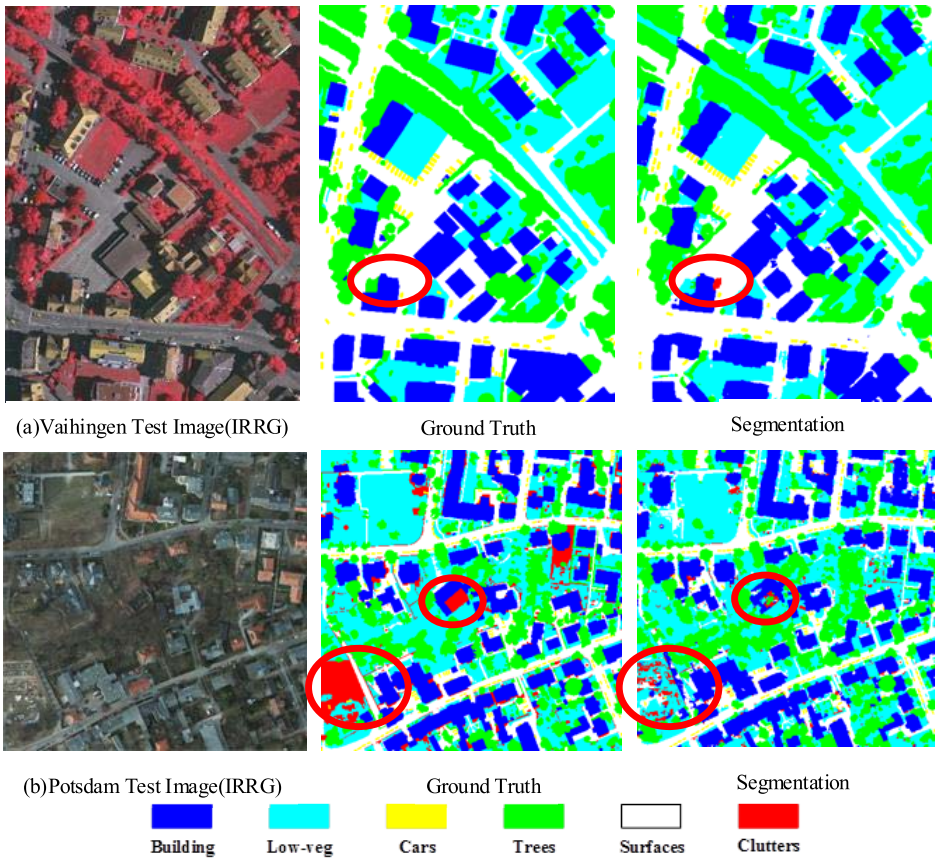


Fig. 5. Segmentation results of MSCA-Net on ISPRS Vaihingen and Potsdam dataset.

was used to calculate the F1 score:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

the true positive TP indicates that the predicted value is 1 and the true value is 1. FP is an example of a false positive indicating the predicted value as 1 and the true value as 0. The False Negative FN indicates that the predicted value is 0 and the true value is 1. P indicates how precise the network is, the ratio of the number of positive cases predicted to be correct, to the total number of positive cases predicted. R is the ratio of correctly predicted positive cases to true positive cases, which means recall. During experimentation, the average value (mF1) of F1 is used to indicate the segmentation accuracy of the network. The performance of the network and the accuracy of the segmentation grows with mF1.

CSAU-Net also used the overall accuracy formula to evaluate the segmentation accuracy of the algorithm. The calculation formula of OA is as follows:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

the True Negative TN (TN) indicates that the predicted value is 0 and the true value is 0.

2) Experimental Results and Analysis: Fig. 5 shows the comparisons between some of the experimental data results obtained by our algorithm (MSCA-Net) and the Ground Truth. Consistent colors are used for objects belonging to the same classes. It can be seen that the overall segmentation results obtained by MSCA-Net are relatively ideal, with only minor segmentation errors highlighted by the red circles. We can see that comparing the segmentation results obtained from the images of the two cities, and our method has segmented the image of the Vaihingen city with fewer errors. The quality and accuracy of ours segmentation results are almost as good as the Ground Truth. The main reason that the quality of the segmentation obtained from some of the small areas in the images of the Potsdam city is inferior is that the Potsdam dataset is more complex. Parts of the ground features of the real labels that have been done manually are blurred, whereas the Vaihingen dataset features are more clearly distributed.

Tab. I shows the segmentation accuracy results obtained by our algorithm when invoked on the test mentioned above sets. From the segmentation accuracy mF1, the overall accuracy OA, and mIOU, obtained from data of the five types of features shown, the segmentation results obtained by this algorithm are excellent.

The model MSCA-Net uses U-Net as the baseline, so we compare it with U-Net in the experiments. Meanwhile, because U-Net++ has a nested skip connection structure, to verify the method's effectiveness in this paper, the MSCA-Net and

TABLE I
SEGMENTATION ACCURACY RESULTS OF MSCA-NET ON ISPRS VAIHINGEN AND POTSDAM DATASET

Dataset	Building (F1-score)	Tree (F1-score)	Low-veg (F1-score)	Surface (F1-score)	Car (F1-score)	OA	mF1	mIOU
Vaihingen	0.961	0.923	0.783	0.948	0.856	0.924	0.894	0.607
Potsdam	0.941	0.892	0.845	0.917	0.918	0.893	0.903	0.761

TABLE II
COMPARISON OF SEGMENTATION ACCURACY BETWEEN MSCA-NET AND RELATED METHODS ON VAIHINGEN DATASET

Models	Building (F1-score)	Tree (F1-score)	Low-veg (F1-score)	Surface (F1-score)	Car (F1-score)	OA	mF1	mIOU
U-Net [16]	0.901	0.801	0.778	0.886	0.687	0.849	0.811	0.569
U-Net++ [1]	0.943	0.912	0.780	0.914	0.725	0.874	0.859	0.584
MSCA-Net	**0.961**	**0.923**	**0.783**	**0.948**	**0.856**	**0.924**	**0.894**	**0.607**
	(+0.018)	(+0.011)	(+0.003)	(+0.034)	(+0.131)	(+0.050)	(+0.035)	(+0.023)

TABLE III
COMPARISON OF SEGMENTATION ACCURACY BETWEEN MSCA-NET AND RELATED METHODS ON POTSDAM DATASET

Models	Building (F1-score)	Tree (F1-score)	Low-veg (F1-score)	Surface (F1-score)	Car (F1-score)	OA	mF1	mIOU
U-Net [16]	0.848	0.749	0.789	0.790	0.875	0.798	0.810	0.694
U-Net++ [1]	0.926	0.857	0.823	0.885	0.893	0.862	0.877	0.746
MSCA-Net	**0.941**	**0.892**	**0.845**	**0.917**	**0.918**	**0.893**	**0.903**	**0.761**
	(+0.015)	(+0.035)	(+0.022)	(+0.032)	(+0.025)	(+0.031)	(+0.026)	(+0.015)

U-Net++ are also compared in the experiments. The comparisons of the result are shown in Table II and Table III. We compare U-Net++ [1] with a dense skip connection network structure on the Vaihingen and Potsdam datasets. It can be seen from Table II and Table III that the segmentation accuracy of MSCA-Net is higher than that of U-Net and U-Net++ on various types of ground objects. The comparative experiments of the two types of data set effectively prove that MSCA-Net has a more vital ability to segment ground objects. The encoder-decoder multi-scale skip connection in the MSCA-Net network can effectively connect the semantic information of feature maps at different scales. The multi-scale skip connection inside the decoder can better integrate the shallow semantic information with the deep semantic information. At the same time, the Atrous convolution introduced in the encoder part can effectively increase the receptive field, so MSCA-Net can effectively improve the classification of ground objects.

This paper evaluates the computational complexity and quantitatively evaluates the efficiency of the U-Net, U-Net++ and MSCA-Net frameworks, including the training and testing time of the network, as shown in Table IV. It can be seen from Table IV that the training time and testing time of MSCA-Net are shorter than U-Net++, but slightly longer than U-Net. Since the skip connection in the U-Net is relatively simple, the features extracted by downsampling are spliced into the upsampling process, and MSCA-Net has a multi-scale skip connection structure compared to the U-Net, the computational complexity is higher. Training and testing time is slightly longer. The U-Net++ has a densely nested skip connection structure, which is more complex than the multi-scale skip connection structure. Therefore, the training and testing time of MSCA-Net is significantly faster than that of U-Net++. At the same time, Table V in the paper has listed the references of detailed structure among MSCA-Net, U-Net and U-Net++. In table V, "$\surd$" means that the network uses this part of the structure, and "–" means that the network does not use this structure. Regarding the connection structure, the connection in the U-Net is a common skip connection. This structure plays a more role of contitation, which plays a more role in concatenation. And the U-Net++ uses the Nested dense connection. This paper proposes that the MSCA-Net has designed a multi-scale skip connection. Therefore, different kinds of connections are the important distinction among various network structures.

This paper also compares the network and other methods on the ISPRS Vaihingen dataset and the Potsdam dataset. The results are shown in Table VI and Table VII. This paper has also compared the results obtained by other methods used on the ISPRS Vaihingen and Potsdam datasets with our completeness method.

The results in Table VI show the comparison between the segmentation accuracy results gained by CSAU-Net and those obtained by the methods in literature works [17], [20]–[25]. MSCA-Net has achieved ideal results from both the segmentation accuracy mF1 score and the overall segmentation

TABLE IV
COMPARISON OF TRAINING AND INFERENCE TIME AMONG MSCA-NET, U-NET AND U-NET++

	U-Net		U-Net++		MSCA-Net	
	Train(hours)	Test(hours)	Train(hours)	Test(hours)	Train(hours)	Test(hours)
Vaihingen	6.589	0.068	7.230	0.092	6.920	0.081
Potsdam	4.965	0.083	5.724	0.146	5.340	0.103

TABLE V
COMPARISON OF DETAILED STRUCTURE AMONG MSCA-NET, U-NET AND U-NET++

Method	Backbone(U-Net)	Dense block	Atrous convolution	Connection
U-Net	√	--	--	Common skip
U-Net++	√	√	--	Nested dense
MSCA-Net	√	√	√	Multi-scale skip

TABLE VI
COMPARISON OF SEGMENTATION ACCURACY BETWEEN MSCA-NET AND OTHER METHODS ON VAIHINGEN DATASET

Models	Building (F1-score)	Tree (F1-score)	Low-veg (F1-score)	Surface (F1-score)	Car (F1-score)	OA	mF1	mIOU
SegNet [17]	0.883	0.791	0.772	0.862	0.543	0.829	0.769	0.581
DeepLabV3 [20]	0.865	0.782	0.778	0.865	0.673	0.833	0.793	0.594
DeepLabV3+ [21]	0.901	0.801	0.802	0.885	0.689	0.852	0.816	0.603
Tiramisu [22]	0.892	0.823	0.761	0.884	0.713	0.842	0.815	0.601
U-Net Att [23]	0.931	0.911	0.748	0.932	0.825	0.898	0.869	0.597
FGC [24]	0.942	0.903	0.759	0.926	0.828	0.904	0.872	0.573
GSN [25]	0.951	0.899	**0.805**	0.922	0.824	0.903	0.881	0.586
MSCA-Net	**0.961** (+0.010)	**0.923** (+0.012)	0.783 (-0.022)	**0.948** (+0.016)	**0.856** (+0.028)	**0.924** (+0.020)	**0.894** (+0.013)	**0.607** (+0.004)

TABLE VII
COMPARISON OF SEGMENTATION ACCURACY BETWEEN MSCA-NET AND OTHER METHODS ON POTSDAM DATASET

Models	Building (F1-score)	Tree (F1-score)	Low-veg (F1-score)	Surface (F1-score)	Car (F1-score)	OA	mF1	mIOU
SegNet [17]	0.864	0.739	0.780	0.811	0.857	0.803	0.810	0.703
Tiramisu [22]	0.867	0.743	0.792	0.841	0.853	0.818	0.819	0.708
DeepLabV3 [20]	0.872	0.751	0.785	0.832	0.862	0.824	0.820	0.712
U-Net Att [23]	0.892	0.785	0.821	0.883	0.875	0.841	0.851	0.714
FGC [24]	0.917	0.782	0.813	0.885	0.872	0.852	0.854	0.701
DeepLabV3+ [21]	0.928	0.784	0.834	0.893	0.882	0.868	0.864	0.723
RiFCN [26]	0.930	0.819	0.837	0.917	**0.937**	0.883	0.888	0.752
MSCA-Net	**0.941** (+0.011)	**0.892** (+0.073)	**0.845** (+0.007)	**0.917** (---)	0.918 (-0.019)	**0.893** (+0.010)	**0.903** (+0.015)	**0.761** (+0.009)

accuracy OA. Except for the Low-veg class, our method's segmentation accuracy is slightly lower than Peng *et al.* [25]. And for the rest of the class, our method's segmentation accuracy is superior to the other algorithms. Especially our algorithm's F1-score of the surface and car categories is higher by 0.16% and 0.28% than the F1-score gained by Oktay *et al.* [23] and Ji *et al.* [24]. And Our algorithm's OA and mF1 surpassed Peng *et al.* [25] by 0.20% and 0.13%, respectively.

The comparison shown in Table VII is between the results gained by CSAU-Net and those acquired by Badrinarayanan *et al.* [17], Chen *et al.* [20], [21], Jégou *et al.* [22], Oktay *et al.* [23], Ji *et al.* [24] and Mou *et al.* [26] from

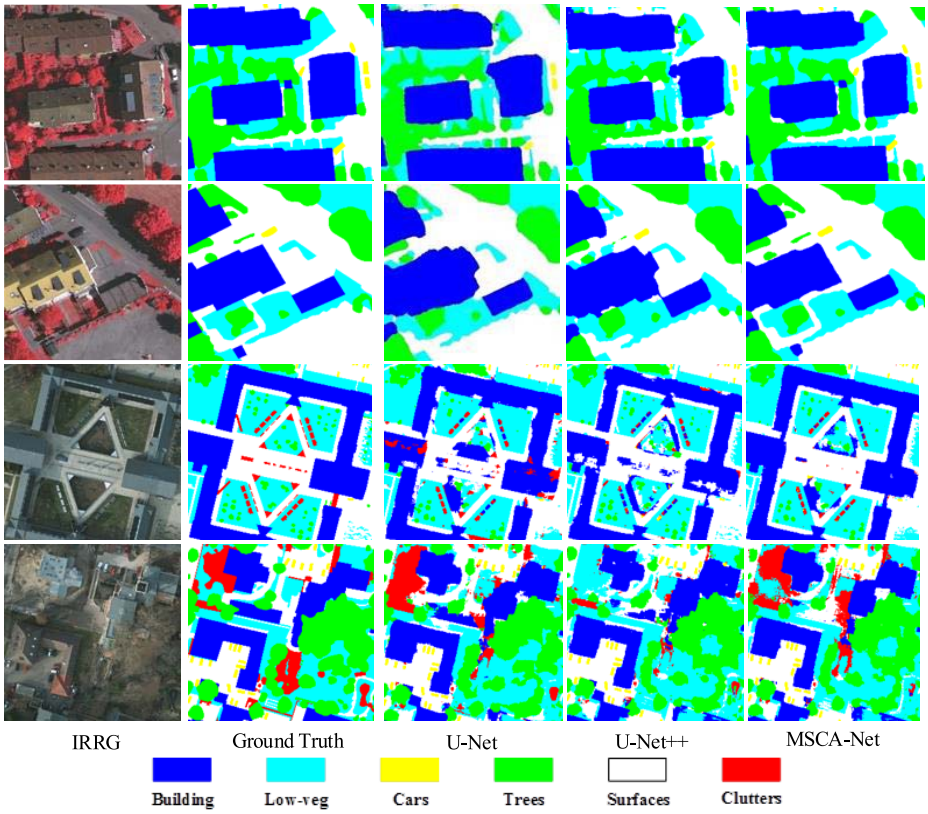


Fig. 6. Comparison of detailed segmentation results on ISPRS dataset.

the Potsdam dataset. Except for the slight decrease in the Car class's segmentation accuracy compared to Mou *et al.* [26], our method is superior to all methods in the other categories. In addition, the overall segmentation accuracy OA and segmentation accuracy mF1 scores of MSCA-Net outperform other algorithms. Compared to Mou *et al.* [26], our algorithm's OA and mF1 growth are higher by 0.10% and 0.15%, respectively. Our algorithm's mIOU are higher by 0.4% and 0.9% than [26]. It proves the effectiveness of our algorithm.

The MSCA-Net and related methods compare the details of image segmentation on the ISPRS Vaihingen dataset and the ISPRS Potsdam dataset, as shown in Figure 6. The first column in the figure is the local details of the input remote sensing images, and the second column is the local true label of the relevant feature images. As illustrated in the figure, compared to U-Net and U-Net++, MSCA-Net has dramatically improved the quality of the segmentation of the ground objects, making the targeted ground objects closer to those of the actual scene, especially on more complex partial images. Our network has a relatively good segmentation ability, which significantly reduces the mis-segmentation of the ground object category, which can be attributed to a series of multi-scale skip connections. MSCA-Net can extract rich information from feature maps of different scales to a large extent, and the network can effectively integrate the semantic features of each layer.

VII. Conclusion

The MSCA-Net deep learning network structure proposed in this paper is used in multi-scale skip connection with Atrous convolution, which achieves a successful semantic segmentation of remote sensing images of a high-resolution nature. Firstly, we introduced Atrous convolution into the encoder to increase the receptive field of the convolution kernel, which provides the necessary conditions for extracting more semantic features. The encoder in the backbone network is used to encode the feature map. Secondly, the multi-scale skip connection between the encoder and decoder can effectively connect the extracted shallow and deep semantic features. The multi-scale skip connection in the decoder transmits features of different scales. The network can better integrate features of different scales. The final stage is to obtain the semantic segmentation results from the input remote sensing image. The experiments show that MSCA-Net can improve the classification accuracy of various objects effectively.

This method also has limitations. Although it has a better segmentation effect for high-resolution remote sensing images, it is not evident for hyperspectral remote sensing images or radar images. So it has certain limitations for selecting data sets. However, although this paper optimizes the connection method of skip connections in the U-Net, designing more effective skip connections to improve the accuracy of segmentation of high-resolution remote sensing image categories is the direction of our subsequent research work.

References

[1] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "U-Net++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop Deep Learn. Med. Image Anal.*, Québec, QC, Canada, Sep. 2017, pp. 3–11.

[2] N. Le, K. Yamazaki, K. G. Quach, D. Truong, and M. Savvides, "A multi-task contextual atrous residual network for brain tumor detection & segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, Jan. 2021, pp. 5943–5950.
[3] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021, *arXiv:2012.15840*.
[4] L. Wang *et al.*, "Design of a precision spraying control system with unmanned aerial vehicle based on image recognition," *J. South China Agricult. Univ.*, vol. 37, no. 6, pp. 23–30, 2016.
[5] J. Wang, Y. Liu, and H. Song, "Counter-unmanned aircraft system(s) (C-UAS): State of the art, challenges, and future trends," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 3, pp. 4–29, Mar. 2021.
[6] Y. Liu, J. Wang, S. Niu, and H. Song, "Deep learning enabled reliable identity verification and spoofing detection," in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl. (WASA)*, Qingdao, China, Sep. 2020, pp. 333–345.
[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Systems*, 2012, pp. 1097–1105.
[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[9] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[10] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4095–4104.
[11] M. S. Ibrahim, A. Vahdat, M. Ranjbar, and W. G. Macready, "Semi-supervised semantic image segmentation with self-correcting networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 12712–12722.
[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[13] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.
[14] S. Niu, M. Liu, Y. Liu, J. Wang, and H. Song, "Distant domain transfer learning for medical imaging," 2020, *arXiv:2012.06346*.
[15] S. Niu, J. Wang, Y. Liu, and H. Song, "Transfer learning based data-efficient machine learning enabled classification," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput., Int. Conf. Pervasive Intell. Comput., Int. Conf. Cloud Big Data Comput., Int. Conf. Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Aug. 2020, pp. 620–626.
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent*, Munich, Germany, Oct. 2015, pp. 234–241.
[17] 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.
[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.
[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and H. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learning (ICML)*, Washington, DC, USA, Jul. 2011, pp. 689–696.
[20] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[21] L.-C. 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)*, Munich, Germany, Sep. 2018, pp. 833–851.
[22] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, Jul. 2017, pp. 1175–1183.
[23] O. Oktay *et al.*, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[24] S. Ji, Z. Zhang, C. Zhang, S. Wei, M. Lu, and Y. Duan, "Learning discriminative spatiotemporal features for precise crop classification from multi-temporal satellite images," *Int. J. Remote Sens.*, vol. 41, no. 8, pp. 3162–3174, Apr. 2020.
[25] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1743–1751.
[26] L. Mou and X. X. Zhu, "RiFCN: Recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images," 2018, *arXiv:1805.02091*.

Bifang Ma received the B.S. degree from the Department of Physics, Fujian Normal University, Fuzhou, China, in 1991, and the M.S. degree in communication and information system from Fuzhou University, Fuzhou, in 2009.

She is currently an Associate Professor with the School of Electronic and Mechanical Engineering, Fujian Polytechnic Normal University, Fuqing, China. Her current research interests include the electronic information, artificial intelligence, and deep learning.

Chih-Yung Chang (Member, IEEE) received the Ph.D. degree in computer science and information engineering from National Central University, Taiwan, in 1995. He is currently a Full Professor with the Department of Computer Science and Information Engineering, Tamkang University, New Taipei, Taiwan. His current research interests include the Internet of Things, wireless sensor networks, artificial intelligence, and deep learning. He has served as an Associate Guest Editor for several SCI-indexed journals, including the *International Journal of Ad Hoc and Ubiquitous Computing* (IJAHUC) from 2011 to 2019, *International Journal of Distributed Sensor Networks* (IJDSN) from 2012 to 2014, *IET Communications* in 2011, *Telecommunication Systems* (TS) in 2010, *Journal of Information Science and Engineering* (JISE) in 2008, and *Journal of Internet Technology* (JIT) from 2004 to 2008.