Land Use Classification of High-Resolution Multispectral Satellite Images With Fine-Grained Multiscale Networks and Superpixel Postprocessing

Yaobin Ma, Xiaohua Deng, and Jingbo Wei

Abstract—**Land use recognition from multispectral satellite images is fundamentally critical for geological applications, but the results are not satisfied. The scale dimension of current multiscale learning is too coarse to account for rich scales in multispectral images, and pixel-wise classification tends to produce "salt-and-pepper" labels due to possible misclassification in heterogeneous regions. In this article, these issues are addressed by proposing a new pixel-wise classification model with finer scales for convolutional neural networks. The model is designed to extract multiscale contextual information using multiscale networks at a fine-grained level, addressing the issue of insufficient multiscale learning for classification. Furthermore, a small-scale segmentation-combination method is introduced as a postprocessing solution to smooth fragmented classification results. The proposed method is tested on GF-1, GF-2, DEIMOS-2, GeoEye-1, and Sentinel-2 satellite images, and compared with six neural-network-based algorithms. The results demonstrate the effectiveness of the proposed model in finding objects of large scale difference, improving classification accuracy, and reducing classified fragments. The discussion also illustrates that convolutional neural networks and pixel-wise inference are more practical than transformer and patch-wise recognition.**

Index Terms—**Classification, convolutional neural network (CNN), multiscale, multispectral, superpixel.**

I. Introduction

CLASSIFICATION of land use is critical in applications, such as agricultural resource survey, crop yield estimation, disaster assessment, and so on. Large-scale classification has been carried out using satellite images in a cost-effective manner. Among the data sources, high-resolution multispectral remote sensing images are the most popular. The classification results are now easier to gather thanks to the rapid advancements in sensor, computer, and aeronautical technology.

There have been numerous methods to classify the land use from remote sensing images, as are categorized into unsupervised classification and supervised classification. Traditional methods, such as k-means clustering, iterative self-organizing data analysis technique (ISODATA) clustering, decision tree, random forest, and support vector machine, perform the pixel-to-pixel classification with the one-dimensional spectral information. New classification algorithms are based on neural networks and utilize both spectral information and two-dimensional structures. Distinguished by the output form, they can be categorized into two groups, namely pixel-wise and patch-wise. The patch-wise classification is essentially a semantic segmentation as the input and the output are both patches. Pixel-wise classification is to use the scene classification pixel by pixel in which the input is a pixel or patch while the output is a class value. The input/output difference of the classification methods are presented in Fig. 1.

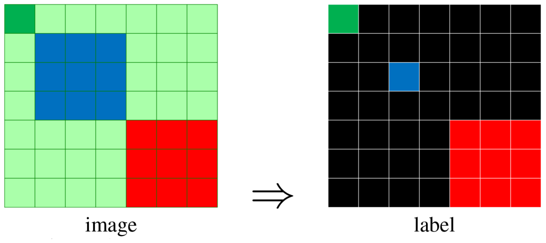


Fig. 1. Land use classification from remote sensing images. Green denotes the pixel-to-pixel classification, blue denotes the patch-to-pixel (pixel-wise) classification, and red denotes the patch-to-patch (patch-wise) classification.

Pixel-wise classification will be carried out in this article. Patch-wise classification, or semantic segmentation, has been carried out in numerous works. For example, Liu et al. [1] used multiple images for training in context aware cascade network and expanded the training set using operations, such as overlapping block taking, rotation, and mirroring. With the context-aware encoder network, Liang et al. [2] presented a patch sampling strategy, in which manual intervention is needed to maximize the separation of training and validation sets. The accuracy of segmentation is usually lower than that of classification. In addition, patch-wise classification uses the comprehensively labeled patches matched with the input patches, which requires the labeled data to be as continuous and complete as possible. In contrast, pixel-wise classification requires only discrete or irregular labels. Therefore, pixel-wise classification is more suitable to the application needs of remote sensing classification.

Manuscript received 14 January 2023; revised 15 March 2023; accepted 19 March 2023. Date of publication 22 March 2023; date of current version 6 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42267070 and in part by the Jiangxi Provincial Institute of Water Sciences under Grant 2021SKTR07 and Grant 202224ZDKT11. *(Corresponding author: Jingbo Wei.)*

Yaobin Ma is with the School of Resources and Environment, Nanchang University, Nanchang 330031, China (e-mail: szj11612@163.com).

Xiaohua Deng and Jingbo Wei are with the Institute of Space Science and Technology, Nanchang University, Nanchang 330031, China (e-mail: dengxhua@gmail.com; wei-jing-bo@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3260448

Pixel-wise classification tends to produce the “salt-and-pepper” results. The derived land cover maps are highly fragmented to be incorporated into a geographic information system database. The reason is from the limited patch size that judges a label locally. Although the small size is effective for homogenous regions, it faces misclassification in heterogeneous regions due to the lack of training data, illegible features, or vague class. Therefore, fragmentation needs to be avoided when the pixel-wise classification is used.

High-resolution multispectral remote sensing images offer valuable spatial contextual information that can enhance classification accuracy through the acquisition of local multiscale features. Derived from the wavelet transform, multiscale learning is the sampling of different scales of a signal or image. Smaller scales can exhibit structures and textures, while larger scales focus more on the spectral features. In convolutional neural networks (CNNs), convolutional layers and residual modules that are cascaded with different kernel sizes can be used for multiscale learning, as has been proposed with different levels of semantic information or various receptive fields. Li et al. [3] presented an adaptive multiscale deep fusion residual network (AMDF) with the purpose of effectively using the useful information contained in shallow features to mine multiscale features with different levels of semantic information. In contrast, Liu et al. [1] and Hua et al. [4] aimed to perceive boundaries, regions, and semantic categories of the target objects by learning features on multiscale receptive fields. Considering that the integration of shallow and deep features faces the differences in size and amount of channels, multiscale features from various receptive fields are of more potential. In addition, Hang et al. [5] proposed a multiscale progressive network for gradually segmenting objects into small scale, large scale, and other scales by cascading three subnetworks.

However, the scale dimension of the current multiscale learning for pixel-wise classification is too coarse to account for rich scales in multispectral images. Detail emerges in high-resolution multispectral images when the spatial resolution approaches one meters, accompanied with more complex spectral features of ground objects [6]. In very high-resolution remote sensing images, land use are larger than a single pixel, and the phenomenon of “same object with different spectrums” and “same spectrum for different objects” becomes more prevalent. Finer multiscale characteristics with multiscale feature learning at smaller granular are, therefore, needed to deal with these challenges.

This article focuses on the two issues mentioned above. A new pixel-wise classification model is designed to extract the multiscale contextual information in images with multiscale networks at a fine-grained level to address the issue of insufficient multiscale learning for the classification of high-resolution multispectral images. A superpixel combination technique is proposed as a postprocessing solution to smooth the fragmented classification results.

The main contributions of this work are summarized as follows.

1) To achieve fine-grained multiscale learning in CNN, new downsampling and residual modules are proposed for the classification task of high-resolution satellite images.
2) To improve the fragmentation of classification results, a small-scale segmentation method is introduced for post-processing to combine labels across class boundaries.

II. Related Work

In this section, classical, neural network-based, and object-oriented classification methods are introduced. Classical and neural network-based classification are pixel-wise that outputs only one label for the center pixel of the input patch. Object-oriented classification is a typical representative of patch-wise classification that outputs all the labels for pixels in a patch.

A. Traditional Pixel-Level Classification

Unsupervised classification methods can directly classify image pixels based on gray-scale spatial features, which are suitable for feature classification scenarios with simple prior knowledge. Currently clustering techniques, such as k-means and ISODATA are often used for unsupervised classification. More and more sophisticated unsupervised classification methods have been developed to extract an appropriate group of features for the classification of remote sensing images in a more efficient way. For example, to achieve accurate classification of remote sensing images, Marinoni and Gamba [7] proposed an unsupervised approach for feature extraction based on data driven discovery, which exploits mutual information maximization to retrieve the most relevant features with respect to information measures. Huang et al. [8] proposed a multiview subspace clustering model, which exploits effectively the rich information from multiple features extracted either from a single data source or from multiple sources. Unsupervised classification methods can only distinguish different categories and not determine the attributes of the categories, since they lack the necessary a priori knowledge.

By learning data relationships from a given training set containing ground truth information, supervised classification methods are more suitable than unsupervised classification methods for remote sensing images with complex ground object types, and usually have higher classification accuracy. There are two primary groups of supervised classification methods. The minimum distance method and the maximum likelihood method are two prominent examples of the first group of supervised classification methods based on statistical models. The minimum distance method, which classifies pixels based on how far they are from the center of each category, is a relatively simplified classification method. The maximum likelihood method is a nonlinear classification based on Bayesian criterion with minimal probability of classification error.

The second group is supervised classification methods based on machine learning, mainly including decision tree, random forest, support vector machine, neural network-based classification, and object-oriented classification. Decision tree classification is a method that compares the eigenvalues of pixels with a set baseline value in a hierarchical manner. The classification and regression tree model was proposed by Breiman et al. in 1984 and is a widely used decision tree classification method. Random forest is an integrated classification model proposed by Breiman

[9] in 2001. To address the problem of noise in training data, Gislason et al. [10] used random forest to classify multispectral images. support vector machine is based on the structural risk minimization criterion to maximize the generalization ability of the classifier with strong nonlinear and high-dimensional data processing capability while making the sample classification error minimal. A fuzzy support vector machine-based multispectral image classification method was put out by Wang and Ma [11] and has higher accuracy than the method that uses an support vector machine directly.

These classical algorithms belong to the pixel-wise classification. The advantages of these methods are fast training or unsupervised classification. These methods only utilize the spectral information of images and cannot utilize the spatial contextual information in images. With the emergence of large amount of details in high-resolution multispectral remote sensing images and the complexity of spectral features, the classification accuracy of these methods is poor. And since the classification is performed pixel by pixel, the categories between neighboring pixels have contingency. Misclassification is prone to occur in the region of ground object category transition or feature ambiguity, resulting in fragmented classification results.

B. *Neural Network-Based Pixel-Wise Classification*

Neural network-based classification contains many different approaches. Multilayer perceptron neural network based on error back propagation is the first algorithm that was introduced for remote sensing image classification. In the latest research, deep learning-based classification methods have been widely used, which can be seen as a development of neural networks. Deep belief network (DBN) achieves image classification by unsupervised pretraining of unlabeled data and supervised fine-tuning of labeled data. Liu et al. [12] calculated the texture features of high-spatial resolution remote sensing images through nonsubsampled contourlet transform, and used DBN to classify images based on spectral and texture features. Subsequently Zhong et al. [13] developed a new diversified DBN through regularizing pretraining and fine-tuning procedures by a diversity promoting prior over latent factors. Chen et al. [14] proposed a SAE to extract the high-level features for remote sensing images using spectral–spatial information. Chen et al. [15] used stacked denoise autoencoder to extract features, and used logistic regression approach in the top layer of the network to perform supervised fine-tuning and classification. However, the inputs of these models are in vectorization form that may ignore the neighborhood structures around pixels.

CNN models allow the use of spatial patches as data input, providing a natural way to integrate spatial contextual information with higher classification accuracy compared to BP, DBN, and SAE. Based on this, Maggiori et al. [16] designed a fully convolutional architecture for the dense classification of remote sensing images and addressed the issue of imperfect training data through a two-step training approach. Ji et al. [17] proposed a novel three-dimensional CNN based method that automatically classifies crops from spatio-temporal remote sensing images. Liu et al. [18] proposed an end-to-end learning framework based on deep multiple instance learning, using CNN and SAE to extract the spatial features of panchromatic images and the spectral features of multispectral images, respectively. In recent years, the development of CNN has made continuous breakthroughs in multispectral image classification. Aiming at the problems of gradient explosion, gradient disappearance, and nonconvergence brought by deeper networks, ResNet [19] used the concepts of residual learning and skip connection to deepen the network complexity in exchange for the higher classification performance. On the other hand, some research has focused on improving the computational efficiency of networks, such as LinkNet [20], which is pretty light but superior in performance.

Recent advancements in deep learning networks have significantly improved the extraction of discriminative features from remote sensing data. However, the performance bottleneck in identifying and recognizing objects of interest when only using single satellite data has become increasingly evident. To overcome this limitation, multimodal networks have been proposed and used for remote sensing images. These networks combine multiple sources of data to improve classification accuracy and obtain better results than only using single satellite data source. For example, Gadiraju et al. [21] developed a multimodal deep learning framework for crop classification using multispectral and multitemporal satellite images. Similarly, Hang et al. [22] proposed an unsupervised feature learning model to extract features by using the relationship between hyperspectral and light detection and ranging data. These studies demonstrate the potential of multimodal networks to improve the accuracy and efficiency of remote sensing image analysis. However, the classification of a single multispectral image still has the greatest universality in terms of the burden of data preparation.

The neural network-based pixel-level classification method can obtain higher classification accuracy compared with the classical method. Furthermore, the CNN-based pixel-level classification adopts the patch-to-pixel classification way, which utilizes both the spatial contextual information and the spectral information of images, and achieves a high classification accuracy. However, patch-to-pixel classification still results in the same fragmented classification results as the classical method and has a slower training speed.

C. *Object-Oriented Classification*

Object-oriented classification is a new remote sensing image classification method that emerged for high-resolution remote sensing image applications. Object-oriented classification takes regions containing similar semantic information as the processing objects for classification, and can use not only the spectral features of images, but also the geometric features, texture features, adjacency relationships, and other spatial features of images. Image segmentation is used in object-oriented classification, where the image to be classified is segmented to generate image objects. Then, the image objects are classified using methods, such as nearest neighbor classification or fuzzy classification.

On the basis of object-oriented classification method, Zhang et al. [23] extracted Zhangjiangkou mangrove communities from

QuickBird images with segmentation, merge, computing, and attributes selection. Mirzapour and Ghassemian [24] proposed an object-based method for multispectral image segmentation and classification. In addition Jin et al. [25] presented a method that combines object-oriented approach with deep CNNs. Baroud et al. [26] also proposed an artificial neural network combined with object-oriented method for land cover classification of high-resolution multispectral remote sensing images.

Object-oriented classification can guarantee continuous classification results and is faster than pixel-wise classification methods. However, the classification accuracy of object-oriented classification is limited by the segmentation accuracy. The existing segmentation algorithms have limited accuracy, which is significantly lower than the classification accuracy, making the overall classification accuracy of object-oriented classification slightly worse.

Although transformer shows advantages in many applications, CNN is more friendly to the training amount and computational burden. CNN focuses mainly on local features, while the cross attention in Transformer can capture global similarity over a larger receptive field. However, the attention mechanism contributes weakly to the pixel-wise classification that uses only local features. Instead, it introduces three disadvantages. First, in order to achieve the same performance as CNN, a very large size of training data is required. Second, far more graphical card memory is used to train a transformer than a CNN. Lastly, the computational effort of a Transformer is usually larger than that of CNN. Since a portion of the image to be classified has to be manually labeled to pursue the best accuracy, pixel-wise classification needs to be trained online which prefers smaller labels. In terms of the computational volume, a remote sensing image commonly has more than 100 million pixels, then it is a huge burden to perform scene classification for each pixel. Global features can also be captured by CNN plus attention module, but the abovementioned disadvantages cannot be avoided. The combination of pretrained transformer and shift learning can reduce the amount of training data, but the consumption of memory and computation is greater than that of CNN.

III. Methodology

In this section, new solutions are proposed for land use classification of high-resolution multispectral satellite images. First, a new pixel-wise classification network is created consisting of downsampling and multiscale residual blocks (MRBs) to capture multiscale contextual information. The MRBs extract multiscale features at a fine-grained level. A postprocessing module is designed that uses superpixels derived from small-scale image segmentation to refine the pixel-wise classification results by stitching the fragmentation. The new model is named as fine-grained multiscale classification network, or FGMCN. The superpixel postprocessing is abbreviated as SPP. The whole method is called FGMCN-SPP.

A. Multiscale Residual Network

There is rich scale diversity in high-resolution satellite images. Woodlands and water can be identified point by point through normalized difference indices. Cultivated land has to be identified over an area. Artificial buildings are identified over larger and more diverse patch sizes. When secondary classification is involved, the diversity of scales is even more complex. However, the multiscale learning in land use classification tasks can only learn features at given scales. Therefore, when plenty of computational resources and training data are given, the classification accuracy can be improved by designing extractors at diverse scales as many as possible.

The proposed network structure is shown in Fig. 2. At the beginning of the network, a batch normalization (BN) layer is first used to normalize the input data. A 3 × 3 convolutional layer is then applied to the normalized image blocks to extract shallow features. Four MRBs are alternatively cascaded with four downsampling blocks (DSBs) to gradually extract high-level features. The separation of downsampling and feature blocks ensures that more features can be learned individually. Multiscale features are learned with the newly designed MRBs. After the feature learning, a high-level feature map is downsampled with a 2 × 2 average pooling with stride 1. Next, a fully connected layer and a softmax activation layer are employed to convert the multichannel feature mapping into a multiclass problem for pixel classification. The dropout operation is used behind the average pooling to reduce possible overfitting.

The network parameter is listed in Fig. 2. A patch with a size of $w \times w$ is created as the feature region centered at each labeled pixel. Therefore, the actual size of the input data is $I \in \mathbb{R}^{w\times w\times B}$, where B is the channel number of input image. The suggested size of the input image for satellite images with spatial resolution greater than 1 m is 27 × 27. The output channels for the first convolutional layer is set to 64, and the output channels for each subsequent block are set to 128, 128, 256, 256, 512, 512, 1024, and 1024, respectively. The dropout probability of dropout is 0.5. The scale of training data has been taken into consideration when setting these settings.

1) Downsampling Block: Existing classification networks use either convolution or pooling for downsampling. Maximum pooling improves the nonlinear representation of the network out of commonplace information. Cascaded downsampling convolution maintains the indistinctive features instead of salient features. The average pooling smoothes out salient features, too. However, there are both distinctive and indistinctive features in high-resolution satellite images as the ground resolutions are roughly between 2 and 30 m. The former is beneficial for identifying artificial facilities, while the latter can be used to identify natural resources such as forest land and water. Since using a single downsampling method may lose features, the complementarity of the two downsampling methods is harnessed to design the downsampling module.

Fig. 3 depicts the three parallel branches that make up the downsampling procedure. The network automatically searches for branches that are appropriate for various structures. After the first convolution layer, the input of the first DSB is $I^1 \in \mathbb{R}^{w\times w\times N}$. The left branch using a 3 × 3 convolution layer with stride 2 to transform I^1 into $I^{D\_L} \in \mathbb{R}^{\frac{w}{2}\times\frac{w}{2}\times\frac{N}{2}}$ for extracting features and downscaling. BN is to avoid the gradient vanishing.

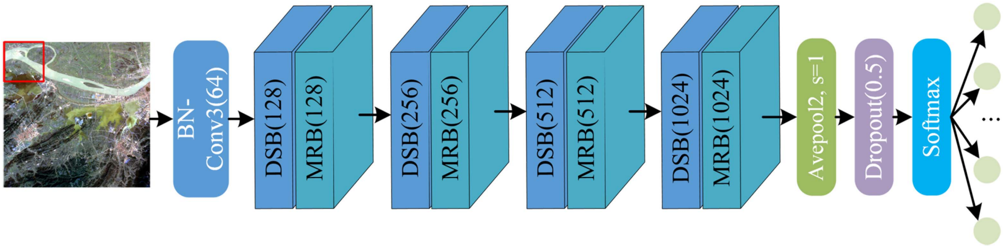


Fig. 2. Network structure of the proposed model. Conv3 denotes the convolution with the kernel size of 3 × 3, BN is the batch normalization, and FC is the fully connected. The number of output channels is appended for each block. The number in dropout brackets indicates the dropout probability, and s indicates the stride.

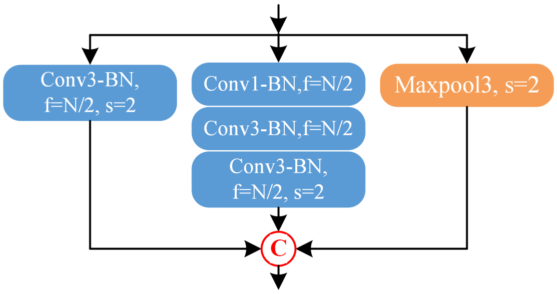


Fig. 3. Structure of a DSB. conv denotes the convolutional operation, BN denotes the batch normalization, f gives the amount of channels, s gives the stride, and ⓒ denotes concatenation.

The calculation can be formalized as

$$I^{D\_L_1} = \text{Conv}\left(I^1\right) = \omega * I^1 + b \tag{1}$$

$$I^{D\_L} = \text{BN}\left(I^{D\_L_1}\right) = \frac{I^{D\_L_1} - \text{E}\left(I^{D\_L_1}\right)}{\sqrt{\text{Var}\left(I^{D\_L_1}\right) + \varepsilon}} \tag{2}$$

where Conv denotes the convolutional operator, ω and b are the weight and bias of the convolution layer, respectively, $\text{E}(I^{D\_L_1})$ and $\text{Var}(I^{D\_L_1})$ are the expectation and variance of $I^{D\_L_1}$, respectively, and ε is a small constant value (i.e., 1e$-$5) to maintain stability.

The middle branch using layers of 1 × 1 convolution, 3 × 3 convolution, and 3 × 3 convolution with stride 2 to transform I^1 into $I^{D\_M} \in \mathbb{R}^{\frac{w}{2} \times \frac{w}{2} \times \frac{N}{2}}$ for enlarging the receptive field and extracting features at a wider scale. Then, a 3 × 3 max pooling with stride 2 is used to obtain the texture detailed information in the right branch. For input feature map $I^1 \in \mathbb{R}^{w \times w \times N}$, the max pooling selects the maximum value of a specific area $R^c_{k,k}$ as its representation

$$I^{D\_R} = \text{Maxpool}\left(I^1\right) = \max_{t \in R^c_{k,k}} I^1 \tag{3}$$

where $1 \leq c \leq N$ and $1 \leq k \leq w$.

Therefore, the output of DSB can be denoted as

$$I^{\text{DSB}} = \text{Concat}\left(I^{D\_L}, I^{D\_M}, I^{D\_R}\right). \tag{4}$$

Our downscaling module is based on the Reduction A module from Inception V4 [27] but deliberately modified to suit for classification applications. The number of output channels in the convolution part is reduced to half the number of input channels. The ReLU activation layer in the convolutional branch is removed to prevent feature loss. Finally, the two convolutional branches are concatenated with the maximum pooling branch, each playing a half role. As a result, the number of output channels is expanded to twice the number of input channels when the scale is reduced to half of the input.

2) Multiscale Residual Block: The proposed MRB is depicted in Fig. 4, which incorporates the multilevel residual connection in [28] and two residual modules similar to Res2Net [29]. Four parallel branches make up the Res2Net-like residual module, which extracts features at four different scales. The network is expected to automatically discover scale features that best suits for input image content. To obtain distinguishable receptive fields at a fine-grained level, the Res2Net module is introduced. By stacking convolutional layers, CNNs may learn coarse-to-fine multiscale features and increase the receptive field. The Res2Net module builds hierarchical residual connections within one single residual block, which can broaden the range of receptive fields. The multiscale residual module has been applied to extract multiscale convolution features of remote sensing images [30].

The structure of our residual block is slightly different from the Res2Net residual module. The input of the first residual module is $I^{\text{DSB}} \in \mathbb{R}^{\frac{w}{2} \times \frac{w}{2} \times 2N}$. The output of the 1 × 1 convolution in the first layer is $I^{M_{1-1}} \in \mathbb{R}^{\frac{w}{2} \times \frac{w}{2} \times 2N}$ and separated into four equal portions to fed into each branch separately. Assume x_i and y_i represent the ith input and output parts, respectively, the abovementioned process can be formulated as

$$y_i = \begin{cases} x_i, & i = 1 \\ \text{BN}\left(\text{Conv}\left(x_i + y_{i-1}\right)\right), & 1 < i \leq 4. \end{cases} \tag{5}$$

These four branches are combined to obtain multiscale features $I^{M_{1-2}} \in \mathbb{R}^{\frac{w}{2} \times \frac{w}{2} \times 2N}$. After that, multiscale features are convolved and connected with the input of first residual module. The output of the first residual module can be expressed as

$$I^{M_{1-2}} = \text{Concat}\left(y_1, y_2, y_3, y_4\right) = \{y_1, y_2, y_3, y_4\} \tag{6}$$

$$I^{M_1} = \text{Add}\left(I^{\text{DSB}}, \text{BN}\left(\text{Conv}\left(I^{M_{1-2}}\right)\right)\right). \tag{7}$$

In addition, the channel quantities of the input and the output are the same to guarantee the feature learning capacity of MRB. To avoid feature loss, the ReLU activation layer in the Res2Net residual block is removed. Finally, the output of the MRB is $I^{\text{MRB}} \in \mathbb{R}^{\frac{w}{2} \times \frac{w}{2} \times 2N}$.

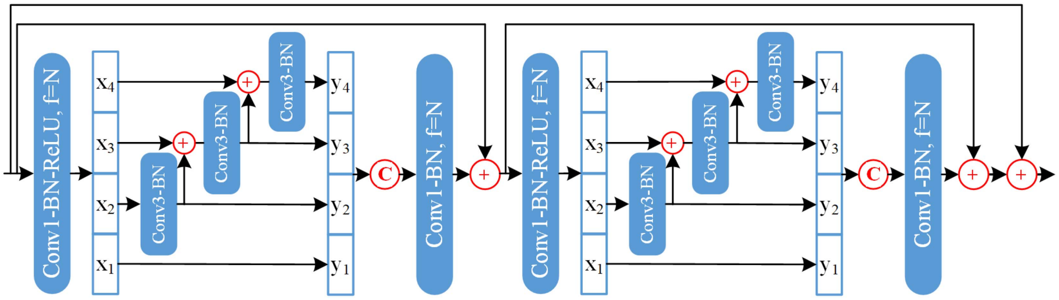


Fig. 4. Structure of a MRB. ⊕ denotes addition and ⓒ denotes concatenation.

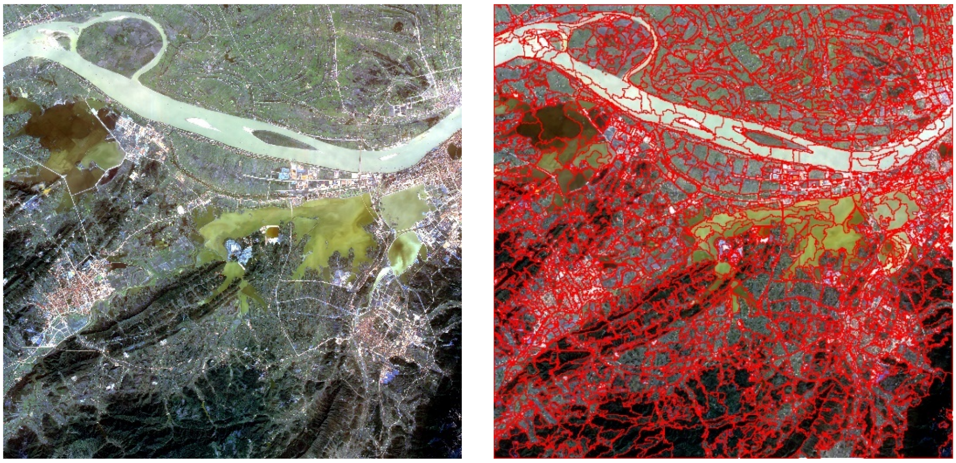


Fig. 5. Display of segmentation results. The left is the original image, and the right is the segmented image.

B. SPP to Smooth Fragmented Labels

Our classification method is pixel-wise, that is, a patch is fed into the network and a label value is output endowing the class of the pixel at the patch center. Broken spots are observed in the results of pixel-wise classification methods. Especially, the classification results based on deep learning are not stable enough as a result of the lack of corresponding training data for the mixed pixel features in the cross-category transition zone. Some pixels in heterogeneous regions show "salt-and-pepper" noise style in the output label images with small proportion.

A postprocessing technique is then suggested using small-scale segmentation to address the label inaccuracy issue for heterogeneous regions. The patch size in pixel-wise classification is fixed, which leads to ambiguous category judgments for pixels in transition regions. In this case, the human vision system is accustomed to searching for salient borders ahead of judging the category. Segmentation is, therefore, incorporated into classification as a processing method to mimic the experience of human eyes.

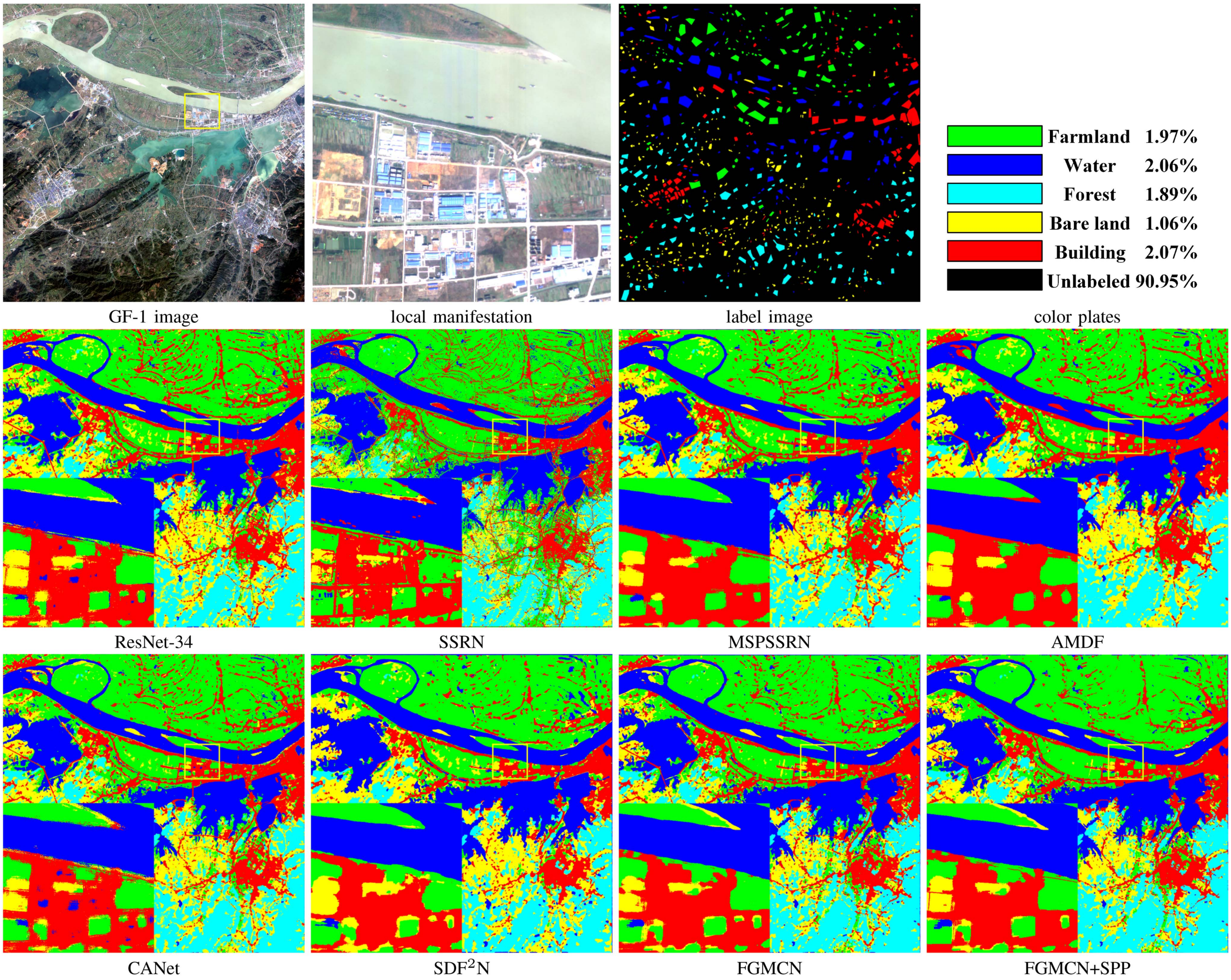
SPP is proposed in light of this. In parallel with the classification, small-scale segmentation is performed on the image to be classified. A superpixel is defined as the pixel set in a segmented image block. An image can be divided into multiple superpixels based on the similarity of feature, shape, or texture.

The superpixel segmentation results will be fused with the classification outcomes. Taking the superpixel S as an example, different categories C_n contain different number of pixels in S, and is defined as

$$P(C_1) + \cdots + P(C_n) = 1 \tag{8}$$

where $P(C_n)$ denotes the proportion of category n in S, that is, the number of pixels of category n contained in S is divided by the total number of pixels in S. The procedure of label correction is analogous to voting. A Superpixel is considered as a uniform class when the fraction of that class is dominant, whereas the remained portion is more likely to be of wrong labels.

An image segmentation method is chosen to meet the demand of our task. Small-scale segmentation results are desired because transition zones are typically tiny in size. Hu et al. [31] proposed a stepwise evolution analysis (SEA) framework. In SEA, the evolution of scale, local variance, and Moran's index are analyzed step-by-step, and four LV- and MI-metrics-based methods are technically integrated for automated scale parameterizations. Later, in [32], they experimentally concluded that the optimal scale of an object-based image classification work is a range rather than a single value, and demonstrated the possibility of the framework to automatically estimate the optimal classification scale. This technique is chosen as our segmentation tool because, with the right parameters, it can produce small block segmentation. Fig. 5 displays the results with the suggested segmentation tool.



GF-1 image local manifestation label image color plates

ResNet-34 SSRN MSPSSRN AMDF

CANet SDF^{2}N FGMCN FGMCN+SPP

Fig. 6. Classification maps of the GF-1 image.

IV. Experimental Scheme

The proposed method is tested on five satellite images. Six existing classification algorithms are compared to validate the performance. The details of competing algorithms are given in this section. The objective metrics are used to evaluate the experimental results.

A. Datasets

Five remote sensing images from GF-1, GF-2, DEIMOS-2, GeoEye-1, and Sentinel-2 satellites are used for the experiment. Images of GF-2 and DEIMOS-2 are from public datasets while others are labeled by us. Their spatial resolutions range from 1 to 10 m. Each image contains blue, green, red, and near infrared bands. The numbers of marked pixels are within the range of 900 000 to 32 000 000.

The GF-1 satellite image has a spatial resolution of 8 m and an image size of 4548 × 4503. This image was taken on December 4, 2014 at Poyang lake, Jiangxi province. The coverage range is $29^\circ 29' 57.84''$–$29^\circ 53' 6''$ N, $115^\circ 39' 50.4''$–$116^\circ 0' 46.8''$ E. The scene was divided into five categories, which consist of farmland, water, forest, bare land, and artificial building. There are 1 854 234 marked pixels in the image, of which 402 730 are farmland, 422 726 are water, 386 300 are forest, 217 864 are bare land, and 424 614 are building.

The GF-2 satellite image has a spatial resolution of 4 m and an image size of 7200 × 6800. This image is extracted from the Gaofen image dataset constructed by Tong et al. [33]. The image was captured on January 23, 2015 in Dongguan City, China. The scene is divided into four categories including building, farmland, forest, and water. There are 31 968 298 marked pixels in the image, of which 12 725 502 are building, 8 299 238 are farmland, 443 836 are forest, and 10 499 722 are water.

The DEIMOS-2 satellite image has a spatial resolution of 4 m and an image size of 2928×3249. This image was provided by the 2016 IEEE GRSS Data Fusion Contest [34]. The image is a level-1B image captured on May 30, 2015 in Vancouver, Canada. The scene was divided into 11 categories including vegetation, four kinds of building areas, boat, road, port, bridge, tree, and water. There are 3 045 244 marked pixels in the image,

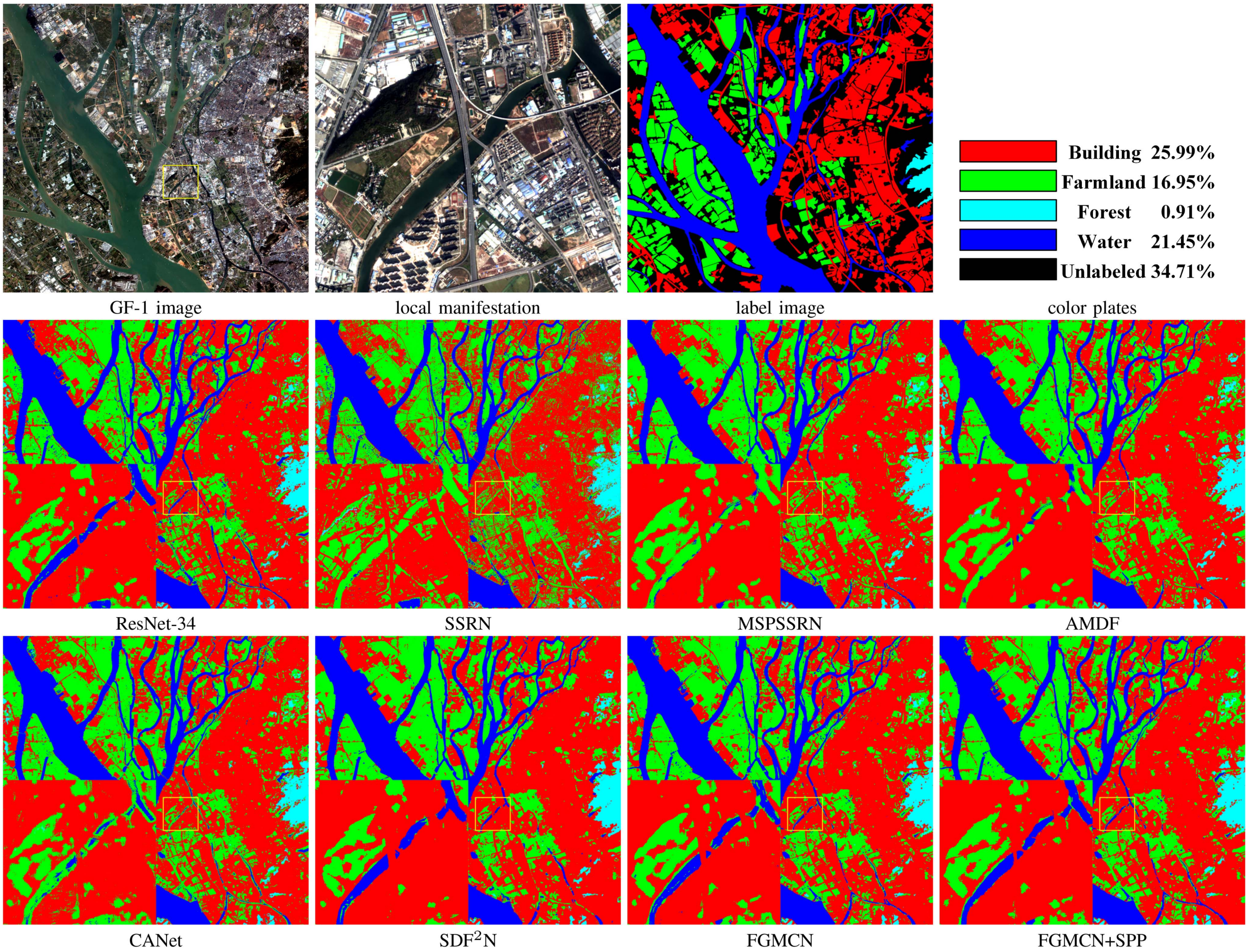


Fig. 7. Classification maps of the GF-2 image.

of which 52 289 are vegetation, 303 374 are building-1, 76 427 are building-2, 513 224 are building-3, 171 289 are building-4, 8 939 are boat, 60 265 are road, 80 905 are port, 6 882 are bridge, 463 384 are tree, and 1 308 267 are water.

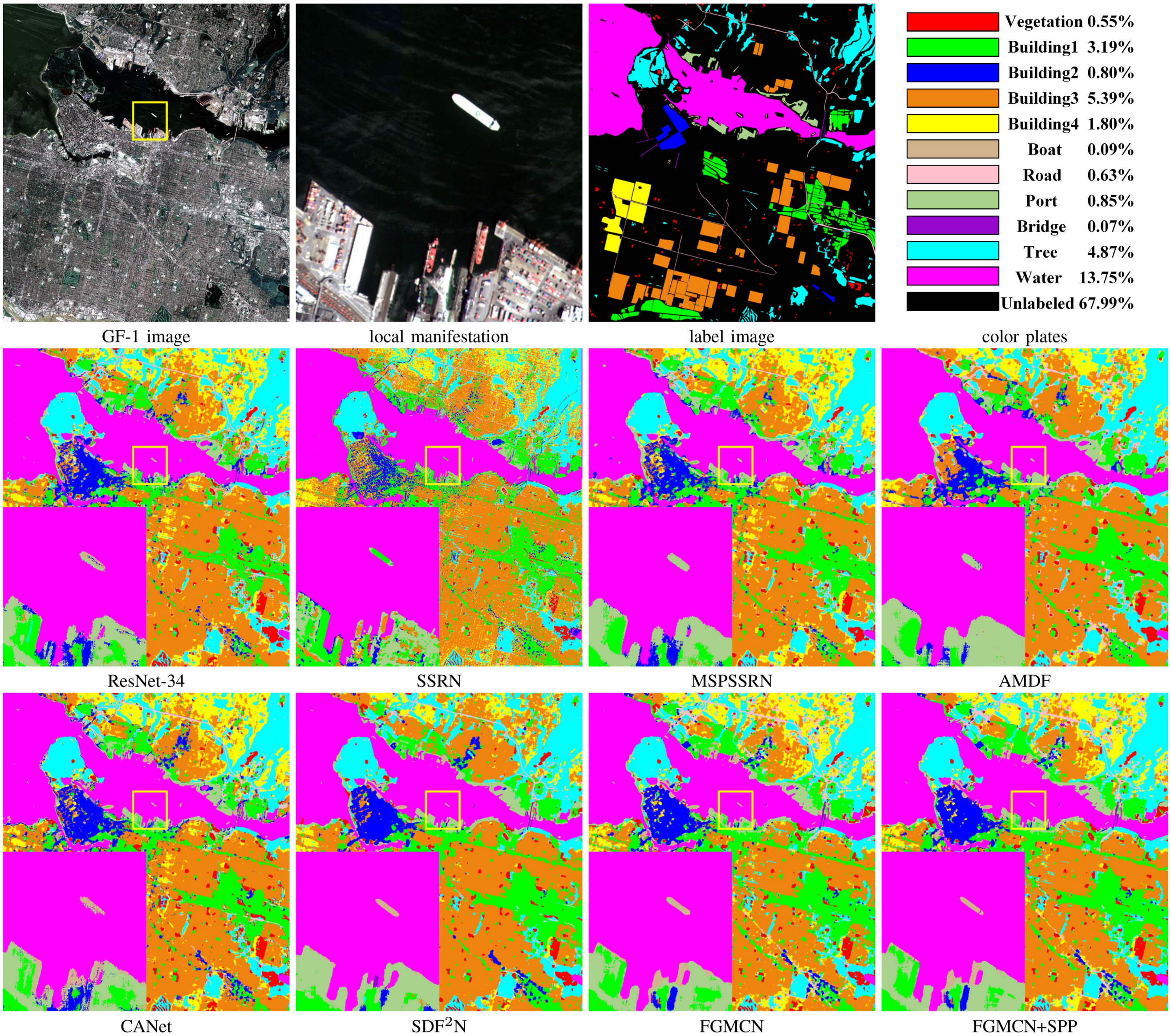
The GeoEye-1 satellite image has a spatial resolution of 1.65 m and an image size of 4399 × 4354. This image was captured on February 7, 2016 in Jingmen City, China. The coverage range is $30^\circ 40' 33.39''$–$30^\circ 45' 19.44''$ N, $112^\circ 15' 12.76''$–$112^\circ 20' 39.61''$ E. The scene was divided into seven categories including farmland, water, road, forest, building, bare land, and grassland. There are 925 749 marked pixels in the image, of which 427 918 are farmland, 272 225 are water, 23 165 are road, 83 584 are forest, 57 154 are building, 37 488 are bare land, and 24 215 are grassland.

The Sentinel-2 satellite image has a spatial resolution of 10 m and an image size of 10980 × 10980. This image was taken on October 3, 2021. The coverage range is $28^\circ 50' 22.27''$–$29^\circ 49' 35.07''$ N, $115^\circ 57' 52.49''$–$117^\circ 5' 59.82''$ E. The scene was divided into eight categories consisting of farmland, forest, artificial building, bare land, sediment in water, norm water, dark water, and sands. There are 2 131 302 marked pixels in the image, of which 184 310 are farmland, 474 531 are forest, 56 652 are building, 24 688 are bare land, 586 139 are sediment in water, 726 709 are normal water, 45 787 are dark water, and 32 486 are sands.

B. Cluster Sampling

Before the experiment, the cluster sampling strategy in [35] is adopted. Sampling strategies dividing the labeled data into train and test sets has a significant impact on the quality and reliability of the estimated generalization error [36], while the cluster sampling strategy can mostly ensure the fairness.

The K-means algorithm partitions all samples into two groups with regard to the spatial coordinates. One group is chosen at random to extract training samples for each category, and the other group are for test. For each category in GF-1, DEIMOS-2, GeoEye-1, and Sentinel-2, 10% of it in a group was randomly chosen for training while all the pixels in the other group are used for test. Because of the vast amount of labeled data in the GF-2 image, 1% of the entire dataset was randomly chosen from one group for training and 5% from the other group for test.



GF-1 image local manifestation label image color plates

ResNet-34 SSRN MSPSSRN AMDF

CANet SDF^{2}N FGMCN FGMCN+SPP

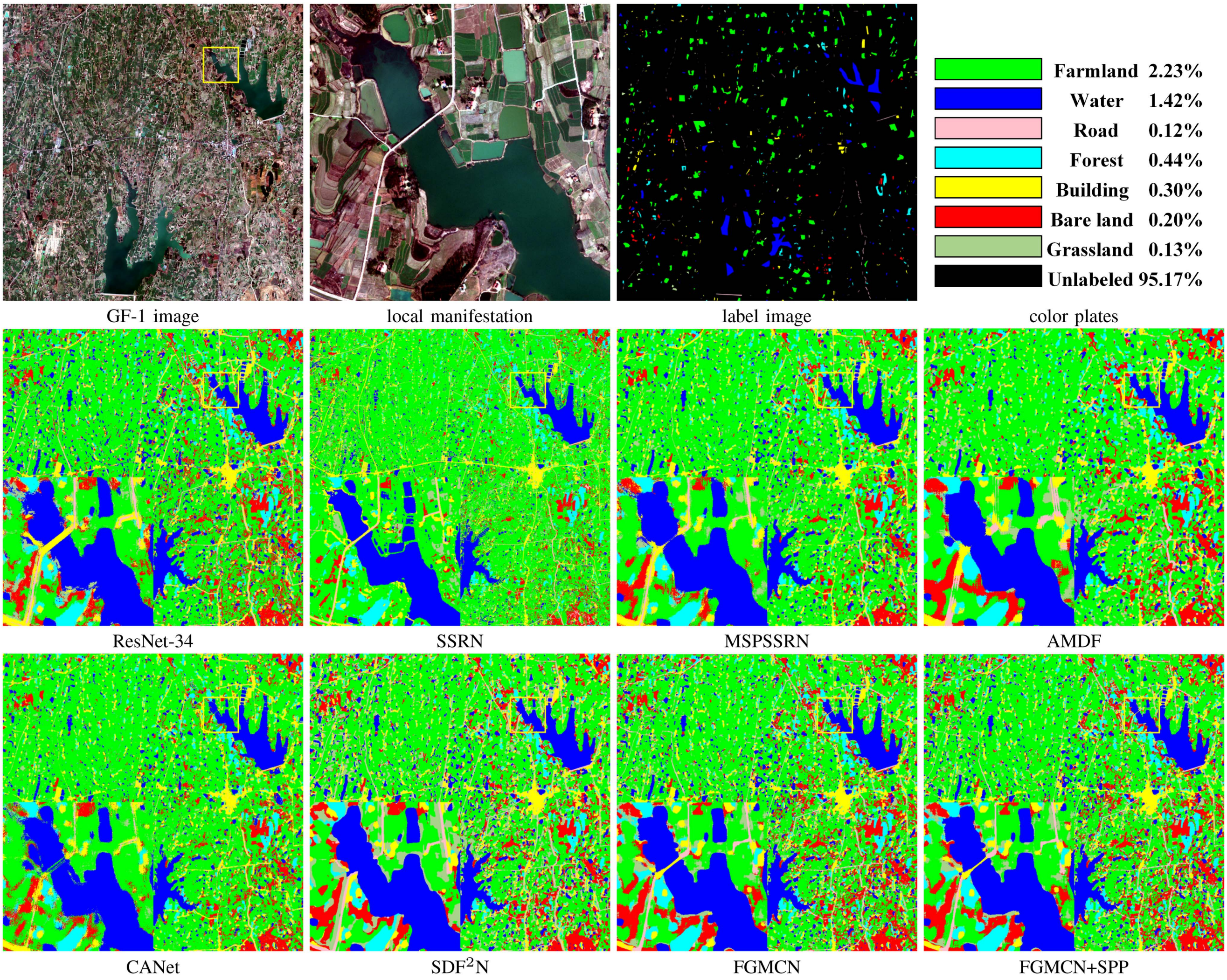
Fig. 8. Classification maps of the DEIMOS-2 image.

C. Parameter Setting

The proposed FGMCN and SPP methods are compared with some algorithms for performance validation, including ResNet-34 [19], SSRN [37], MSPSSRN [38], AMDF [3], CANet [39], and SDF^{2}N [40] to ascertain the efficacy of the proposed approach, which are all CNN-based. Among them, ResNet-34 is a standard residual network, CANet is a residual network with an attention mechanism, SSRN is for hyperspectral images, while MSPSSRN, AMDF, and SDF^{2}N are specifically designed for multispectral images. The programming language is PYTHON with the KERAS framework for deep learning.

Training parameters are set for the algorithms. The categorical cross entropy loss is used in all algorithms. The batch sizes of all algorithms are set to 64 to ensure fairness. ResNet-34, CANet, and the proposed FGMCN algorithm share the same parameter settings. The network optimization algorithm uses stochastic gradient descent (SGD) optimizer and trains 200 epochs. The learning rate is 0.001 in the 101–200 cycles and 0.0005 in the 100–200 cycles. The input patch has 27 × 27 pixels. The 7 × 7 average pooling layer at the end of ResNet-34 is removed to adapt to the input size. As for SSRN, MSPSSRN, AMDF, and SDF^{2}N, they use the original parameters. SSRN uses RMSProp optimizer and trains 200 epochs. The initial learning rate is 0.0003 and the input patch has 7 × 7 pixels. MSPSSRN uses the SGD optimizer and trains 200 epochs. The learning rate is 0.0003 in the 1–100 cycles and 0.00015 in the 101–200 cycles. The input patch has 27 × 27 pixels. AMDF uses the SGD optimizer and trains 200 epochs. The initial learning rate is 0.001 and the input patch has 31 × 31 pixels. SDF^{2}N uses the Adam optimizer and trains 100 epochs. The initial learning rate is 0.001 and the input patch has 33 × 33 pixels.

Metrics are used to evaluate the classification results. Accuracy is the most commonly used rule for this topic. In addition to



GF-1 image | local manifestation | label image | color plates

ResNet-34 | SSRN | MSPSSRN | AMDF

CANet | SDF^{2}N | FGMCN | FGMCN+SPP

Fig. 9. Classification maps of the GeoEye-1 image.

accuracy, recall, and F1-score are also evaluated for category by category. To give a general conclusion related to the accuracy, overall accuracy (OA), average accuracy (AA), and the Kappa coefficient are also used to measure the classification quality. All these metrics output scores within the range [0,1] where the higher gives the better.

V. Experimental Results

The experimental results are visually presented in this section. They are also quantitatively evaluated with metrics. The proposed FGMCN model and the SPP solution are validated separately.

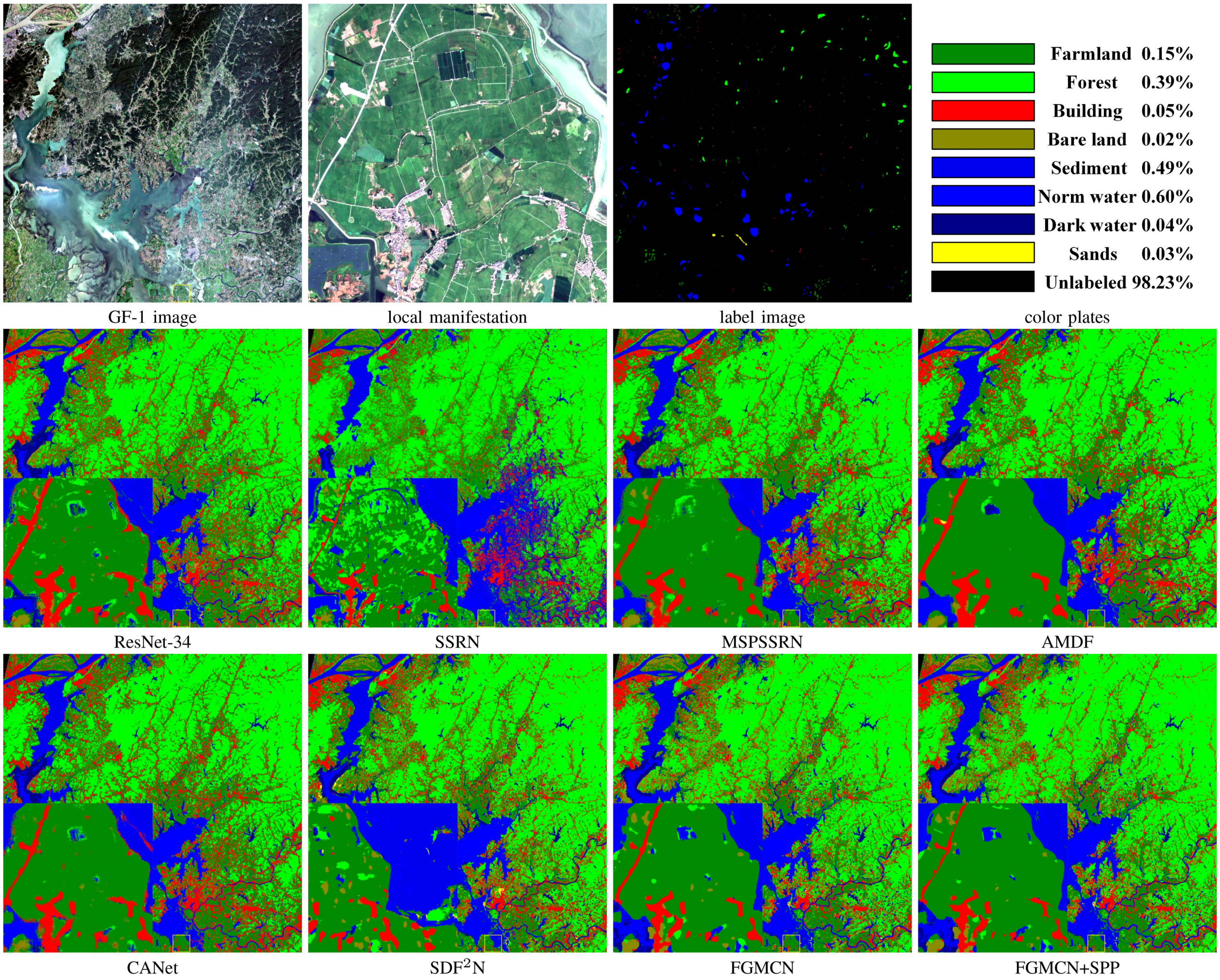
A. *Quantitative Comparison on Test Sets*

The test sets of the five images were tested using FGMCN and the competition algorithms. To avoid the effect of random clustering, each algorithm runs five times with random training and test set, and the average scores are recorded as the final result, which are presented in Tables I–V. The proposed SPP was not used at this stage because of the discrete test samples. Bold digits represent the best scores.

TABLE I
Assessment for the Test Dataset of the GF-1 Image

class	metric	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN
Farmland	accuracy	0.960	0.874	0.959	0.969	0.942	0.947	0.922
	recall	0.795	0.866	0.804	0.838	0.817	0.840	0.956
	F1-score	0.870	0.870	0.875	0.898	0.875	0.888	0.938
Water	accuracy	0.987	0.988	0.995	0.998	0.993	0.991	0.993
	recall	0.998	0.995	0.999	0.998	0.997	0.999	1.000
	F1-score	0.992	0.991	0.997	0.998	0.995	0.995	0.997
Forest	accuracy	0.999	0.995	0.999	0.999	0.998	0.999	0.999
	recall	0.988	0.986	0.989	0.997	0.989	0.996	0.999
	F1-score	0.994	0.991	0.994	0.998	0.994	0.998	0.999
Bareland	accuracy	0.529	0.530	0.557	0.552	0.538	0.626	0.817
	recall	0.884	0.506	0.882	0.727	0.854	0.864	0.659
	F1-score	0.662	0.518	0.683	0.628	0.660	0.720	0.729
Building	accuracy	0.994	0.955	0.997	0.957	0.993	0.995	0.983
	recall	0.971	0.974	0.992	0.991	0.967	0.983	0.995
	F1-score	0.983	0.964	0.995	0.974	0.980	0.989	0.989
	OA	0.931	0.921	0.940	0.937	0.931	0.945	**0.962**
	AA	0.894	0.868	0.902	0.895	0.893	0.912	**0.943**
	Kappa	0.911	0.897	0.922	0.919	0.911	0.930	**0.951**

In Tables I–V, OA, AA, and the Kappa coefficient indicate that the newly proposed FGMCN method performs better than



GF-1 image
local manifestation
label image
color plates
ResNet-34
SSRN
MSPSSRN
AMDF
CANet
SDF^{2}N
FGMCN
FGMCN+SPP

Fig. 10. Classification maps of the Sentinel-2 image.

competing algorithms. SSRN performs poorly, which is overweighed by ResNet-34 and ADMF. MSPSSRN, CANet, and SDF^{2}N are even better, but the proposed FGMCN algorithm gives the best results steadily.

The DEIMOS-2 classification is the most challenging because of the largest number of classes. There are large class imbalance between the different categories. Bridges are only identified by ResNet-34, CANet, SDF^{2}N, and our model. In this scene, our method has the highest stability, which can be partially explained by fine-grained multiscale learning.

Table IV shows that grassland is difficult to be identified for all algorithms because of insufficient labeled pixels and the similarity to forest or farmland. The confusion matrix shows that grassland is tend to be misclassified as farmland. The accuracy of other algorithms is around poor 45%. The proposed FGMCN algorithm gives the highest recall rate and F1-score.

TABLE II
Assessment for the Test Dataset of the GF-2 Image

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN
Building	accuracy	0.883	0.837	0.863	0.874	0.891	0.880	0.926
	recall	0.914	0.846	0.933	0.935	0.909	0.941	0.908
	F1-score	0.898	0.842	0.897	0.904	0.900	0.909	0.917
Farmland	accuracy	0.823	0.728	0.820	0.833	0.829	0.872	0.857
	recall	0.824	0.747	0.805	0.829	0.874	0.808	0.909
	F1-score	0.823	0.737	0.812	0.831	0.851	0.839	0.882
Forest	accuracy	0.924	0.889	0.823	0.885	0.875	0.877	0.877
	recall	0.926	0.914	0.966	0.962	0.917	0.969	0.959
	F1-score	0.925	0.901	0.889	0.922	0.895	0.921	0.916
Water	accuracy	0.970	0.954	0.984	0.984	0.978	0.964	0.967
	recall	0.924	0.910	0.899	0.903	0.907	0.935	0.945
	F1-score	0.946	0.931	0.940	0.942	0.942	0.949	0.956
	OA	0.895	0.846	0.887	0.897	0.901	0.905	**0.920**
	AA	0.900	0.852	0.873	0.894	0.893	0.898	**0.907**
	Kappa	0.842	0.768	0.831	0.845	0.851	0.857	**0.880**

B. *Quantitative Comparison on Full Images*

To verify the effectiveness of SPP, classification was performed on the full images using the trained models. The complete labeled pixels are evaluated and the results are presented in Tables VI–X. The last columns of these tables give the scores where SPP is used to correct the FGMCN results through the full images which are abbreviated as "+SPP."

Tables VI–X show that the classification accuracy is improved for all algorithms, which indirectly indicate the effectiveness of cluster sampling strategy for generalization comparison. The

TABLE III
ASSESSMENT FOR THE TEST DATASET OF THE DEIMOS-2 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN
Vegetation	accuracy	0.677	0.920	0.832	0.728	0.726	0.768	0.837
	recall	0.738	0.926	0.746	0.719	0.774	0.888	0.906
	F1-score	0.706	0.923	0.787	0.724	0.749	0.824	0.870
Building1	accuracy	0.680	0.516	0.667	0.672	0.697	0.628	0.668
	recall	0.825	0.688	0.881	0.910	0.906	0.941	0.963
	F1-score	0.746	0.589	0.759	0.773	0.788	0.752	0.789
Building2	accuracy	0.230	0.168	0.188	0.108	0.204	0.362	0.429
	recall	0.426	0.343	0.416	0.180	0.502	0.562	0.707
	F1-score	0.299	0.226	0.259	0.135	0.291	0.429	0.534
Building3	accuracy	0.792	0.723	0.874	0.825	0.872	0.842	0.867
	recall	0.987	0.937	0.986	0.973	0.982	0.929	0.966
	F1-score	0.879	0.816	0.927	0.893	0.923	0.883	0.914
Building4	accuracy	0.920	0.951	0.927	0.943	0.927	0.942	0.941
	recall	0.913	0.679	0.916	0.840	0.929	0.859	0.927
	F1-score	0.916	0.792	0.922	0.889	0.928	0.898	0.934
Boat	accuracy	0.653	0.839	0.603	0.367	0.569	0.558	0.273
	recall	0.294	0.004	0.121	0.073	0.323	0.532	0.626
	F1-score	0.406	0.009	0.202	0.122	0.412	0.542	0.380
Road	accuracy	0.525	0.566	0.498	0.172	0.482	0.478	0.683
	recall	0.108	0.298	0.100	0.048	0.184	0.107	0.093
	F1-score	0.179	0.390	0.167	0.075	0.267	0.172	0.164
Port	accuracy	0.581	0.801	0.555	0.687	0.644	0.694	0.666
	recall	0.548	0.440	0.664	0.715	0.583	0.724	0.546
	F1-score	0.564	0.568	0.605	0.700	0.612	0.708	0.600
Bridge	accuracy	0.123	0	0	0	0.091	0.356	0.343
	recall	0.014	0	0	0	0.027	0.093	0.093
	F1-score	0.025	0	0	0	0.042	0.147	0.146
Tree	accuracy	0.581	0.801	0.555	0.941	0.962	0.970	0.666
	recall	0.913	0.969	0.931	0.904	0.891	0.906	0.965
	F1-score	0.710	0.877	0.695	0.922	0.925	0.936	0.788
Water	accuracy	0.999	1.000	0.998	0.998	0.999	0.999	0.999
	recall	0.971	0.985	0.973	0.984	0.984	0.983	0.986
	F1-score	0.985	0.992	0.986	0.991	0.991	0.991	0.992
	OA	0.905	0.886	0.914	0.909	0.915	0.914	**0.926**
	AA	0.648	0.672	0.645	0.586	0.652	0.691	**0.698**
	Kappa	0.856	0.826	0.870	0.862	0.871	0.870	**0.887**

TABLE IV
ASSESSMENT FOR THE TEST DATASET OF THE GEOEYE-1 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN
Farmland	accuracy	0.886	0.869	0.906	0.908	0.884	0.921	0.937
	recall	0.975	0.962	0.968	0.957	0.984	0.937	0.936
	F1-score	0.928	0.913	0.936	0.932	0.931	0.929	0.936
Water	accuracy	0.982	0.947	0.988	0.990	0.987	0.984	0.981
	recall	0.991	0.981	0.989	0.994	0.996	0.982	0.991
	F1-score	0.986	0.964	0.988	0.992	0.991	0.983	0.986
Road	accuracy	0.629	0.645	0.745	0.626	0.739	0.967	0.852
	recall	0.652	0.714	0.552	0.245	0.422	0.752	0.768
	F1-score	0.640	0.677	0.634	0.352	0.537	0.846	0.808
Forest	accuracy	0.939	0.973	0.934	0.897	0.951	0.980	0.952
	recall	0.809	0.816	0.849	0.850	0.810	0.803	0.879
	F1-score	0.869	0.888	0.889	0.873	0.875	0.883	0.914
Building	accuracy	0.907	0.841	0.877	0.874	0.858	0.911	0.947
	recall	0.796	0.743	0.882	0.873	0.887	0.938	0.887
	F1-score	0.848	0.789	0.880	0.874	0.872	0.924	0.916
Bareland	accuracy	0.806	0.830	0.949	0.926	0.886	0.936	0.830
	recall	0.478	0.545	0.656	0.718	0.533	0.651	0.808
	F1-score	0.600	0.658	0.776	0.809	0.666	0.767	0.819
Grassland	accuracy	0.425	0.458	0.475	0.412	0.480	0.382	0.451
	recall	0.248	0.064	0.382	0.409	0.140	0.587	0.587
	F1-score	0.314	0.112	0.423	0.411	0.217	0.463	0.510
	OA	0.893	0.880	0.912	0.905	0.900	0.908	**0.920**
	AA	0.796	0.795	0.839	0.805	0.826	**0.869**	0.850
	Kappa	0.834	0.814	0.865	0.855	0.845	0.862	**0.880**

TABLE V
ASSESSMENT FOR THE TEST DATASET OF THE SENTINEL-2 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN
Farmland	accuracy	0.916	0.857	0.861	0.873	0.877	0.854	0.905
	recall	0.892	0.714	0.898	0.910	0.912	0.941	0.938
	F1-score	0.904	0.779	0.879	0.891	0.894	0.892	0.921
Forest	accuracy	0.944	0.596	0.961	0.987	0.986	0.972	0.957
	recall	0.955	0.958	0.973	0.966	0.961	0.947	0.957
	F1-score	0.949	0.735	0.967	0.976	0.973	0.959	0.957
Building	accuracy	0.907	0.949	0.961	0.972	0.903	0.951	0.987
	recall	0.890	0.895	0.895	0.838	0.920	0.899	0.909
	F1-score	0.898	0.921	0.927	0.900	0.911	0.923	0.946
Bareland	accuracy	0.789	0.911	0.943	0.852	0.789	0.900	0.847
	recall	0.901	0.958	0.941	0.836	0.705	0.947	0.987
	F1-score	0.841	0.934	0.942	0.844	0.745	0.922	0.912
Sediment	accuracy	0.849	0.842	0.819	0.857	0.868	0.849	0.857
	recall	0.893	0.972	0.910	0.921	0.965	0.918	0.960
	F1-score	0.870	0.902	0.862	0.888	0.914	0.882	0.906
NormWater	accuracy	0.822	0.887	0.813	0.855	0.899	0.858	0.920
	recall	0.605	0.549	0.550	0.640	0.633	0.692	0.755
	F1-score	0.697	0.678	0.657	0.732	0.743	0.765	0.830
DarkWater	accuracy	0.113	0.030	0.088	0.054	0.049	0.213	0.131
	recall	0.741	0.065	0.519	0.303	0.267	0.597	0.319
	F1-score	0.196	0.041	0.151	0.092	0.083	0.311	0.186
Sands	accuracy	0.890	0	0.167	0	0.969	0.628	0.940
	recall	0.072	0	0	0	0.145	0.011	0.021
	F1-score	0.133	0	0	0	0.252	0.021	0.042
	OA	0.802	0.777	0.791	0.818	0.833	0.841	**0.874**
	AA	0.779	0.634	0.702	0.681	0.793	0.778	**0.818**
	Kappa	0.740	0.703	0.725	0.760	0.779	0.789	**0.830**

TABLE VI
ASSESSMENT FOR THE WHOLE GF-1 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN	+SPP
Farmland	accuracy	0.977	0.841	0.972	0.977	0.942	0.974	0.935	0.940
	recall	0.883	0.914	0.881	0.901	0.926	0.894	0.962	0.967
	F1-score	0.928	0.876	0.924	0.937	0.934	0.932	0.948	0.953
Water	accuracy	0.996	0.988	0.997	0.998	0.993	0.995	0.998	0.999
	recall	0.999	0.984	0.998	0.996	0.999	0.999	0.999	0.999
	F1-score	0.997	0.986	0.998	0.997	0.996	0.997	0.999	0.999
Forest	accuracy	0.999	0.974	0.999	0.997	0.999	0.999	0.999	0.997
	recall	0.993	0.958	0.995	0.997	0.995	0.996	0.998	0.998
	F1-score	0.996	0.966	0.997	0.997	0.997	0.998	0.999	0.998
Bareland	accuracy	0.798	0.755	0.800	0.815	0.851	0.816	0.918	0.926
	recall	0.957	0.663	0.944	0.886	0.891	0.964	0.886	0.894
	F1-score	0.870	0.706	0.866	0.849	0.871	0.884	0.902	0.910
Building	accuracy	0.995	0.940	0.994	0.961	0.996	0.997	0.998	0.997
	recall	0.986	0.938	0.993	0.992	0.985	0.981	0.988	0.987
	F1-score	0.990	0.939	0.993	0.976	0.990	0.989	0.993	0.992
	OA	0.965	0.915	0.964	0.962	0.966	0.967	0.975	**0.977**
	AA	0.953	0.899	0.952	0.950	0.956	0.956	0.970	**0.972**
	Kappa	0.956	0.893	0.955	0.952	0.958	0.959	0.969	**0.970**

TABLE VII
ASSESSMENT FOR THE WHOLE GF-2 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN	+SPP
Building	accuracy	0.925	0.871	0.919	0.919	0.928	0.924	0.952	0.957
	recall	0.942	0.871	0.955	0.958	0.952	0.965	0.957	0.958
	F1-score	0.933	0.871	0.936	0.938	0.940	0.944	0.954	0.957
Farmland	accuracy	0.898	0.757	0.889	0.894	0.878	0.925	0.917	0.922
	recall	0.861	0.803	0.883	0.884	0.881	0.878	0.920	0.931
	F1-score	0.879	0.779	0.886	0.889	0.880	0.901	0.919	0.927
Forest	accuracy	0.932	0.888	0.895	0.923	0.913	0.923	0.959	0.962
	recall	0.966	0.917	0.979	0.979	0.965	0.985	0.969	0.972
	F1-score	0.949	0.902	0.935	0.950	0.938	0.953	0.964	0.967
Water	accuracy	0.968	0.956	0.989	0.988	0.986	0.978	0.982	0.985
	recall	0.977	0.908	0.944	0.944	0.950	0.963	0.973	0.975
	F1-score	0.973	0.931	0.966	0.965	0.968	0.971	0.978	0.980
	OA	0.933	0.866	0.933	0.934	0.933	0.942	0.953	**0.957**
	AA	0.931	0.868	0.923	0.931	0.926	0.938	0.952	**0.956**
	Kappa	0.899	0.800	0.899	0.901	0.899	0.913	0.929	**0.935**

scores of OA, AA, and the Kappa coefficient indicate that our method performs far better than competing methods. When SPP was added for classification refinement, OA, AA, and the Kappa coefficient are slightly improved for almost all images.

In addition, the value of the optimal SPP threshold varies depending on the number of annotated pixels. Based on the

TABLE VIII
ASSESSMENT FOR THE WHOLE DEIMOS-2 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN	+SPP
	accuracy	0.833	0.609	0.941	0.893	0.895	0.911	0.956	0.971
Vegetation	recall	0.908	0.426	0.883	0.863	0.937	0.948	0.972	0.971
	F1-score	0.869	0.501	0.911	0.878	0.915	0.929	0.964	0.971
	accuracy	0.894	0.701	0.853	0.861	0.901	0.899	0.884	0.891
Building1	recall	0.877	0.788	0.936	0.961	0.935	0.953	0.982	0.987
	F1-score	0.885	0.742	0.892	0.908	0.918	0.925	0.931	0.936
	accuracy	0.843	0.685	0.802	0.840	0.854	0.921	0.904	0.904
Building2	recall	0.922	0.603	0.863	0.837	0.950	0.983	0.975	0.979
	F1-score	0.881	0.641	0.831	0.839	0.899	0.951	0.938	0.940
	accuracy	0.902	0.766	0.937	0.916	0.933	0.910	0.954	0.957
Building3	recall	0.984	0.908	0.974	0.984	0.987	0.996	0.988	0.995
	F1-score	0.941	0.831	0.955	0.949	0.960	0.951	0.971	0.976
	accuracy	0.929	0.720	0.915	0.942	0.930	0.975	0.961	0.965
Building4	recall	0.912	0.757	0.926	0.894	0.939	0.874	0.948	0.968
	F1-score	0.921	0.738	0.921	0.917	0.935	0.922	0.955	0.966
	accuracy	0.795	0.222	0.751	0.705	0.449	0.799	0.744	0.787
Boat	recall	0.503	0.001	0.285	0.248	0.824	0.590	0.795	0.776
	F1-score	0.617	0.003	0.413	0.367	0.581	0.679	0.768	0.781
	accuracy	0.702	0.187	0.701	0.405	0.763	0.836	0.911	0.929
Road	recall	0.394	0.140	0.314	0.161	0.426	0.328	0.374	0.344
	F1-score	0.505	0.160	0.433	0.230	0.546	0.472	0.531	0.503
	accuracy	0.773	0.766	0.677	0.797	0.877	0.812	0.825	0.846
Port	recall	0.749	0.437	0.772	0.791	0.780	0.786	0.753	0.764
	F1-score	0.761	0.557	0.721	0.794	0.826	0.799	0.787	0.803
	accuracy	0.832	0	0.970	0	0.638	0.863	0.804	0.798
Bridge	recall	0.285	0	0.019	0	0.344	0.329	0.437	0.420
	F1-score	0.425	0	0.037	0	0.447	0.476	0.566	0.550
	accuracy	0.973	0.916	0.983	0.974	0.994	0.982	0.992	0.993
Tree	recall	0.967	0.832	0.977	0.963	0.975	0.980	0.990	0.990
	F1-score	0.970	0.872	0.980	0.968	0.985	0.981	0.991	0.992
	accuracy	0.999	0.986	0.998	0.997	0.999	0.997	0.999	0.999
Water	recall	0.991	0.979	0.984	0.991	0.988	0.990	0.991	0.992
	F1-score	0.995	0.982	0.991	0.994	0.994	0.994	0.995	0.996
	OA	0.946	0.858	0.945	0.945	0.958	0.957	0.966	**0.969**
	AA	0.861	0.596	0.866	0.757	0.839	0.900	0.903	**0.913**
	Kappa	0.928	0.810	0.927	0.926	0.944	0.943	0.955	**0.958**

TABLE IX
ASSESSMENT FOR THE WHOLE GEOEYE-1 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN	+SPP
	accuracy	0.939	0.817	0.936	0.935	0.928	0.951	0.962	0.962
Farmland	recall	0.983	0.936	0.981	0.973	0.981	0.966	0.979	0.980
	F1-score	0.960	0.872	0.958	0.954	0.954	0.959	0.970	0.971
	accuracy	0.992	0.971	0.995	0.993	0.994	0.993	0.993	0.993
Water	recall	0.995	0.918	0.994	0.994	0.997	0.992	0.995	0.995
	F1-score	0.994	0.944	0.994	0.993	0.996	0.993	0.994	0.994
	accuracy	0.896	0.295	0.881	0.879	0.896	0.917	0.983	0.984
Road	recall	0.821	0.142	0.766	0.581	0.641	0.708	0.828	0.828
	F1-score	0.857	0.191	0.819	0.699	0.748	0.799	0.899	0.899
	accuracy	0.987	0.916	0.977	0.962	0.985	0.985	0.986	0.986
Forest	recall	0.928	0.692	0.948	0.945	0.941	0.926	0.959	0.959
	F1-score	0.956	0.788	0.962	0.954	0.963	0.954	0.972	0.972
	accuracy	0.933	0.593	0.916	0.908	0.889	0.919	0.961	0.961
Building	recall	0.901	0.641	0.913	0.912	0.904	0.941	0.961	0.962
	F1-score	0.917	0.616	0.914	0.910	0.896	0.930	0.961	0.961
	accuracy	0.830	0.566	0.902	0.860	0.812	0.938	0.886	0.887
Bareland	recall	0.744	0.435	0.726	0.776	0.735	0.749	0.873	0.874
	F1-score	0.784	0.492	0.805	0.816	0.771	0.833	0.879	0.880
	accuracy	0.651	0.376	0.568	0.467	0.606	0.476	0.641	0.658
Grassland	recall	0.429	0.205	0.389	0.377	0.300	0.673	0.599	0.597
	F1-score	0.517	0.265	0.462	0.417	0.402	0.557	0.619	0.626
	OA	0.948	0.831	0.947	0.939	0.941	0.946	0.963	**0.964**
	AA	0.890	0.648	0.882	0.858	0.873	0.883	0.916	**0.919**
	Kappa	0.923	0.745	0.921	0.910	0.913	0.920	0.945	**0.947**

experiments, it is recommended to set the threshold to 0.90 for less than one million annotated pixels, 0.75 for millions of annotated pixels, and 0.60 for tens of millions of annotated pixels. To pursue the best combination performance, various thresholds can be set for separate categories.

TABLE X
ASSESSMENT FOR THE WHOLE SENTINEL-2 IMAGE

class	metrics	ResNet-34	SSRN	MSPSSRN	AMDF	CANet	SDF^{2}N	FGMCN	+SPP
	accuracy	0.932	0.789	0.883	0.893	0.888	0.874	0.926	0.933
Farmland	recall	0.916	0.749	0.912	0.923	0.924	0.925	0.953	0.962
	F1-score	0.924	0.768	0.897	0.908	0.906	0.899	0.939	0.947
	accuracy	0.984	0.829	0.987	0.995	0.995	0.977	0.988	0.991
Forest	recall	0.986	0.977	0.991	0.989	0.985	0.985	0.988	0.990
	F1-score	0.985	0.897	0.989	0.992	0.990	0.981	0.988	0.990
	accuracy	0.957	0.893	0.961	0.962	0.919	0.989	0.995	0.995
Building	recall	0.949	0.740	0.946	0.913	0.957	0.956	0.963	0.960
	F1-score	0.953	0.809	0.954	0.937	0.937	0.973	0.979	0.977
	accuracy	0.908	0.779	0.950	0.913	0.899	0.912	0.941	0.934
Bareland	recall	0.956	0.527	0.950	0.906	0.789	0.991	0.995	0.994
	F1-score	0.931	0.629	0.950	0.910	0.840	0.949	0.967	0.963
	accuracy	0.911	0.861	0.885	0.896	0.903	0.888	0.916	0.916
Sediment	recall	0.934	0.979	0.937	0.946	0.974	0.999	0.978	0.987
	F1-score	0.922	0.916	0.911	0.920	0.937	0.940	0.946	0.950
	accuracy	0.936	0.952	0.928	0.940	0.956	0.995	0.972	0.980
NormWater	recall	0.839	0.775	0.811	0.833	0.841	0.857	0.904	0.902
	F1-score	0.885	0.855	0.865	0.883	0.895	0.921	0.937	0.940
	accuracy	0.355	0.486	0.339	0.332	0.325	0.532	0.584	0.579
DarkWater	recall	0.912	0.584	0.804	0.768	0.705	0.809	0.812	0.823
	F1-score	0.512	0.530	0.477	0.464	0.445	0.642	0.679	0.680
	accuracy	0.993	0.999	1.000	0.941	0.994	0.997	0.999	0.999
Sands	recall	0.690	0.644	0.626	0.633	0.573	0.666	0.673	0.670
	F1-score	0.814	0.783	0.770	0.757	0.727	0.799	0.805	0.802
	OA	0.908	0.864	0.897	0.905	0.912	0.931	0.944	**0.947**
	AA	0.872	0.824	0.867	0.859	0.860	0.896	0.915	**0.916**
	Kappa	0.879	0.819	0.864	0.875	0.884	0.908	0.926	**0.930**

C. Visual Comparison

The classification results for the whole images are shown in Figs. 6–10 where local blocks are magnified to compare the details. Fig. 7 shows that the river in the city is tend to be recognized as farmland except for ResNet-34, SDF^{2}N, and our method. In Fig. 8, the boats can be identified only by SDF^{2}N and our method.

The results confirm that SPP can improve accuracy and reduce fragmentation. For example, roads in Fig. 6 are unrecognizable in the FGMCN result but are recovered by SPP. The fragments of the river areas in Fig. 7 are also effectively removed by SPP. As can be seen in these figures, SPP can outline ground types at a small scale, which smooths out minor classification errors in the category transition regions.

VI. ABLATION STUDY

The key to the performance improvement achieved by our FGMCN model is the DSB and MRB modules designed for the classification problem. Distinguishing from existing deep learning-based classification models, our FGMCN model first uses the DSB to automatically search for branches that fit various structures while performing downscaling. Then, MRB is used for feature extraction to obtain fine-grained distinguishable perceptual fields.

To illustrate the contribution of these two modules, an ablation study was conducted. In this experiment, the DSB and MRB components were selectively removed to understand the contribution to the overall classification model. Specifically, DSB was replaced by a 3 × 3 convolution with the stride of 2, and MRB was replaced by a common residual block. The modified FGMCN model was then trained, and the convergence is presented in Fig. 11 by assessing the variation of the validation

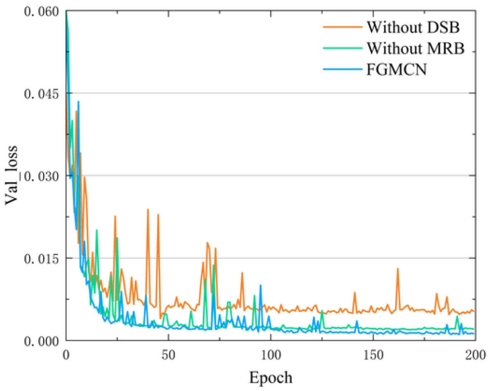


Fig. 11. Validation errors with regard to training epochs on the GF-1 image.

TABLE XI
ASSESSMENT THE ABLATION RESULTS ON THE GF-1 IMAGE (F1-SCORE FOR SUB CLASSES)

	Farmland	Water	Forest	Bareland	Building	OA	AA	Kappa
without DSB	0.852	0.999	0.995	0.660	0.990	0.930	0.894	0.911
without MRB	0.910	0.996	0.998	0.754	0.982	0.953	0.924	0.940
DSB+MRB	0.938	0.997	0.999	0.729	0.989	0.962	0.943	0.951

error during the training of the model. The assessment results are presented in Table XI. In general, error increases significantly when DSB is removed, and removing MRB contributes to incur slight accuracy loss, too. In contrast, FGMCN has the fastest convergence and the lowest errors on the validation dataset.

VII. DISCUSSION

As mentioned in the first section, our classification method is pixel-wise and CNN-based. The motivations are from the practical purpose targeting the best performance as well as convenience. In this section, the advantages of the two strategies are discussed.

To compare the performance difference between pixel-wise classification and patch-wise classification, an additional experiment is conducted by introducing a patch-wise classification algorithm for comparison. Since patch-wise training requires complete patch labels, it is tested only on the GF-2 image. CAEN [2] is chosen as the competing algorithm, which is patch-wise-based. To train CAEN, the Adam optimizer is used iterating 80 epoches with a learning rate of 0.001. The input patch size is set to 56 × 56 pixels.

Two partition ratios are used on the GF-2 data to verify the performance difference. In the first test, 10% of the dataset is used for training and 50% for testing, as is the same to the proportion used in the experimental section. In this case, the OA score is 0.899 and the Kappa score is 0.850. By comparing the results with the scores in Table II, it is concluded that CAEN performs poorly than CANet, SDF^{2}N, and the proposed FGMCN model. In order to explore the performance boundary of the patch-wise classification method, 50% of the dataset is used for training and 50% for test, which is consistent with the ratio in this article [2]. In this case, the OA score is 0.907 and the Kappa score is 0.861. The new results are getting better and only lower than FGMCN. However, patches in the training and test datasets are much similar as the high ratio makes the random clustering not feasible for a fair comparison. In other words, pursuing the extreme performance of patch-wise classification methods are in the cost of huge consistently labeled samples that are difficult to be satisfied. In contrast, our pixel-level classification method achieves higher accuracy with far less tags that are discrete and irregular.

TABLE XII
CLASSIFICATION ACCURACY CITED FROM [40]

image	metrics	SVM	RF	SSRN	SpectralFormer	SDF^{2}N
QuickBird	OA	0.903	0.910	0.938	0.912	0.971
	Kappa	0.974	0.883	0.919	0.885	0.963
GF-2	OA	0.888	0.885	0.946	0.910	0.966
	Kappa	0.834	0.827	0.920	0.865	0.948
Hyperspectral	OA	0.946	0.950	0.922	0.949	0.971
	Kappa	0.929	0.934	0.896	0.932	0.961

The poor classification effect of transformer can be indirectly demonstrated by the results of the existing literature. In the SDF^{2}N study, SpectralFormer [41], a transformer-based classification method, was tested on three images, including two multispectral images and an airborne hyperspectral image. The evaluation results of SDF^{2}N and SpectralFormer are partly shown in Table XII, where the Transformer-based SpectralFormer algorithm is not as good as the CNN-based SDF^{2}N method. On the other hand, Tables I–V show that the proposed FGMCN method performs better than SDF^{2}N for multispectral classification, and the scores of MSPSSRN and CANet are similar to SDF^{2}N, which indicate the superiority of CNN-based methods than Transformer-based methods.

VIII. CONCLUSION

This study introduces a novel classification method that utilizes FGMCN and SPP to enhance the quality of multispectral classification in high-resolution images. To improve the effectiveness of learning multiscale information images, the proposed method constructs a multiscale residual network at a finer scale. The proposed method is compared with six widely used image classification algorithms on five remote sensing images acquired by GF-1, GF-2, DEIMOS-2, GeoEye-1, and Sentinel-2 satellites. The experimental results demonstrate that the proposed method performs well in terms of OA, AA, and the Kappa coefficient, with good classification accuracy for high-resolution multispectral images. Additionally, SPP can reduce the speckles for pixel-wise classification results substantially, thereby improving both accuracy and visual acceptance of the proposed method.

REFERENCES

[1] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, “Context-aware cascade network for semantic labeling in VHR image,” in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 575–579.

[2] W. Liang, Y. Wu, M. Li, and Y. Cao, “High-resolution SAR image classification using context-aware encoder network and hybrid conditional random field model,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5317–5335, Aug. 2020.

[3] G. Li, L. Li, H. Zhu, X. Liu, and L. Jiao, “Adaptive multiscale deep fusion residual network for remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8506–8521, Nov. 2019.

[4] W. Hua, S. Wang, W. Xie, Y. Guo, and X. Jin, “Dual-channel convolutional neural network for polarimetric SAR images classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3201–3204.

[5] R. Hang, P. Yang, F. Zhou, and Q. Liu, “Multiscale progressive segmentation network for high-resolution remote sensing imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2022, Art. no. 5412012.

[6] B. Luo and L. Zhang, “Robust autodual morphological profiles for the classification of high-resolution satellite images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1451–1462, Feb. 2014.

[7] A. Marinoni and P. Gamba, “Unsupervised data driven feature extraction by means of mutual information maximization,” *IEEE Trans. Comput. Imag.*, vol. 3, no. 2, pp. 243–253, Jun. 2017.

[8] S. Huang, H. Zhang, and A. PižUrica, “Hybrid-hypergraph regularized multiview subspace clustering for hyperspectral images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5505816.

[9] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[10] P. Gislason, J. Benediktsson, and J. Sveinsson, “Random forests for land cover classification,” *Pattern Recognit. Lett.*, vol. 27, pp. 294–300, 2006.

[11] H.-B. Wang and J.-H. Ma, “A classification method of multispectral images which is based on fuzzy SVM,” in *Proc. Int. Conf. Comput. Sci. Softw. Eng.*, 2008, vol. 1, pp. 815–818.

[12] D. Liu, L. Han, and X. Han, “High spatial resolution remote sensing image classification based on deep learning,” *Acta Optica Sinica*, vol. 36, 2016, Art. no. 0428001.

[13] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, “Learning to diversify deep belief networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[14] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[15] X. Chen, L. Ma, and X. Yang, “Stacked denoise autoencoder based feature extraction and classification for hyperspectral images,” *J. Sensors*, vol. 2016, 2016, Art. no. 3632943.

[16] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Convolutional neural networks for large-scale remote-sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[17] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, “3D convolutional neural networks for crop classification with multi-temporal remote sensing images,” *Remote. Sens.*, vol. 10, 2018, Art. no. 75.

[18] X. Liu et al., “Deep multiple instance learning-based spatial–spectral classification for PAN and MS imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 461–473, Jan. 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[20] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representations for efficient semantic segmentation,” in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[21] K. K. Gadiraju, B. Ramachandra, Z. Chen, and R. R. Vatsavai, “Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2020, pp. 3234–3242.

[22] R. Hang, X. Qian, and Q. Liu, “Cross-modality contrastive learning for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jul. 2022, Art. no. 5532812.

[23] D. S. Zhang, L. X. Cong, Z. Q. Wang, H. Chen, and F. Wang, “Object-oriented Zhangjiangkou mangrove communities classification using quickbird imagery,” *Adv. Mater. Res.*, vol. 605–607, pp. 2274–2278, 2012.

[24] F. Mirzapour and H. Ghassemian, “Object-based multispectral image segmentation and classification,” in *Proc. 7th Int. Symp. Telecommun.*, 2014, pp. 430–435.

[25] B. Jin, P. Ye, X. Zhang, W. Song, and S. Li, “Object-oriented method combined with deep convolutional neural networks for land-use-type classification of remote sensing images,” *J. Indian Soc. Remote Sens.*, vol. 47, pp. 951–965, 2019.

[26] S. Baroud, S. Chokri, S. Belhaous, Z. Hidila, and M. Mestari, “An artificial neural network combined to object oriented method for land cover classification of high resolution RGB remote sensing images,” in *Proc. Smart Appl. Data Anal., 3rd Int. Conf. Proc.*, 2020, pp. 221–232.

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, pp. 4278–4284.

[28] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, “Residual networks of residual networks: Multilevel residual networks,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1303–1314, Jun. 2018.

[29] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2Net: A new multi-scale backbone architecture,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[30] R. Hang, Q. Liu, and Z. Li, “Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery,” *IEEE Trans. Image Process.*, vol. 30, pp. 7256–7265, Aug. 2022.

[31] Z. Hu, Q. Zhang, Q. Zou, Q. Li, and G. Wu, “Stepwise evolution analysis of the region-merging segmentation for scale parameterization,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2461–2472, Jul. 2018.

[32] Z. Hu, T. Shi, C. Wang, Q. Li, and G. Wu, “Scale-sets image classification with hierarchical sample enriching and automatic scale selection,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 105, 2021, Art. no. 102605.

[33] X.-Y. Tong et al., “Land-cover classification with high-resolution remote sensing images using transferable deep models,” *Remote Sens. Environ.*, vol. 237, 2018, Art. no. 111322.

[34] L. Mou et al., “Multitemporal very high resolution from space: Outcome of the 2016 IEEE GRSS data fusion contest,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3435–3447, Aug. 2017.

[35] R. Hänsch, A. Ley, and O. Hellwich, “Correct and still wrong: The relationship between sampling strategies and the estimation of the generalization error,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3672–3675.

[36] J. Liang, J. Zhou, Y.-T. Qian, L. Wen, X. Bai, and Y. Gao, “On the sampling strategy for evaluation of spectral–spatial methods in hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, Feb. 2017.

[37] Z. Zhong, J. Li, Z. Luo, and M. Chapman, “Spectral spatial residual network for hyperspectral image classification: A 3-D deep learning framework,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[38] Y. Huang, J. Wei, W. Tang, and C. He, “Pyramid convolutional neural networks and bottleneck residual modules for classification of multispectral images,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1949–1952.

[39] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.

[40] S. Liu et al., “A shallow-to-deep feature fusion network for VHR remote sensing image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, Jun. 2022.

[41] D. Hong et al., “SpectralFormer: Rethinking hyperspectral image classification with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Feb. 2022.