Object-Scale Adaptive Convolutional Neural Networks for High-Spatial Resolution Remote Sensing Image Classification

Jie Wang, Yalan Zheng, Min Wang, Qian Shen, and Jiru Huang

***Abstract*—Object-based image analysis (OBIA) is regarded as an effective technology for high-spatial resolution (HSR) image classification due to its clear and intuitive technical process. However, OBIA depends on the manual tuning of image classification features, which is a tricky job. Deep learning (DL) technology autolearns image features from massive images and obtains higher image classification accuracy than traditional techniques. In this article, a novel method called object-scale adaptive convolutional neural network (OSA-CNN), which combines OBIA with CNN, is proposed for HSR image classification. First, OSA-CNN collects the image patches along the main axes of the object primitives obtained through image segmentation; the size of the former is automatically determined through the axis widths of the latter. This step generates the input units required for the CNN classification. Second, the Squeeze-and-Excitation block is extracted from the SE network into the network structure of GoogleNet, which realizes the weighted fusion of the multiscale convolutional features, enhances useful features, and suppresses useless ones. In the classification stage, multiscale image segmentation and CNN classification are fused using an object-scale adaptive mechanism. Finally, object primitives are classified through majority voting on the image patches. The network structure modifications, multiscale classification fusion, and other improvements are verified by gradually incorporating these steps into the original GoogleNet. The experiments show that these improvements effectively enhance the image classification accuracy. This article presents an effective way of combining OBIA and DL techniques to utilize the advantages of both approaches and facilitate HSR image classification.**

***Index Terms*—Convolutional neural network (CNN), deep learning, high-spatial resolution remote sensing, image classification, image segmentation, multiscale, object primitive, object-based image analysis (OBIA).**

Manuscript received August 18, 2020; revised September 28, 2020; accepted November 26, 2020. Date of publication December 2, 2020; date of current version January 6, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0503902, in part by the National Natural Science Foundation of China under Grant 42071301 and Grant 41671341, in part by the Major Science and Technology Program for Water Pollution Control and Treatment under Grant 2017ZX07302003, and in part by the Project Supported by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF-2019-04-008. *(Corresponding author: Min Wang.)*

Jie Wang, Yalan Zheng, Qian Shen, and Jiru Huang are with the School of Geography, Nanjing Normal University, Nanjing 210023, China, with the Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China, with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China, and also with the State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China (e-mail: wangjieyl@sina.com; zyl340311@126.com; shenqian_gis@163.com; huangjiru0907@163.com).

Min Wang is with the School of Geography, Nanjing Normal University, Nanjing 210023, China, with the Key Laboratory of Virtual Geographic Environment (Nanjing Normal University), Ministry of Education, Nanjing 210023, China, with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China, with the State Key Laboratory Cultivation Base of Geographical Environment Evolution (Jiangsu Province), Nanjing 210023, China, and also with the Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518034, China (e-mail: sysj0918@126.com).

Digital Object Identifier 10.1109/JSTARS.2020.3041859

I. Introduction

OBJECT-BASED image analysis (OBIA) is an effective technology for high-spatial resolution (HSR) image classification, which is conducted following the technical route of "segment and then classify." OBIA obtains homogeneous segments through image segmentation, extracts features, and then conducts information extraction, such as object recognition/classification in a multidimensional feature space. In such a scenario, the segments in OBIA are called object primitives because they replace pixels as fundamental units for the subsequent feature extraction and image analysis [1]–[4]. OBIA utilizes more abundant features than pixel-based image analysis, improves the interpretability of the information extraction results, and facilitates information fusion between remote sensing image processing and geographical information system. These advantages make OBIA well-recognized and utilized in remote sensing applications, especially when dealing with information extraction from HSR images [5].

Despite its clear and intuitive technical process, OBIA requires the manual tuning of image classification features, which is sometimes tricky. Meanwhile, OBIA has the so-called scale issue. The object scale in OBIA refers to the average size of the segments. Given the variety and complexity of remote sensing images, over- and under-segmentation errors coexist when images are segmented at a single scale, which is not conducive to the subsequent feature extraction and analysis. To address the scale issue, many studies used multiscale analyses (MSA), which involves multilevel segmentation from small scales to large ones. Superpixel segmentation is commonly used to obtain multiscale spectral–spatial features in HSR/hyperspectral image classification [6]–[8]. Such studies confirmed that MSA improves the classification accuracy.

Unlike OBIA, deep learning (DL) technology autolearns image features from massive images, avoids manual feature extraction, and obtains high accuracy in the image classification. DL-based image classifications that utilize convolutional neural

networks (CNNs) are research hotspots in the field of remote sensing image processing [9]–[12]. The typical applications include scene classification [13]–[18] and land use/land cover classification [19]–[24]. CNN uses patch-based or pixel-to-pixel schemes for image classification. The patch-based scheme contains a full connection layer and uses sliding window [25] or superpixel segmentation [11] to obtain image patches as net inputs. The pixel-to-pixel scheme utilizes a full convolutional network (FCN) with an encoder–decoder structure to realize end-to-end learning [26]–[28]. This scheme achieves pixel-level image classification (i.e., semantic segmentation) and shows a strong capability in precise object delineation. However, pixel-to-pixel scheme often requires tedious image annotation (i.e., accurate manual delineation of ground objects). Considering the specific DL demands for sample quantity and sample balance, the pixel-to-pixel scheme is more suitable for binary classification problems (i.e., extracting targets from background) than for multiclassification ones.

Combining OBIA and DL, which involves automatically segmenting images into object primitives and conducting CNN sample collection and classification based on the object primitives, offers a convenient alternative for the HSR image classification. CNN realizes the autolearning of classification features and therefore avoids the manual tuning of OBIA features. Therefore, object-based CNN classification is a promising OBIA scheme for the HSR image classification. However, CNN requires fixed image patches as net inputs, which does not match the OBIA's inputs of object primitives. Hence, a conversion between the two input interfaces is necessary to fulfill object-based CNN classification.

Superpixel segmentation approaches, such as simple linear iterative clustering [29], are commonly used in object-based CNN classification [30], [31]. This kind of segmentation generates averaged object primitives, which facilitates the conversion of object primitives into image patches using the former's boundary boxes. However, over- and under-segmentation errors coexist in the superpixel segmentation because the object primitive sizes are averaged. Some studies obtained object primitives using other segmentation methods, including multiresolution segmentation [2], generated randomly positioned image patches within the object primitives [32], mapped one object primitive into many image patches, and conducted CNN classification. In such a scenario, the image patches were shrunk and sometimes positioned close to the object primitive boundaries. Such low-quality sampling lowered the classification quality.

In the previous studies [33]–[37], the hard-boundary-constrained segmentation (HBC-SEG) method was proposed and applied to OBIA. HBC-SEG is a hybrid segmentation method wherein the image edges are embedded into a merging region to obtain high segmentation accuracy. HBC-SEG was used in this article as the segmentation module for the object-based CNN classification. A novel object primitive/image patch conversion scheme was designed to adapt to the more natural segmentation configuration of HBC-SEG than superpixel segmentation. MSA with a novel object-scale adaptive mechanism was applied to flexibly fuse the multiscale classification results. CNN network structure modification and majority voting mechanism were also adopted to improve the method performance. The experiments showed that the proposed method called object-scale adaptive convolutional neural network (OSA-CNN) effectively combined OBIA and DL technologies and achieved good HSR image classification results.

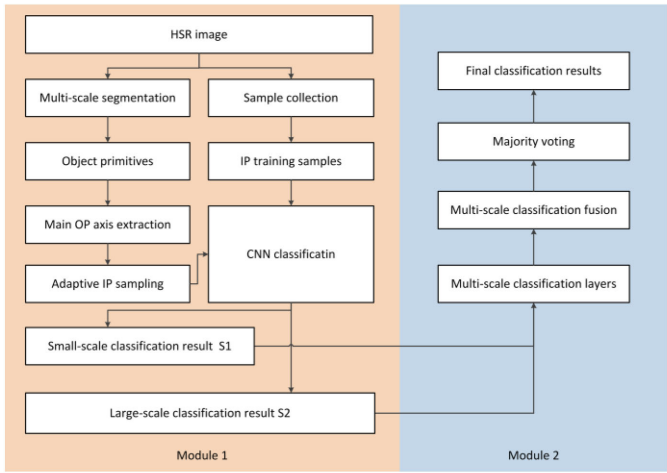


Fig. 1. OSA-CNN technical process.

The remainder of this article is organized as follows. Section II introduces the methodology of OSA-CNN. Section III presents the experimental analyses and discusses the method characteristics. Section IV provides the conclusions and future work directions.

II. Methodology

OSA-CNN is divided into two main modules (see Fig. 1). The first one is the image segmentation and classification module. The object primitives at a set of scales are obtained through HBC-SEG. Then, the object primitives are mapped into image patches using adaptive patch sampling along the object primitive axes, and the results are classified using CNN. These processes are conducted at all scales to obtain a set of classification results. The second one is the multiscale classification fusion module. Each image patch's maximum classification probability scale (MCPS) is determined according to the classification probability comparisons at different scales, and the multiscale classifications are fused into one result. Finally, each object primitive is classified through majority voting on the image patches' classification results to obtain the entire image classification result. The object primitive/image patch conversion, CNN network modification, and multiscale classification fusion are the key links for fulfilling the proposed method.

A. Object Primitive/Image Patch Conversion

The object primitive/image patch conversion involves two steps: main object primitive axis extraction and adaptive image patch sampling along the axes. The method proposed by [38] is adopted and modified for the main axis extraction. As illustrated in Fig. 2, the main skeletons of the object primitives are kept and the burrs (short axes) are removed during the extraction. Axis

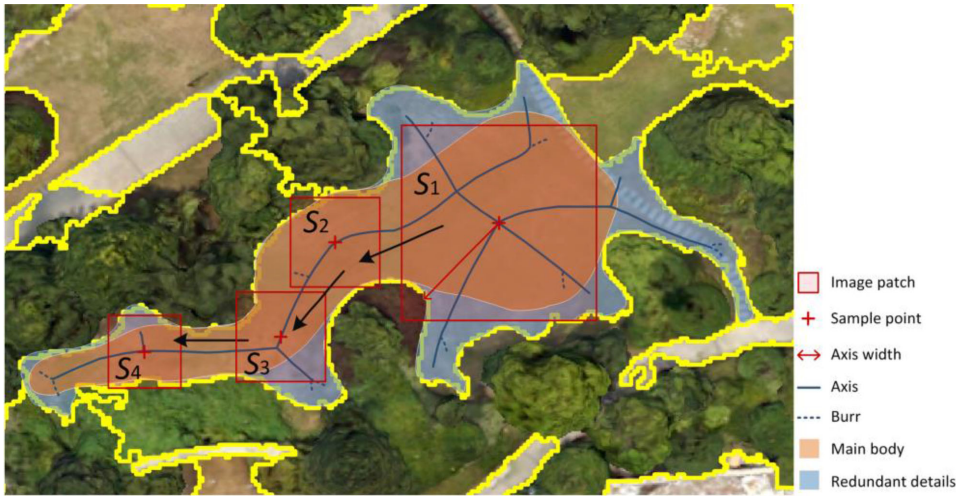


Fig. 2. Extracted main axes of an object primitive and adaptive image patch sampling. After the main axis extraction, the main body (brown) is kept for sampling and the redundant parts (blue) are discarded.

width is defined as the shortest distance from an axis point to the object primitive's boundary. After the axis extraction, the image patches are sampled along the object primitive's axes, and the change in their position and widths is in accordance with that of the axis. A suitable sampling interval is required to reduce the image patch overlap and increase the sample representativeness, reduce redundancy, and improve the efficiency. On this basis, the following adaptive sampling scheme is designed.

Step 1: Position the first image patch sample at the widest axis point and set the square width twice of the axis width.

Step 2: Excluding the axes covered by the sample generated in Step 1, search for the second widest axis point from the remaining axes and generate the candidate sample. Discard the candidate if its overlap rate with the previous samples exceeds 30%; otherwise, consider the candidate as a new image patch sample.

Step 3: Repeat Step 2 until no axis point is available or the number of samples reaches a certain threshold (e.g., maximum of five samples).

The number of image patches is dependent on the size and shape of the object primitives. One to two image patches are often enough to cover simple object primitives in block shapes. Complex object primitives with forks generally produce complex axes and require more image patches than simple object primitives. In such scenarios, the image patch number is limited to five in maximum considering the method efficiency.

Fig. 2 shows a set of image patches generated using the above sampling scheme. These image patches cover majority of the parts of the object primitive, accurately represent the object primitive's configuration, and facilitate the succeeding CNN training and classification.

B. Network Design

Previous studies verified that the CNN performance can be improved not only by deepening the network by stacking layers, but also by adopting a delicate net design, such as depthwise separable convolution [39], dense connectivity [40], and channel-wise feature recalibration [41]. The object primitives obtained through segmentation cover a small piece of the image, and its features may disappear after several rounds of convolutions in a deep CNN; deep CNNs do not necessarily yield accurate classification in OBIA. An extremely shallow network limits its capabilities in learning image features. Considering the above factors, GoogleNet [42] is selected as the skeleton network in this article. The squeeze-and-excitation (SE) block, which executes the weighted fusion of the CNN channel features [41], is embedded into GoogleNet to enhance the classification performance.

1) SE Block Extraction: The SE network [41] learns the importance of the features in image classification. This network suppresses the useless features and utilizes the useful ones. SENet, which is originally based on ResNet, has a deep network structure and a concise key component (i.e., SE block). To control the network size, the SE block is extracted and embedded into the inception module of GoogleNet to realize the weighted fusion of the channel features.

The SE block is composed of three operations (see Fig. 3). The first part is the squeeze operation. Given an input feature map u with dimensions of $H \times W \times C$, a statistic z for the channels is generated as

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j). \tag{1}$$

In (1), z represents the global spatial information distribution obtained through global pooling and u_c denotes the cth feature map. The second is the excitation operation, in which the weight s of the channels is obtained as

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{2}$$

where δ is the ReLU function [43], W_1 and W_2 are two fully connected layers, and σ is the sigmoid activation. The last is the

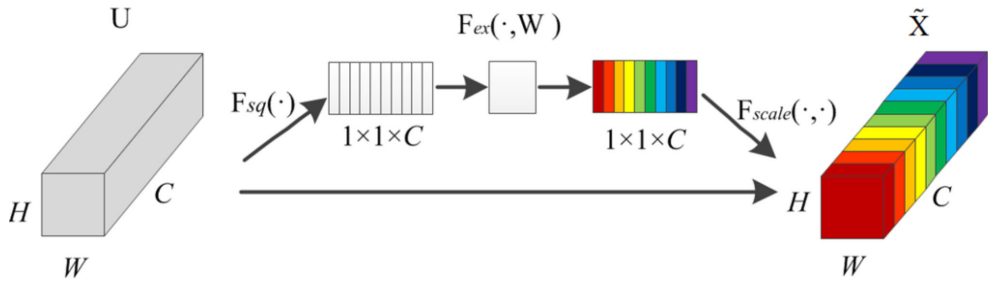


Fig. 3. SE block.

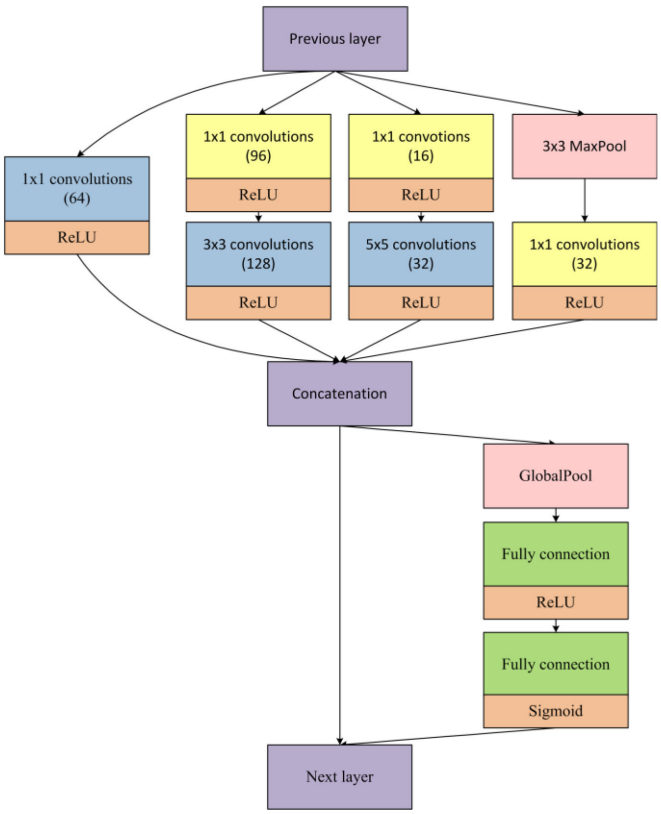


Fig. 4. Inception-SE module.

scale operation, in which the features are reweighted as

$$\tilde{x}_c = F_{\text{scale}}(u_c, s_c) = s_c \cdot u_c. \tag{3}$$

In (3), $\tilde{x}$ represents the output feature map after the weighted fusion and F_{scale} is the channelwise multiplication.

SE block can be flexibly embedded into other networks to improve the performance because its inputs can be any arbitrary convolution features with any number of channels. Global pooling introduces a small amount of network parameters, which facilitates network training. In this article, the SE block is applied for the CNN channel weighting.

2) Network Structure: OSA-CNN embeds the SE block into the inception module to form the inception-SE module (see Fig. 5). All inception modules of GoogleNet are replaced with inception-SE modules. For example, the details of the first inception-SE structure are shown in Table I and Fig. 4.

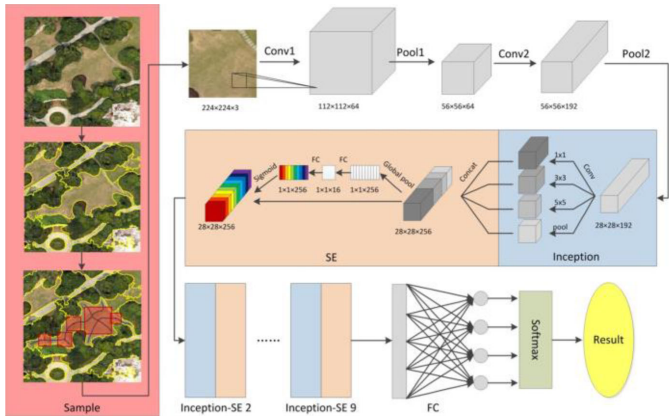


Fig. 5. OSA-CNN network structure.

TABLE I
INCEPTION-SE ARCHITECTURE

Input: A 28×28×192 feature map Output: A 28×28×256 feature map
1. Divide the net into four branches and process each branch with different convolutions, as follows: (1) A 1×1×64 convolution and a ReLU, which outputs a 28×28×64 feature map. (2) A 1×1×96 convolution and a ReLU, followed by a 3×3×128 convolution and a ReLU, which output a 28×28×128 feature map. (3) A 1×1×16 convolution layers and a ReLU, followed by a 5×5×32 convolution and a ReLU, which output a 28×28×32 feature map. (4) A 3×3 max pooling with stride 1, followed by a 1×1×32 convolution and a ReLU, which outputs a 28×28×32 feature map.
2. Concatenate the above four branches, which outputs a 28×28×256 feature map.
3. Apply the SE block, with the following succeeding steps: (1) A global pooling, which outputs a 1×1×256 feature map. (2) A fully connected layer and a ReLU, which outputs a 1×1×16 feature. (3) A fully connected layer and a Sigmod, which outputs a 1×1×256 feature.
4. Multiply the output of the step 3(3) with the output of step 2, which obtains the final 28×28×256 feature map.

The entire network structure of OSA-CNN is displayed in Fig. 5. When object primitives are converted into image patches, they are resized uniformly to 224 × 224 × 3 as the network inputs. Subsequently, deep features are extracted using two convolution and two pooling layers. Each convolution is appended

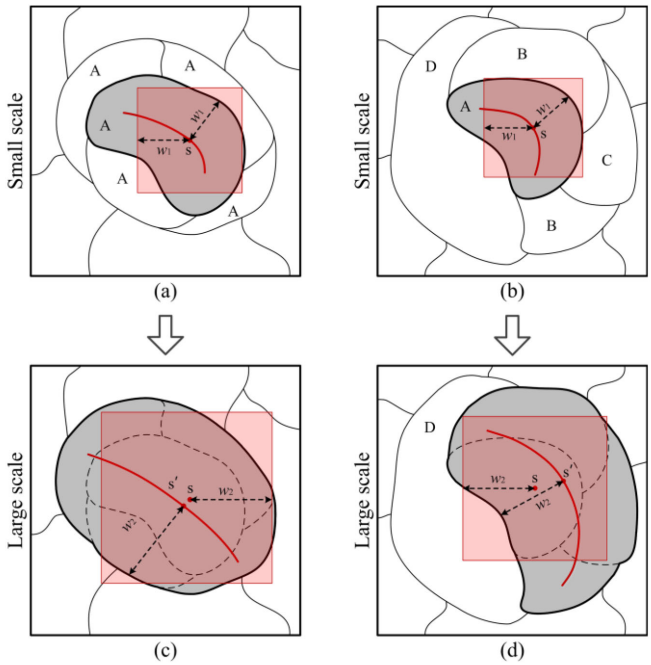


Fig. 6. Multiscale analysis for classification fusion.

by a ReLU activation. Normalization is performed after the first pooling layer to ensure the uniform distribution of the input data and generate a $28 \times 28 \times 192$ feature map. This feature map is then processed using nine inception-SE modules, followed by a fully connected layer. A softmax classifier obtains the final classification and finishes the entire process.

C. Multiscale Classification Fusion

1) Problem Description: A small segmentation scale in OBIA can avoid undersegmentation errors and obtain pure object primitives, which facilitates the succeeding classification. However, DL requires sufficiently large image patches to capture adequate features for classification. Meanwhile, object primitives may be extremely fragmented in strong textured regions (i.e., seriously oversegmented). The segmentation scale in such areas must be increased to allow the CNN to capture the texture patterns. An intuitive scheme increases the segmentation scale, which results in large object primitives and subsequently yields large image patches. However, increasing the segmentation scale leads to undersegmentation errors, which produces impure object primitives that cannot be classified correctly. An alternative approach is to uniformly increase the image patch size. However, such scheme lacks an objective criterion in determining the appropriate image patch size. In highly heterogeneous areas, directly increasing the image patch sizes might result in hybrid image patches that cannot be classified accurately. Single-scale analysis cannot appropriately handle the scale issue; thus, a novel MSA scheme is designed and introduced to object-based CNN classification.

2) Method Design: Fig. 6 illustrates the principles of MSA for classification fusion. An oversegmented area in small scale, which is possibly caused by the strong texture, is shown in Fig. 6(a). The shadowed object primitive has the same category "A" as its neighbors. Fig. 6(c) displays a large-scale segmentation configuration, in which the object primitives in the same category are merged into one and the corresponding image patch width changes from w_1 (small scale) to w_2 (large scale). The image patch at the large scale, which remains homogeneous and covers abundant image features, is better than the image patch at the small scale. The classification probability at the large scale may increase compared with that at the small scale. In this case, the MSA selects the classification result at the large scale.

Fig. 6(b) illustrates a correctly segmented area at the small scale, where the shadowed object primitive's categories are different from those of its neighbors. In this scenario, increasing the segmentation scale results in the overmerging of the object primitive and induces undersegmentation errors. The image patch at the large scale becomes heterogeneous and less satisfactory than that at the small scale in terms of classification. Consequently, the classification probability in large scale may decrease compared with that in small scale. In such case, the MSA selects the classification result at the small scale.

The design of the probability-based multiscale classification fusion is explained as follows. Let $\{Q(s)s\epsilon N\}$ be a child-to-parent object primitive hierarchy obtained through multiscale segmentation, where s represents the segmentation scale and $Q(0)$ is the child object primitive at the smallest scale. A parent object primitive is merged from small scales

$$Q(s) \subseteq Q(s+1). \tag{4}$$

$P(Q, v)$ is an image patch of Q at position v, and the size of P is determined by the axis width at v. Let $\{P(Q(s), v)s\epsilon N\}$ be the image patch hierarchy of $\{Q(s)s\epsilon N\}$ at position v

$$P(Q(s), v) \subseteq P(Q(s+1), v). \tag{5}$$

Given an object primitive at a certain scale with an image patch at position v, let $cls(P(Q(\bullet), v))$ be the image patch's category and $\mathrm{prob}(cls(P(\bullet), v))$ as the corresponding classification probability. The classification probability of the image patch is calculated by the CNN's softmax layer, which is formulated as follows:

$$p(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \tag{6}$$

where x_i represents the output of the fully connected layer for class 'i', K is the total number of classification categories, and p represents the corresponding classification probability. The CNN classifier obtains the classification probability of each category and takes the maximum classification probability and corresponding category as the final outputs.

The MCPS of $P(Q(\bullet), v)$ is expressed as

$$s^* = \arg_s \max \mathrm{prob}\left(\mathrm{cls}\left(P\left(Q(s), v\right)\right)\right). \tag{7}$$

The multiscale classification fusion rules are defined to classify $Q(0)$ (i.e., object primitive at the smallest scale)

Rule 1: If $cls\left(P\left(Q(0), v\right)\right)$

$$= cls\left(P(Q(s^*), v)\right),\ output\ cls\left(P\left(Q(0), v\right)\right). \tag{8}$$

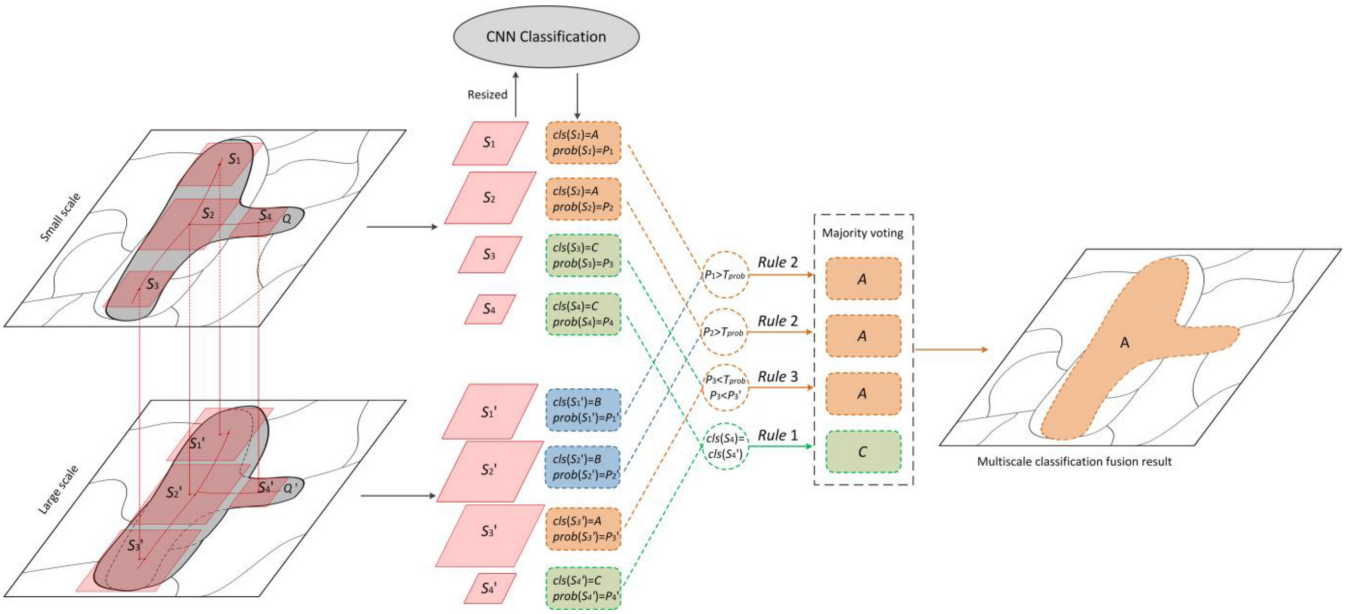


Fig. 7. Multiscale classification processing.

Rule 2: else, if $prob\left(cls\left(P\left(Q(0),v\right)\right)\right.$

$$\geq T_{prob},\ output\ cls\left(Q\left(0\right),v\right). \tag{9}$$

Rule 3: else, $output\ cls\left(P(Q\left(s^{*}\right),v)\right).$ (10)

In the above rules, T_{prob} denotes the classification probability threshold.

Rule 1 suggests that the classification results are consistent between the smallest scale 0 and the MCPS $s*$. This rule also indicates that as the scale increases, either the classification results in this region remain stable or the classification certainty decreases. The classification result of the smallest scale is selected as the MSA output.

Rule 2 addresses the case when the classifications from the small scales to the large ones are inconsistent and the small-scale probability is larger than the threshold. Increasing the scale may increase the heterogeneity of the object primitives; thus, the small-scale classification result is selected regardless whether the classification probabilities at large scales are higher than that at small ones. This rule jointly considers the classification probability and small scale priority.

Rule 3 holds when the classifications at small and large scales are inconsistent and the small-scale classification probability is low. As demonstrated in the strong textured regions, object primitives at small scales may be fragmented and the classification certainty maybe low. Therefore, the classification is reinvestigated at large scales and the rule selects the classification result at the MCPS.

Fig. 7 illustrates a typical decision process of the multiscale classification fusion and majority voting for an object primitive. At the small scale, four image patches (S_1-S_4) with small sizes are collected along the object axes. At the large scale, the object is merged with surrounding objects. Thus, four new image patches ($S_1{}'-S_4{}'$) are collected around the centers of S_1-S_4 in which

Fig. 8. Experimental area 1.

$S_1{}'-S_3{}'$ are enlarged. All the eight image patches are resized to $224 \times 224 \times 3$ and fed to the CNN to predict their categories.

At the multiscale classification fusion stage, the categories of each image patch pair at the same location from the small and large scale are compared. Assume $cls(S_1)$ and $cls(S_1{}')$ are inconsistent and the $prob(S_1)$ is larger than the classification probability threshold T_{prob}. In such case, rule 2 is activated and the small scale classification result is selected as the multiscale fusion result for image patch S_1. The same rule is also applied to image patch S_2. The classification results of S_3 and $S_3{}'$ are

Fig. 9. Experimental area 2.

Fig. 10. Experimental area 3.

inconsistent, and the classification probability of S_3 is less than T_{prob}. Rule 3 is thus activated and the large scale classification result is chosen as the multiscale fusion result of S_3. S_4 and $S_4{}'$ are in the same class, so rule 1 is activated. In the end, three "*A*" classes and one "*C*" class are obtained by this MSA module. The object primitive is finally classified as "*A*" by the succeeding majority voting module.

On the basis of the object primitive/image patch conversion and MSA rules, the image patch sizes are automatically adjusted and selected in accordance to the object primitive axis widths and multiscale classification probability comparison. Such schemes organically integrate OBIA and CNN, utilize the advantage of MSA, and improve the OBIA classification. To ensure the efficiency, OSA-CNN selects two scales for probability comparison and information fusion.

TABLE II
INFORMATION OF THE THREE EXPERIMENTAL DATA

Image	Imaging date	Size (pixel)	Location	Spatial Resolution
Area 1	2007.6.14	4000×4000	81°39'0"W 41°17'37"N	0.3 m
Area 2	2007.6.14	3000×3000	81°40'45"W 41°17'5"N	0.3 m
Area 3	2018.6.18	2000×2000	119°21'28"E 32°23'18"N	0.1 m

III. EXPERIMENTS

The OSA-CNN method was developed and integrated into the RSFinder software developed by the research team using Visual C++. Caffe, which was developed in C++ [44], was selected as the DL platform to seamlessly integrate the proposed CNN classification module into the RSFinder. The proposed method was tested on a Windows 7 operation system using different HSR aerial images. The central processing unit was Intel Core i7-6850k, the graphics processing unit was NVIDIA GeForce GTX1080Ti, the video RAM size was 11 GB, and the RAM was 32 GB.

A. Experimental Scheme

1) Experimental Data: The experimental data consisted of aerial images collected by the Ohio Statewide Imagery Program and the unmanned aerial vehicle imagery located in Yangzhou City, Jiangsu Province, China. The specific imaging parameters are presented in Table II. The land cover types were divided into six categories according to the ground feature distributions in the experimental areas: water, grassland, bare land, forest, building, and road.

2) Evaluation Scheme: A set of gradually enhanced methods starting from GoogleNet (GN-CNN) was adopted to validate the improvements induced by OSA-CNN. These methods evolved to GNSE-CNN, GNSEM-CNN, and OSA-CNN, respectively. GN-CNN is the object-based CNN classification that uses the original GoogleNet. Each object primitive was classified using a single image patch located at the object primitive's center, and the classification was conducted at a single scale without MSA. GN-CNN is the most basic combination of OBIA and CNN classification. GNSE-CNN embedded the SE block into GN-CNN to verify the effectiveness of the structure. GNSEM-CNN further adopted MSA, in which each object primitive was classified using a single image patch located at the object primitive's center without majority voting. OSA-CNN added image patch sampling along the object primitive's axes and majority voting

TABLE III
EXPERIMENTAL SCHEME

Method	GN-CNN	GNSE-CNN	GNSEM-CNN	OSA-CNN
IP number and location	Single image patch at object primitive's center	Single image patch at object primitive's center	Single image patch at object primitive's center	Multi image patches along the object primitive's axes
Segmentation scale	80	80	80, 150	80, 150
Classification	Single-scale	Single-scale	Multiscale	Multiscale

to GNSEM-CNN, which is the final form of the improvements. The details of the parameters are as listed in Table III.

A set of training samples were collected and uniformly used to train the compared methods. Approximately 120–140 samples were collected for each category, and 850–900 samples were collected from each experimental area. For relatively small training datasets, transfer learning was performed on the basis of the ImageNet pretrained model. One thousand validation points were randomly generated in each experimental area and compared with the classification results to evaluate the accuracy of the proposed method.

The CNN training parameters are elaborated as follows. Stochastic gradient descent was used for the CNN model training. The initial learning rate for GN-CNN was 0.001, whereas those for configurations with SE blocks were set as 0.01, which provided fast convergence at the early training stage. The batch size was 32 and the number of iterations was 3000.

3) Accuracy Evaluation Measures: A confusion matrix was established to evaluate the accuracy of the compared methods. The accuracy measures used in this article include user accuracy (UA), producer accuracy (PA), overall accuracy (OA), and kappa

$$UA(a) = \frac{X_{aa}}{\sum_{i=1}^{n} X_{ai}} \tag{11}$$

$$PA(a) = \frac{X_{aa}}{\sum_{i=1}^{n} X_{ia}} \tag{12}$$

$$OA = \frac{\sum_{i=1}^{n} X_{ii}}{N} \tag{13}$$

$$\text{Kappa} = \frac{N\sum_{a=1}^{n} X_{aa} - \sum_{a=1}^{n}(\sum_{i=1}^{n} X_{ai} \times \sum_{i=1}^{n} X_{ia})}{N^2 - \sum_{a=1}^{n}(\sum_{i=1}^{n} X_{ai} \times \sum_{i=1}^{n} X_{ia})} \tag{14}$$

where n represents the number of classes, N denotes the total number of test samples, and X_{ai} represents the number of samples whose actual and predicted classes are a and i, respectively.

B. Experimental Analysis

The influences of the parameters of the two most important inputs, namely, image segmentation scale and classification probability threshold (T_{prob}), on the method performance were investigated. Comparative analyses of the four classification schemes were conducted to validate the features and advantages of OSA-CNN.

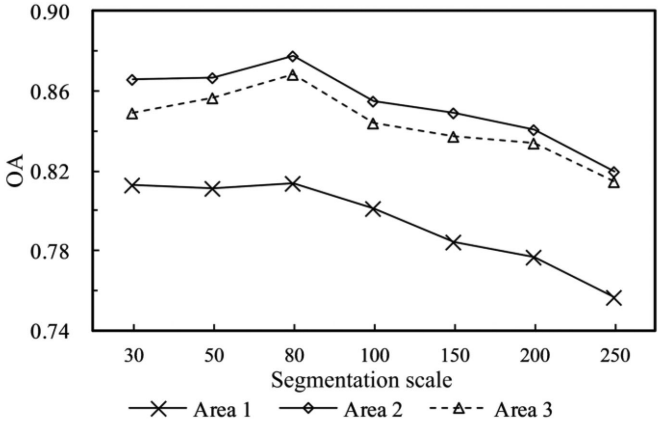


Fig. 11. OA curves with the changed scales.

1) Parameter Influence:

a) Segmentation scale: The influence of segmentation scales on the overall accuracy was investigated by varying the segmentation scale from small to large. Fig. 11 show the OA of OSA-CNN at seven segmentation scales in the three experimental areas. Within a scale range of 30–80, the OA curve in the first experimental area was flat, whereas those in the second and third experimental areas rose. At a segmentation scale range of 80–250, the OA of the three experimental areas showed a decreasing trend. Increasing the segmentation scale beyond 80 included the neighbor objects in the object primitives, which reduced the classification quality. The OAs of three experimental areas reached the peak at a scale value of 80, which implies that this scale is appropriate for segmenting the selected HSR images. The influence of the scale on the performances of the compared methods is the same. Therefore, a scale value of 80 was used as the optimal threshold for the subsequent analyses. GNSEM-CNN and OSA-CNN required two scales for MSA. A scale value of 150 was selected as the large scale because the strong textured regions were fully merged without obvious undersegmentation errors in the other regions.

b) Classification probability threshold: The influence of the classification probability threshold on OA is presented in Fig. 12. The influence was minimal and had a maximum

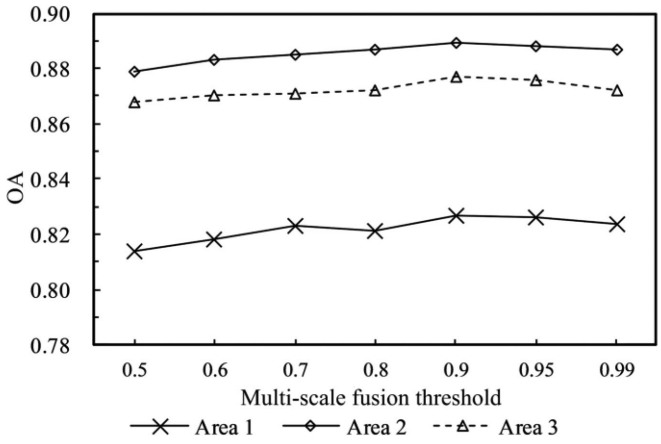


Fig. 12. Classification probability threshold.

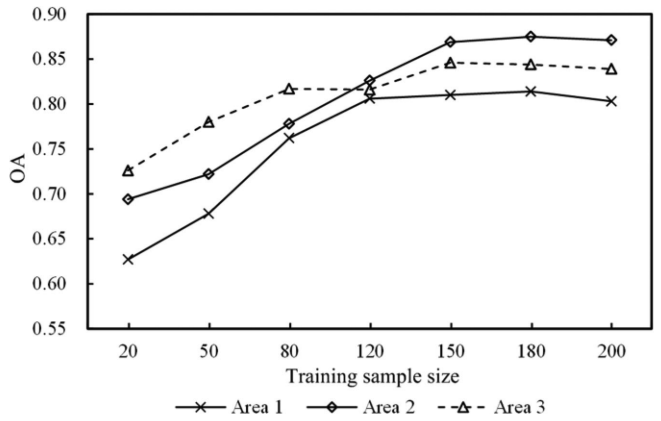


Fig. 13. OA curves with the changed sample sizes.

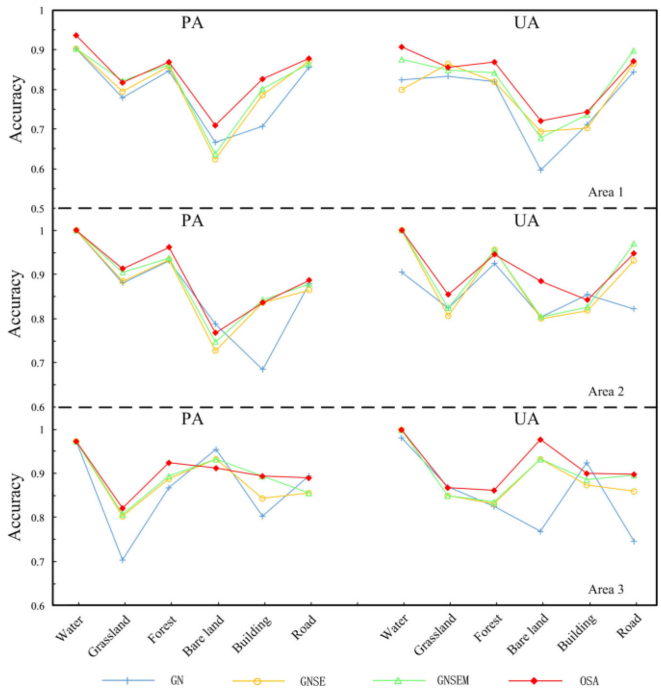


Fig. 14. PA and UA statistics with different categories.

TABLE IV
OA AND KAPPA STATISTICS OF THE DIFFERENT METHODS

Method	Area 1		Area 2		Area 3	
	OA	Kappa	OA	Kappa	OA	Kappa
GN-CNN	0.797	0.732	0.862	0.820	0.841	0.799
GNSE-CNN	0.814	0.754	0.878	0.841	0.868	0.832
GNSEM-CNN	0.827	0.771	0.889	0.855	0.877	0.844
OSA-CNN	**0.839**	**0.788**	**0.902**	**0.872**	**0.894**	**0.865**

difference of less than 2% when the threshold changed from 0.5 to 0.99. This threshold signifies the trust on small-scale segmentation. Low or high threshold makes OSA-CNN tend to small- or large-scale classification results, respectively. The OA curves in the three experimental areas reached their peaks (0.827, 0.889, and 0.877) at a threshold of approximately 0.9. Therefore, 0.9 was selected as the optimal parameter for the succeeding analyses.

c) Influence of the sample size: Inaccurate segmentation and inappropriate sampling position influence the quality of image patches generated along the object axes, and thus affect the model training. Thus, image patch samples for model training are manually collected at suitable image positions through visual interpretation of the experimental areas. The CNN models training and classification were then conducted for method accuracy evaluation. Sample sizes were balanced, i.e., evenly distributed, among the classes. The influence of sample size on measure OA is presented in Fig. 13. OA increased obviously before the sample size reached 150, and remained even after that. The peaks of the OA curves were 0.814, 0.875, and 0.846 with the sample sizes of 180, 180, and 150 in the three experimental areas, respectively. Thus, in our experiments, 150–180 training samples were collected for every class for the transfer learning of the CNN pretrained models.

2) Accuracy Evaluation: The four classification schemes were compared and analyzed with the optimal parameters. OSA-CNN obtained the highest UA and PA among the four. However, it demonstrated minimal lag in PA in bare land and in UA in building and road (see Fig. 14). The SE block, MSA, and majority voting gradually improved the OA (see Table IV). OSA-CNN achieved the highest OA (0.839, 0.902, and 0.894) and kappa (0.788, 0.872, and 0.865) in the three experimental areas. As compared with GN-CNN, the OA was increased by more than 4%, 4%, and 5% and kappa was by more than 5%, 5%, and 6% for the three experimental areas, respectively. These findings showed that the applied improvements effectively improved the method accuracy.

3) Visual Interpretation: The visual interpretation results are consistent with the quantitative analyses of the classification findings (see Figs. 15 –20). In experimental area 1, several small grassland object primitives were misclassified as building object primitives by GN-CNN at spot A (see Fig. 16). At spots B and C, many grassland and forest object primitives were misclassified as road and building object primitives. The GNSE-CNN reduced the misclassifications but the small object primitives still induced classification errors. GNSEM-CNN further reduced the misclassifications through MSA, which corrected most of

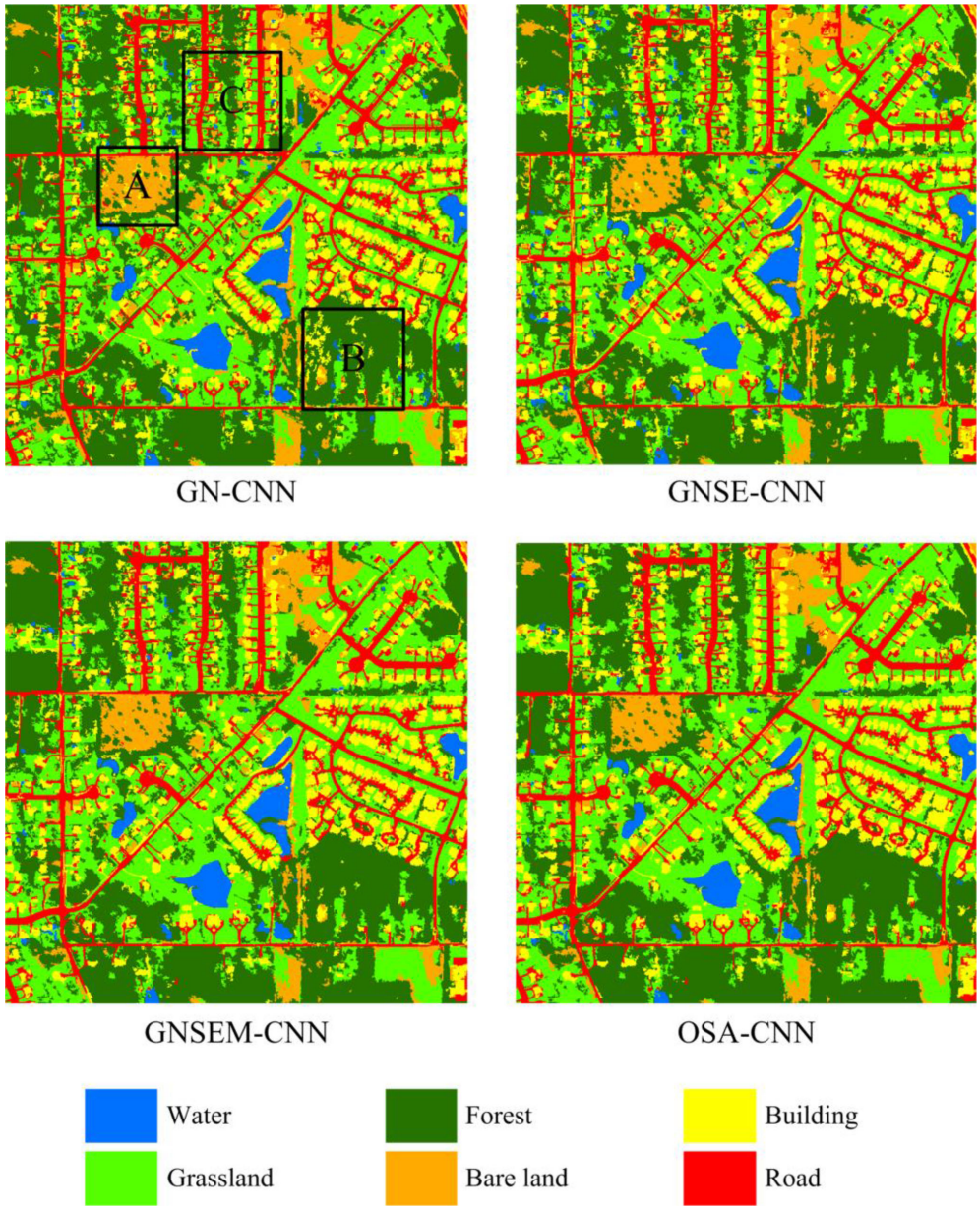


Fig. 15. Classification results in Area 1. The black boxes mark some remarkably changed spots (A, B, and C) and the method enhancements.

the misclassified small object primitives. Similarly, OSA-CNN corrected most of the misclassifications and obtained the best classification results among the four methods.

GN-CNN displayed obvious misclassifications between buildings and roads in experimental area 2, which might be because of their closed spectral features (see Fig. 17 and spot A in Fig. 18). Some forests were misclassified by GN-SNN as bare lands in spot B, whereas in spot C, obvious misclassifications for building areas were committed by GN-CNN. GNSE-CNN remarkably reduced the misclassifications of the roads, buildings, and bare lands as shown in the three spots. Similarly, the MSA of GNSEM-CNN further reduced these errors. OSA-CNN, which adopted the majority voting mechanism, exhibited the most satisfactory performance among the compared methods as it removed most of the misclassifications caused by small object primitives and spectral confusion.

GN-CNN misclassified roads, bare lands, and buildings in experimental area 3 (see Figs. 19 and 20). For instance, part of the forest was misclassified as water in spot C. As illustrated in the fourth row of Fig. 20, GNSE-CNN partially solved these misclassification errors and GNSEM-CNN reduced the errors through MSA. OSA-CNN solved most of the classification deficiencies, and thus demonstrated a superior result over its counterparts (fifth row of Fig. 20).

On the basis of the results of the comparative analyses, the object primitive/image patch conversion, CNN structure enhancement, MSA, and majority voting strategies can effectively improve the object-based CNN classification accuracy.

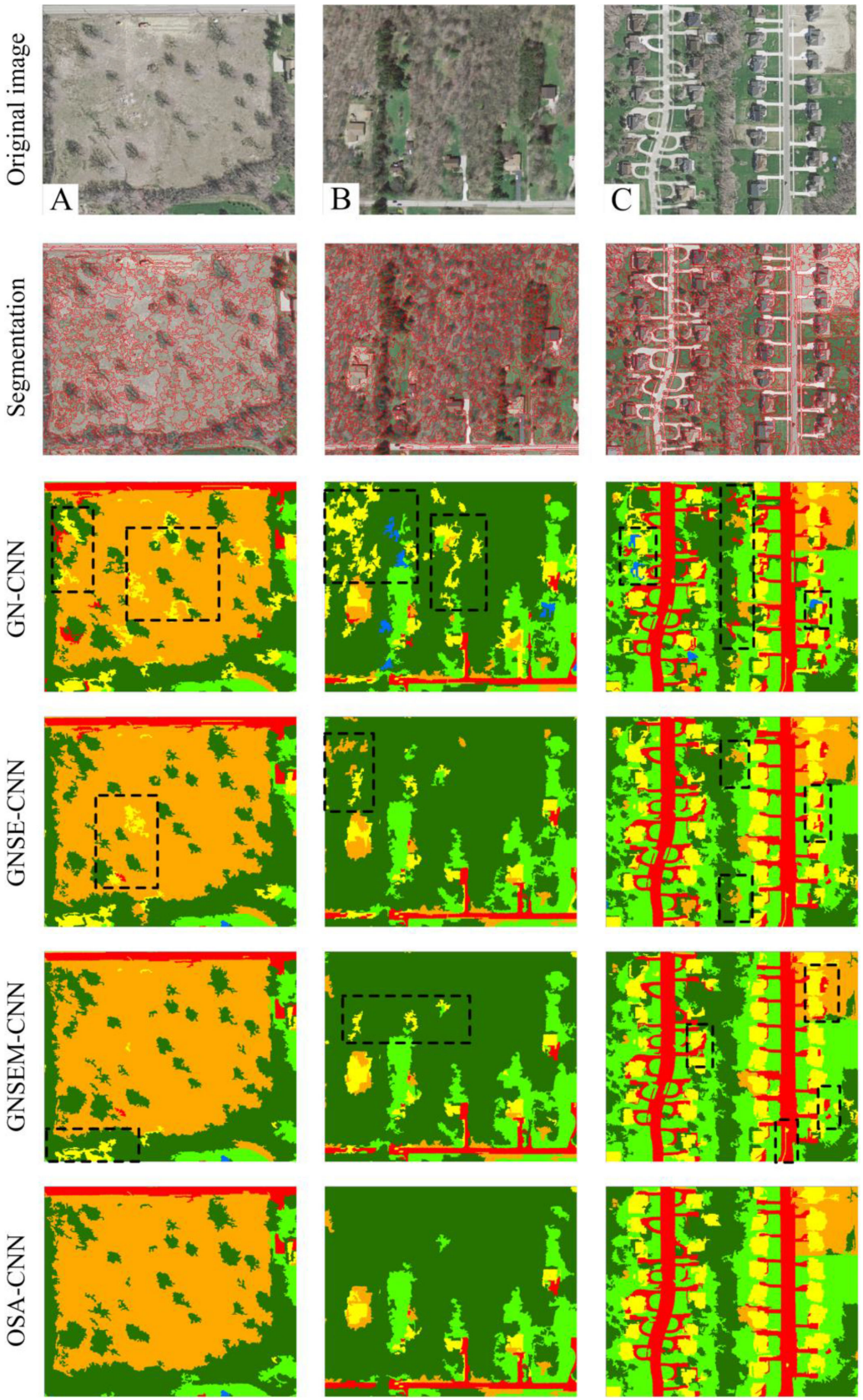


Fig. 16. Some spots in Area 1. The black dashed boxes mark some classification deficiencies, which are solved by the succeeding methods.

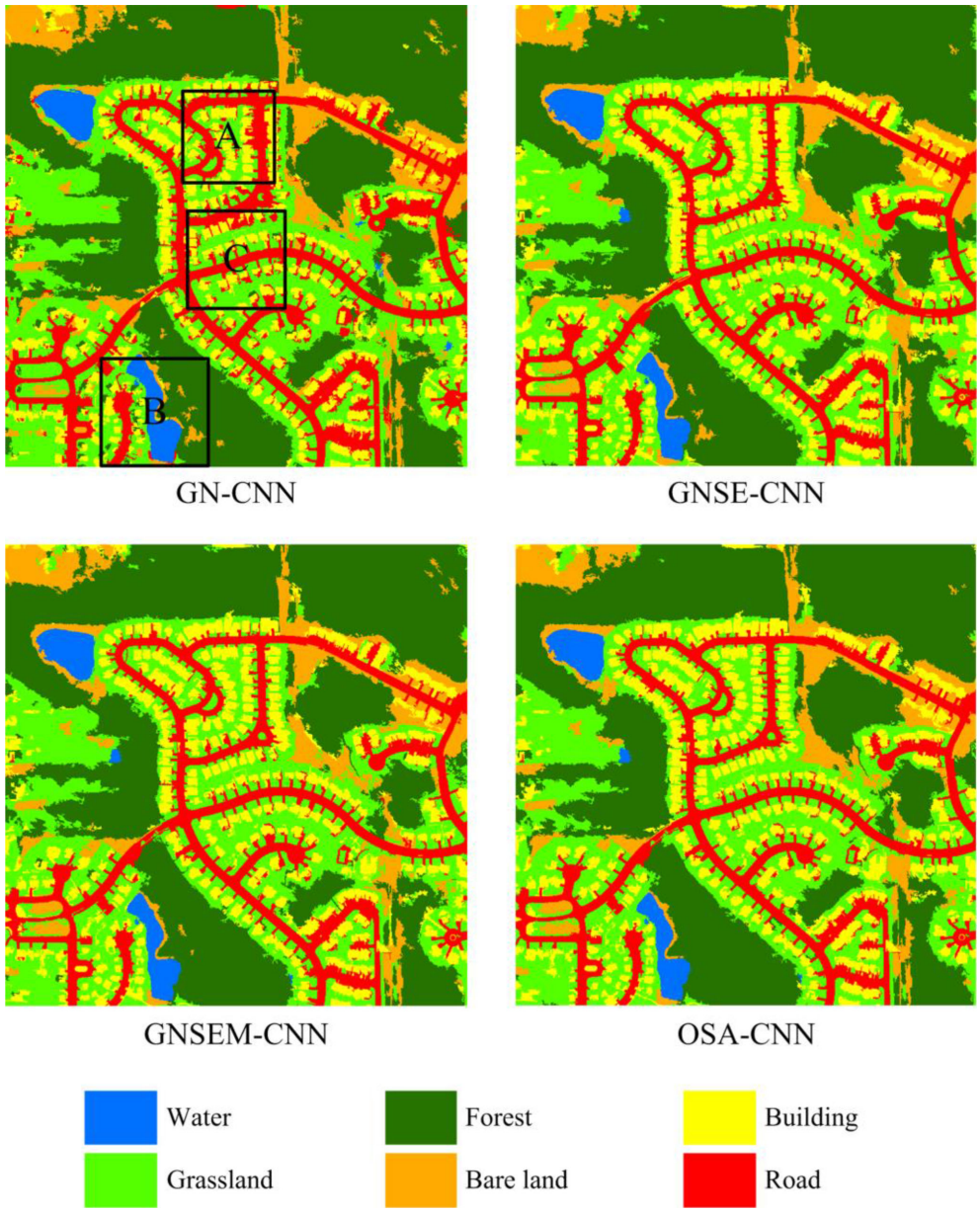


Fig. 17. Classification results in Area 2.

4) Discussion: Routine OBIA requires users to explicitly specify image features and limits the classification performance. DL autonomously learns image features from training data, avoids artificial feature selection, and facilitates classification processes. Therefore, introducing DL (e.g., CNN) as an OBIA classifier can eliminate the manual feature tuning. However, the conversion between object primitives and image patches must be carefully addressed because of its substantial influence on image classification accuracy. The results of this article show that the adaptive image patch sample collection along the object primitive axes produces appropriate inputs for the CNN classification.

Image classification can also be conducted using the FCN-based semantic segmentation. However, semantic segmentation requires manual object boundary delineation for the annotation collection for model training, which is tedious and time-consuming. Considering the workload of image annotation and the sample balance among the classes, semantic segmentation is suitable for two-class classification problems (i.e., separating targets from the surrounding background). Through image segmentation, OSA-CNN collects annotation by single clicking the object primitives, which improves the sample collection efficiency and facilitates CNN applications.

The segmentation in OSA-CNN significantly influences the succeeding classification quality. Segmentation errors will reflect in the classification if the object primitives are formed inaccurately, which is an inherent deficiency of OBIA. Multiscale classification fusion can alleviate several classification errors,

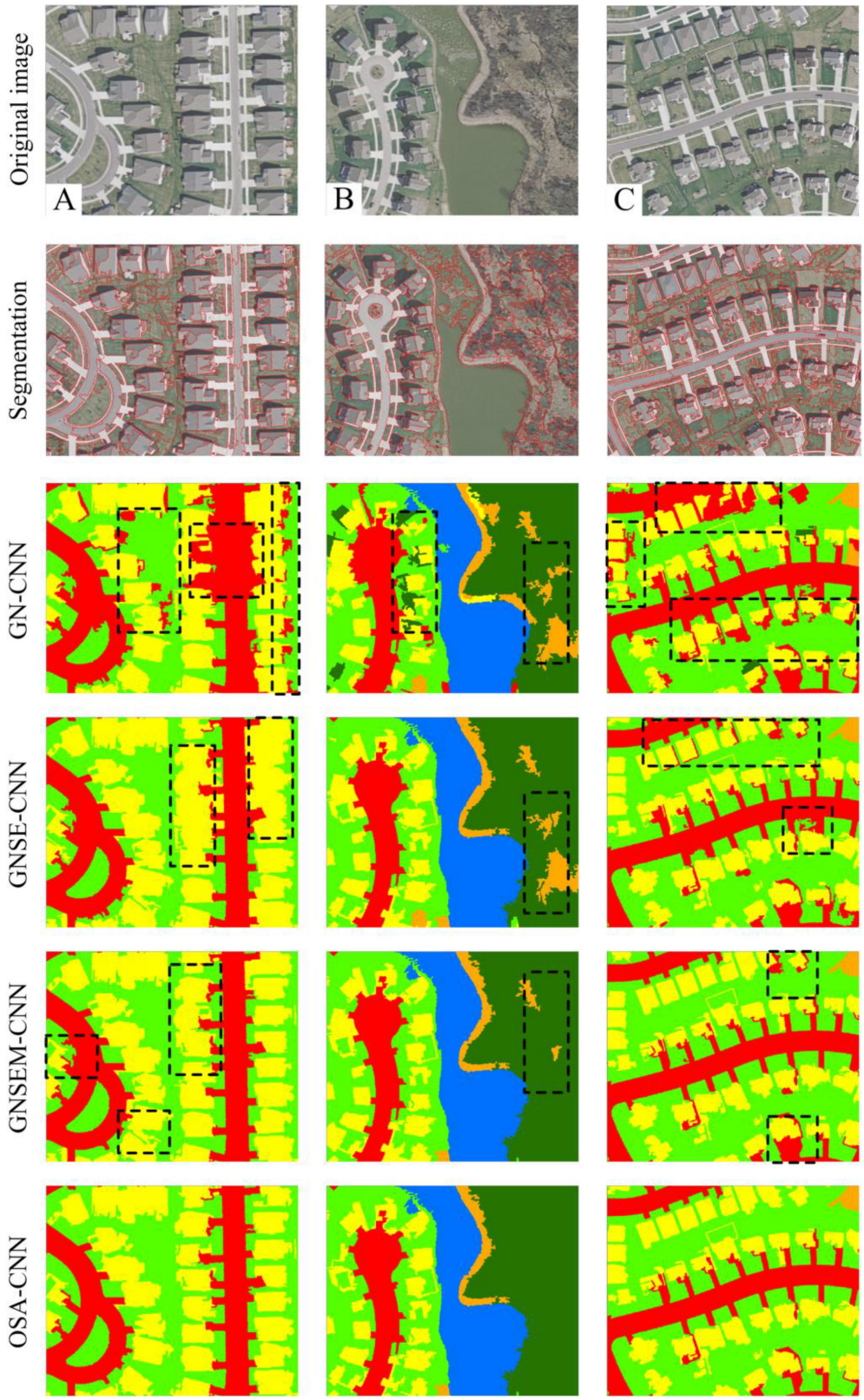


Fig. 18. Some spots in Area 2.

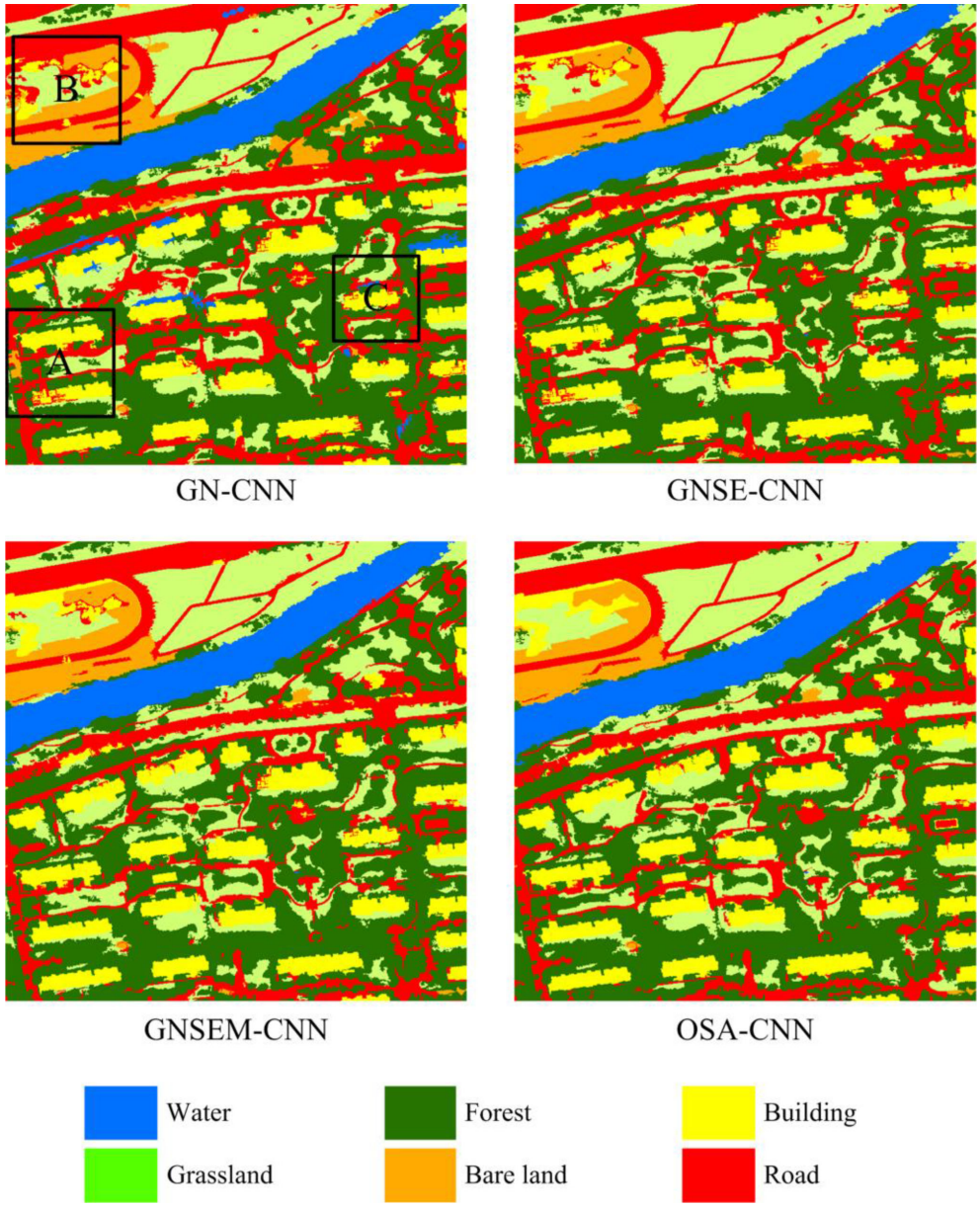


Fig. 19. Classification results in Area 3.

but it cannot resolve the segmentation ones. However, shape and category errors can be possibly fixed by an additional object primitive refinement process, which will be investigated in the future work.

The introduction of image segmentation, object primitive/image patch conversion, and MSA in OSA-CNN increases the additional run time compared with routine patch-based CNN classification. As shown in Table V, image segmentation and image patch generation occupy large parts of the run time. Meanwhile, the run time of the segmentation is affected by the image complexity, which is indicated by the relationship of run time and segment number in the three experimental areas.

From the viewpoint of method utilization, adding multiscale mechanism into the object-based CNN increases method complexity to some degree, as users need to conduct several (usually two) rounds of image segmentation. On the other hand, CNN training and classification in the proposed method are not different with other common CNN-based methods. In total, the proposed method can be easily applied for the HSR image classification. Meanwhile, the classification process does not involve global feature extraction. Thus, image partitioning and

TABLE V
TIME CONSUMING OF OSA-CNN

Image	Segment number	Time consuming (s)			
		Segmentation	Sample generation	Classification	Total
Area 1	16623	213	332	360	905
Area 2	6959	95	125	192	412
Area 3	2251	57	44	87	188

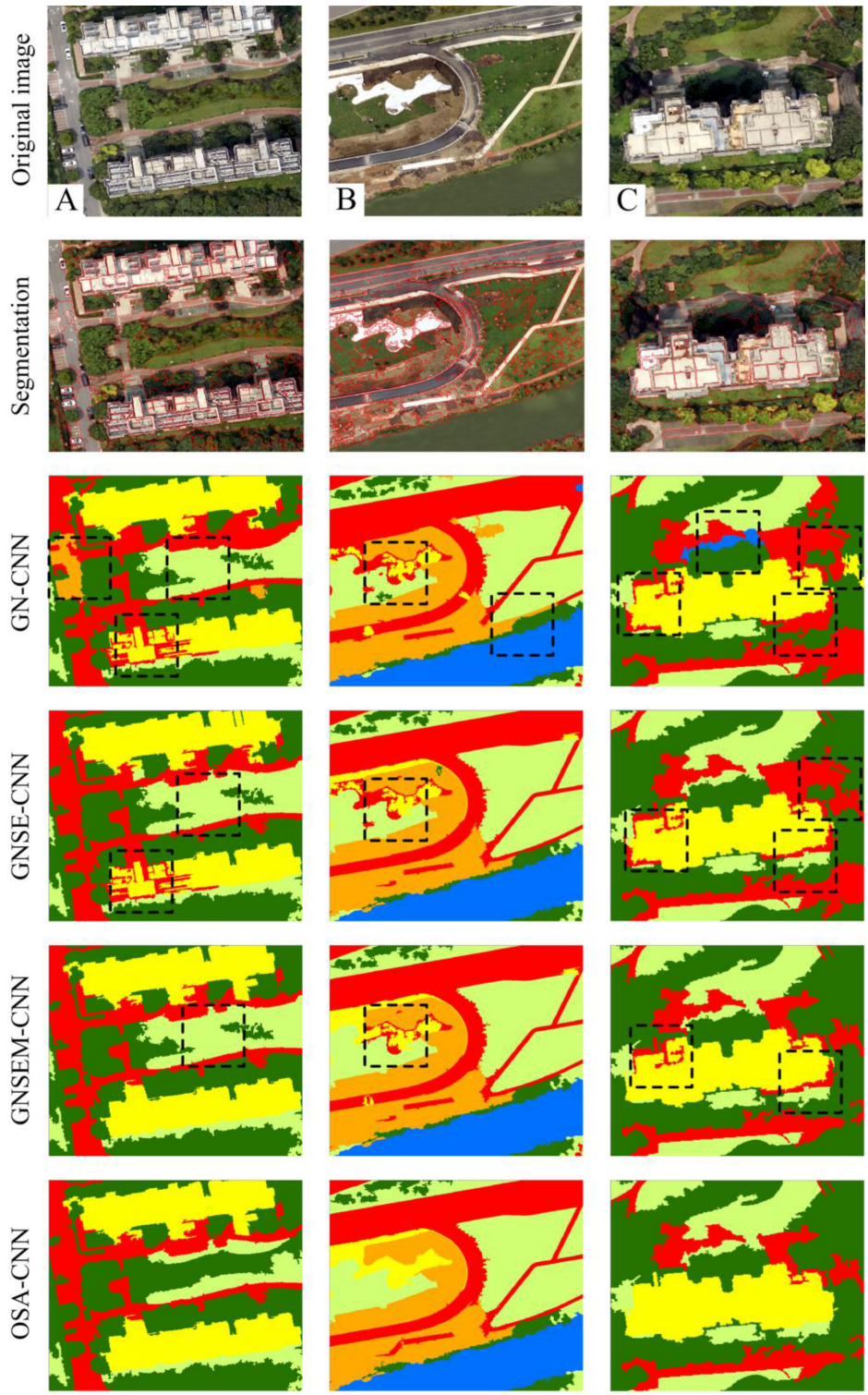


Fig. 20. Some spots in Area 3.

parallel processing can be easily applied to further improve the method efficiency on classifying large images.

IV. Conclusion

The combination of OBIA and DL techniques exhibits great potential in the HSR image classification. However, many problems still need to be addressed. OSA-CNN solves two issues of object-based CNN, namely, the interface transformation between object primitive (inputs for OBIA) and image patch (inputs for CNN) and the fusion of multiscale classification information. The features of the proposed method include an adaptive image patch sample generation based on object primitive axes, a feature-weighted network structure, and a multiscale

OBIA classification scheme based on the probability fusion. The experiments show that the proposed technical route is feasible and achieves high accuracy in the HSR image classification. In addition, OSA-CNN adopts the traditional OBIA technical framework of "segment and then classify." Once an image is segmented, the object primitives are fixed and difficult to modify during the feature extraction and classification, which limits the OBIA performance. This limitation can be addressed by introducing a flexible interaction mechanism between the segmentation and classification within the OBIA framework. Moreover, the algorithm efficiency should be improved by introducing some optimization methods, such as parallel computation.

REFERENCES

[1] M. Baatz and A. Schäpe, "Object-oriented and multi-scale image analysis in semantic networks," in *Proc. 2nd Int. Symp. Operationalization Remote Sens.*, 1999, pp. 16–20.

[2] M. Baatz and A. Schäpe, "Multiresolution segmentation: An optimization approach for high quality multi-scale image segmentation," in Angewandte Geographische Informationsverarbeitung XI. Salzburg, Austria: Herbert Wichmann Verlag, 2000, pp. 12–23.

[3] T. Blaschke and J. Strobl, "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS," *Zeitschrift für Geoinformationssysteme*, vol. 6, no. 1, pp. 12–17, 2001.

[4] U. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS J. Photogrammetry Remote Sens.*, vol. 58, pp. 239–258, 2004.

[5] T. Blaschke, "Object-based image analysis for remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.

[6] S. Li, L. Ni, X. Jia, L. Gao, B. Zhang, and M. Peng, "Multi-scale superpixel spectral-spatial classification of hyperspectral images," *Int. J. remote Sens.*, vol. 37, no. 19–20, pp. 4905–4922, 2016.

[7] A. Zhang, S. Zhang, G. Sun, F. Li, and Z. Wang, "Mapping of coastal cities using optimized spectral–spatial features based multi-scale superpixel classification," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 998.

[8] X. Cao, D. Wang, X. Wang, J. Zhao, and L. Jiao, "Hyperspectral imagery classification with cascaded support vector machines and multi-scale superpixel segmentation," *Int. J. Remote Sens.*, vol. 41, no. 12, pp. 4530–4550, 2020.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[10] J. E. Ball, D.T. Anderson, and C.S. Chan, "A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042609.

[11] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[12] X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, and F. Xu, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[13] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state-of-the-art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[14] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[15] Q. Zhou, Y. Zhou, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.

[16] E. Li, J. Xia, P. Du, C. Ling, and A. Samat, "Integrating multi-layer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[17] P. Du, E. Li, J. Xia, Alim. Samat, and X. Bai, "Feature and model level fusion of pretrained CNN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2600–2611, Oct. 2019.

[18] F. Hu, G. Xia, W. Yang, and L. Zhang, "Mining deep semantic representations for scene classification of high-resolution remote sensing imagery," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 522–536, Sep. 2019.

[19] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*. vol. 10, no. 7, pp. 3386–3396, Jul. 2017.

[20] X. Tong, G. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Learning transferable deep models for land-use classification with high-resolution remote sensing images," 2018, *arXiv:1807.05713*.

[21] Y. Xu, B. Du, F. Zhang, and L. Zhang, "Hyperspectral image classification via a random patches network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 142, pp. 344–357, 2018.

[22] B. Huang, B. Zhao, and Y. Song, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery," *Remote Sens. Environ.*, vol. 214, pp. 73–86, 2018.

[23] K. Zhou, D. Ming, X. Lv, J. Fang, and M. Wang, "CNN-based land cover classification combining stratified segmentation and fusion of point cloud and very high-spatial resolution remote sensing image data," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 2065.

[24] S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, Jun. 2019.

[25] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[26] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[27] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks," *Int. Archives Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, no. W1, pp. 481–486, 2017.

[28] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in VHR Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[30] C. Gonzalo-Martin, A. Garcia-Pedrero, M. Lillo-Saavedra, and E. Menasalvas, "Deep learning for superpixel-based classification of remote sensing images," in *Proc. GEOBIA, Solutions Synergies*, 2016, pp. 14–16.

[31] X. Lv, D. Ming, Y. Y. Chen, and M. Wang, "Very high resolution remote sensing image classification with seeds-CNN and scale effect analysis for superpixel CNN classification," *Int. J. Remote Sens.*, vol. 40, no. 1–2, pp. 506–531, 2019.

[32] X. Lv, D. Ming, T. Lu, K. Zhou, M. Wang, and H. Bao, "A new method for region-based majority voting CNNs for very high resolution image classification," *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 1946.

[33] M. Wang and R. Li, "Segmentation of high spatial resolution remote sensing imagery based on hard-boundary constraint and two-stage merging," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5712–2725, Sep. 2014.

[34] M. Wang and J. Wang, "A region-line primitive association framework for object-based remote sensing image analysis," *Photogrammetric Eng. Remote Sens.*, vol. 82, no. 2, pp. 149–159, 2016.

[35] M. Wang, Q. Cui, J. Wang, D. Ming, and G. Lv, "Raft cultivation area extraction from high resolution remote sensing imagery by fusing multi-scale region-line primitive association features," *ISPRS J. Photogrammetry Remote Sens.*, vol. 123, pp. 104–113, 2017.

[36] M. Wang, J. Xing, J. Wang, and G. Lv, "Technical design and system implementation of region-line primitive association framework," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 202–216, 2017.

[37] M. Wang, Q. Cui, Y. Sun, and Q. Wang, "Photovoltaic panel extraction from very high-resolution aerial imagery using region-line primitive association analysis and template matching," *ISPRS J. Photogrammetry Remote Sens.*, vol. 141, pp. 100–111, 2018.

[38] T. Zhang and C. Suen, "A fast thinning algorithm for thinning digital patterns," *Commun. ACM*, vol. 27, no. 3, pp. 236–239, 1984.

[39] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[40] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[42] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9.
[43] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

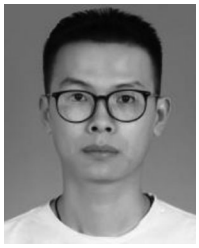
Jie Wang received the B.S. degree in surveying and mapping engineering from the School of Geosciences and Info-Physics, Central South University, Changsha, China, in 2014. He is currently working toward the Ph.D. degree with School of Geography, Nanjing Normal University, Nanjing, China.

His research interests include remote sensing and computer vision, deep learning, and their applications in remote sensing.

Yalan Zheng received the B.S. degree in cartography and geographic information system from Chuzhou College, Chuzhou, China, in 2015, and the M.S. degree in cartography and geographic information system from Southwest Forestry University, Kunming, China, in 2018. She is currently working toward the Ph.D. degree with the School of Geography at Nanjing Normal University, in the field of remote sensing image processing, deep learning, and their applications on remote sensing.

Min Wang received the Ph.D. degree in cartography and geographic information system from Chinese Academy of Sciences, Beijing, China, in 2003.

He is currently a Professor with the Key Laboratory of Virtual Geographic Environment, Nanjing Normal University, Ministry of Education of China, Nanjing, China. His research interests include remote-sensing image processing, segmentation, feature extraction, classification, and remote-sensing image mining. He has authored more than 70 peer-reviewed scientific manuscripts in journals including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, AND PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING.



Qian Shen is currently working toward the Ph.D. degree with the School of Geographical Sciences, Nanjing Normal University, majoring in cartography and geographic information systems.

His research interests include remote sensing image processing and computer vision, deep learning, and their applications in remote sensing.

Jiru Huang is currently working toward the Ph.D. degree with the Key Laboratory of Virtual Geographic Environment, Ministry of Education of China, Nanjing Normal University, Nanjing, China.

His research interests include remote sensing image processing and information extraction.