Object-Based Convolutional Neural Network for High-Resolution Imagery Classification

Wenzhi Zhao, Shihong Du, and William J. Emery, *Fellow, IEEE*

***Abstract*—Timely and accurate classification and interpretation of high-resolution images are very important for urban planning and disaster rescue. However, as spatial resolution gets finer, it is increasingly difficult to recognize complex patterns in high-resolution remote sensing images. Deep learning offers an efficient strategy to fill the gap between complex image patterns and their semantic labels. However, due to the hierarchical abstract nature of deep learning methods, it is difficult to capture the precise outline of different objects at the pixel level. To further reduce this problem, we propose an object-based deep learning method to accurately classify the high-resolution imagery without intensive human involvement. In this study, high-resolution images were used to accurately classify three different urban scenes: Beijing (China), Pavia (Italy), and Vaihingen (Germany). The proposed method is built on a combination of a deep feature learning strategy and an object-based classification for the interpretation of high-resolution images. Specifically, high-level feature representations extracted through the convolutional neural networks framework have been systematically investigated over five different layer configurations. Furthermore, to improve the classification accuracy, an object-based classification method also has been integrated with the deep learning strategy for more efficient image classification. Experimental results indicate that with the combination of deep learning and object-based classification, it is possible to discriminate different building types in Beijing Scene, such as commercial buildings and residential buildings with classification accuracies above 90%.**

***Index Terms*—Convolutional neural network (CNN), deep learning, high-resolution image, image classification.**

I. Introduction

REMOTE sensing imagery has provided a real-time and low-cost means to map urban land cover during the last few decades. The recent availability of submeter resolution imagery from advanced satellite sensors, such as WorldView-3 and GaoFen, can provide new opportunities for detailed urban land cover mapping at the object level (such as commercial buildings and residential buildings). However, the complexity of urban imagery increases sharply as the observation scale gets finer. More efficient high-resolution image classification methods need to be proposed in order to efficiently perform urban high-resolution imagery classification.

Remote sensing imagery acquired by advanced satellite sensors has the potential for detailed mapping of these urban images. However, rich spatial information presented in high-resolution images also hinders accurate interpretation [1], [2]. In fact, objects in urban areas are commonly composed of different construction materials, which can produce confused feature representations in the spectral and spatial domains. For one thing, small objects are cojointly distributed around target objects (such as antennas or chimneys on building roofs) which results in a strong intraclass variation as the spatial resolution gets finer. For another, man-made objects share similar spectral properties (such as parking lots and roofs with similar construction materials), which makes it even harder to accurately classify high-resolution images [3], [4]. As a consequence, the spectral and spatial responses from ground objects in urban scenes exhibit complex patterns, especially for high-resolution images.

The main purpose of this study is to transform the inefficient process of manually designed features into automatic feature learning by employing deep learning [5], for the efficient classification of high-resolution satellite imagery. However, due to the hierarchical structure of deep learning, deeply learned features suffer greatly from abstraction and often fail to capture the objects' contours at the pixel level. Therefore, pixel-level segments are combined to further improve the performances of interpretation in complex urban scenes [6]–[8].

The proposed method is based on the analysis of deep features extracted from high-resolution images with convolutional neural networks (CNN). To account for the boundary information of urban scenes, an object-based classification method has been used for imagery interpretation for the combination of deeply learned features. Specifically, deep features have been computed over the fixed receptive window with a five-layer CNN framework. In addition, three different segment scales have been employed in evaluating the effectiveness of object-based classification with extracted deep features.

The remainder of this paper is organized as follows. Related work on high-resolution image feature exploration and classification is outlined in Section II, and the datasets are described in Section III. Detailed information about CNN-based deep learning feature exploration and the object-based classification are introduced in Section IV. Experimental results and analysis of the deep features as well as segmentation scales of

Manuscript received November 23, 2016; revised January 11, 2017 and February 13, 2017; accepted March 5, 2017. Date of publication March 29, 2017; date of current version July 26, 2017. This work was supported by the National Natural Science Foundation of China under Grant 41471315. *(Corresponding author: Shihong Du.)*

W. Zhao and S. Du are with the Beijing Key Laboratory of Spatial Information Integration and Its Applications, Institute of Remote Sensing and Geographic Information System, Peking University, Beijing 100871, China (e-mail: w.zhao@pku.edu.cn; dshgis@hotmail.com).

W. J. Emery is with the Colorado Center for Astrodynamics Research, University of Colorado, Boulder CO 80303 USA (e-mail: emery@colorado.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2017.2680324

1939-1404 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

objects are discussed in Section V, followed by the conclusion in Section VI.

II. Related Work

A. Feature Design Versus Feature Learning

Over the past few decades, intensive studies have focused on high-resolution image classification with handcrafted features elaborated from both spectral and spatial domains. For spectral information, the brightness of each image band is usually taken as the primary feature for recognition of various objects in remote sensing imagery. Later, semantic spectral features which contain physical meanings have been proposed for efficient classification of certain objects, where it strengthened the reflectance discrepancy of different objects on specific wavelengths, such as the normalized difference vegetation index (NDVI). Further exploration of image texture demonstrated that texture-based descriptors characterize spectral variations that can provide supplementary information for efficient image classification, such as statistical descriptors based on the gray-level co-occurrence matrix (GLCM) [9], [10]. Compared with their spectral properties, high-resolution imagery contains much richer information in the spatial domain. To characterize these spatial features, Benediktsson *et al.* proposed extended morphological profiles for high-resolution imagery classification in urban areas [11]. It efficiently captures spatial information by implementing morphological operations (open and close) on high-resolution images. Similarly, spatial filters (such as Gabor filters [12]) and wavelet analysis [13] were also proposed for the extraction of spatial features in the context of high-resolution images. However, it usually requires lots of experience and expert knowledge for end-users to define such elaborate features. Moreover, even after the complicated design process of various features, it is still difficult to find the most effective features for the recognition of different objects. To illustrate the complexity of high-resolution imagery, a typical Worldview-2 image is presented here as Fig. 1. Due to the strong intraclass variation of buildings (caused by a sunlit wall and chimney), traditional feature representations have shown much more mixture and variation than the deep learning ones. For this reason, it is necessary to explore more efficient and representative image features from high-resolution images in order to accurately recognize objects with complex patterns, such as complex buildings in an urban area.

How to develop automatic classification schemes for feature learning has become one of the most significant topics in image classification over the last few years. Instead of handcrafted feature design, Cheriyadat [14] introduced a sparse coding scheme for learning of high-resolution image features with predefined filter banks. It indicated that handcrafted features have a certain level of redundancy which leads to lower interpretation accuracies. Also, Tuia *et al.* [15] proposed a feature learning model by using sparse-constrained support vector machine (SVM). Regardless of the compulsory step to select for predefined features of conventional methods, this method can automatically discover the relevant features in the potentially infinite space of image features and choose them in a heuristic way. It concluded that the automatic learning scheme can keep image features more compact, discriminative, and robust than human visual selection methods, thus resulting in better classification accuracies. In practical applications, it demonstrates that nonlinear features are more effective for class discrimination due to the existence of nonlinear class boundaries. Therefore, besides using a sparse coding strategy, the exploration of image features through nonlinear transformations also has been proven to be effective [16]. Furthermore, Tuia *et al.* [17] proposed a sparse and hierarchical feature learning model to find efficient data representations for a good classification. In this model, the selected features could be repeatedly used for filtering at the next feature generation step, thus, producing more representative features with higher nonlinearity. However, the above methods still require a certain level of experience for end-users, such as the definition of the possible feature space and a corresponding selection criteria.

Deep learning [18], [19] is one of the significant advances in artificial intelligence, and it has shown great potential for discriminative feature learning without human intervention. Inspired by the hierarchical cognition process of the human brain, deep learning can automatically generate robust and representative features layer by layer in neural networks [20], [21]. Distinctly different from low-level feature representations, deeply learned features are generally more general and robust, and it has demonstrated great effectiveness in image classification, such as face recognition [22] and scene classification [23]. In the remote sensing field, several studies have focused on imagery classification using deep learning models, such as stacked autoencoder (SAE) and convolutional neural network (CNN). But, the original SAE focused on extracting one-dimensional spectral features which probably is not sufficient for high-resolution image interpretation. Therefore, Chen *et al.* [24] improved the SAE model by introducing spatial features for the efficient classification of hyperspectral images. At the same time, the CNN algorithm became popular for high-resolution image classification due to its effectiveness in spatial feature exploration [3], [25]–[27]. Furthermore, to fit the character of remote sensing imagery, Zhao and Du [28] proposed a multiscale strategy based on the CNN model to retrieve the information in high-resolution imagery. Although the CNN-based method is efficient for robust spatial feature extraction, two major flaws of the basic CNN framework for high-resolution image classification also need to be clarified. On one hand, the CNN framework feeds on labeled image patches of fixed sizes and outputs feature vectors using layer-wise activation and abstraction. Since the output features are highly abstract and usually without specific spatial arrangement, it is difficult to predict the precise contour of target objects in image and producing blurred edge prediction results. On the other hand, deep features are learned from local images patches, which, regardless of the contextual image information, could produce misclassification results, as shown here in Fig. 2.

B. Per-Pixel Versus Object-Based Classification

As the spatial resolution gets finer, the detailed image data exhibit complicated urban patterns in both spectral and spatial domains. Using the per-pixel classification of high-resolution

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

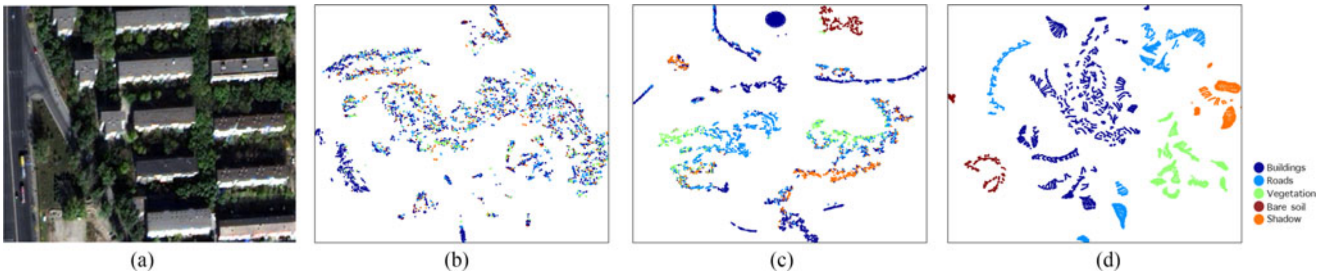


Fig. 1. Measurements of different feature representatives on a Worldview-2 study site. The high-dimensional feature representations were projected into two-dimensional space for illustration. (a) Worldview-2 image; (b) GLCM feature representation; (c) EMAP feature representation; and (d) CNN feature representation (generated by five-layer CNN).

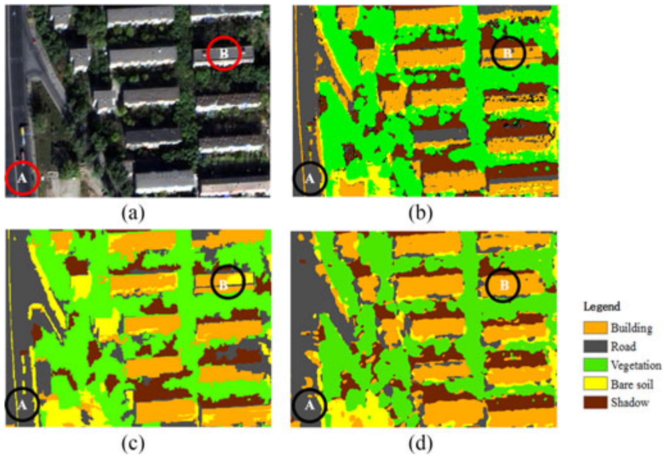


Fig. 2. Classification results of WorldView-2 image (after pan sharpening) with different strategies.

images may lead to poorer interpretation results due to the "salt and pepper effect" [29], as shown in Fig. 2(b). To reduce the pixel-level spectral heterogeneity and improve the classification performance, an object-based approach can be used to efficiently delineate and classify high-resolution imagery [30]. Object-based classification method takes image segments as building blocks for the overall image analysis. In contrast to pixel-based approaches, image objects mainly have two characteristics: 1) They are relatively homogeneous and 2) they can provide rich image features. On one hand, objects integrated with contextual information can greatly reduce local spectral variation and overcome the so-called "pepper & salt" effect. On the other hand, objects in images opens a new gate to access the target of interest, thus more discriminant semantic features can be defined, such as an objects' shape and sizes. So far, object-based classification method has shown great potential for efficient mapping of high-resolution images, especially for complex urban scenes. However, the process of object-based classification is usually very complicated and needs a certain level of expert knowledge to produce satisfactory results. Specifically, three factors weigh heavily in terms of object-based classification accuracy: 1) the segmentation scale, 2) feature extraction and selection, and 3) classification rules. A possible solution to these problems is to incorporate deep learning with the object-based method for feature self-learning and automatic classification of complex urban high-resolution imagery without much supervision. The general outline of this process is presented in Fig. 3.

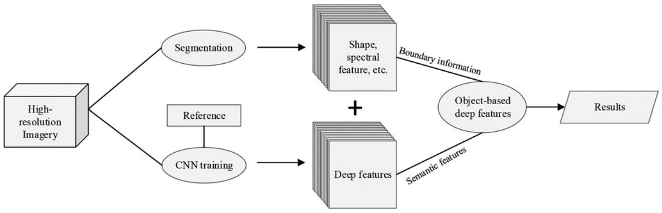


Fig. 3. Outline of the proposed method. High-resolution images and their reference maps are combined to train CNN frameworks. Segmentation results are used to improve the performance of pixel-level boundary prediction.

III. Methodology

A. *Deep Feature Learning Through CNN*

The complexity of high-resolution images causes traditional human-dependent classification methods to fail due to the limited representation power of handcrafted features. CNN as the core of deep learning has shown great potential for robust automatic feature extraction and complex object recognition in high-resolution images [31]–[33].

To obtain deep feature representations, two parts of trainable parameters for a CNN framework should be determined, i.e., the filters $\mathbf{W}$ and the biases $\mathbf{b}$, which are collectively denoted θ. During the training stage, a CNN framework f with L layers feeds with training samples $\mathbf{X}_i$, $i \in \{1, \ldots, N\}$ and is formulated

$$f(\mathbf{X};\theta) = \mathbf{W}_L \mathbf{h}_{L-1} + \mathbf{b}_l \tag{1}$$

where $\mathbf{h}_l$, $l \in \{1, \ldots, L-1\}$ denotes the vector of hidden units at the lth layer. Particularly, $\mathbf{h}_0$ represents the original input data, as shown in Fig. 4.

More specifically, when training a CNN, the input image is first connected with convolution layers by a set of convolution kernel banks. For the convolutional layers, the kernels were denoted as $\mathbf{W}_l$ and are combined with the bias terms $\mathbf{b}_l$ to convolute the input image. After that, a point-wise nonlinear activation function $g(\cdot)$ (typically the tan$h$ function) is deployed before the final output of this layer. Then, the spatial pooling layer usually follows to generate the dominant features over

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

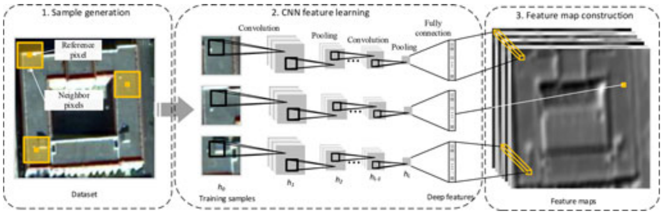


Fig. 4. Framework of a traditional CNN. Convolution layer interspersed with pooling layer in this hierarchical structure. Full connection is found at the last layer and linked with the classifier.

nonoverlapping windows for each feature map. To formulate the feed-forward process, we have

$$\mathbf{h}_l = \text{pool}(g(\mathbf{h}_{l-1} * \mathbf{W}_l + \mathbf{b}_l)). \tag{2}$$

Once the parameters θ are trained, the unlabeled datasets $\mathbf{Y}_j$, $j \in \{1, 2, \ldots, N\}$ can be encoded by

$$\mathbf{F}_j = f(\mathbf{Y}_j, \theta). \tag{3}$$

As mentioned above, deep features extracted by the CNN framework are generally robust and effective for complex image pattern descriptions, especially for the case of high-resolution urban scenes. Unlike the traditional CNN-based image classification methods, we focus on exploring deep features as local patch descriptors. However, deep features with high-level abstractions naturally fail to detect the edges of complex objects on the pixel level. Object-based classification methods interpret high-resolution images with segmented objects which can preserve objects' edges and reduce spectral variation effects. Therefore, it is widely suggested that deep features should combine with object-based image analysis methods for better classification performances.

B. Object-Based Classification With Deep Features

Although, the CNN-based methods are efficient for complex pattern description, the deep features generated from multilayer activations are usually highly abstracted, and thus, failed to predict objects' contours [34], [35]. In contrast to the patch-based image classification methods (e.g., CNN-based method), the object-based methods can effectively classify image objects using image segments with precise boundaries. Segments are homogeneous regions which are generated by one or more homogeneity measurements in feature space (typically, spectral space). Therefore, image segments can transform the whole image into meaningful regions while preserving the precise edges of targets of interest. Furthermore, image segments have additional information compared to pixels or image patches in terms of spatial and contextual extensions (objects' shape and topologies). Thus, it is natural to combine the deep features with object-based classification for the accurate interpretation of high-resolution images.

In this paper, highly abstracted deep features are combined with image segments for precise mapping of complex remote sensing images. Specifically, the multiresolution segmentation algorithm is used to generate image objects with precise edges.

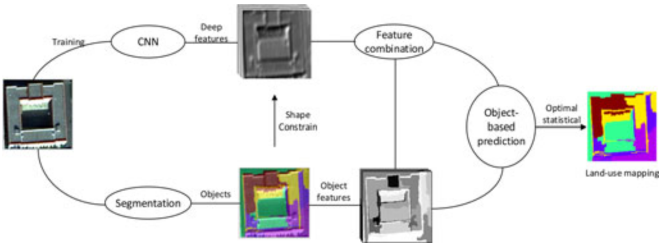


Fig. 5. Flowchart of object-based high-resolution image classification combining with deep features.

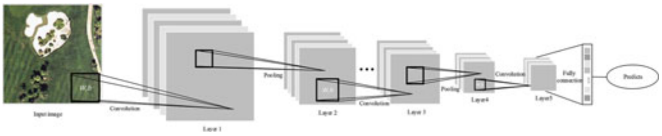


Fig. 6. Illustration of the CNN-based deep learning framework.

Then, with the shape constraint of the image objects, deep features are combined with objects' features on the pixel level and used for image classification. Finally, for each image object, the optimal statistical method was applied to determine the land cover class for mapping.

Suppose an image I contains N image objects $O_i, i \in \{1, 2, \ldots, N\}$, and, there are M pixels $I_j, j \in \{1, 2, \ldots, M\}$ inside object O_i. For each pixel I_j, the deep feature representation is denoted as F_j. Similarly, the object-based features (such as, NDVI) are represented as R_j. For feature combination purpose, the deep features are stacked with object-based properties $U_j = [F_j, R_j]$ for joint feature classification of the original imagery at the pixel level. For this method, the label of an object O_i is predicted from the majority statistics of feature vectors U_j using a two-layer neural network:

$$\mathbf{p}_j = \mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{U}_j + \mathbf{b}_1) \tag{4}$$

$$\mathbf{d}_{i,a} = \frac{1}{t(O_i)} \sum_{j \in [1,M]} \text{count}(\mathbf{p}_j == a). \tag{5}$$

Matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ are the trainable parameters of the two-layer neural network classifier. $\mathbf{p}_j$ is the predicted label at location j, and $t(O_i)$ is the pixel number of the object. The final label for each image object O_i is given by

$$l_i = \arg\max_{a \in \text{classes}} \mathbf{d}_{i,a}. \tag{6}$$

The flowchart of object-based deep learning classification is depicted in Fig. 5.

IV. Datasets

The datasets analyzed include three very different cities with complex urban conditions: Beijing, Pavia, and Vaihingen. The first Beijing scene was acquired by Worldview-2 in 2010. The second scene was acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy, and was distributed in 2008 as part of a test of the airborne Intergraph/ZI Digital Mapping

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

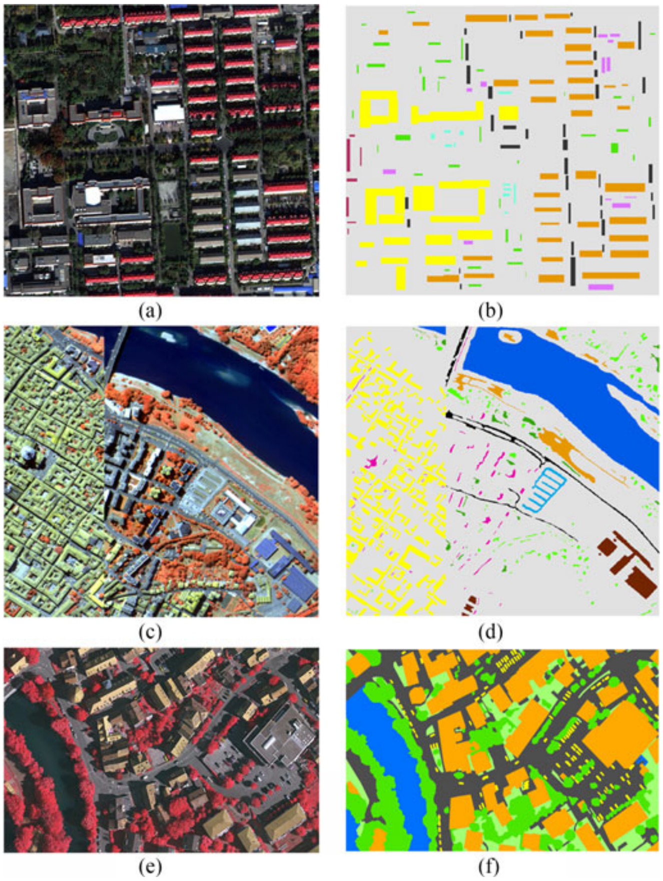


Fig. 7. Study datasets and ground truth reference maps.

Camera Instrument. The Vaihingen dataset was provided by the German Association of Photogrammetry and Remote Sensing.

A. Description of Scenes

The Beijing scene, as shown in Fig. 7(a), contains 1036 × 1146 pixels, and it has eight multispectral bands with a spatial resolution of 0.5 m (after pan sharpening). It contains typical residential buildings and examples of commercial structures with similar heights (about five or six floors) but of different sizes. The high-resolution Beijing image was chosen for two reasons. First, it embodies the main difficulties of high-resolution classification, especially for the detailed mapping of complex buildings. Second, the Beijing scene represents a common Chinese urban landscape, including flat building and narrow streets, which are obviously different from western style of modern cities. The other two study areas, Pavia and Vaihingen are shown in Fig. 7(c) and (e). The Pavia scene, contains 1096 × 715 pixels and 102 hyperspectral bands, with a spatial resolution of 1.3 m. It is composed of an elaborate urban lattice with complex structures showing a great variety in dimension, shapes, and heights. The image size of Vaihingen dataset is 892 × 1498 (reduced half resolution for efficient computation), with very high spatial resolution of 0.09 m per pixel. For better discrimination, the band combination of Vaihingen dataset was set to infrared (IR), red (R), and green (G). The urban scene of Vaihingen contains elements that differ from the other two, since it is mainly composed of small residential houses (two or three floors), roads, some sparse trees, and vegetated areas.

B. Classes, Training, and Validation Set Definition

To efficiently map different scenes, several land cover types of interest have been identified. For the Beijing case, the primary goal was to discriminate between residential buildings and commercial buildings based on the differences in construction style and size. These two kinds of buildings appear with complex elements, such as a small chimney on the top of a building and sunlit wall with great spectral variation. Moreover, roofs of commercial buildings are similar to roads in terms of spectral reflectance and shapes, which further increased the difficulty of performing a successful image classification. A further exploration was made to distinguish the different uses of the asphalt surfaces, which included parking lots and roads. Other classes such as shadow, impervious, trees, and bare soil were also added for a total of eight classes. Since vegetation occupies a large portion of urban scenes, three bands with the combining IR, R, and G were chosen to discriminate vegetation from structures.

Similar to the previous dataset, the Pavia center also has crowded residential buildings. Generally, the buildings change rapidly in terms of sizes and spectral reflectance where the left part of Pavia dataset contains with old constructions, while the eastern part of the city new buildings are mainly made of concrete. It is possible to discriminate between asphalt (roads) and bitumen (newly developed roofs). Other classes of interest were also identified, including self-blocking bricks, water, trees, shadows, meadows, and bare soil for a total of nine classes. In order to reduce the computational cost and preserve the substantial information across spectral bands, we reduced the dimensionality of Pavia dataset by using principle component (PC) analysis. Following the previous studies [28], the first three PCs hold more than 95% of the original spectral information. Thus, similar to the IR, R, and G bands, we selected the first three PCs to feed the CNN model.

The Vaihingen scene has features different from the previous two. It has a finer spatial resolution that can resolve more complex and challenging urban patterns, such as road markings and even car windows. Buildings with large varieties in shapes and sizes are interspersed with roads. It is also possible to distinguish short vegetation and trees at this level of resolution. Other common classes like water, roads, and buildings were also included for classification.

The reference map for each scene were reported in Fig. 7(b), (d), and (f), respectively. Ground reference labels have been acquired by careful visual interpretation. Both the Vaihingen scene and its reference map were distributed by the International Society for Photogrammetry and Remote Sensing (ISPRS) and the two-dimensional semantic labeling contest [36].[1] Instead of using the selected pixels, we have extended them into image patches with the labeled pixels in the center before feeding them to the CNN with its deep features. In the experiments,

[1] The Vaihingen dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [36], http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

TABLE I
BEIJING SCENE: CLASSES, COLOR LEGEND, TRAINING, AND VALIDATION SAMPLES

Beijing classes	Legend	TR	TE	NOS
Commercial buildings	🟨	838	418	8387
Residential buildings	🟧	899	447	8990
Roads	⬛	220	108	2200
Vegetation	🟩	137	67	1379
Parking lots	🟪	94	45	943
Shadow	🟥	35	16	355
Other imprevious	🟦	34	16	348
Bare soil	🟫	31	14	315

TABLE II
PAVIA SCENE: CLASSES, COLOR LEGEND, TRAINING, AND VALIDATION SAMPLES

Pavia classes	Legend	TR	TE	NOS
Water	🟦	711	353	7113
Trees	🟩	82	39	821
Meadows	🟩	29	13	296
Blocking bricks	🟦	29	13	299
Bare soil	🟧	73	35	737
Asphalt	⬛	101	49	1017
Bitumen	🟫	80	38	806
Tiles	🟨	462	229	4625
Shadow	🟪	31	14	316

TABLE III
VAIHINGEN SCENE: CLASSES, COLOR LEGEND, TRAINING, AND VALIDATION SAMPLES

Vaihingen classes	Legend	TR	TE	NOS
Buildings	🟧	17059	8527	170590
Water	🟦	4893	2444	48932
Trees	🟩	11897	5947	118978
Grass	🟩	5431	2713	54311
Cars	🟨	1060	528	10604
Road	⬛	18146	9071	181462

we have subsampled the original image in order to reduce the computation complexity.

To demonstrate the robustness of the proposed method, for each dataset, we randomly select 10% nonoverlapping samples (NOS) for training and another 5% for the test. In order to avoid the overlap phenomenon between samples, the overlap threshold is set to 80% (rejected if the overlap area is more than 80% between two samples). Detailed information about the training samples (TR), test samples (TE), and total NOS are reported in Tables I, II, and III for the Beijing, Pavia, and Vaihingen cases, respectively. The overall accuracy and Kappa coefficient were used to evaluate the classification performance.

V. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the capability of the object-based deep learning method to classify the complex urban scenes described previously. As mentioned above, we combine the powerful representations of deep features with shape-preserving object-based classification method for accurate high-resolution image interpretation. More specifically, the CNN-based deep feature learning algorithm has been adopted to describe complex urban patterns. The depth of the CNN can directly affect the abstraction level of our deep features. A detailed analysis of this depth effect has been carried out by assigning different layer configurations on the CNN framework, as described in Section V-A. Then, the effect of segmentation scales and classification accuracies is illustrated in Section V-B. In addition, deep features are combined with image objects for efficient urban mapping, the classification results are reported in Section V-C.

A. *Depth Effect on Deep Feature Learning*

The CNN-based deep learning framework has been used to extract complex urban patterns in this work. Instead of handcrafted features, the CNN network automatically learn effective feature representations from the hierarchical activation structure. Two parameters, the input size of the training samples and the depth of the CNN play important roles in terms of classification accuracies.

For the input training samples (also known as the receptive field), it should have an appropriate spatial coverage for the objects of interest. To achieve this goal, we set the window size of training sample extraction to 18 × 18 pixels (about 9, 23.4, and 3.24 m, for Beijing, Pavia, and Vaihingen scene, respectively). The majority of urban geographical objects have the dominant scales between about 3.0 and 24.0 m [37].

The depth configuration of the CNN framework controls the robustness of deep features in terms of their abstraction level. The representation power increases as the CNN involves more layers. However, due to the constraint of input training sample sizes, the deepest configuration of the CNN should be no more than five layers with the 3 × 3 kernels in our network. Following to the previous studies [3], [28], we set the number of features to 20 for each convolutional layer (except the last layer). Fig. 6 presents the detailed configuration of our CNN-based deep learning framework. The learning rate was set to 0.001.

In order to illustrate the importance of CNN depth when classifying high-resolution images, we separately trained CNN models with different depth that varies from 1 to 5. That is, one-layer CNN only has a single convolution layer, teo-layer CNN is combined with two convolution layers, three-layer CNN has the combination of two convolution layers and one pooling layer, four-layer CNN constitutes by three convolution layers and one pooling layer, and five-layer CNN is with three convolution layers and two pooling layers. The classification results using different CNN models are reported in Tables IV–VI. The overall classification accuracy is calculated to measure the classification performance of the CNN networks.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
CLASSIFICATION ACCURACIES (IN PERCENTAGE) OF WORLDVIEW DATASET WITH DIFFERENT CONFIGURATION CNNS

CNNs	CB	RB	RD	VG	PL	SD	IM	BS	OA
One-layer	60.64	95.27	35.40	81.19	13.68	4.43	0	0	68.45
Two-layers	97.38	91.02	76.46	99.68	64.90	17.65	76.18	46.26	89.30
Three-layers	99.09	97.48	87.65	100	76.86	97.44	78.97	78.91	95.89
Four-layers	99.04	98.79	88.73	99.98	77.94	97.58	66.58	59.31	96.06
Five-layers	99.40	98.10	88.80	99.98	79.47	96.08	78.32	69.53	96.29

CB: Commercial building, RB: Residential building, RD: Road, VG: Vegetation, PL: Parking lot, SD: Shadow, IM: Impervious, BS: Bare soil.

TABLE V
CLASSIFICATION ACCURACIES (IN PERCENTAGE) OF PAVIA DATASET WITH DIFFERENT CONFIGURATION CNNS

CNNs	WA	TR	ME	BB	BS	AS	BI	TI	SD	OA
One-layer	100	98.66	85.67	44.77	72.61	93.83	87.39	97.59	99.22	95.64
Two-layers	99.96	95.60	92.28	53.45	80.17	95.16	84.96	96.82	93.58	95.72
Three-layers	99.98	96.86	93.88	82.72	89.92	95.64	89.06	99.14	95.77	97.77
Four-layers	99.96	96.74	92.32	85.81	90.18	92.29	84.71	98.31	94.46	97.10
Five-layers	99.95	97.50	90.18	90.69	92.90	93.91	89.41	98.42	94.74	97.69

WA: Water, TR: Tree, ME: Meadow, BB: Bricks, BS: Bare soil, AS: Asphalt, BI: Bitumen, TI: Tiles, SD: Shadow.

TABLE VI
CLASSIFICATION ACCURACIES (IN PERCENTAGE) OF PAVIA DATASET WITH DIFFERENT CONFIGURATION CNNS

CNNs	BD	WA	TR	GR	CA	RD	OA
One-layer	67.29	0.01	88.29	43.66	1.49	69.23	63.27
Two-layers	75.50	90.40	90.39	42.36	3.64	75.98	75.64
Three-layers	86.52	94.84	91.92	66.75	38.46	78.47	83.14
Four-layers	82.68	75.15	90.13	54.10	26.48	89.53	82.13
Five-layers	86.60	95.99	94.29	62.58	40.49	90.38	87.14

BD: Buildings, WA: Water, TR: Trees, GR: Grass, CA: Cars, RD: Road.

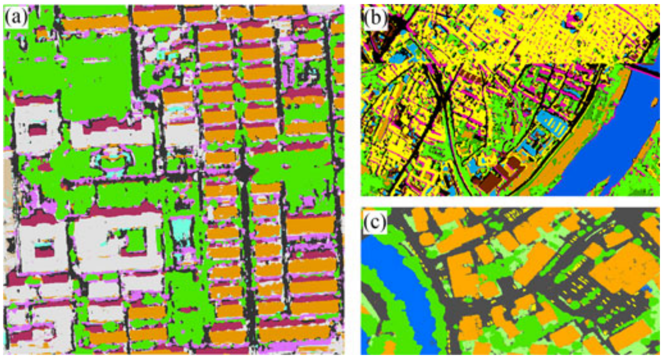


Fig. 8. In (A), (B), and (C) is shown the classification results by using five-layer CNN network over Worldview-2, Pavia, and Vaihingen complex urban scene, respectively.

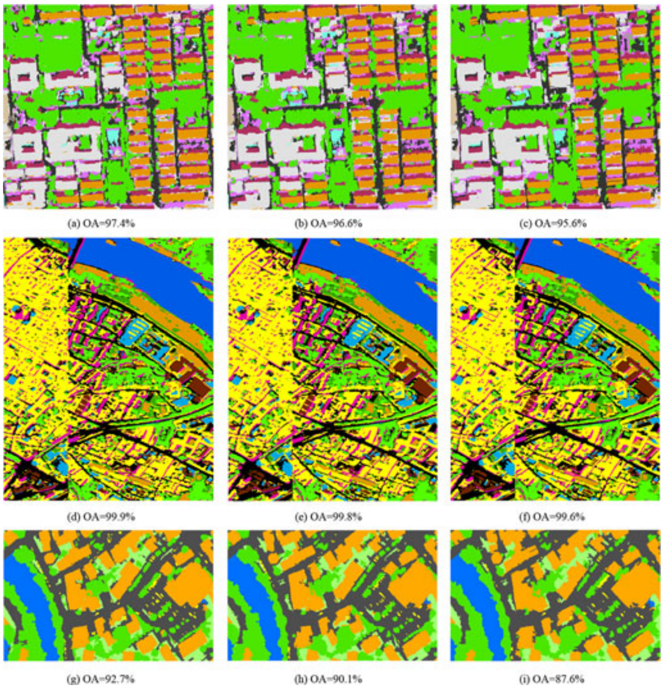


Fig. 9. Classification results with 10, 20, 30 segmentation scales for Worldview-2 (a)–(c), Pavia center (d)–(f), and Vaihingen (g)–(i).

The depth parameter of the CNN framework significantly impacts the classification performance in terms of accuracy. As stated above, classification accuracies become greater if feature representations get deeper and deeper. Therefore, the deepest CNN configuration will obtain the best classification results. The classification results using a five-layer configuration of CNN are presented in Fig. 8. To quantitatively evaluate the accuracy of the classification results for these three datasets, we reported the classification accuracies and their variations over different depth, as shown in Tables IV–VI. In general, these tables indicate that the overall classification accuracy increase as the CNN gets deeper. However, for the cars in Vaihingen dataset and bare soil in the scene of Worldview-2, there is a significant drop with the deeper configuration in terms of classification accuracies. The reason for this phenomenon is probably that the bare soil does not have a fixed spatial pattern but mainly depends on spectral reflectance for discrimination. The other classes with fixed spatial patterns become well classified as the network gets

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

deeper. We also noticed that the classification results of the Vaihingen dataset suffer from the "pepper & salt" affect, which also reduced the classification accuracy. Therefore, to further increase the classification accuracy, it is advisable to combine image segments with deeply abstracted features by introducing accurate boundary information.

B. Segmentation Scale Effects

Although deep features are efficient in terms of complex pattern description, they fail to predict image objects edges due to the highly abstracted feature representations. To improve the classification accuracy through deep learning, an object-based complementary strategy is necessary. The object-based method overcomes local spectral variation and provides accurate edge realizations with the image segments. Furthermore, the image objects have access to geographic entities, thus object-based image features with more semantic meanings could be utilized for better discrimination of complex urban scenes.

The multiresolution segmentation algorithm [38] is applied to generate image objects through experiments. Three key parameters need to be determined before the segmentation algorithm is deployed, namely they are scale (S_{sc}), shape (S_{sh}), and compactness (S_{cm}). The scale parameter directly controls the size of segmented objects by matching the required level of detail. In the experiments, we set the shape parameter S_{sh} and compactness parameter S_{cm} to 0.1 and 0.5, respectively. To evaluate the segmentation scale effects on classification results, we visually checked segmented objects by considering scale parameter S_{sc} varieties from 10 to 30. As reported in [39], the lowest scale level presented the best classification performance. Therefore, to provide accurate edges for deep learning-based classification, we set oversegmentation scales 10, 20, and 30 for each image dataset. For each object, the mean brightness of each band and the NDVI are selected as object-based features.

With three different segmentation scales, the classification results of three datasets are reported in Fig. 9. To quantitatively evaluate the performance of the object-based classification results, we present the detailed information about segmented image objects and the variation of classification accuracies at different segmentation scales. The number of segmented objects for each dataset is presented in Fig. 10(a). Classification results of each dataset are reported in Fig. 10(b)–(d). These results indicate that the classification accuracy significantly drops as the segmentation scale gets larger. Therefore, we can conclude that oversegmentation is more suitable for deep feature-based image interpretation, especially for high-resolution complex urban scenes.

C. Method Comparison

To further illustrate the effectiveness of the object-based CNN, we compared this proposed classification method with several traditional classification methods by classifying all three datasets. More specifically, we compared the object-CNN (OCNN) with SVM (spectral features), extended morphological attribute profiles method (EMAP), spectral+EMAP, pixel-based CNN (PCNN), spectral and spatial feature classification (SSFC)[3], and the multiscale CNN (MCNN)[28]. During the comparison, the EMAP features are built using the area (related to the size of the regions) and standard deviation (which measures the homogeneity of the pixels enclosed by the regions) attributes. The threshold values of area are chosen in the range of $\{50{,}500\}$ and standard deviation ranging from 2.5% to 20%. The classification maps are shown in Fig. 11, and the classification accuracies are reported in Tables VII–IX. For the traditional methods, we investigated the pure spectral information, EMAP features and spectral-EMAP features for image classification. The default classifier for traditional methods is linear SVM, the parameter of SVM classifier selected by five-cross validation. However, due to the low-level feature similarities between different classes (such as roads and roofs), the predicted maps with traditional methods suffer a lot from misclassification and noises. Different from using the traditional feature descriptors, the PCNN explored high-level feature representations with hierarchical framework. Therefore, the PCNN-based classification results are more robust and accurate. However, it fails to capture the spectral information over different image bands. The SSFC method was proposed to integrate the rich spectral information and high-level spatial features. Thus, the SSFC method achieved better classification results, in terms of classification accuracies. The geographical objects in remote sensing images are commonly displayed in different scales (e.g., roofs with different sizes). But, both PCNN and SSFC are not able to capture the useful information over scales. The MCNN was designed to model image objects over different scales. As a result, the classification results of MCNN are better than the previous two models. The PCNN, SSFC, and MCNN explored deep features from image patches, which have overlooked the boundary information of image objects. To overcome this problem, the object-based CNN is presented. It combined the robust deep features and accurate boundaries from image segments for better classification of high-resolution images. In this work, we chose a five-layer CNN for deep feature extraction and the scale parameter for image objects generation was set to 10. The classification results of OCNN demonstrated that the strategy of combining deep features and image objects is effective, in terms of classification accuracies.

As shown in Fig. 11, the classification maps directly classified by using spectral information or EMAP features suffer greatly from the "pepper and salt" effect. The CNN-based methods produce smoothing and accurate results, especially for the class building. Although the previous CNN methods are effective for complex spatial feature extraction, it shows less strength in capturing boundary information. For the complex classes (such as buildings and roads), the classification results of using PCNN, SSFC, and MCNN are less promising. However, by integrating image objects and deep features, the classification accuracies for all the datasets are increased sharply.

VI. Discussion

In this study, we propose an object-based CNN for the classification of high-resolution images. Feature extraction is one of the biggest challenges for the analysis of remote sensing images. Traditional image classification methods require numerous image features to be empirically designed and linked to

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

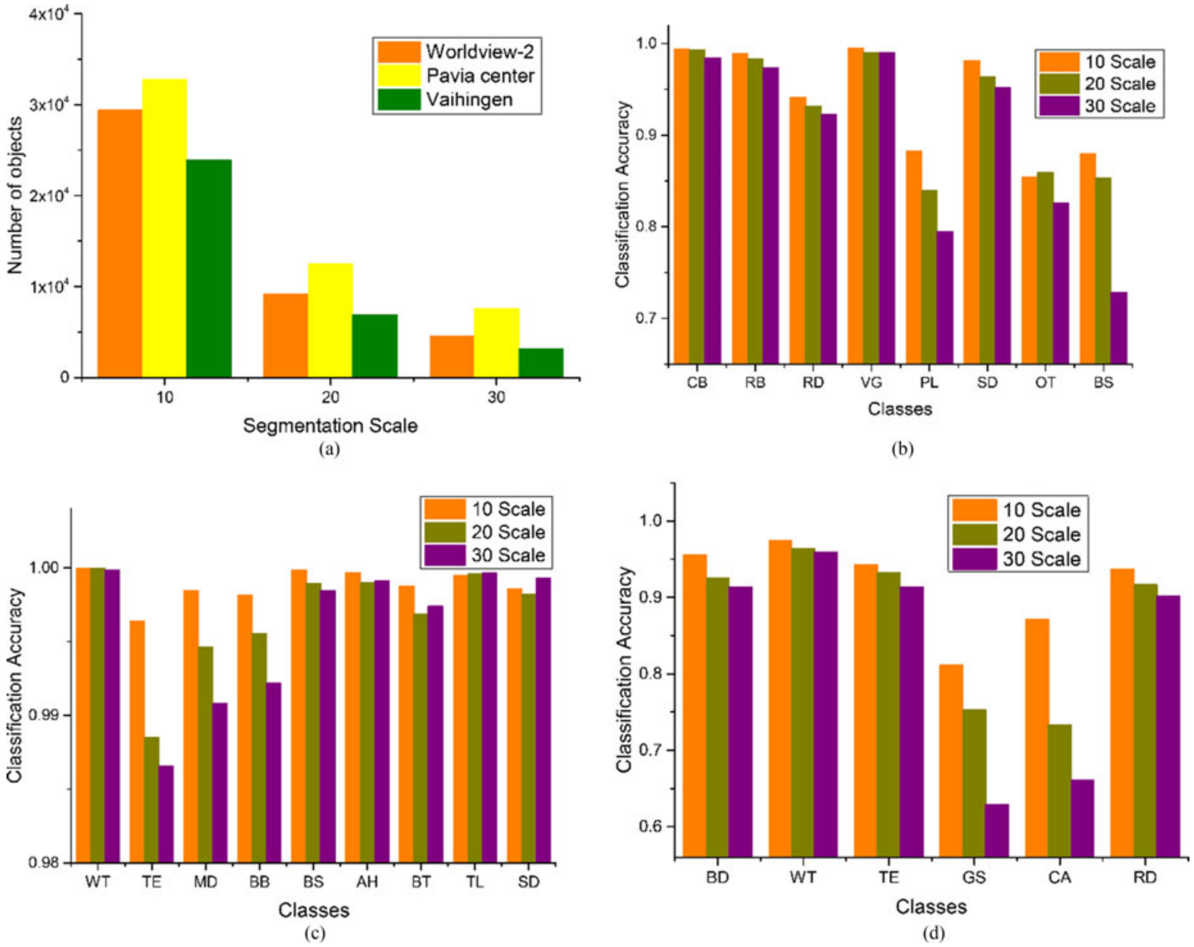


Fig. 10. Quantitative evaluation of segmentation scale on classification accuracies. (a) Number of objects. (b) Worldview-2 classification results. (c) Pavia center classification results. (d) Vaihingen classification results.

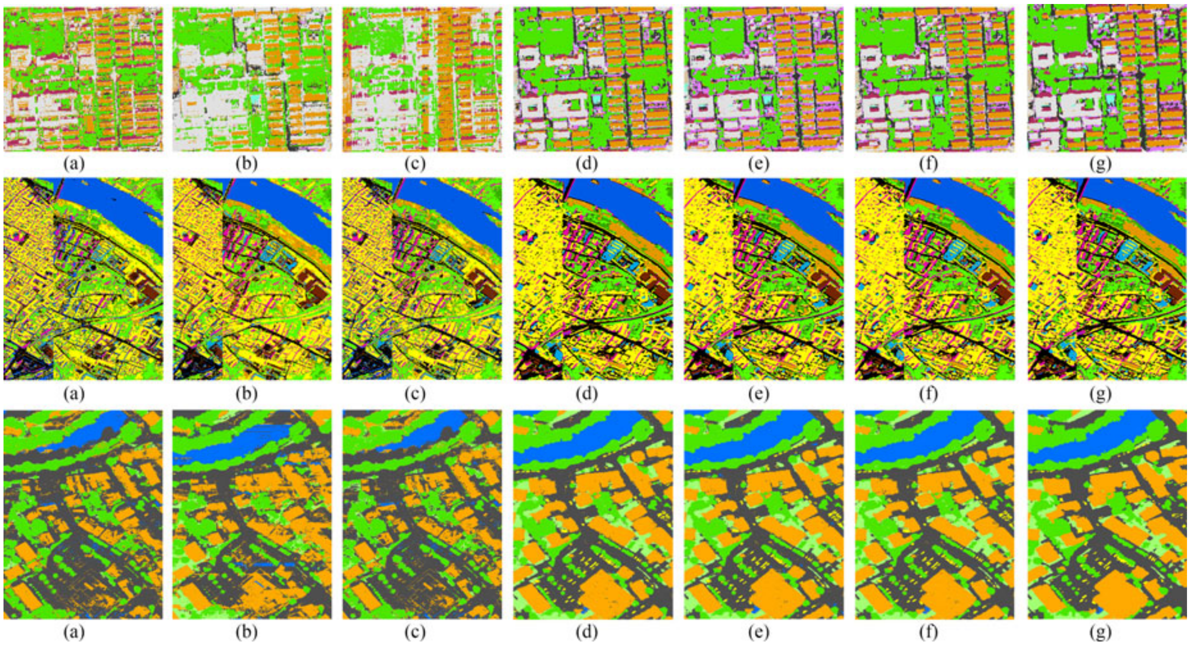


Fig. 11. Classification results for the Worldview-2, Pavia center, and Vaihingen datasets, respectively. (a) SVM classification, (b) EMAP-based classification, (c) spectral and the EMAP combined classification, (d) pixel-based CNN classification, (e) SSFC classification, (f) MCNN classification, and (g) object-based classification.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
CLASSIFICATION ACCURACIES OF THE WORLDVIEW-2 DATASET

	SVM	EMAP	SEMAP	PCNN	SSFC	MCNN	OCNN
CB	91.74	86.70	91.84	98.15	98.36	99.56	99.34
RB	67.68	91.23	94.17	95.66	96.37	97.92	98.71
RD	29.61	7.12	47.43	86.52	90.73	93.69	94.56
VG	33.28	83.81	94.45	99.36	98.13	99.97	99.67
PL	7.37	8.37	14.75	80.95	91.67	84.66	88.67
SD	62.68	45.57	4.83	76.83	89.20	98.81	99.03
IM	19.91	28.43	32.80	62.28	74.14	80.27	86.62
BS	5.27	18.48	23.85	46.03	45.52	74.90	66.58
AA	39.69	46.21	50.51	80.72	85.52	91.22	91.65
OA	66.47	74.97	82.16	93.78	95.29	97.11	97.45
KP	48.72	61.52	73.16	91.03	93.26	95.86	96.49

Both spectral and EMAP feature representations were classified by the SVM classifier. OCNN refers to the object-based CNN method.

TABLE VIII
CLASSIFICATION ACCURACIES OF THE PAVIA CENTER DATASET

	SVM	EMAP	SEMAP	PCNN	SSFC	MCNN	OCNN
WA	99.79	100	99.98	100	100	100	100
TR	92.95	98.99	93.18	97.50	95.56	99.10	99.53
ME	79.06	24.40	82.78	97.09	99.58	98.70	99.81
BB	53.59	88.34	70.24	98.51	99.18	99.78	99.93
BS	41.18	85.27	65.42	97.94	99.94	99.98	99.97
AS	91.71	95.49	96.40	99.72	99.94	99.98	100
BI	81.47	90.30	81.62	96.25	99.20	99.89	99.82
TI	97.74	99.47	98.76	99.53	99.66	99.93	99.98
SD	69.58	92.21	96.23	99.37	99.93	100	100
AA	78.56	86.05	87.18	98.32	99.13	99.67	99.88
OA	92.98	96.44	95.65	99.34	99.61	99.90	99.95
KP	89.94	94.94	93.80	99.07	99.44	99.86	99.94

Both spectral and EMAP feature representations were classified by the SVM classifier. OCNN refers to the object-based CNN method.

TABLE IX
CLASSIFICATION ACCURACIES OF THE VAIHINGEN DATASET

	SVM	EMAP	SEMAP	PCNN	SSFC	MCNN	OCNN
BD	58.41	78.36	94.50	96.00	96.37	96.20	96.40
WA	43.61	68.64	79.80	98.15	98.85	94.69	97.73
TR	90.01	87.83	86.86	93.30	95.48	94.68	95.70
GR	11.99	35.51	12.94	83.25	86.55	88.02	81.87
CA	2.84	11.81	13.27	62.06	70.81	68.78	66.05
RD	85.70	80.30	74.37	92.05	91.16	93.23	94.27
AA	48.33	58.64	52.82	87.54	89.87	89.27	88.67
OA	66.60	74.64	75.51	92.77	92.41	92.60	93.84
KP	54.60	65.95	62.81	90.47	91.35	91.57	91.87

Both spectral and EMAP feature representations were classified by the SVM classifier. OCNN refers to the object-based CNN method.

the characteristics of different images, which is time-consuming and often fails to achieve accurate interpretation. Unlike these traditional methods, in this paper, the CNN as the characteristic deep learning method was chosen for automatic feature learning. With the hierarchical structure of the CNN, image features at higher levels can be automatically extracted and the method has shown robustness and good accuracy in the presence of complex targets. As the abstraction level increased, the extracted deep features demonstrated strong invariance in terms of semantic content. However, the method often fails to capture boundary information of the target and suffers from the well-known “pepper & salt” effect.

To further improve the quality of our CNN-based classification results, we suggested to combine the CNN-based features with the low-level image segments for better descriptions of targets’ boundary and to reduce the “pepper & salt” effect. As demonstrated in our experiments, the combination of image objects and deep features is quite effective. For one thing, it alleviates people from the time-consuming process of feature selection. For another, the combination of object-based image interpretation and deep learning method provides accurate targets’ boundary information as well as semantic labels.

VII. CONCLUSION

In this paper, we propose an effective way to classify high-resolution images by combining deep features and image objects. Compared to the traditional classification methods, the proposed procedure utilizes deep CNN framework to automatically extract robust and discriminative features for complex urban objects classification (such as building roofs and cars). To evaluate the effectiveness of deep features, we tested the CNN framework with five different layer configurations for the classification of high-resolution imagery. However, as the CNN framework gets deeper, the generated features become more and more robust but often too abstract (overlooked the shape of the target objects) to describe boundary information. Complementary, the object-based classification method can preserve edge information in complex urban scenes which can be integrated with the highly abstracted deep features. As a solution, an object-based classification method is combined with deep features to promote the interpretation accuracy of high-resolution images. Experimental results indicate that the combination of deep learning and object-based classification method is effective for mapping complex urban image datasets.

However, the proposed method has no access to the contextual information at the global level. The relationships between image objects are also important to promote the classification results. Therefore, we focus on modeling the contextual information with deep feature-based image objects and for the improved image classification.

REFERENCES

[1] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, “Active learning methods for remote sensing image classification,” *IEEE Trans. Geoscience Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[2] L. Bruzzone and L. Carlin, “A multilevel context-based system for classification of very high spatial resolution images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2587–2600, Sep. 2006.

[3] W. Zhao and S. Du, “Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[4] F. Zhang, B. Du, L. Zhang, and M. Xu, “Weakly supervised learning based on coupled convolutional neural networks for aircraft detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.

[5] Y. Bengio, “Learning deep architectures for ai,” *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[6] U. C. 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, no. 3, pp. 239–258, 2004.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.

[7] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[8] W. Zhao and S. Du, "Scene classification using multi-scale deeply described visual words," *Int. J. Remote Sens.*, vol. 37, no. 17, pp. 4119–4131, 2016. [Online]. Available: http://dx.doi.org/10.1080/01431161.2016.1207266

[9] A. Puissant , J. Hirsch, and C. Weber, "The utility of texture analysis to improve perpixel classification for high to very high spatial resolution imagery," *Int. J. Remote Sens.*, vol. 26, no. 4, pp. 733–745, 2005. [Online]. Available: http://dx.doi.org/10.1080/01431160512331316838

[10] W. Zhao, L. Luo, Z. Guo, J. Yue, X. Yu, H. Liu, and J. Wei, "Road extraction in remote sensing images based on spectral and edge analysis," *Spectrosc. Spectral Anal.*, vol. 35, no. 10, pp. 2814–2819, 2015.

[11] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[12] T. C. Bau, S. Sarkar, and G. Healey, "Hyperspectral region classification using a three-dimensional gabor filterbank," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3457–3464, Sep. 2010.

[13] X. Huang, L. Zhang, and P. Li, "A multiscale feature fusion approach for classification of very high resolution satellite imagery based on wavelet transform," *Int. J. Remote Sens.*, vol. 29, no. 20, pp. 5923–5941, 2008. [Online]. Available: http://dx.doi.org/10.1080/01431160802139922

[14] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[15] D. Tuia, M. Volpi, M. D. Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse svm," *IEEE Trans. Geosci.Remote Sens.*, vol. 52, no. 10, pp. 6062–6074, Oct. 2014.

[16] J. Li *et al.*, "Multiple feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1592–1606, Mar. 2015.

[17] D. Tuia, R. Flamary, and N. Courty, "Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions," *J. Photogrammetry Remote Sens.*, vol. 105, pp. 272–285, 2015.

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 8595–8598.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2012, pp. 1097–1105.

[22] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2014, pp. 1988–1996.

[23] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2010, pp. 1378–1386.

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

[25] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3368–3379, 2015. [Online]. Available: http://dx.doi.org/10.1080/2150704X.2015.1062157

[26] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[27] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015.

[28] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 113, pp. 155–165, 2016.

[29] T. Blaschke, S. Lang, E. Lorup, J. Strobl, and P. Zeil, "Object-oriented image processing in an integrated gis/remote sensing environment and perspectives for environmental applications," *Environ. Inf. Planning Politics Public*, vol. 2, pp. 555–570, 2000.

[30] G. Duveiller, P. Defourny, B. Desclée, and P. Mayaux, "Deforestation in central africa: Estimates at regional, national and landscape levels by advanced processing of systematically-distributed landsat extracts," *Remote Sens. Environ.*, vol. 112, no. 5, pp. 1969–1981, 2008.

[31] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[32] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[33] M. A. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[34] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[36] M. Cramer, "The dgpf-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie Fernerkundung Geoinf.*, vol. 2010, no. 2, pp. 73–82, 2010.

[37] C. Small, "High spatial resolution spectral mixture analysis of urban reflectance," *Remote Sens. Environ.*, vol. 88, nos. 1/2, pp. 170–186, 2003.

[38] A. Darwish, K. Leukert, and W. Reinhardt, "Image segmentation for the purpose of object-based classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2003, vol. 3, pp. 2039–2041.

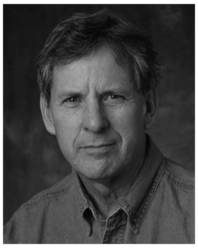
[39] S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng, "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery," *Remote Sens. Environ.*, vol. 115, no. 5, pp. 1145–1161, 2011.

Wenzhi Zhao was born in Shandong, China, in 1990. He is currently working toward the Ph.D. degree in the Institution of Remote Sensing and Geographic Information System, Peking University, Beijing, China.

His research interests include hyperspectral data analysis, high-resolution image processing, deep learning techniques, and computational intelligence in remote sensing images.

Shihong Du received the B.S. and M.S. degrees in cartography and geographic information system from the Wuhan University, Hubei, China, and the Ph.D. degree in cartography and geographic information system from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 1998, 2001, and 2004, respectively.

He is currently an Associate Professor with the Peking University, Beijing. His research interests include qualitative knowledge representation, reasoning and its applications, and semantic understanding of spatial data including GIS and remote sensing data.



William J. Emery (F'02) received the Ph.D. degree in physical oceanography from the University of Hawaii, Honolulu, HI, USA, in 1975.

After working at Texas A&M University, he moved to the University of British Columbia, in 1978, where he created a Satellite Oceanography facility/education/research program. He was an Appointed Professor in Aerospace Engineering Sciences at the University of Colorado in 1987. He is an Adjunct Professor of informatics at Tor Vergata University, Rome, Italy. He has authored more than 182-refereed publications and 2 textbooks in addition to having given 91 conference papers. Dr. Emery is the VP for publications of the IEEE Geoscience and Remote Sensing Society (GRSS). He received the 2004 GRSS Educational Award and the 2009 GRSS Outstanding Service Award. He is a Fellow of the American Meteorological Society (2010), the American Astronautical Society (2011), and the American Geophysical Union (2012). He is the new Chair of the IEEE Periodicals Committee of the IEEE Technical Advisory Board.

Authorized licensed use limited to: b-on: ISCTE. Downloaded on December 03,2023 at 19:35:37 UTC from IEEE Xplore. Restrictions apply.