A Multispectral and Multiangle 3-D Convolutional Neural Network for the Classification of ZY-3 Satellite Images Over Urban Areas

Xin Huang, *Senior Member, IEEE*, Shuang Li, Jiayi Li, *Member, IEEE*, Xiuping Jia, *Senior Member, IEEE*, Jun Li, *Senior Member, IEEE*, Xiao Xiang Zhu, *Senior Member, IEEE*, and Jón Atli Benediktsson, *Fellow, IEEE*

***Abstract*— The recent availability of high-resolution multiview ZY-3 satellite images, with angular information, can provide an opportunity to capture 3-D structural features for classification. In high-resolution image classification over urban areas, objects with diverse vertical structures make urban landscape more heterogeneous in 3-D space and consequently can make the classification challenging. In this article, a novel multiangle gray-level cooccurrence tensor feature is proposed based on the multiview bands of the ZY-3 imagery, namely, GLCM$^{\mathrm{MA-T}}$. The GLCM$^{\mathrm{MA-T}}$ feature captures the distributions of the gray-level spatial variation under different viewing angles, which can depict the 3-D textures and structures of urban objects. The spectral and GLCM$^{\mathrm{MA-T}}$ tensor features are interpreted by two 3-D convolutional neural network (CNN) streams and then concatenated as the input to the fully connected layer. This novel multispectral and multiangle 3-D convolutional neural network (M^2-3-DCNN) combines the spectral and angular information, and the fused feature has the potential to provide a comprehensive description of urban objects with complex vertical structures. The experimental results on ZY-3 multiview images from four test areas indicate that the proposed method can significantly improve the classification accuracy when compared with several state-of-the-art multiangle features and deep-learning-based image classification methods.**

***Index Terms*— Convolutional neural network (CNN), gray-level cooccurrence matrix (GLCM), high-resolution image classification, multiangle (MA), tensor.**

I. Introduction

CURRENTLY, with the ongoing development of remote sensors and platforms, more and more high-resolution remote sensing (HRRS) images, including monoscopic and multiview images, are becoming available. The improved resolution and additional viewing angles can provide abundant details [1] in 3-D space, which are of great benefit to urban land-cover classification. However, new challenges also arise with the improved spatial resolution. Specifically, intraclass variance increases, while interclass variation decreases [2], [3]. Consequently, the interpretation in the spectral domain can be difficult. To tackle this issue, a large number of studies have worked toward developing spatial features, including texture [4]–[7], shape [8], and morphological profiles [9]–[11], which are effective ways to improve the classification accuracy [4]–[11]. For example, application of the gray-level cooccurrence matrix (GLCM) [4], which is a texture descriptor that counts how often two pixels of certain gray levels appear in predefined directions and distances, has resulted in an improvement in the classification accuracy of high-resolution images [12].

Meanwhile, although the use of spatial features has been important in HRRS image classification, these features are often extracted from monoscopic images that only contain information of a single viewing angle, and they are unable to describe the 3-D structure in urban image scenes [13]. Furthermore, the angular information in high-resolution images such as those acquired by the WorldView-2/3 (WV-2/3) and ZY-3 satellites has not been fully exploited.

During the process of urbanization, buildings in urban areas become more diversified, with different heights and materials, resulting in more complex 3-D structure and increasing heterogeneity [14]. Due to the limitations of 2-D planar features

Manuscript received August 30, 2020; revised October 29, 2020; accepted November 7, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0501403, in part by the National Natural Science Foundation of China under Grant 41701382 and Grant 41771360, in part by the National Program for Support of Top-notch Young Professionals, and in part by the Hubei Provincial Natural Science Foundation of China under Grant 2017CFA029. *(Corresponding author: Jiayi Li.)*

Xin Huang is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: xhuang@whu.edu.cn).

Shuang Li and Jiayi Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: lishuang_789@163.com; zjjercia@whu.edu.cn).

Xiuping Jia is with the School of Engineering and Information Technology, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: x.jia@adfa.edu.au).

Jun Li is with the Guangdong Provincial Key Laboratory of Urbanization and Geo-Simulation, Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: lijun48@mail.sysu.edu.cn).

Xiao Xiang Zhu is with the German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), 82234 Weßling, Germany, and also with the Signal Processing in Earth Observation, Technische Universität Munich (TUM), 80333 Munich, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Jón Atli Benediktsson is with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (e-mail: benedikt@hi.is).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2020.3037211.

Digital Object Identifier 10.1109/TGRS.2020.3037211

in representing vertical structures, it is essential to design features that can capture 3-D information using multiview high-resolution images.

The ZY-3 satellite, which was launched in January 2012, is the first civilian stereo mapping satellite of China. The satellite is equipped with three-line cameras and has the ability to simultaneously collect the triple-view panchromatic images, that is, nadir, forward, and backward (NFB), with inclination angles of ± 22°. The multiview images are obtained along-track, together with the multispectral (MS) images, and there is almost no interval for the acquisition time between them. The short interval between the acquisition time of multiview images can minimize the land cover change and the variations in atmosphere and light conditions in the study area. In this situation, the variations in gray level in the multiview images are mainly caused by the difference of viewing angles. In addition, an appropriate inclination angle (e.g., 22° for the ZY-3 satellite) is beneficial to detecting the man-made classes like buildings [15]. These characteristics make the ZY-3 imagery a potential data source for capturing urban vertical structure information.

There are two traditional strategies when using high-resolution multiview images for classification. One is to generate a digital surface model (DSM) from stereo images as additional elevation information by using an image matching method [16]. For example, Li *et al.* [17] used the normalized DSM (nDSM) generated from GeoEye images and OpenStreetMap to distinguish building types. The obtained accuracy improvement (3%) confirmed the effectiveness of the stereo images; nevertheless, the performance of this kind of method relies on the quality of the DSM, which can be affected by the image matching accuracy, image occlusions, shadows, and so on [18]. The other traditional strategy concentrates on exploring the multiangle (MA) reflectance information. The MA reflectance is obtained by converting the original raw digital numbers (DN) of multiview images into the surface reflectance values that contain discriminative information of different objects [19], [20]. For example, Yan *et al.* [21] constructed a bidirectional reflectance distribution function (BRDF) model based on the multiview observations of an unmanned aircraft vehicle (UAV) platform. Compared with the base case of digital orthophoto maps (DOM), the reflectance based on BRDF extrapolation can provide a 24% improvement in overall accuracy (OA). Tao and Amr [22] exploited the multiview information of UAV images to classify wetland land covers, obtaining a classification result that was superior to that of the object-based and BRDF-based methods. However, the above-mentioned studies mainly focused on the spectral reflectance of stereo images, while neglecting the abundant MA spatial information that can be described by the features within and between multiview images on the basis of the texture, shape, structure, and spatial position (e.g., MA texture patterns).

Under different viewing angles, urban objects with various 3-D structures exhibit different spatial variation characteristics. For instance, the roof of a building is usually presented in the nadir-view image and its lateral sides are often shown in the forward- or backward-view image. This can be used to boost the interpretation performance for urban scenes but has not yet been fully investigated. Hence, there is a need to develop novel and effective features to exploit the angular information in multiview imagery. For example, Huang *et al.* [23] considered the differences of multiview images as additional information that can reveal urban structures and materials. By analyzing the differences from the pixel level, feature level, and label level, the angular difference features (ADFs) were extracted. The results showed that ADFs can perform well in classifying urban scenes, especially complex man-made classes. However, the ADFs were processed and classified as vectors, leading to the loss of the 3-D contextual structure inherent in MA images.

The above-mentioned features are, however, low-level or mid-level features (i.e., ones that focus on image details), and they lack the ability to capture semantic information from the images [24]. Recently, deep-learning-based methods, particularly convolutional neural networks (CNNs), have shown the ability to learn features hierarchically from low levels to high levels [25]. These deep features, compared with handcraft low-level or mid-level ones, are more abstract and robust and have shown powerful discriminative ability [26] in HRRS image interpretation. In recent research, 3-D-CNN models have been used to represent and learn different data/features from remote sensing (RS) images [27]–[31]. For example, by simultaneously extracting the spectral–spatial features, Fang *et al.* [32] built a depthwise-pointwise (DP) block, which is composed of a 3-D depthwise convolutional layer and two pointwise convolutional layers [33]. Ji *et al.* [34] classified crop types using a 3-D-CNN model to exploit the temporal information (such as the full crop growth cycle) of multitemporal data. However, to the best of our knowledge, there have been few studies that have used 3-D-CNN models to exploit angular information in multiview satellite images, which is an approach that has the potential to explicitly depict 3-D urban structures. Moreover, the integration of hand-crafted features and deep learning features has been demonstrated to be an effective way to improve the classification accuracy [35], [36]. For example, a two-stream network [35] was constructed by integrating spectral information and the local binary pattern feature (LBP, a man-made texture feature proposed in [37]). Compared with a one-steam network with only spectral images, the addition of the LBP stream achieves a better classification performance.

In this article, a multiangle gray-level cooccurrence tensor feature ($\text{GLCM}^{\text{MA}-\text{T}}$) and a multispectral and multiangle 3-D convolutional neural network (M^2-3-DCNN) are proposed for urban area classification based on the ZY-3 imagery. The flowchart of the proposed approach is shown in Fig. 1. The major contributions of this article are summed up as follows.

First, the proposed $\text{GLCM}^{\text{MA}-\text{T}}$ feature converts ZY-3 stereo images into the multiview gray-level cooccurrence space to capture the variation characteristics of the gray-level spatial distributions across viewing angles. Second, in the feature interpretation stage, the novel M^2-3-DCNN method is designed to exploit the spectral and angular information in the ZY-3 images. The MS and MA GLCM tensor features are separately interpreted by the spectral and angular 3-D-CNN streams, respectively, and are fused in the following layers. The deep fusion of spectral, spatial,

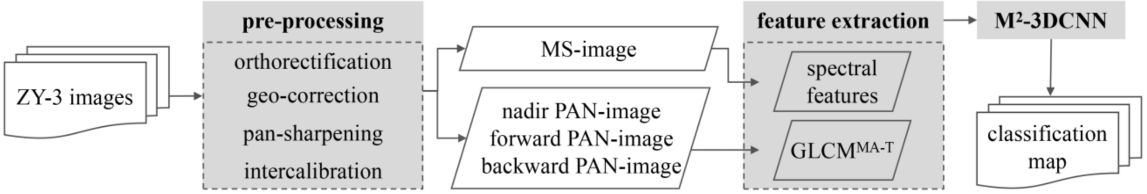

Fig. 1. Processing flow of the proposed approach for the classification of urban ZY-3 images.

and angular information achieved by M^2-3-DCNN can provide a comprehensive description for urban areas in 3-D space. Overall, the proposed GLCM$^{\text{MA}-\text{T}}$ feature and M^2-3-DCNN method are able to fully use and interpret ZY-3 multiview satellite images over urban areas.

The rest of this article is organized as follows. A description of the GLCM$^{\text{MA}-\text{T}}$ feature is provided in Section II. The architecture of the proposed M^2-3-DCNN method is presented in Section III. The experimental results of the proposed method and in-depth comparisons with other state-of-the-art multiview features and deep-learning-based methods for HRRS image classification are reported in Section IV. This is followed by a discussion on the parameters and the three components of the proposed method in Section V. Finally, the conclusions of this article are given in Section VI.

II. MA Tensor Texture Feature

A. Traditional GLCM

The GLCM is commonly used to analyze the textures in remote sensing images. It describes the planar spatial relationship within a local area of an image by measuring the correlation between the gray levels of two pixels appearing in a certain distance r and direction θ. Before calculating the GLCM, the gray tones appearing in the image are usually quantized as N_g, which determines the size of the cooccurrence matrices. In this article, the relative position is defined by a displacement vector $\vec{\Delta} = (\Delta_x, \ \Delta_y)$ that represents the separation of the two neighboring pixels in the row and column directions. The corresponding distance and direction between the pixel pairs can be defined as $r = \sqrt{\Delta_x^2 + \Delta_y^2}$ and $\theta = \tan^{-1}(\Delta_y/\Delta_x)$, respectively. Given the displacement vector $\vec{\Delta}$, the element $(i, \ j)$ of the GLCM is obtained by counting the frequency of the cooccurrence between the gray values for pixel pairs within a sliding window, as follows [4]:

$$P(i, j, \vec{\Delta}) = \#\Big\{(x_1, y_1), (x_2, y_2) \in S|[x_2 - x_1, \\ y_2 - y_1] = \vec{\Delta}, I(x_1, y_1) = i, I(x_2, y_2) = j\Big]\Big\} \quad (1)$$

where # denotes the number of elements contained in the set in (1). W_x and W_y are the size of the moving window. We let $D_x = \{0, 1, \ldots, W_x - 1\}$ and $D_y = \{0, 1, \ldots, W_y - 1\}$ be the horizontal and vertical spatial domains, respectively. The positions of the $W_x \times W_y$ pixels contained in the local area are represented by $S = \{(x, \ y)| \ x \in D_x, \ y \in D_y\}$, and $I \ (x_1, \ y_1)$, $I \ (x_2, \ y_2)$ are the gray levels of two pixels at positions $(x_1, \ y_1)$, $(x_2, \ y_2) \in S$. Fig. 2 shows an example of the cooccurrence matrices generated from a region of a ZY-3 nadir-view image. Typically, the GLCMs are normalized by dividing each element by the total number of pixel pairs that satisfy the predefined spatial relationship.

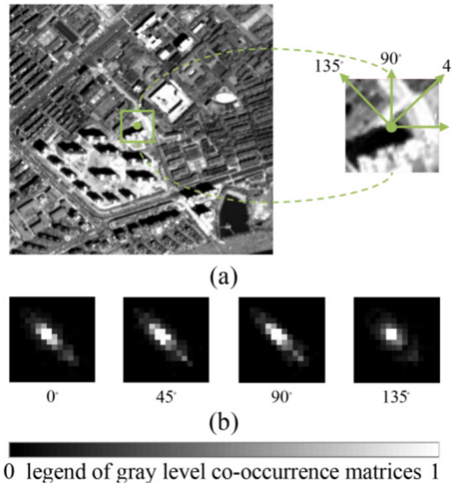

Fig. 2. Example of the traditional GLCM computation. (a) ZY-3 nadir image and the local area within a sliding window. (b) Corresponding GLCM matrices in four directions.

B. MA GLCM Tensor

The ZY-3 imagery provides an opportunity to apply GLCM on multiview images to capture the contextual relationship in 3-D. In particulars, we introduce both inter- and intraangle GLCM. Under different viewing angles, the differences can be apparently observed among multiview images, due to the viewing angles, solar observational cross section [19], the presence of lateral sides [38], surface anisotropy, and the occlusions from other objects and shadows [39]. These factors can all affect the gray-level spatial distributions of multiview imagery and reflect the 3-D structures of different land-cover classes, which are essential to capture the vertical structural information of urban scenes.

The method proposed here aims to describe the spatial dependence of gray values in multiview images on the basis of the GLCM conceptual framework, and hence make full advantage of the MA textural information. To reduce the computational cost, the gray values of the MA images are first quantized to N_g gray levels. A sliding window with the size of $W_x \times W_y \times N_a$ is used to extract the data cube I that contains N_a multiview panchromatic bands and $W_x \times W_y$ spatial pixels. We define $D_x = \{0, 1, \ldots, W_x - 1\}$, $D_y = \{0, 1, \ldots, W_y - 1\}$, and $D_a = \{0, 1, \ldots, N_a - 1\}$ as the spatial and angular domains, and $S = \{(x, \ y, \ a)| \ x \in D_x, \ y \in D_y, \ a \in D_a\}$ denotes the row, column, and angle coordinates of

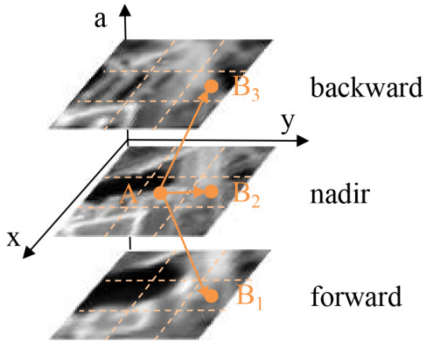

Fig. 3. Example of the spatial relationship between the neighboring pixels in multiview images.

all the pixels in data cube I. In our study, since the ZY-3 multiview images are arranged into a 3-D tensor structure, the position of pixels in the data cube is expressed by $(x, y, a) \in S$, where $a = [0, -1, 1]$ represents the nadir, forward, and backward images, respectively. The vector $\vec{\Delta} = (\Delta_x, \Delta_y, \Delta_a)$ is used to describe the displacement between pixel pairs in the spatial and angular domains, and then the distance r and direction θ between the pixel pairs can be defined as $r = \sqrt{\Delta_x^2 + \Delta_y^2 + \Delta_a^2}$, $\theta = \tan^{-1}\left(\Delta_y/\Delta_x\right)$. Fig. 3 shows an example of the spatial relationship between pixel pairs in a 3-D data cube. The displacement vectors represented by $\vec{\Delta}$ between pixel A and its neighboring pixels B_1, B_2, and B_3 are (0, 1, −1), (0, 1, 0), and (0, 1, 1), respectively. In the figure, it is shown that pixels A and B_2 are located in the nadir image ($\Delta_a = 0$ for B_2), while pixels B_1 and B_3 are located in the forward and backward images, with Δ_a equaling −1 and 1, respectively. Each element (i, j) of the MA GLCM tensor (GLCM$^{\text{MA}-\text{T}}$) records the frequency of the cooccurrence between gray values for pixel pairs separated by a displacement vector $\vec{\Delta}$ within the 3-D data cube I, which can be described as

$$\begin{aligned} P\left(i, j, \vec{\Delta}\right) &= \#\Big\{(x_1, y_1, a_1), (x_2, y_2, a_2) \in S \big| [x_2 - x_1, y_2 \\ &\quad - y_1, a_2 - a_1] = \vec{\Delta}, I(x_1, y_1, a_1) = i, I(x_2, y_2, a_2) = j\Big\} \end{aligned} \tag{2}$$

where # indicates the number of elements, and the gray values of the two neighboring pixels at positions (x_1, y_1, a_1) and (x_2, y_2, a_2) of the data cube are represented by $i, j \in \{1, 2, \ldots, N_g\}$, respectively.

Fig. 4 demonstrates the computation of GLCM$^{\text{MA}-\text{T}}$ using a series of displacement vectors $(x, y) = [(1, 0); (1, 1); (0, 1); (-1, 1)]$, $a = [0, -1, 1]$ and the corresponding directions [0°, 45°, 90°, 135°]. The proposed features are divided into two groups, according to whether the pixel pairs are from the same image, which are hereafter named the intraangle and interangle texture, respectively. These two kinds of textures provide an effective description of both the planar and vertical structures in urban areas. The intraangle textures are calculated based on a mono-view image (i.e., $\Delta_a = 0$). At the same viewing angle, the intraangle textures can capture the different characteristics of the gray-level cooccurrences in four directions (0°, 45°, 90°, and 135°). Moreover, the differences between the intraangle textures among different viewing angles are small for low-lying objects [e.g., a road with no evident variation in its shape and spatial position in the multiview images in Fig. 4(b)] and slightly larger for high-rise objects [e.g., the building shown in Fig. 4(d) and (f)].

On the other hand, the interangle textures, calculated using a combination of multiview images ($\Delta_a \neq 0$), can directly indicate the variation characteristics of the gray-level spatial distributions under different viewing angles that are more closely related to the vertical structures of urban objects. In addition, the differences between the interangle textures obtained from different combinations of MA images are more apparent for high-rise objects than for low-lying ones. Taking the buildings as example, areas F and B contain more pixels of shadows in the forward image than those in the nadir image, while the same area in the backward image is more affected by the pixels of the lateral sides of the buildings. Therefore, it can be observed that the differences of the interangle textures among different multiview image pairs are more apparent for tall buildings than low-lying ones [Fig. 4(e) and (g)]. This can be explained by the fact that the variations in the observational cross sections [19] of low buildings at different viewing angles are less apparent than those of tall buildings. In general, the above comparison shows that the GLCM$^{\text{MA}-\text{T}}$ feature can provide an effective description of the variation characteristics of urban objects from the multiview images and can improve the separability of urban land-cover classes. Please note that if the urban objects have relatively small variations in the positions and gray values among the viewing angles (e.g., low objects), the interangle textures will be similar to the intraangle ones of the nadir image [see Fig. 4(c)].

III. Multispectral and Multiangle 3-D Convolutional Neural Network

The GLCM$^{\text{MA}-\text{T}}$ feature intrinsically constitutes a third-order tensor [Fig. 4(h)] with row, column, and angle modes, which is more likely to preserve the local contextual relationship of the textures obtained from multiview images. The traditional methods for processing tensor features usually involve vectorization, leading to the loss of the spatial context among the multiview images. This problem can be addressed by the tensor-based algorithms that are capable of effectively exploiting the structures and correlations in the MA textures. In this section, a tensorial interpretation framework named M^2-3-DCNN is further proposed to extract the 3-D structural information encoded in the GLCM$^{\text{MA}-\text{T}}$ feature and to learn high-level features in the spectral and spatial-angular domains jointly through a two-stream structure.

A. Multiview 3-D-CNN Filter

CNNs are powerful tools for feature learning and classification that take the spatial correlation into consideration. Nevertheless, when the input data inherently have tensor structures, for example, MA images, it is desirable to exploit the spatial correlations within and between the MA images.

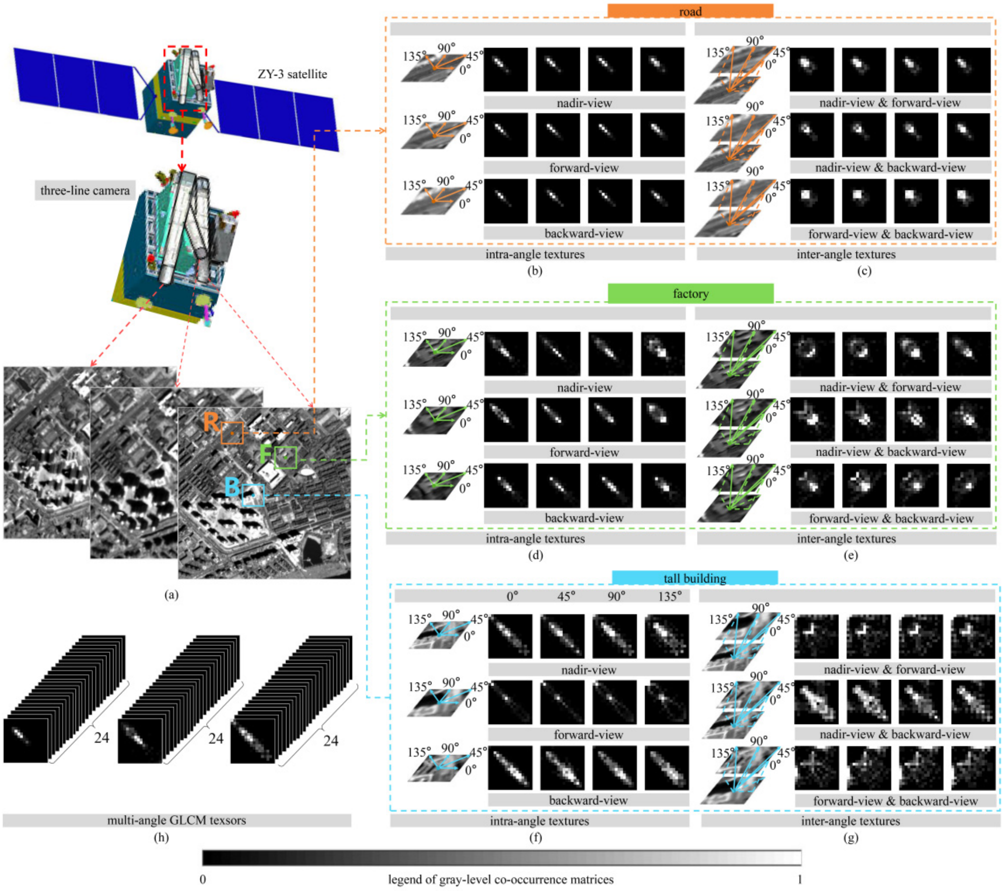

Fig. 4. ZY-3 MA images (a) and the derived GLCM$^{\text{MA}-\text{T}}$ features calculated from three areas belonging to the road [(b) and (c) for intraangle and interangle textures, respectively], factory [(d) and (e) for interangle and interangle textures, respectively], and tall building [(f) and (g) for interangle and interangle textures, respectively]. The corresponding MA GLCM tensors are shown in (h).

Compared with the traditional 1-D or 2-D networks, the convolution operations and the weights in a 3-D-CNN model [28] are 3-D, to maintain the intrinsic data structures. With the third-order tensor data as input, the 3-D kernels are applied to the convolution stage to generate the 3-D feature maps. We define i, j, and a as the positions of the elements in the 3-D feature map, where i and j are the coordinates of the planar spatial domain, and a represents the position in the angular domain. Given the element y_{lm}^{ija} at the position (i, j, a) of the feature map m in the lth layer, the element of the feature map n in the following $(l+1)$th layer can be obtained by performing the convolution operation with kernel $w \in R^{W\times W\times W}$, which is denoted as

$$y_{(l+1)n}^{ija} = F\left(b_{(l+1)n} + \sum_{m}\sum_{w_i,w_j,w_a} \omega_{(l+1)mn}^{w_i w_j w_a} y_{lm}^{(i+w_i)(j+w_j)(a+w_a)}\right) \tag{3}$$

where the size of the convolutional kernel is $W \times W \times W$. $\omega_{(l+1)mn}^{w_i w_j w_a}$ represents the (w_i, w_j, w_a)th value of the kernel in the $(l+1)$th layer between the input feature map m and the output feature map n, and $b_{(l+1)n}$ is the corresponding bias tensor. Unlike the 2-D convolution operators, the 3-D ones can reveal the correlation and difference of images from the adjacent viewing angles.

B. Proposed M^2-3-DCNN Framework

Diverse urban objects with similar spectral properties and complex vertical structures make it a challenging task to classify high-resolution images of urban areas, especially for images containing man-made architectures. Hence, it is essential to jointly consider the spectral, spatial, and angular information contained in the multiview images to improve the separability of urban objects with complex 3-D structures. The ZY-3 multiview satellite simultaneously collects nadir, forward, and backward panchromatic bands, as well as MS bands. To make better use of the ZY-3 imagery for urban classification, we propose the M^2-3-DCNN framework. In the two-stream structure, the 3-D-CNN model was used as the basic unit to interpret the MS data cubes and GLCM$^{\text{MA}-\text{T}}$ tensor textures, respectively, and both features have a 3-D structure. As shown in Fig. 5, an MS stream is designed to consider the spectral tensor features. Meanwhile, the proposed GLCM$^{\text{MA}-\text{T}}$ feature is used as input of the MA stream, which describes both the intraangle and interangle textures. By regarding the input data as third-order tensors, the network preserves the spectral–spatial–angular structure inherent to the features. The fusion of the final learned high-level features, which involves stacking the outputs of the MS and MA streams, can effectively describe the urban area in a joint spectral–spatial–angular manner. These MS and MA samples

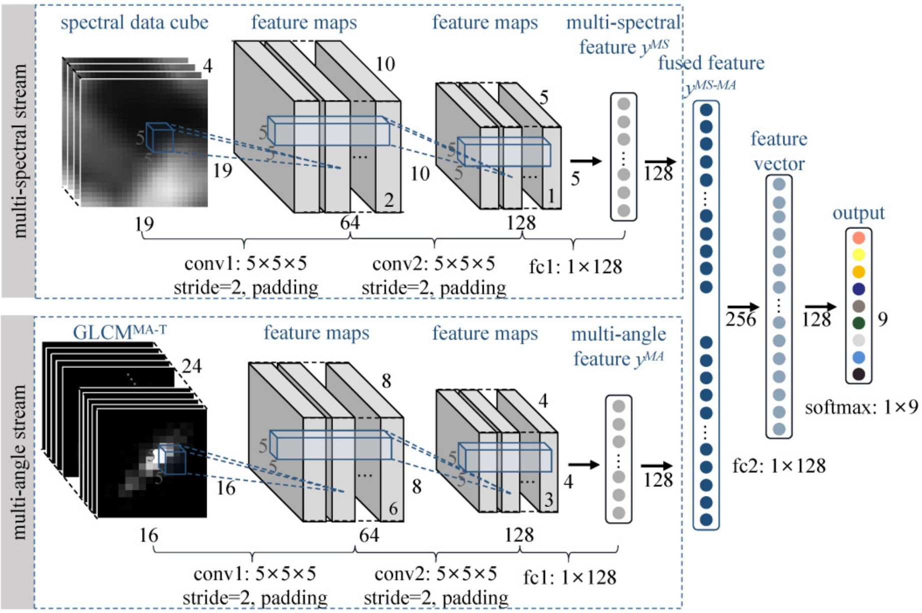

Fig. 5. Proposed M^2-3-DCNN framework.

are fed into the two 3-D-CNN streams, each of which contains two 3-D convolutional layers and one fully connected layer. The kernel size of the convolutional layers is empirically set as $5 \times 5 \times 5$ with a stride of 2, and the zero-padding strategy [40] is also applied. The number of kernels in the two convolutional layers and the following fully connected layer is 64, 128, and 128, respectively. The feature maps that pass through the two convolutional layers and one fully connected layer are then concatenated. It should be noted that the outputs of the fully connected layers in the MS and MA streams are the feature vectors denoted as y^{MS} and y^{MA}, whose lengths are N^{MS} and N^{MA}, respectively. These two feature vectors are concatenated as a whole feature vector $y^{\mathrm{MS-MA}} = [y^{\mathrm{MS}}, y^{\mathrm{MA}}]$ with the length of $(N^{\mathrm{MS}}, N^{\mathrm{MA}})$. $y^{\mathrm{MS-MA}}$ is used as the input to the following fully connected layer, which can generate an appropriate description of the implicit correlations between the two streams from the fused feature vectors. The output feature vector of this fully connected layer is

$$y_{(l+1)n} = F\left(b_{(l+1)n} + \sum_{m=1}^{N^{\mathrm{MS}}+N^{\mathrm{MA}}} \omega_{(l+1)mn}\, y_{lm}^{\mathrm{MS-MA}}\right) \tag{4}$$

where $w_{(l+1)mn}$ and $b_{(l+1)n}$ denote the weight and bias vector, respectively, which can be regarded as descriptors of the correspondence between the MS and MA streams. $F(\cdot)$ is the rectified linear unit (ReLU) activation function. The output of the fully connected layer (i.e., $y_{(l+1)n}$) has the potential to capture the complementary information from the MS and MA streams. Finally, a softmax layer is used to predict the class label for each sample.

IV. Experiments

A. Data Sets and Study Areas

The images used in this study were acquired by the ZY-3 satellite, composing of the nadir, forward, and backward panchromatic images and an MS image (red, green, blue, near-infrared) in the nadir view. The spatial resolution of ZY-3 is 2.1 m for the nadir panchromatic *m* for the MS image with four bands and 3.5 m for the forward- and backward-view images. As the first step of preprocessing, DSMs were acquired from the stereo pairs by the semiglobal matching (SGM) approach [41], and all the ZY-3 images were orthorectified with the digital elevation models (DEMs) derived from the DSMs. The calculation of GLCM$^{\mathrm{MA-T}}$ is related to the spatial position of pixels. Without orthorectification, it is difficult to determine whether the variations in gray-level spatial distributions are caused by topography or different viewing angles, especially in areas with complex terrain. The forward and backward images were registered to the nadir image using a polynomial transformation, with a root-mean-square value of less than 0.5 pixels for each study area and were resampled to the spatial resolution of 2.1 m, so that their coordinates can be aligned. This step is essential for the calculation of GLCM$^{\mathrm{MA-T}}$ and fusion of spectral–spatial–angular information. Subsequently, the histogram matching method was then used to perform the relative radiometric calibration between the two off-nadir images and the nadir image. This step is able to eliminate the radiation difference between multiview panchromatic images and ensure that the variations in gray levels are mainly derived from different viewing angles. With the aim of improving the spatial resolution of the MS images, Gram–Schmidt pan-sharpening [42] was conducted to fuse the MS images with the nadir panchromatic image.

The study areas were chosen from four representative cities of China, that is, Wuhan, Hefei, Shanghai, and Xi'an. It should be noted that these cities have different degrees of urban development and scene complexities, which allowed us to investigate the effectiveness and robustness of the proposed method in urban area classification. The test images were acquired at different times, with sizes of 2971 × 3612,

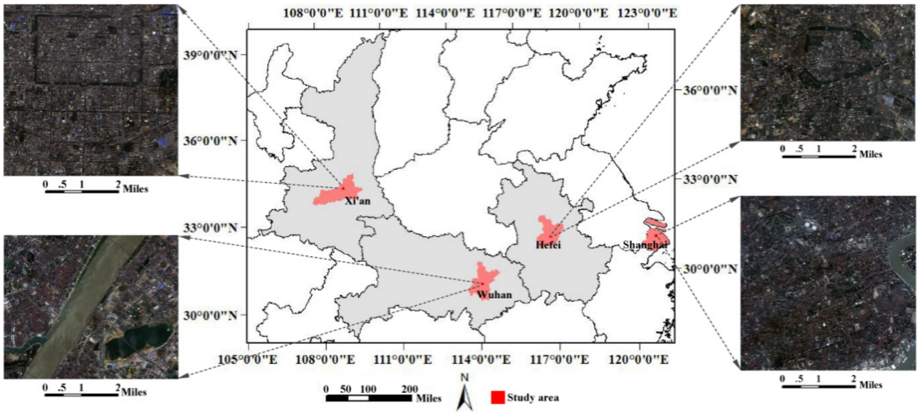
108°0'0"E 111°0'0"E 114°0'0"E 117°0'0"E 120°0'0"E 123°0'0"E
39°0'0"N
36°0'0"N
33°0'0"N
30°0'0"N
105°0'0"E 108°0'0"E 111°0'0"E 114°0'0"E 117°0'0"E 120°0'0"E
Xi'an
Hefei
Shanghai
Wuhan
0 .5 1 2 Miles
0 50 100 200 Miles
N
Study area

Fig. 6. Study areas and the corresponding RGB composites of the ZY-3 images of Wuhan, Hefei, Shanghai, and Xi'an.

TABLE I

NUMBERS OF TRAINING (TR) AND TEST (TT) SAMPLES PER CLASS FOR THE WUHAN, HEFEI, SHANGHAI, AND XI'AN DATA SETS

	Classes	Numbers of samples							
		WH		HF		SH		XA	
		TR	TT	TR	TT	TR	TT	TR	TT
🟧	Low-rise residential buildings (LRB)	121	136	49	55	109	118	237	248
🟧	Middle-rise residential buildings (MRB)	103	112	401	538	221	221	287	417
🟨	High-rise residential buildings (HRB)	84	92	150	193	600	612	185	223
🟦	Industrial buildings (IB)	67	106	61	109	169	194	130	181
⬛	Road	40	77	26	87	56	152	46	150
🟩	Vegetation (Veg.)	57	215	53	189	98	140	54	147
⬜	Bare soil (B. S.)	43	116	23	43	39	54	35	67
🟦	Water	84	213	31	74	120	213	111	104
⬛	Shadows (Sha.)	117	119	224	233	262	270	117	121
	Total	716	1,186	1,018	1,521	1,674	1,974	1,202	1,658

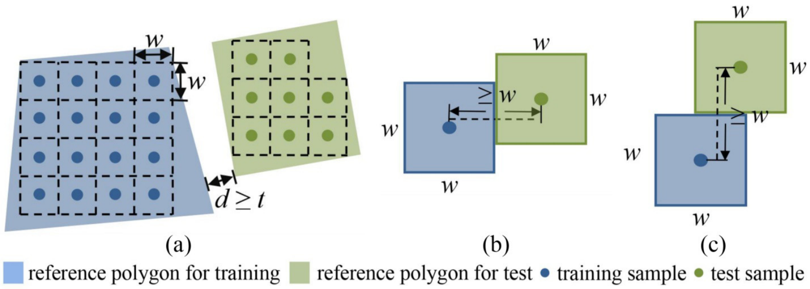

Fig. 7. (a) Example of the training and test sample selection. Constraint of distance between them in the (b) horizontal direction and the (c) vertical direction. (d is the minimum distance between two independent polygons and w represents the window size.)

2835 × 3775, 3559 × 3559, and 3367 × 3429, respectively, as shown in Fig. 6.

In this study, nine urban land-cover classes were considered (see Table I). The reference polygons for the four study areas were first manually delineated through visual inspection of the ZY-3 nadir images and Google Earth high-resolution images. These polygons were composed of pixels belonging to a certain class. To eliminate the spatial autocorrelation, for the reference polygons of a certain class, the semivariance analysis [43] was applied to quantitatively measure their autocorrelation level and determine the distance threshold t. The minimum distance d between two independent polygons should be larger than t and only the polygons satisfying this constraint were preserved. Stratified random sampling [44] was then adopted to select 50% of the polygons for training and the other 50% for testing. Subsequently, as shown in Fig. 7, each polygon was divided into a set of square windows, and the size of each window is the same as the one of the spatial window used for extracting GLCM$^{\mathrm{MA-T}}$. The training and test samples were collected by selecting the central pixel of each block (see Fig. 7). The numbers of training and test samples for the four data sets are provided in Table I.

B. Experimental Setting

1) Parameter Setting of GLCM$^{\mathrm{MA-T}}$: The proposed GLCM$^{\mathrm{MA-T}}$ feature was generated from the multiview images, with the window size of $W_x = W_y = 19$, and the displacement vectors were described by the interpixel distance $r = 1$ (for a detailed discussion of the window size and distance value, see Section V-E) and directions $\theta = [0°, 45°, 90°, 135°]$. These parameters for GLCM$^{\mathrm{MA-T}}$ calculation were determined according to the spatial resolution of the

TABLE II
DESCRIPTIONS OF THE COMPARABLE METHODS AND THE PROPOSED METHOD

Abbreviation	Description	Multi-view or not	Deep learning or hand-crafted	Spectral-spatial-angular or not	Manner of fusion
RF$_{S+nDSM}$	Uses the combination of spectral feature and nDSM as input to the random forest (RF) [45] classifier, as nDSM describes the height information extracted from the multi-view images (MV).	nDSM	Hand-crafted	MS + nDSM	Data-level
RF$_{S+ADF}$	Feeds the combination of the spectral feature and ADF [23] into RF. ADF describes the angular differences at the pixel, feature, and label levels for urban classification.	ADF	Hand-crafted	MS + ADF	Data-level
AMDF-ResNet$_S$	AMDF-ResNet [46] adaptively fuses contextual features at adjacent scales (i.e., adjacent residual blocks). By using three residual blocks, it captures urban land covers at varied scales. As a comparison, we took the MS images (i.e., AMDF-ResNet_F$_S$) and the stacked MS and MV images (i.e., AMDF-ResNet_F$_{S+NFB}$) as input.	-	DL	-	-
AMDF-ResNet$_{S+NFB}$		MV	DL	MS + MV	Data-level
ResUNet-a$_S$	Using U-Net [47] as the architecture, ResUNet-a [48] integrates recent techniques, including residual blocks [49], atrous convolutions [50], pyramid scene parsing pooling [51], and a dice loss-like function, and it models the HRRS image classification as semantic segmentation. Here, two versions of ResUNet-a were analyzed.	-	DL	-	-
ResUNet-a$_{S+NFB}$		MV	DL	MS + MV	Data-level
M^2-3DCNN$_{S+MA}$	Employs two 3D-CNN streams to interpret the spectral feature (MS images) and GLCM$^{MA\text{-}T}$ feature (MA) respectively, and deeply fuses these two features by fully connected layers.	GLCM$^{MA\text{-}T}$	DL + Hand-crafted	MS + MA	Feature-level

ZY-3 images and the sizes of the objects in the study areas. The original images were quantized to 16 gray levels by linear transformation, as suggested in [4]. Hence, the GLCM matrices calculated using the multiview images from the four directions formed a three-mode tensor feature with a size of 16 × 16 × 24, including both the intraangle and interangle textures.

2) Hyperparameter Setting for the Proposed Neural Network: First, to alleviate the over-fitting problem, the dropout strategy [52] was applied to the second convolutional layer in each stream, with the dropout rate set to 0.5, according to [52]. All the networks were run on a desktop computer with TensorFlow-GPU-1.13.1, an Intel Core i9-9900X CPU (at 3.50 GHz), 128 GB RAM, and an 11 GB GeForce RTX 2080 Ti GPU. For the training of M^2-3-DCNN, the learning rate was initially set to 0.001 and was reduced by 1/e (where e is the Euler number) when the loss value did not decrease in two consecutive iterations. The Adam optimization method [53] was adopted to optimize the cross-entropy loss function, and the batch size was 64. The hyperparameters of the deep-learning-based methods (see Table II) used in the comparison were set according to the original articles. It should also be noted that a lot of labeled samples are required in the training procedure of a deep network. However, the availability of spatially independent and well-annotated training samples is limited due to the high cost of manual labeling [54], [55]. To solve this dilemma, data augmentation [56] was used to increase the amount of training data. The new samples were created by applying deformations to the annotated samples, which is an approach that requires little additional computation and does not alter the original labels. Specifically, a series of image transformations were applied in the spatial dimension for the data augmentation, that is, rotation, horizontal and vertical flipping, and copying [57], so that the network can be more robust to spatial deformations. The training samples were empirically augmented to 25 000 per category.

3) Comparable Methods: In the experiments, a number of state-of-the-art algorithms were taken as benchmarks, as listed in Table II. The compared methods included two novel hand-crafted multiview features and two of the most recent CNN-based algorithms designed for HRRS image classification.

C. Experimental Results

The results of all the experiments for the four data sets are presented in Tables III–VI, respectively, where the numbers in bold represent the highest accuracy for each class. The corresponding classification maps are shown in Figs. 8–11, and the zoomed-in classification maps are provided in Fig. 12. In addition to the overall and class-specific accuracies, McNemar's test [58] is also used to indicate whether the differences between the classification results of the two methods are statistically significant (Table VII). The following conclusions can be drawn:

1) The classification performance using the three features extracted from the multiview images, that is, nDSM, ADF, and GLCM^{MA-T}, is first analyzed. It can be seen that M2-3-DCNN$_{S+MA}$ gives better results than

TABLE III

CLASSIFICATION RESULTS FOR THE WUHAN DATA SET

Class	$\text{RF}_{\text{S+nDSM}}$	$\text{RF}_{\text{S+ADF}}$	AMDF-$\text{ResNets}_{\text{S}}$	AMDF-$\text{ResNets}_{\text{S+NFB}}$	ResUNet-a_{S}	ResUNet-$\text{a}_{\text{S+NFB}}$	M^2-$\text{3DCNNs}_{\text{S+MA}}$
LRB	0.721	0.816	0.912	0.912	0.919	**0.926**	**0.926**
MRB	0.661	0.491	0.821	0.893	0.723	0.902	**0.938**
HRB	0.261	0.674	0.761	0.783	0.739	0.739	**0.804**
IB	0.604	0.528	0.670	0.660	0.623	0.660	**0.679**
Roads	0.494	0.364	0.831	0.766	0.779	0.649	**0.883**
Veg.	0.916	0.926	**0.977**	**0.977**	0.949	0.926	0.972
B. S.	0.612	0.621	0.716	0.802	0.819	**0.974**	0.776
Water	0.967	0.986	0.962	0.981	0.981	**0.995**	**0.995**
Sha.	0.975	**0.983**	0.950	0.941	0.975	0.882	0.958
OA	0.749	0.767	0.870	0.885	0.864	0.880	0.902

TABLE IV

CLASSIFICATION RESULTS FOR THE HEFEI DATA SET

Class	$\text{RF}_{\text{S+nDSM}}$	$\text{RF}_{\text{S+ADF}}$	AMDF-$\text{ResNets}_{\text{S}}$	AMDF-$\text{ResNets}_{\text{S+NFB}}$	ResUNet-a_{S}	ResUNet-$\text{a}_{\text{S+NFB}}$	M^2-$\text{3DCNNs}_{\text{S+MA}}$
LRB	0.145	0.073	0.509	0.655	0.473	0.655	**0.691**
MRB	0.879	0.954	0.965	0.959	0.844	0.909	**0.968**
HRB	0.373	0.751	0.772	0.870	0.819	0.891	**0.912**
IB	0.523	0.569	0.752	0.752	0.771	0.758	**0.789**
Roads	0.402	0.644	0.862	0.851	0.920	0.931	**0.943**
Veg.	0.963	0.873	0.968	0.942	**0.989**	0.947	0.979
B. S.	0.558	0.535	0.860	0.837	0.930	0.930	**0.953**
Water	0.757	0.662	0.824	0.878	**0.959**	0.838	0.919
Sha.	0.970	0.966	0.974	**0.979**	0.966	0.953	**0.979**
OA	0.745	0.817	0.894	0.909	0.873	0.896	0.936

TABLE V

CLASSIFICATION RESULTS FOR THE SHANGHAI DATA SET

Class	$\text{RF}_{\text{S+nDSM}}$	$\text{RF}_{\text{S+ADF}}$	AMDF-$\text{ResNets}_{\text{S}}$	AMDF-$\text{ResNets}_{\text{S+NFB}}$	ResUNet-a_{S}	ResUNet-$\text{a}_{\text{S+NFB}}$	M^2-$\text{3DCNNs}_{\text{S+MA}}$
LRB	0.254	0.356	0.669	0.771	0.712	0.822	**0.873**
MRB	0.683	0.864	0.900	0.900	0.905	0.905	**0.932**
HRB	0.861	0.933	0.933	0.944	0.889	0.954	**0.956**
IB	0.763	0.820	0.861	0.881	0.835	0.845	**0.912**
Roads	0.138	0.467	0.724	0.809	0.829	0.803	**0.842**
Veg.	**0.993**	0.907	**0.993**	0.936	0.964	0.936	0.964
B. S.	0.426	0.519	0.833	0.852	0.852	**0.926**	0.889
Water	0.972	0.972	**0.995**	0.972	0.995	0.986	0.991
Sha.	0.974	**0.978**	**0.978**	0.970	0.981	0.937	0.974
OA	0.764	0.841	0.905	0.915	0.899	0.917	0.940

TABLE VI

CLASSIFICATION RESULTS FOR THE XI'AN DATA SET

Class	$\text{RF}_{\text{S+nDSM}}$	$\text{RF}_{\text{S+ADF}}$	AMDF-$\text{ResNets}_{\text{S}}$	AMDF-$\text{ResNets}_{\text{S+NFB}}$	ResUNet-a_{S}	ResUNet-$\text{a}_{\text{S+NFB}}$	M^2-$\text{3DCNNs}_{\text{S+MA}}$
LRB	0.649	0.718	0.770	0.802	0.794	0.802	**0.806**
MRB	0.875	0.890	0.921	0.923	0.859	0.856	**0.952**
HRB	0.498	0.852	0.852	0.861	0.870	**0.892**	**0.892**
IB	0.674	0.630	0.851	0.851	0.867	**0.879**	0.878
Roads	0.120	0.120	0.613	0.760	0.720	0.713	**0.853**
Veg.	0.966	0.939	**0.973**	0.891	0.850	0.966	**0.973**
B. S.	0.567	0.343	0.701	0.642	0.776	0.776	**0.821**
Water	0.913	0.923	0.962	0.942	**0.990**	**0.990**	0.971
Sha.	0.760	0.893	0.967	**0.975**	0.950	0.926	0.934
OA	0.690	0.746	0.855	0.865	0.854	0.864	0.902

$\text{RF}_{\text{S+nDSM}}$ and $\text{RF}_{\text{S+ADF}}$, with significant increments of about 18.3% and 12.7% in OA, respectively. The reason why nDSM is inferior to the proposed method is that in some situations, the height information affected by mismatched pixels is not accurate enough to distinguish urban objects with complex vertical structures [23], as in the confusion between residential buildings with different heights and factories in Fig. 12 [b2 (as shown in the red rectangles)]. In contrast to nDSM, $\text{GLCM}^{\text{MA}-\text{T}}$ can more adequately exploit the implicit angular information and obtain a superior classification performance (e.g., the 38.2% increase in the accuracy for low-rise residential buildings, denoted as LRB for convenience in Tables III–VII). Using another feature extracted from multiview images, $\text{RF}_{\text{S+ADF}}$, gives significantly better results than $\text{RF}_{\text{S+nDSM}}$, with an average increment of 5.6% in OA. However, $\text{RF}_{\text{S+ADF}}$ is still inferior to the proposed work. The OA improvements achieved by M2-3-$\text{DCNN}_{\text{S+MA}}$ over $\text{RF}_{\text{S+ADF}}$ are partly due to the fact that the tensor-based

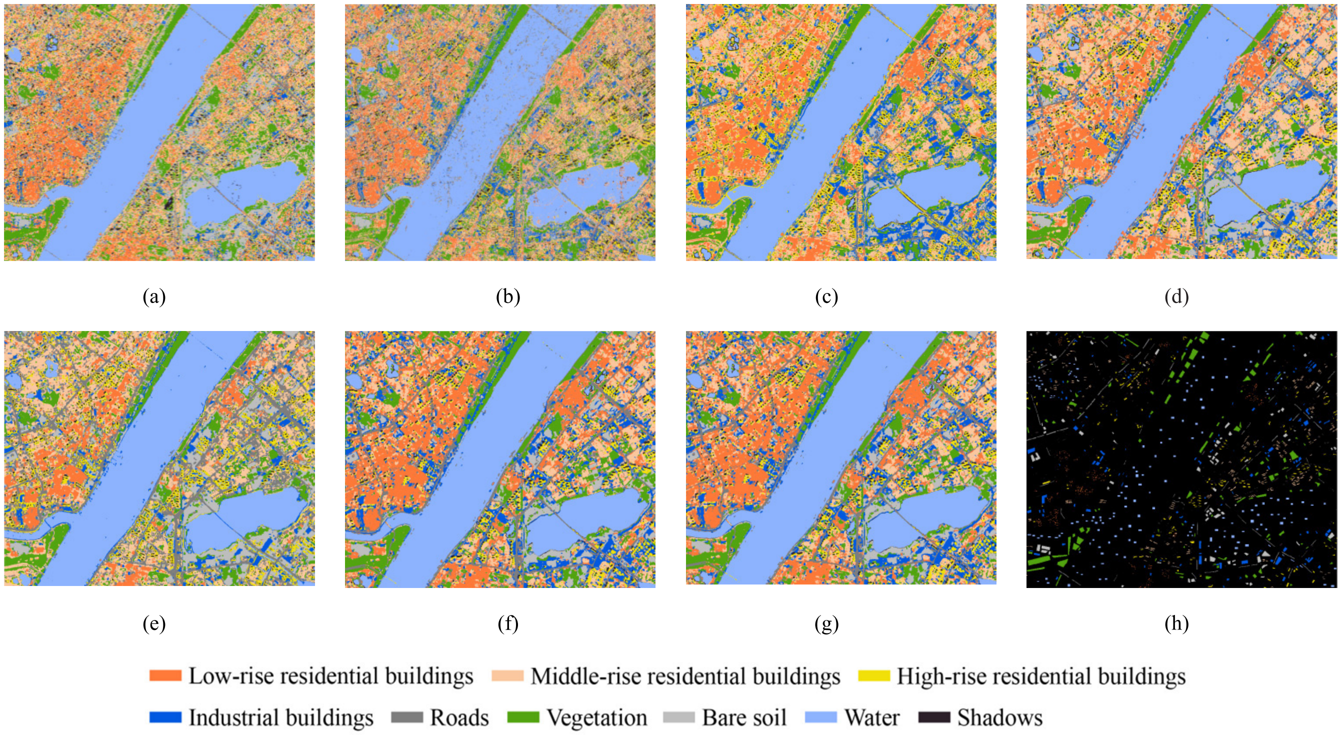
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Low-rise residential buildings
Middle-rise residential buildings
High-rise residential buildings
Industrial buildings
Roads
Vegetation
Bare soil
Water
Shadows

Fig. 8. Classification maps of (a) $\text{RF}_{\text{S+nDSM}}$, (b) $\text{RF}_{\text{S+ADF}}$, (c) AMDF-$\text{ResNets}_{\text{S}}$, (d) AMDF-$\text{ResNets}_{\text{S+NFB}}$, (e) ResUNet-$\text{a}_{\text{S}}$, (f) ResUNet-$\text{a}_{\text{S+NFB}}$, and (g) M^2-3-$\text{DCNN}_{\text{S+MA}}$ for the Wuhan data set.

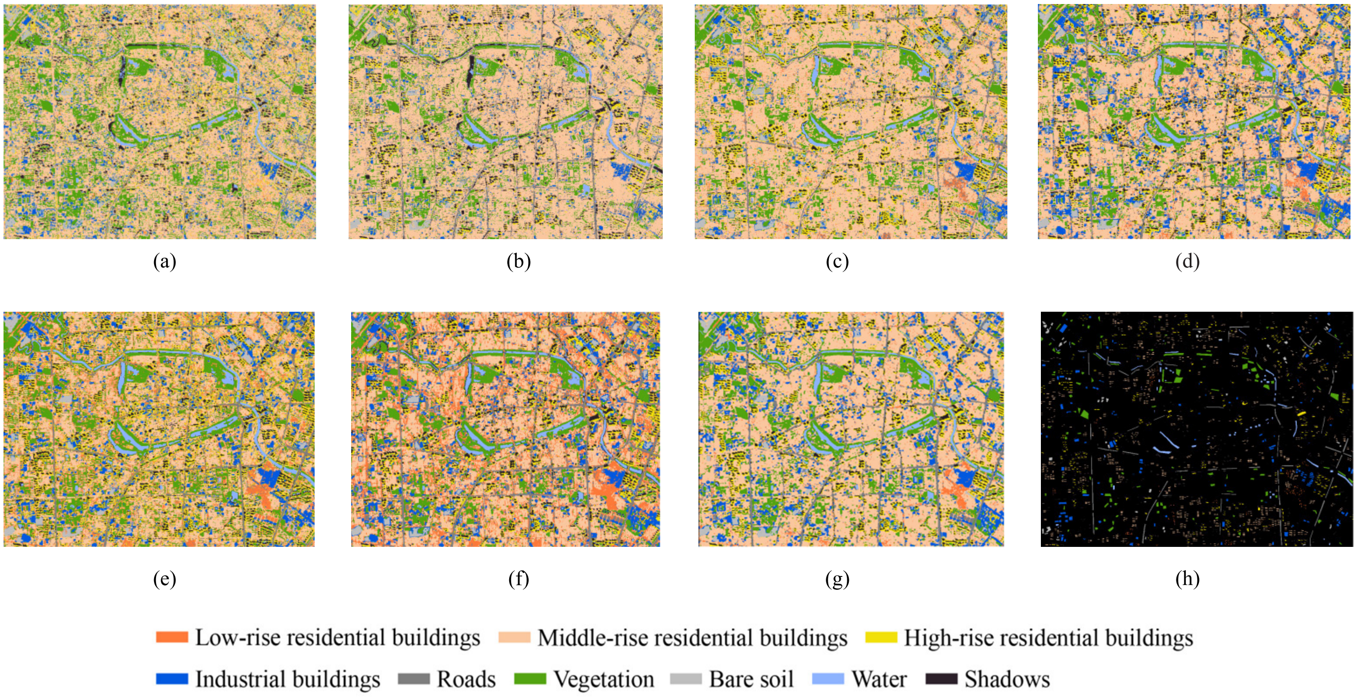
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Low-rise residential buildings
Middle-rise residential buildings
High-rise residential buildings
Industrial buildings
Roads
Vegetation
Bare soil
Water
Shadows

Fig. 9. Classification maps of (a) $\text{RF}_{\text{S+nDSM}}$, (b) $\text{RF}_{\text{S+ADF}}$, (c) AMDF-$\text{ResNets}_{\text{S}}$, (d) AMDF-$\text{ResNets}_{\text{S+NFB}}$, (e) ResUNet-$\text{a}_{\text{S}}$, (f) ResUNet-$\text{a}_{\text{S+NFB}}$, (g) M^2-3-$\text{DCNN}_{\text{S+MA}}$, and (h) ground-truth reference for the Hefei data set.

$\text{GLCM}^{\text{MA-T}}$ feature captures both the interangle and intraangle information (i.e., the angular–spatial information), but the vector-based ADF feature may result in the loss of spatial structures and incomplete utilization of the information in the multiview images. Taking the low-rise residential buildings and roads as examples, the accuracy improvements achieved by M2-3-$\text{DCNN}_{\text{S+MA}}$ are 33.3% and 48.2% on average for the four data sets, respectively, compared with $\text{RF}_{\text{S+ADF}}$. The misclassifications of these two categories using $\text{RF}_{\text{S+ADF}}$ are marked by the red rectangles in Fig. 12 (a3). In short, the proposed method which

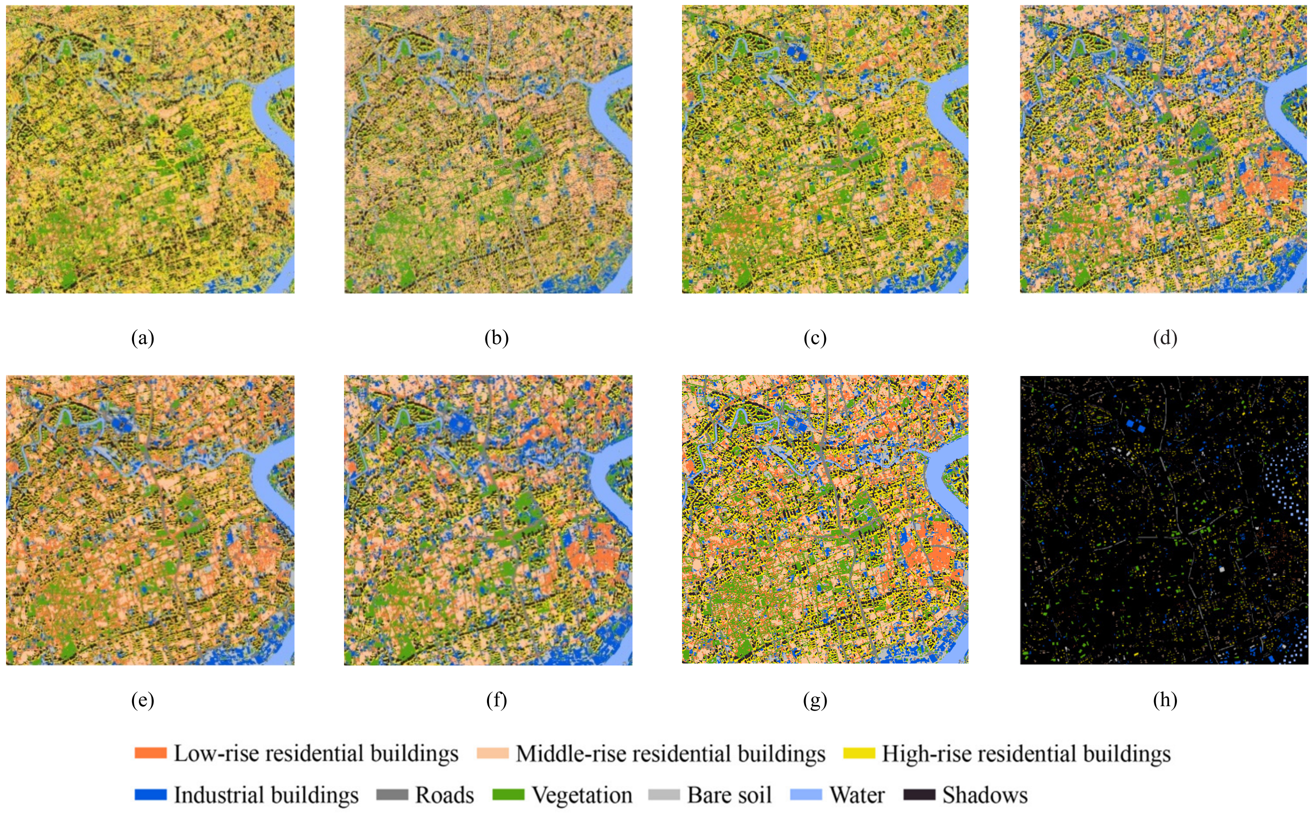
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Low-rise residential buildings
Middle-rise residential buildings
High-rise residential buildings
Industrial buildings
Roads
Vegetation
Bare soil
Water
Shadows

Fig. 10. Classification maps of (a) RF$_{S+nDSM}$, (b) RF$_{S+ADF}$, (c) AMDF-ResNet$_S$, (d) AMDF-ResNet$_{S+NFB}$, (e) ResUNet-a$_S$, (f) ResUNet-a$_{S+NFB}$, (g) M^2-3-DCNN$_{S+MA}$, and (h) ground-truth reference for the Shanghai data set.

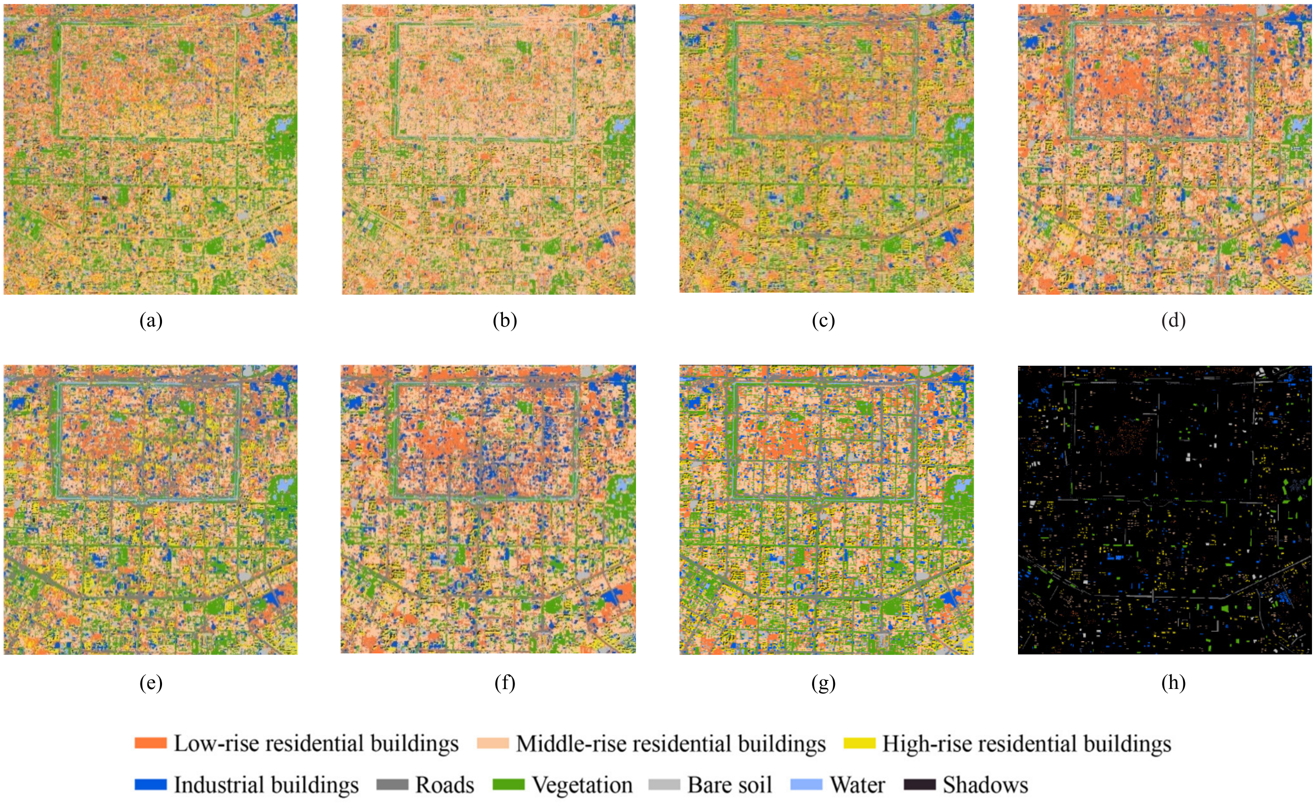
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Low-rise residential buildings
Middle-rise residential buildings
High-rise residential buildings
Industrial buildings
Roads
Vegetation
Bare soil
Water
Shadows

Fig. 11. Classification maps of (a) RF$_{S+nDSM}$, (b) RF$_{S+ADF}$, (c) AMDF-ResNet$_S$, (d) AMDF-ResNet$_{S+NFB}$, (e) ResUNet-a$_S$, (f) ResUNet-a$_{S+NFB}$, (g) M^2-3-DCNN$_{S+MA}$, and (h) ground-truth reference for the Xi'an data set.

jointly uses spectral–spatial–angular information is superior to the recently developed hand-crafted spectral–angular feature extraction methods, from both the perspective of ground and above-ground objects.

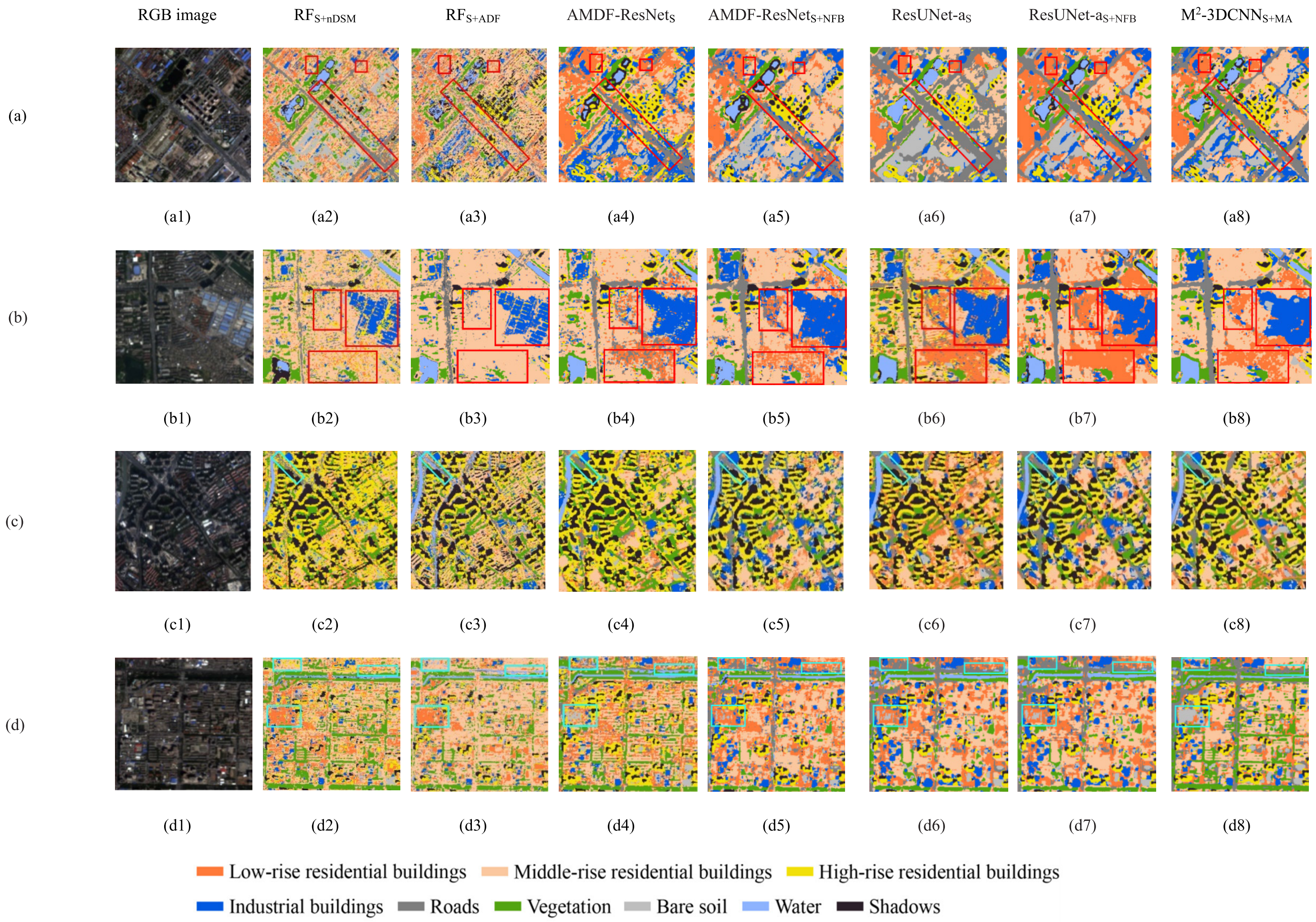

Fig. 12. Zoomed-in classification maps for (a) Wuhan, (b) Hefei, (c) Shanghai, and (d) Xi'an data sets.

TABLE VII

McNemar's Test Between M2-3-DCNNS+MA and the Compared Methods. The Significantly Different Methods Are Indicated as ** With $\gamma > 3.84$ at a 95% Confidence Level and * for $\gamma > 2.71$ at a 90% Level

	M^2-3DCNN$_{S+MA}$ vs.					
datasets	RF$_{S+nDSM}$	RF$_{S+ADF}$	AMDF-ResNet$_S$	AMDF-ResNet$_{S+NFB}$	ResUNet-a$_S$	ResUNet-a$_{S+NFB}$
Wuhan	149.512**	145.312**	11.221**	4.889**	15.488**	5.040**
Hefei	244.803**	171.235**	26.912**	9.752**	57.962**	12.545**
Shanghai	267.384**	216.790**	19.929**	7.955**	28.298**	7.358**
Xi'an	251.540**	235.253**	17.552**	14.954**	25.245**	17.326**

2) Two recently proposed deep learning methods that learn the spectral–spatial structure of the land covers for HRRS image classification were also tested in this study. For a fair comparison, inputs with the spectral feature alone and a stack of spectral and multiview bands were considered. By importing the spatial features learned by the deep networks, it can be observed that the accuracies of AMDF-ResNet and ResUNet-a are higher than those of the hand-crafted feature methods described earlier (Tables III–VI), especially for roads (with a 27.6%–60.0% increment in accuracy), as shown by the cyan rectangles in Fig. 12 (c2–4 and c6). In addition, for both the AMDF-ResNet and ResUNet-a architectures, the addition of multiview features can improve the classification accuracy slightly, which therefore encourages us to further exploit MA information. Compared with AMDF-ResNet$_{S+NFB}$ and ResUNet-a$_{S+NFB}$, the proposed M2-3-DCNN$_{S+MA}$ shows a significantly better classification performance [see Table VII and the cyan rectangles in Fig. 12 (d4, d7, and d8)], which demonstrates the advantage of deep fusion of the spectral–spatial–angular information. In summary, it can be said that the proposed method is significantly superior to the recent state-of-the-art methods. Detailed descriptions and analysis of each component in M2-3-DCNN$_{S+MA}$ are provided in Section V.

V. Discussion

In the following, the three components of the proposed method are discussed, that is, the excavation of spatial–angular information from multiview images, the fusion of the

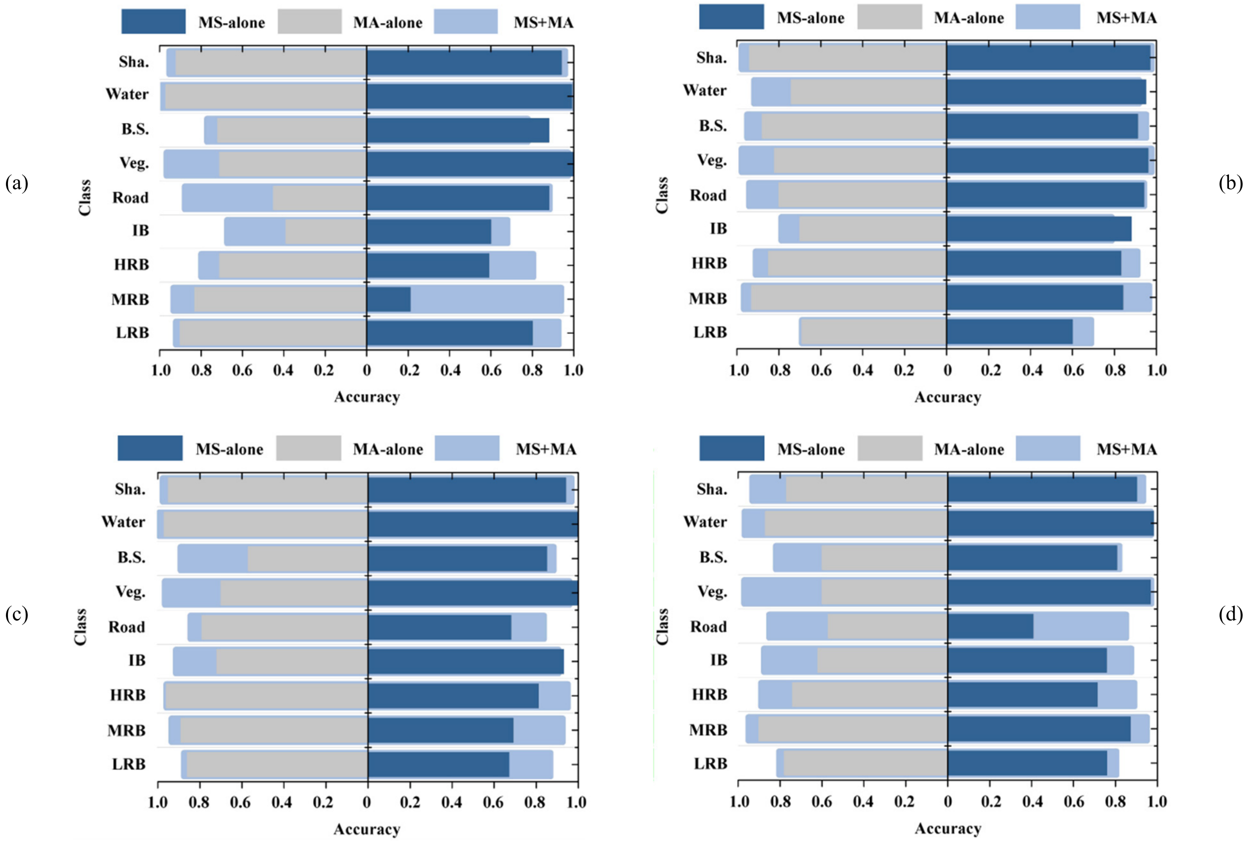

Fig. 13. Accuracy of each category when using the MS or MA stream alone and when using both these streams for (a) Wuhan, (b) Hefei, (c) Shanghai, and (d) Xi'an data sets.

multisource information, and the effect of the 3-D convolution. The robustness of the parameter of GLCM$^{\mathrm{MA-T}}$ is also analyzed.

A. Comparison Between GLCM^{MA-T} and GLCM

To demonstrate the advantage of the MA-image-derived planar and vertical textures, the traditional GLCM [4] calculated from four directions of the nadir image was compared. The parameters for generating the traditional GLCM, including the window size and displacement vector, were the same as those for GLCM$^{\mathrm{MA-T}}$. For a fair comparison, the design of M^2-3-DCNN$_{\mathrm{S+N}}$ was the same as that of M^2-3-DCNN$_{\mathrm{S+MA}}$, except that GLCM$^{\mathrm{MA-T}}$ was replaced by the traditional GLCM. As seen in Table VIII, the performance differences can be mainly attributed to the better classification of GLCM$^{\mathrm{MA-T}}$ on the above-ground objects (with accuracy increases of 5.5%, 9.6%, 7.0%, and 2.8%, on average, for low-rise residential buildings, middle-rise residential buildings, high-rise residential buildings, and industrial buildings (IBs), respectively).

TABLE VIII

ACCURACY INCREMENTS ACHIEVED BY M^2-3-DCNN$_{\mathrm{S+MA}}$ OVER M^2-3-DCNN$_{\mathrm{S+N}}$

Class	Wuhan	Hefei	Shanghai	Xi'an
LRB	+0.044	+0.018	+0.110	+0.048
MRB	+0.099	+0.119	+0.127	+0.039
HRB	+0.043	+0.036	+0.116	+0.086
IB	+0.056	+0.009	+0.031	+0.015
Roads	+0.000	−0.011	+0.026	+0.033
Veg.	0.000	0.000	+0.014	+0.007
B. S.	−0.060	0.000	0.000	+0.015
Water	+0.000	−0.013	+0.010	+0.009
Sha.	+0.017	+0.009	+0.004	+0.015
OA	+0.019	+0.047	+0.065	+0.036

B. Effect of Fusing MS and MA Information

The proposed M^2-3-DCNN framework simultaneously extracts the spectral and spatial–angular information with a two-stream architecture. To demonstrate the function of this architecture, three benchmark methods were designed. The first two used only one stream with either the MS or MA feature as input, and the third method stacked the MS and MA features as input (i.e., a data-fusion approach) and fed this into one stream. In Fig. 13, it can be clearly seen that the MA stream performs better in classifying the buildings with different heights, while the MS stream is better at identifying the natural land covers (e.g., water, soil, vegetation) that have apparent spectral characteristics. In particular, an interesting example is the IBs, which are better classified by the MS stream. This phenomenon can be attributed to two factors: 1) the IBs usually have a low height, leading to a small angular difference; and 2) the color of the roofs for most of the IBs is blue or red. These factors make the MS features more appropriate than the MA features for describing the characteristics of the IBs. By courtesy of the concatenation-based fusion [59]

TABLE IX

OA of the Data-Level and Feature-Level Fusion-Based AMDF-ResNets$_{S+NFB}$, ResUNet-as$_{S+NFB}$, and M^2-3-DCNNs$_{S+MA}$, As Well As the Results of McNemar's Test Between Them, Where ** Indicates $\gamma > 3.84$ at the 95% Confidence Level and * for $\gamma > 2.71$ at a 90% Level

	AMDF-ResNets$_{S+NFB}$			ResUNet-as$_{S+NFB}$			M^2-3DCNNs$_{S+MA}$		
Study area	Data-level	Feature-level	γ value	Data-level	Feature-level	γ value	Data-level	Feature-level	γ value
Wuhan	0.885	0.890	2.734*	0.880	0.888	2.784*	0.884	0.902	5.142**
Hefei	0.909	0.923	9.035**	0.896	0.898	1.026	0.907	0.936	19.959**
Shanghai	0.915	0.924	5.850**	0.917	0.925	5.657**	0.921	0.940	12.342**
Xi'an	0.865	0.879	21.347**	0.864	0.879	7.418**	0.855	0.902	27.102**

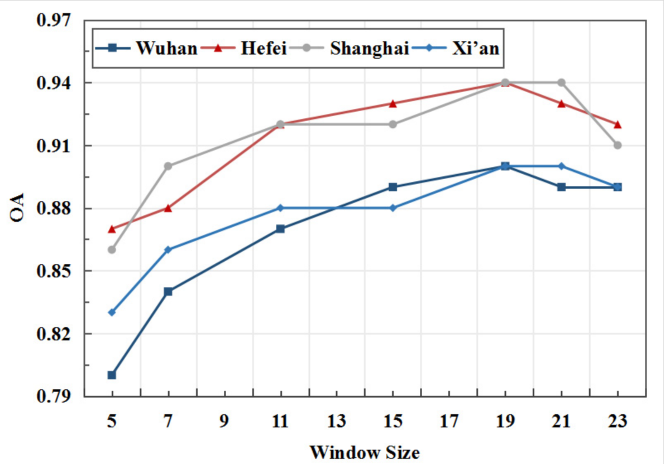

Fig. 14. OA of M2-3-DCNN with a series of window sizes for the four study areas.

of the MS and MA information, M^2-3-DCNN$_{S+MA}$ is superior to the MA single-stream network in almost all the categories, and it also presents advantages over the MS stream in the three classes: residential buildings and roads (for the Shanghai and Xi'an data sets). Hence, it is of value to jointly exploit the complementary multisource information for complex urban area classification.

The advantage of fusing the complementary MS and MA information has been underlined by the experimental results. However, the fusion of the spectral cubes and the GLCM^{MA-T} tensor textures by M^2-3-DCNN can be carried out at both the feature level (i.e., the proposed method) and the data level (i.e., the fourth and sixth benchmark methods). For the purpose of comparison, we achieved the data-level M^2-3-DCNN, that is, the spectral data cubes and GLCM^{MA-T} were stacked and input into the single 3-D-CNN stream, and the feature-level AMDF-ResNets$_{S+NFB}$ or ResUNet-as$_{S+NFB}$ method, that is, the 3-D-CNNs were replaced by AMDF-ResNet or ResUNet-a (without the softmax layer) in the two-stream architecture, and the input of these two streams were MS and multiview data cubes, respectively.

In Tables IX and X, the OAs of these two fusion approaches are presented, as well as the results of McNemar's test between them. It can be found that for all the data sets, the feature-level fusion of M^2-3-DCNN performs significantly better than the data-level fusion and improves the OA by 2.8% on average. For AMDF-ResNets$_{S+NFB}$ and ResUNet-as$_{S+NFB}$, using feature-level fusion can improve the OA by 1.1% and 0.8%, respectively, compared with their data-level fusion approach. These results demonstrate that the deep feature learning in a separate manner from two independent 3-D-CNNs (i.e., MA and MS) is more capable of distinguishing complicated urban objects than one-time learning in a feature stacking manner. In addition, it should be noted that our results are also consistent with the finding in [35] and [59], in that late fusion (feature-level) is better than early fusion (data-level), and our research further extends this to the semantic classification of MS and MA images. Furthermore, our M^2-3-DCNN$_{S+MA}$ can still significantly outperform AMDF-ResNet and ResUNet-a base on feature-level fusion, with increment of 1.6% and 2.3% in OA, respectively. Based on the above analysis, it can be said that the superior accuracy achieved by our proposed method mainly benefits from the proposed GLCM^{MA-T} multiview tensor features and the two-stream 3-D-CNN network. Among the multiview images, high-rise objects will show apparent variations in textures, while low-lying ones usually present similar textural patterns. In this regard, the intraangle and interangle textures defined in GLCM^{MA-T} can describe these planar and vertical characteristics of urban objects more effectively. On the other hand, in contrast to the 2-D networks, for example, AMDF-ResNet or ResUNet-a, 3-D-CNN is able to maintain the local contexts and structures of spectral data cubes and multiview tensor textures, which, therefore, fully exploits the spectral–spatial–angular information contained in the ZY-3 images.

TABLE X

McNemar's Test Between M^2-3-DCNN$_{S+MA}$ and Data-Level and Feature-Level Fusion-Based ResUNet-as$_{S+NFB}$, and ResUNet-as$_{S+NFB}$. The Significantly Different Methods Are Indicated as ** With $\gamma > 3.84$ at a 95% Confidence Level and * for $\gamma > 2.71$ at a 90% Level

M^2-3DCNN$_{S+MA}$ vs.	AMDF-ResNets$_{S+NFB}$		ResUNet-as$_{S+NFB}$	
Study area	Data-level	Feature-level	Data-level	Feature-level
Wuhan	4.889**	3.907**	5.040**	4.005**
Hefei	9.752**	5.007**	12.545**	11.918**
Shanghai	7.955**	5.556**	7.358**	4.663**
Xi'an	14.954**	9.197**	16.326**	9.986**

C. Advantages of 3-D-CNN Over Other Popular 2-D Networks

To further mine the cross-channel information (i.e., spectral bands for the MS images and MA texture for the multiview images), we embedded the 3-D convolution into the network structure, so as to maintain the advantage of 3-D convolution

TABLE XI

THE CLASSIFICATION RESULTS OF THE FOUR TWO-STREAM NETWORKS, WITH EACH STREAM USING 2D-CNN, VGG-19, RESNET-50, AND 3D-CNN, RESPECTIVELY, FOR THE FOUR DATASETS (THE BOLD NUMBERS REPRESENT THE HIGHEST ACCURACY FOR EACH CLASS)

Datasets	Class	LRB	MRB	HRB	IB	Roads	Veg.	B. S.	Water	Sha.	OA
Wuhan	M^2-2DCNN$_{S+MA}$	0.890	0.893	0.750	0.632	0.779	0.967	0.776	**0.995**	0.933	0.875
	M^2-VGG-19$_{S+MA}$	0.809	0.902	0.565	0.509	0.831	0.940	**0.871**	0.981	0.950	0.848
	M^2-Resnet-50$_{S+MA}$	0.912	0.911	0.674	0.538	0.701	0.926	0.853	0.962	0.924	0.853
	M^2-3DCNN$_{S+MA}$	**0.926**	**0.938**	**0.804**	**0.679**	**0.883**	**0.972**	0.776	**0.995**	**0.958**	0.902
Hefei	M^2-2DCNN$_{S+MA}$	0.673	0.799	0.886	0.761	0.920	**0.984**	0.930	0.838	**0.979**	0.866
	M^2-VGG-19$_{S+MA}$	0.618	0.859	0.777	**0.862**	0.897	0.921	0.884	**0.946**	0.931	0.866
	M^2-Resnet-50$_{S+MA}$	0.564	0.922	0.865	0.752	0.920	0.873	**0.953**	0.878	0.948	0.886
	M^2-3DCNN$_{S+MA}$	**0.691**	**0.968**	**0.912**	0.789	**0.943**	0.979	**0.953**	0.919	**0.979**	0.936
Shanghai	M^2-2DCNN$_{S+MA}$	0.814	0.833	0.895	0.876	0.829	0.871	0.889	0.977	**0.985**	0.896
	M^2-VGG-19$_{S+MA}$	0.568	0.891	0.940	0.835	0.809	**0.971**	**0.944**	0.981	0.967	0.902
	M^2-Resnet-50$_{S+MA}$	0.805	0.900	0.935	0.825	0.770	0.907	0.907	0.958	0.952	0.902
	M^2-3DCNN$_{S+MA}$	**0.873**	**0.932**	**0.956**	**0.912**	**0.842**	0.964	0.889	**0.991**	0.974	0.940
Xi'an	M^2-2DCNN$_{S+MA}$	0.782	0.859	0.870	0.856	0.833	**0.980**	0.791	0.913	**0.967**	0.866
	M^2-VGG-19$_{S+MA}$	0.782	0.909	0.874	0.856	0.720	0.912	0.776	0.894	0.942	0.859
	M^2-Resnet-50$_{S+MA}$	0.794	0.861	0.879	0.876	0.793	0.886	0.765	0.929	0.909	0.855
	M^2-3DCNN$_{S+MA}$	**0.806**	**0.952**	**0.892**	**0.878**	**0.853**	0.973	**0.821**	**0.971**	0.934	0.902

in excavating the spectral–spatial–angular information. Considering the burdensome parameters of the 3-D convolution filter and the use of the oversized receptive field in deep networks (e.g., 224 × 224 for ResNet-50 [49] with 50 layers) for HRRS image classification, the proposed method uses a lightweight architecture with two 3-D convolution blocks. To analyze the rationality of the proposed method, several popular 2-D networks with more complex architectures were used for comparison, that is, VGG-19 [60] and ResNet-50. In this section, the three two-stream networks that fuse the MS images and $\mathrm{GLCM}^{\mathrm{MA-T}}$ are denoted as M^2-2-DCNN$_{\mathrm{S+MA}}$, M^2-VGG-19$_{\mathrm{S+MA}}$, and M^2-Resnet-50$_{\mathrm{S+MA}}$, respectively:

1) M^2-2-DCNN$_{\mathrm{S+MA}}$ has the same architecture as M^2-3-DCNN$_{\mathrm{S+MA}}$, except that the convolutional kernels are 2-D filters with the same spatial size;
2) M^2-VGG-19$_{\mathrm{S+MA}}$ uses the same fusion strategy as M^2-3-DCNN$_{\mathrm{S+MA}}$, except that each stream has been replaced by VGG-19;
3) M^2-Resnet-50$_{\mathrm{S+MA}}$ is built in a similar fashion to M^2-VGG-19$_{\mathrm{S+MA}}$. The inputs were spatially interpolated to fit the requirements of VGG-19 and ResNet-50.

The classification accuracies and the results of McNemar's test are provided in Tables XI and XII. For the different deep networks, the OA improvements achieved by M^2-3-DCNN$_{\mathrm{S+MA}}$ are 2.7%–7.0%, compared with M^2-2-DCNN$_{\mathrm{S+MA}}$, M^2-VGG-19$_{\mathrm{S+MA}}$, and M^2-Resnet-50$_{\mathrm{S+MA}}$. The 3-D-CNN combined with $\mathrm{GLCM}^{\mathrm{MA-T}}$ is particularly effective in classifying buildings and roads, showing a 7.3% increase in the accuracy for low-rise residential buildings. Compared with the networks with more complex architectures, that is, VGG-19 and Resnet-50, the 3-D-CNN used in each stream has fewer parameters and a lower complexity. In addition, the floating-point operations (FLOPs) of the proposed M^2-3-DCNN$_{\mathrm{S+MA}}$ total 0.13 billion, which is 0.3% and 1.4% of the FLOPs of M^2-VGG-19$_{\mathrm{S+MA}}$ (39.84 billion) and M^2-Resnet-50$_{\mathrm{S+MA}}$ (9.43 billion), respectively. Hence, the proposed M^2-3-DCNN$_{\mathrm{S+MA}}$ requires fewer training samples and can save on the computational cost.

TABLE XII

MCNEMAR'S TEST BETWEEN M^2-3-DCNN$_{\mathrm{S+MA}}$ AND THREE 2-D NETWORKS (M^2-2-DCNN$_{\mathrm{S+MA}}$, M^2-VGG-19$_{\mathrm{S+MA}}$, AND M^2-RESNET-50$_{\mathrm{S+MA}}$). THE SIGNIFICANTLY DIFFERENT METHODS ARE INDICATED AS $**$ WITH $\gamma > 3.84$ AT 95% CONFIDENCE LEVEL AND $*$ FOR $\gamma > 2.71$ AT 90% LEVEL, RESPECTIVELY

	M^2-3DCNN$_{S+MA}$ vs.		
datasets	M^2-2DCNN$_{S+M}$	M^2-VGG-19$_{S+M}$	M^2-Resnet-50$_{S+M}$
Wuhan	10.417**	32.008**	27.068**
Hefei	63.787**	65.162**	36.790**
Shangha	33.306**	27.698**	25.219**
Xi'an	16.445**	21.571**	23.626**

D. Advantages of M^2-3-DCNN Over Other Tensor Classifier

Tensor classifiers, such as the support tensor machine (STM) [61], can be adopted to classify the combined spectral data cubes and $\mathrm{GLCM}^{\mathrm{MA-T}}$ (denoted as $\mathrm{STM}_{\mathrm{S+MA}}$) while maintaining their tensorial structures. The classification accuracies are compared in Table XIII. It can be seen that M^2-3-DCNN can significantly improve the OA by 3.0%–5.6% compared with STM. The improvement of OA is mainly attributed to better classification of middle-rise residential buildings, factory buildings, roads, and bare soil, with average accuracy increments by 8.1%, 4.9%, 9.9%, and 5.7%, respectively. These results also demonstrate the advantage of deep fusion of spatial and angular features.

E. Influence of the Window Size, Displacement Value, and Proportion of Training Samples

The spectral data cubes and the $\mathrm{GLCM}^{\mathrm{MA-T}}$ feature are obtained using a local window, and its size is related to the spatial resolution of the image and the objects of interest.

TABLE XIII

Classification Results of STMs$_{S+MA}$ and M^2-3-DCNN and McNemar's Test Between Them for Four Data Sets. The Significantly Different Methods Are Indicated as ** with $\gamma > 3.84$ at a 95% Confidence Level and * for $\gamma > 2.71$ at a 90% Level

	Wuhan		Hefei		Shanghai		Xi'an	
Class	STMs$_{S+MA}$	M^2-3DCNNs$_{S+MA}$	STMs$_{S+MA}$	M^2-3DCNNs$_{S+MA}$	STMs$_{S+MA}$	M^2-3DCNNs$_{S+MA}$	STMs$_{S+MA}$	M^2-3DCNNs$_{S+MA}$
LRB	0.915	**0.926**	0.614	**0.691**	0.823	**0.873**	0.796	**0.806**
MRB	0.861	**0.938**	0.880	**0.968**	0.834	**0.932**	0.892	**0.952**
HRB	0.791	**0.804**	0.879	**0.912**	0.939	**0.956**	0.864	**0.892**
IB	0.654	**0.679**	0.705	**0.789**	0.865	**0.912**	0.839	**0.878**
Roads	0.714	**0.883**	0.880	**0.943**	0.726	**0.842**	0.804	**0.853**
Veg.	0.944	**0.972**	0.961	**0.979**	0.961	**0.964**	0.902	**0.973**
B. S.	0.760	**0.776**	0.828	**0.953**	0.865	**0.889**	0.759	**0.821**
Water	0.977	**0.995**	0.837	**0.919**	0.973	**0.991**	0.939	**0.971**
Sha.	0.955	**0.958**	0.971	**0.979**	0.971	**0.974**	0.900	**0.934**
OA	0.872	0.902	0.880	0.936	0.908	0.940	0.857	0.902
γ value	7.481**		51.186**		14.679**		17.523**	

TABLE XIV

Experimental Results Using Different Distance Values for the HEFEI Dataset

Class	r=1	r=3	r=5
LRB	**0.691**	0.620	0.688
MRB	**0.968**	0.954	0.911
HRB	**0.912**	0.908	0.906
IB	**0.789**	0.725	0.737
Roads	**0.943**	0.927	0.925
Veg.	**0.979**	0.975	0.956
B. S.	**0.953**	0.919	0.861
Water	0.919	**0.959**	0.925
Sha.	**0.979**	0.978	**0.979**
OA	0.936	0.920	0.905

To analyze the impact of the window size on the classification performance, a series of experiments were implemented with the following values of $W_x \times W_y$: 5×5, 7×7, 11×11, 15×15, 19×19, and 23×23. Fig. 14 gives an illustration of the relationship between the window size and the classification accuracy. It can be observed that for all the study areas, the OA values of M^2-3-DCNN$_{S+MA}$ increase with the increase in parameters W_x and W_y until the size is larger than 19×19. This effect of the window size can be explained by the fact that small windows cannot provide sufficient spatial and angular information, while large sizes may result in reduced separability of samples with the inclusion of more pixels from the neighboring objects.

The experimental results using different spatial distances are shown in Table XIV. It can be seen that the classification accuracies of M^2-3-DCNN$_{S+MA}$ using the three displacement values (i.e., $r = 1$, 3, or 5) are quite similar, and the results obtained by $r = 1$ are slightly superior to those obtained by $r = 3$ and $r = 5$, by 1.6% and 3.1% in OA, respectively. Furthermore, the CNN also has the ability to capture the spatial relationships in a neighborhood [62]. Therefore, in this study, we only used one spatial distance for computing the GLCM$^{\text{MA}-\text{T}}$ feature ($r = 1$).

In addition, we have investigated the impact of the number of training samples on the classification results. As shown in Fig. 15, with the proportion of training polygons varying from 25% to 75%, the OA of M^2-3-DCNN$_{S+MA}$ has been improved by 5.9%–7.0%. It is a common sense that a larger proportion of training samples can lead to higher classification accuracies.

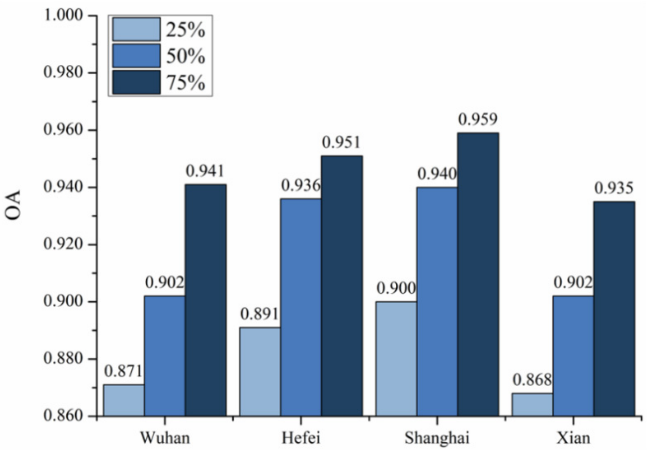

Fig. 15. Classification results of M^2-3-DCNN$_{S+MA}$ with 25%, 50%, and 75% of the reference polygons for training samples.

However, it can be observed that our method also has satisfactory performance with small number of training samples, for example, M^2-3-DCNN$_{S+MA}$ with 25% training samples can still outperform RF$_{S+nDSM}$ and RF$_{S+ADF}$ with 50% training samples.

F. Additional Experiments on Worldview-2 (WV-2) Images

We have conducted an additional experiment using WV-2 images of Wuhan with two viewing angles. The RGB composites and the ground-truth reference are shown in Fig. 16. The classification result of our proposed M^2-3-DCNN$_{S+MA}$ is provided in Table XV. We also compare our method with the state-of-the-art multiview features, for example, the nDSM and ADFs, and two deep-learning-based methods: AMDF and ResUNet. The input of the two networks includes the spectral and multiview images, denoted as AMDF-ResNet$_{S+NF}$ and ResUNet-a$_{S+NF}$, respectively. It can be seen that our M^2-3-DCNN$_{S+MA}$ can also be adapted to the WV-2 images and improve the OA by 3.1%–16.5% compared with other methods.

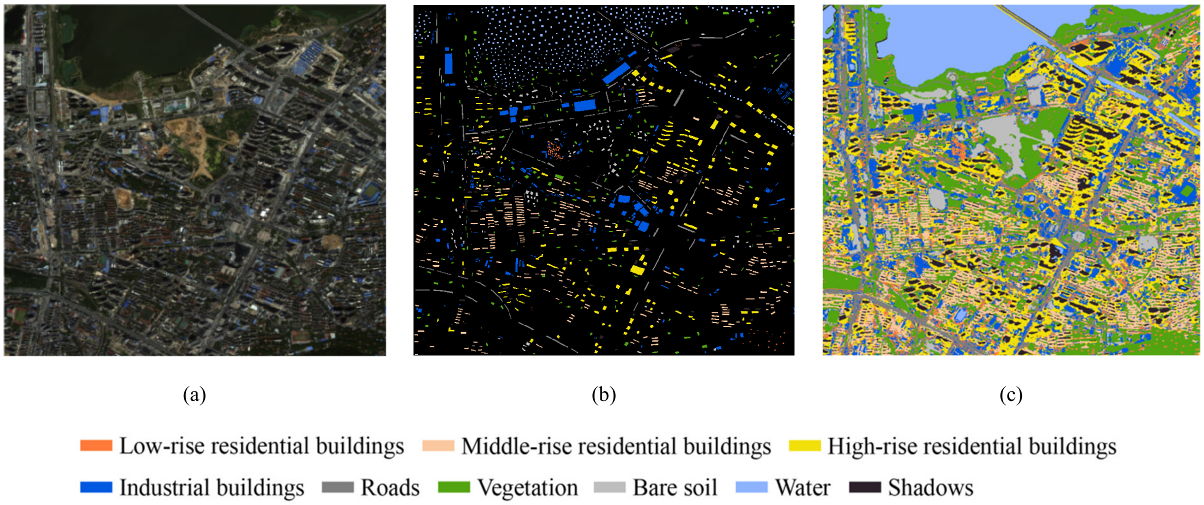
(a)
(b)
(c)
Low-rise residential buildings
Middle-rise residential buildings
High-rise residential buildings
Industrial buildings
Roads
Vegetation
Bare soil
Water
Shadows

Fig. 16. (a) RGB composites of the WV-2 image of Wuhan. (b) Ground-truth reference.

TABLE XV

CLASSIFICATION RESULTS FOR THE WORLDVIEW-2 DATA SET

Class	RF_{S+nDSM}	RF_{S+ADF}	AMDF-$ResNets_{S+NF}$	ResUNet-a_{S+NF}	M^2-$3DCNN_{S+MA}$
LRB	0.294	0.153	0.583	0.667	**0.691**
MRB	0.745	0.763	**0.874**	0.863	**0.874**
HRB	0.616	0.679	0.671	0.707	**0.832**
IB	0.515	0.541	0.835	0.807	**0.859**
Roads	0.597	0.673	**0.840**	0.837	**0.840**
Veg.	**0.982**	0.980	0.970	0.955	0.970
B. S.	0.757	0.808	0.923	0.923	**0.931**
Water	0.992	**0.994**	0.993	0.971	0.986
Sha.	0.969	0.954	0.968	0.968	**0.979**
OA	0.731	0.754	0.865	0.861	0.896

VI. CONCLUSION

The ZY-3 imagery, with MS and multiview bands, has the potential to achieve accurate classification of urban areas. In this article, a novel GLCM^{MA-T} has been proposed, which consists of using intraangle and interangle textures. The intraangle textures are obtained from a single-view band, while the interangle textures are calculated from different combinations of multiview bands from the ZY-3 images. The GLCM^{MA-T} feature can depict the vertical structures in urban areas by capturing the variation characteristics of the gray tones under different viewing angles. Furthermore, to fully exploit the abundant information in the ZY-3 multiview satellite images, the M^2-3-DCNN$_{S+MA}$ framework has been designed. The spectral tensor feature is the input to the MS stream, and the GLCM^{MA-T} feature is interpreted by the MA stream. The deep features learned from the two streams are concatenated and used as inputs for the fully connected layer to exploit the implicit information in the spectral–spatial–angular domain.

The classification performance of the proposed M^2-3-DCNN$_{S+MA}$ when applied on the ZY-3 images over four study areas was compared with the performance of other state-of-the-art methods. A series of experiments were carried out to further verify the effectiveness of each component in the proposed method. The GLCM^{MA-T}, as well as the deep excavation and fusion of the joint spectral–spatial–angular information, can improve the separability of urban objects. We explored the potential of angular information in multiview images to describe the 3-D urban structures by designing an MA feature extraction and interpretation framework. In the future, we will work toward on the proposed multiangular tensor for the classification and change detection of time-series RS images. When applied to multitemporal images, the GLCM^{MA-T} could describe the multitemporal texture characteristics or phonological variations.

ACKNOWLEDGMENT

The authors would also like to thank the editors and anonymous reviewers for the insightful suggestions, which significantly improved the quality of this article.

REFERENCES

[1] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise MRF optimization," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.

[2] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[3] H. Albanwan and R. Qin, "A novel spectrum enhancement technique for multi-temporal, multi-spectral data using spatial-temporal filtering," *ISPRS J. Photogramm. Remote Sens.*, vol. 142, pp. 51–63, Aug. 2018.

[4] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[5] S. Jia, L. Shen, J. Zhu, and Q. Li, "A 3-D Gabor phase-based coding and matching framework for hyperspectral imagery classification," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1176–1188, Apr. 2018.

[6] C. Zhu and X. Yang, "Study of remote sensing image texture analysis and classification using wavelet," *Int. J. Remote Sens.*, vol. 19, no. 16, pp. 3197–3203, Jan. 1998.

[7] O. Regniers, L. Bombrun, V. Lafon, and C. Germain, "Supervised classification of very high resolution optical images using wavelet-based textural features," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3722–3735, Jun. 2016.

[8] L. Zhang, X. Huang, B. Huang, and P. Li, "A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2950–2961, Oct. 2006.

[9] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[10] Minh-Tan, Pham, Sébastien, Lefèvre, and E. Aptoula, "Local feature-based attribute profiles for optical remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1199–1212, Feb. 2018.

[11] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[12] P. Kupidura, "The Comparison of different methods of texture analysis for their efficacy for land use classification in satellite imagery," *Remote Sens.*, vol. 11, no. 10, pp. 1233–1253, May 2019.

[13] A. J. Mathews, A. E. Frazier, S. V. Nghiem, G. Neumann, and Y. Zhao, "Satellite scatterometer estimation of urban built-up volume: Validation with airborne lidar data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 77, pp. 100–107, May 2019.

[14] P. Misra, R. Avtar, and W. Takeuchi, "Comparison of digital building height models extracted from AW3D, TanDEM-X, ASTER, and SRTM digital surface models over Yangon City," *Remote Sens.*, vol. 10, no. 12, pp. 2008–2033, Dec. 2018.

[15] C. Liu, X. Huang, D. Wen, H. Chen, and J. Gong, "Assessing the quality of building height extraction from ZiYuan-3 multi-view imagery," *Remote Sens. Lett.*, vol. 8, no. 9, pp. 907–916, Jun. 2017.

[16] W. Yang, X. Li, B. Yang, and Y. Fu, "A novel stereo matching algorithm for digital surface model (DSM) generation in water areas," *Remote Sens.*, vol. 12, no. 5, pp. 870–891, Mar. 2020.

[17] M. Li, K. M. de Beurs, A. Stein, and W. Bijker, "Incorporating open source data for Bayesian classification of urban land use from VHR stereo images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4930–4943, Nov. 2017.

[18] R. Qin, "A critical analysis of satellite stereo pairs for digital surface model generation and a matching quality prediction model," *ISPRS J. Photogramm. Remote Sens.*, vol. 154, pp. 139–150, Aug. 2019.

[19] G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia, "Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: A study of two multi-angle in-track image sequences," *ISPRS J. Photogramm. Remote Sens.*, vol. 107, pp. 99–111, Sep. 2015.

[20] J. Ma *et al.*, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.

[21] Y. Yan, L. Deng, X. Liu, and L. Zhu, "Application of UAV-based multi-angle hyperspectral remote sensing in fine vegetation classification," *Remote Sens.*, vol. 11, no. 23, pp. 2753–2771, Dec. 2019.

[22] T. Liu and A. Abd-Elrahman, "Multi-view object-based classification of Wetland land covers using unmanned aircraft system images," *Remote Sens. Environ.*, vol. 216, pp. 122–138, Oct. 2018.

[23] X. Huang, H. Chen, and J. Gong, "Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 127–141, Jan. 2018.

[24] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogramm. Remote Sens.*, vol. 132, pp. 48–60, Oct. 2017.

[25] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019.

[26] B. Rasti *et al.*, "Feature extraction for hyperspectral imagery: The evolution from shallow to deep (Overview and toolbox)," *IEEE Geosci. Remote Sens. Mag.*, early access, Apr. 29, 2020, doi: 10.1109/MGRS.2020.2979764.

[27] W. Hu, H. Li, L. Pan, W. Li, R. Tao, and Q. Du, "Spatial–spectral feature extraction via deep ConvLSTM neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4237–4250, Jun. 2020.

[28] R. Li, Z. Pan, Y. Wang, and P. Wang, "A convolutional neural network with mapping layers for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3136–3147, May 2020.

[29] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[30] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.

[31] J.-H. Lee, S. S. Lee, H. G. Kim, S.-K. Song, S. Kim, and Y. M. Ro, "MCSIP net: Multichannel satellite image prediction via deep neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2212–2224, Mar. 2020.

[32] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 164–178, Mar. 2020.

[33] L. Sifre and S. Mallat, "Rigid-motion scattering for texture classification," *Int. J. Comput. Vis.*, vol. 3559, pp. 501–515, Mar. 2014.

[34] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3D convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sens.*, vol. 10, no. 1, pp. 75–92, Jan. 2018.

[35] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.

[36] J. Feng *et al.*, "CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019.

[37] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Proc. Eur. Conf. Comput. Vis.*, Dublin, Ireland, 2000, pp. 404–420.

[38] J. Xiao, M. Gerke, and G. Vosselman, "Building extraction from oblique airborne imagery based on robust façade detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 68, pp. 56–68, Mar. 2012.

[39] F. Pacifici, N. Longbotham, and W. J. Emery, "The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6241–6256, Oct. 2014.

[40] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[41] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[42] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875 A, Jan. 4, 2000.

[43] D. Chen and D. Stow, "The effect of training strategies on supervised classification at different spatial resolutions," *Photogramm. Eng. Remote Sens.*, vol. 68, no. 11, pp. 1155–1161, Nov. 2002.

[44] J. L. Van Genderen, B. F. Lock, and P. A. Vass, "Remote sensing: Statistical testing of thematic map accuracy," *Remote Sens. Environ.*, vol. 7, no. 1, pp. 3–14, Jan. 1978.

[45] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

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

[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 2015, pp. 234–241.

[48] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet–A: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2016, pp. 770–778.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6230–6239.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[53] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 12–24.

[54] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[55] J. Li, X. Huang, and J. Gong, "Deep neural network for remote-sensing image interpretation: Status and perspectives," *Nat. Sci. Rev.*, vol. 6, no. 6, pp. 1082–1086, Nov. 2019.

[56] M. Nieto-Hidalgo, A. J. Gallego, P. Gil, and A. Pertusa, "Two-stage convolutional neural network for ship and spill detection using SLAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5217–5230, Sep. 2018.

[57] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621

[58] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.

[59] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–15, 2020, doi: 10.1109/TGRS.2020.3016820.

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–14.

[61] X. Guo, X. Huang, L. Zhang, L. Zhang, A. Plaza, and J. A. Benediktsson, "Support tensor machines for classification of hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3248–3264, Jun. 2016.

[62] S. S. Heydari and G. Mountrakis, "Meta-analysis of deep neural networks in remote sensing: A comparative study of mono-temporal classification to support vector machines," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 192–210, Jun. 2019.

Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2009.

He is a Luojia Distinguished Professor with Wuhan University, where he teaches remote sensing, photogrammetry, image interpretation, and so on. He is the Founder and the Director of the School of Remote Sensing and Information Engineering, Institute of Remote Sensing Information Processing (IRSIP), Wuhan University. He has published more than 150 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. His research interests include remote sensing image processing methods and applications. He has been supported by The National Program for Support of Top-notch Young Professionals in 2017, the China National Science Fund for Excellent Young Scholars in 2015, and the New Century Excellent Talents in University from the Ministry of Education of China in 2011.

Dr. Huang was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the second place recipient for the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the winner of the IEEE GRSS 2014 Data Fusion Contest. He was the Lead Guest Editor of the Special Issue for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (May 2015 and August 2019), the *Journal of Applied Remote Sensing* (October 2016), *Photogrammetric Engineering and Remote Sensing* (November 2018), and *Remote Sensing* (November 2019). He was an Associate Editor of the *Photogrammetric Engineering and Remote Sensing* from 2016 to 2019. He has been serving as an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2014 and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2018. He has also been an Editorial Board Member of the *Remote Sensing of Environment* since 2019, the *Science of Remote Sensing* since 2020, and the *Remote Sensing* since 2018.

Shuang Li received the B.S. degree from Wuhan University, Wuhan, China, in 2018, where she is pursuing the M.S. degree with the School of Remote Sensing and Information Engineering.

Her research interests include multiview imagery, land cover/land use classification, change detection, and machine learning.

Jiayi Li (Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 30 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision and pattern recognition, and remote sensing images.

Dr. Li is a Reviewer for more than ten international journals, including the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE SIGNAL PROCESSING LETTERS, and the *International Journal of Remote Sensing*. She is the Guest Editor of the Special Issue on *Change Detection Using MultiSource Remotely Sensed Imagery for the Remote Sensing* (an open access journal from MDPI).

Xiuping Jia (Senior Member, IEEE) received the B.Eng. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 1982, and the Ph.D. degree in electrical engineering from the University of New South Wales, Canberra, ACT, Australia, in 1996.

Since 1988, she has been with the School of Engineering and Information Technology, University of New South Wales, where she is an Associate Professor. She has more than 200 publications, including more than 100 articles in leading technical journals. She has coauthored the remote sensing textbook titled *Remote Sensing Digital Image Analysis* [Springer-Verlag, Third (1999) and Fourth (2006) Editions]. Her research interests include remote sensing, machine learning, and spatial data analysis.

Dr. Jia is a Subject Editor of the *Journal of Soils and Sediments* in recent years and has been an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2005.

Jun Li (Senior Member, IEEE) received the B.S. degree in geographic information systems from Hunan Normal University, Changsha, China, in 2004, the M.E. degree in remote sensing from Peking University, Beijing, China, in 2007, and the Ph.D. degree in electrical engineering from the Instituto de Telecomunicações, Instituto Superior Técnico (IST), Universidade Técnica de Lisbon, Lisbon, Portugal, in 2011.

She is a Full Professor with Sun Yat-sen University, Guangzhou, China. Her main research interests comprise remotely sensed hyperspectral image analysis, signal processing, supervised/semisupervised learning, and active learning.

Dr. Li is the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. She has been a Guest Editor for several journals, including the PROCEEDINGS OF THE IEEE and the *ISPRS Journal of Photogrammetry and Remote Sensing*.

Xiao Xiang Zhu (Senior Member, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is a Professor with the Signal Processing in Earth Observation, TUM, and the German Aerospace Center (DLR), Remote Sensing Technology Institute, Weßling, Germany, where she is also the Head of the Department "Earth Observation (EO) Data Science," DLR's Earth Observation Center and the Head of Helmholtz Young Investigator Group "SiPEO" at DLR and TUM. Since 2019, she has been coordinating the Munich Data Science Research School, Munich. She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Research Field "Aeronautics, Space and Transport." She was a Guest Scientist or a Visiting Professor with the CNR, IREA, Italian National Research Council, Naples, Italy; Fudan University, Shanghai, China; The University of Tokyo, Tokyo, Japan; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg), Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Jón Atli Benediktsson (Fellow, IEEE) received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, Iceland, in 1984, and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, USA, in 1987 and 1990, respectively.

He is a Pro-Rector for Academic Affairs and a Professor of electrical and computer engineering with the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in these fields.

Dr. Benediktsson is a Fellow of SPIE. He is a member of the Association of Chartered Engineers in Iceland (VFI), the Societas Scinetiarum Islandica, and Tau Beta Pi. He was the 2011–2012 President of the IEEE Geoscience and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. He received the Stevan J. Kristof Award from Purdue University in 1991 as an Outstanding Graduate Student in remote sensing, the Icelandic Research Council's Outstanding Young Researcher Award in 1997, the IEEE Third Millennium Medal in 2000, the Yearly Research Award from the Engineering Research Institute, University of Iceland, in 2006, the Outstanding Service Award from the IEEE GRSS in 2007, and the IEEE/VFI Electrical Engineer of the Year Award in 2013. He was a corecipient of the University of Iceland's Technology Innovation Award in 2004, the 2012 IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING Paper Award, and the IEEE GRSS Highest Impact Paper Award in 2013. He was an Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) from 2003 to 2008 and has been serving as an Associate Editor for TGRS since 1999, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2003, and the IEEE ACCESS since 2013. He is on the International Editorial Board of the *International Journal of Image and Data Fusion* and was the Chairman of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (J-STARS) from 2007 to 2010. He is a Co-Founder of the biomedical startup company Oxymap (www.oxymap.com).