A Shallow-to-Deep Feature Fusion Network for VHR Remote Sensing Image Classification

Sicong Liu, *Senior Member, IEEE*, Yongjie Zheng, Qian Du, *Fellow, IEEE*, Lorenzo Bruzzone, *Fellow, IEEE*, Alim Samat, *Member, IEEE*, Xiaohua Tong, *Senior Member, IEEE*, Yanmin Jin, and Chao Wang

***Abstract*— With more detailed spatial information being represented in very-high-resolution (VHR) remote sensing images, stringent requirements are imposed on accurate image classification. Due to the diverse land objects with intraclass variation and interclass similarity, efficient and fine classification of VHR images especially in complex scenes are challenging. Even for some popular deep learning (DL) frameworks, geometric details of land objects may be lost in deep feature levels, so it is difficult to maintain the highly detailed spatial information (e.g., edges, small objects) only relying on the last high-level layer. Moreover, many of the newly developed DL methods require massive well-labeled samples, which inevitably deteriorates the model generalization ability under the few-shot learning. Therefore, in this article, a lightweight shallow-to-deep feature fusion network (SDF^{2}N) is proposed for VHR image classification, where the traditional machine learning (ML) and DL schemes are integrated to learn rich and representative information to improve the classification accuracy. In particular, the shallow spectral–spatial features are first extracted and then a novel triple-stage fusion (TSF) module is designed to learn the saliency and discriminative information at different levels for classification. The TSF module includes three feature fusion stages, that is, low-level spectral–spatial feature fusion, middle-level multiscale feature fusion, and high-level multilayer feature fusion. The proposed SDF^{2}N takes the advantage of the shallow-to-deep features, which can extract representative and complementary information from crossing layers. It is important to note that even with limited training samples, the SDF^{2}N still can achieve satisfying classification performance. Experimental results obtained on three real VHR remote sensing datasets including two multispectral and one airborne hyperspectral images covering complex urban scenarios confirm the effectiveness of the proposed approach compared with the state-of-the-art methods.**

***Index Terms*— Extended multiattribute profiles (EMAP), shallow-to-deep feature fusion, spectral–spatial feature extraction, squeeze-excitation (SE) attention mechanism, very-high-resolution (VHR) image classification.**

Manuscript received February 15, 2022; revised May 20, 2022; accepted May 25, 2022. Date of publication May 30, 2022; date of current version June 10, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505000 and Grant 2018YFB0505400, in part by the National Natural Science Foundation of China under Grant 42071324 and Grant 42001387, and in part by the Shanghai Rising-Star Program under Grant 21QA1409100. *(Corresponding authors: Xiaohua Tong; Yongjie Zheng.)*

Sicong Liu, Yongjie Zheng, Xiaohua Tong, Yanmin Jin, and Chao Wang are with the College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China (e-mail: sicong.liu@tongji.edu.cn; yongjie.zheng@outlook.com; xhtong@tongji.edu.cn; jinyanmin@tongji.edu.cn; wangchao2019@tongji.edu.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA (e-mail: du@ece.msstate.edu).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

Alim Samat is with the Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Ürümqi 830011, China (e-mail: alim.smt@gmail.com).

Digital Object Identifier 10.1109/TGRS.2022.3179288

I. Introduction

THE rapid development of new-generation earth-observation (EO) satellites allows the acquisition of an increasing number of high-resolution (HR) and very-high-resolution (VHR) remote sensing images. This results in VHR multispectral (VHR-MS) images with very high spatial resolution and HR hyperspectral (HR-HS) images with a high spectral–spatial resolution, which makes it possible to analyze land surface objects at an unprecedented detailed scale [1], [2]. Accurate and robust identification of multiclass objects in VHR remote sensing images is of great significance in various applications [3], [4]. However, such high spatial and spectral resolutions lead to many issues and challenges in image processing and applications. For example, the limited number of spectral bands but an over-rich spatial representation of objects in VHR-MS images; the rich but redundant spectral–spatial information in HR-HS images; and the difficult feature extraction for classification. Moreover, the lack of a sufficient number of training samples is the primary cause of low classification performance, especially for complex scenarios (e.g., urban land-use) [5]–[8]. Therefore, it is still a very challenging task in real applications to obtain high-quality and reliable semantic land-cover mapping results from VHR remote sensing images with a limited number of samples.

In the past decades, researchers have made great efforts to exploit effective spectral–spatial joint methods for VHR image classification [9]–[16]. According to the feature extraction and fusion strategies in the literature, such classification methods can be divided into two main groups based on the use of traditional machine learning (ML) and on advanced deep learning (DL) models.

For traditional ML models, many spectral–spatial feature extraction and fusion approaches have been proposed to fully exploit the properties of VHR images, thus improving the classification performance. These approaches include filtering-based methods (e.g., Gabor filter [9], guided filter [10]), morphology-based methods (e.g., attribute profiles (AP) [17],

extended attribute profiles (EAP) [18], and extended multi-attribute profiles (EMAP) [19]), sparse-based methods [12], [20], multiple kernel-based methods [13], [21], and other integrated learning strategies [22]. ML-based classification methods have clear advantages due to their high flexibility, low time consumption, and low training data requirement, which lead to their successful application to VHR image classification [11]. The concepts of morphological profile (MP) and extended MP (EMP) to model the spatial information were presented in [23] and [24]. As an extension of MP, AP was proposed in [17], which models the spatial information more precisely than MP since more attributes of the input image can be considered. In [25], the K-means and principal component analysis (PCA) were utilized to learn the spatial feature. Then, spectral–spatial features were generated for HS image classification by concatenating the spatial feature representations in all or some principal components (PCs). In [26], a spectral–spatial multiple kernel learning method was proposed for HS image classification. Unlike the direct stacking methods, it used the spectral–spatial weighted composite kernel structure to better integrate spectral–spatial information. However, the performances of these methods are still far to be satisfactory due to the limited handcrafted feature representation, sensitive parameter settings, and possible poor generalization abilities.

Thanks to the discriminative feature representations and end-to-end learning capabilities, recently many DL-based frameworks have achieved remarkable success in the remote sensing image classification (e.g., convolutional neural networks (CNNs) [14], [27], recurrent neural networks (RNNs) [15], [28], generative adversarial networks (GANs) [16], [29], and hybrid networks [30]). The existing DL-based methods are usually constructed with multiple layers, which consider the low-level features as the input and produce the output for the middle-level or high-level features. However, neurons and layers in different models have their unique connections, and the effective integration of spectral–spatial information is mainly achieved through the specific design of forward and backward response units. This fusion strategy is limited by the characteristics of the network, which results in poor interpretability. Therefore, except for the internal fusion within each unit or layer, external fusion operations including the concatenate (concat), add, multiply, and attention weighting mechanism operations are usually applied. For instance, in [31], a novel cross-resolution hidden layer feature fusion (CRHFF) approach was proposed for the joint classification of multiresolution MS and PAN images. Hence, the latent information is extracted and fused according to an autoencoder-like deep network. In [32], a novel fast dense spectral–spatial convolution network (FDSSC) was proposed based on 3-D densely connected structures for the accurate classification of HS images. In [33], a spectral–spatial unified network (SSUN) was developed to extract spatial and spectral features according to a multiscale CNN and a long short-term memory (LSTM) network, respectively, and then features were cascaded together for image classification. In [34], a spectral–spatial residual network (SSRN) was proposed for HS image classification to learn deep discriminative features from abundant spectral features and spatial contexts based on consecutive specific residual blocks. In addition to the above CNN-based frameworks, there are some other advanced frameworks such as graph convolutional network (GCN) and transformer. By considering the rich spectral–spatial information of HS images, in [35], a novel miniGCN was proposed to train large-scale graph networks in a minibatch fashion. In [36], a new transformer-based backbone network, named SpectralFormer, was proposed to extract more spectral information from HS images. Although these DL-based methods have shown promising progress in VHR remote sensing image classification, there are still some open issues that require further investigations. They consist in.

1) Most spectral–spatial fusion deep networks are designed for HS image classification, only a few studies in the literature focus on VHR-MS image classification, with the result and the model generalization remain poor when simultaneously considering the two tasks.
2) The spectral–spatial fusion frameworks, on the one hand, often do not integrate multilevel features, thus they have shortcomings for accurately identifying the interior and edges of high-detailed objects. On the other hand, some existing fusion strategies are not able to properly utilize the shallow-to-deep features and the saliency information presented at different scales.
3) Most of the DL-based networks require a large number of training samples to support an effective model learning. Thus, their accuracy and stability in the few-shot learning cases are relatively poor.

Inspired by the aforementioned classification methods and also motivated to overcome the existing open issues, in this article, we propose a novel shallow-to-deep feature fusion network (named SDF^{2}N) for VHR remote sensing image classification. The main contributions of this network can be summarized as follows.

1) Based on the joint fusion of shallow spectral–spatial features and the corresponding deep multiscale features, it is capable of better capturing the detailed spectral–spatial and shallow-to-deep information. Accordingly, the classification performance of highly detailed land objects in VHR images is enhanced step by step by following a hierarchical feature fusion process.
2) A novel triple-stage fusion (TSF) strategy with three core feature fusion stages is designed. It can sequentially capture and fuse the specific discriminative and representative spectral–spatial features presented in VHR images at low, middle, and high levels. Therefore, the identification ability especially for the edge details of complex objects or small objects is significantly improved.
3) Differently from other advanced DL-based methods, the proposed SDF^{2}N approach shows its stability and excellent classification performance, especially in the small-sample cases, and is in general suitable for different types of VHR datasets such as MS, HS, and unmanned aerial vehicle (UAV)-RGB camera images. This greatly

increases the potential use of the proposed approach to deal with complex scenes in practical multisensor applications.

Experimental results obtained on three real VHR remote sensing datasets including two MS datasets and one HS dataset confirmed the effectiveness of the proposed approach compared with the state-of-the-art methods.

The reminder of this article is organized as follows. Section II briefly introduces the related work. The proposed SDF^{2}N approach is described in detail in Section III. Experimental results and the related analysis are presented in Section IV. Finally, Section V draws the conclusions and provides future directions.

II. Related Work

A. Extended Multi-Attribute Profiles

As one of the most popular shallow feature extraction techniques, the morphology-based methods are proved to be useful to enhance the classification performance of VHR images. The popular algorithms include MP [23], EMP [24], AP [17], EAP [18], and EMAP [19]. Their rigorous mathematical foundation and inherent ability to capture spectral–spatial information have led to the rapid development of ML-based feature extraction and fusion strategies. It is worth noting that they also have the advantage to be effective also with a limited number of training data [11]. Among them, the AP and its expansions, that is, EAP and EMAP, are the most widely used approaches owning to the stronger capability to model local spatial context information [25].

Let $\boldsymbol{X} \in \mathbb{R}^{H\times W\times M}$ be a VHR image, where H, W, and M represent height, width, and the number of bands, respectively. Let ϕ_λ and γ_λ be the attribute thickening and attribute thinning, respectively. AP calculated on a given band $\boldsymbol{x}_m$ ($m \subseteq[0, M]$) of $\boldsymbol{X}$ can be defined as

$$\mathrm{AP}(\boldsymbol{x}_m)=\{\phi_{\lambda_N}(\boldsymbol{x}_m),\ldots,\phi_{\lambda_1}(\boldsymbol{x}_m),\boldsymbol{x}_m,\gamma_{\lambda_1}(\boldsymbol{x}_m),\ldots,\gamma_{\lambda_N}(\boldsymbol{x}_m)\} \tag{1}$$

where N is the number of attribute thinning and thickening operations.

As an extension of AP, the EAP is achieved by consecutively applying the thinning and thickening filters to the original spectral bands or on their PCs

$$\mathrm{EAP}(\boldsymbol{X}) = \{\mathrm{AP}(\boldsymbol{x}_1),\ldots,\mathrm{AP}(\boldsymbol{x}_M)\} \tag{2}$$

The EMAP is designed by stacking different types of EAP generated according to different attribute parameters [37], to comprehensively model complex objects in an image [11]. In this work, three attributes including area (a), diagonal box (db), and standard deviation (sd) are selected to generate the EMAP features, which can be formulated as

$$\mathrm{EMAP}(\boldsymbol{X}) = \{\mathrm{EAP}_a(\boldsymbol{X}),\mathrm{EAP}_{db}(\boldsymbol{X}),\mathrm{EAP}_{sd}(\boldsymbol{X})\} \tag{3}$$

where the thresholds for the above attributes are $a = 150$, $db = 50$, and $sd = 20$.

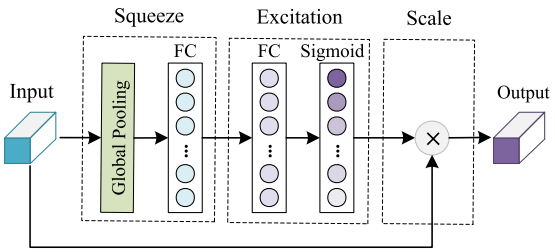


Fig. 1. SE module [42].

B. CNN

CNN is a feed-forward neural network containing at least one convolution layer. In the field of remote sensing image processing, it is popular for pixel-wise classification [38]. In general, four types of layers are included in a CNN architecture: 1) the convolution layer; 2) the pooling layer; 3) the batch normalization (BN) layer; and 4) the fully connected (FC) layer [39]. According to the processing dimensions, the convolution can be 1-D, 2-D, and 3-D. For the most common 2-D convolution, a 2-D kernel moves along the height and width directions of an image, which extracts deep features within a specified local neighborhood. In order to better model the spectral–spatial characteristics of VHR images, the 2-D convolution is selected as the base module in this work. The mathematical formulation of the 2-D convolution can be expressed as

$$\boldsymbol{X}^{l+1} = F(\boldsymbol{X}^l) = f_\delta(\omega^l * (\boldsymbol{X}^l) + b^l) \tag{4}$$

where $\boldsymbol{X}^l$ represents the input feature maps of the lth convolution layer, $\boldsymbol{X}^{l+1}$ is the output feature maps of the lth layer [also the input set of the $(l+1)$th layer] [40], ω^l and b^l are the weights and bias of the lth layer, respectively, and f_δ represents the ReLU activation function.

C. Squeeze-Excitation (SE) Attention Mechanism

Attention has arguably become one of the most important concepts in the DL field. It is inspired by the biological systems of humans that tend to focus on distinctive parts when processing large amounts of information [41]. Among the popular attention mechanisms, one widely used module is the SE attention [42]. It can adaptively re-calibrate channel-wise feature responses by explicitly modeling interdependencies between channels [41]. In particular, the SE module mainly uses global average-pooled features and FC features to compute channel-wise attention. As shown in Fig. 1, the structure of SE mainly consists of the squeeze, excitation, and scale steps [6], [43]. Let us assume $\boldsymbol{U} \in \mathbb{R}^{H\times W\times M}$ represent a given input feature vector.

1) Squeeze: The global average pooling (AvgPool) and FC operations are selected to build the squeeze transform F_{sq}, and the input feature vector $\boldsymbol{U}$ is squeezed into a global spatial 1-D feature vector (i.e., channel descriptor), which can be formulated as

$$z_m = F_{\mathrm{sq}}(\boldsymbol{u}_m) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{u}_m(i,j) \tag{5}$$

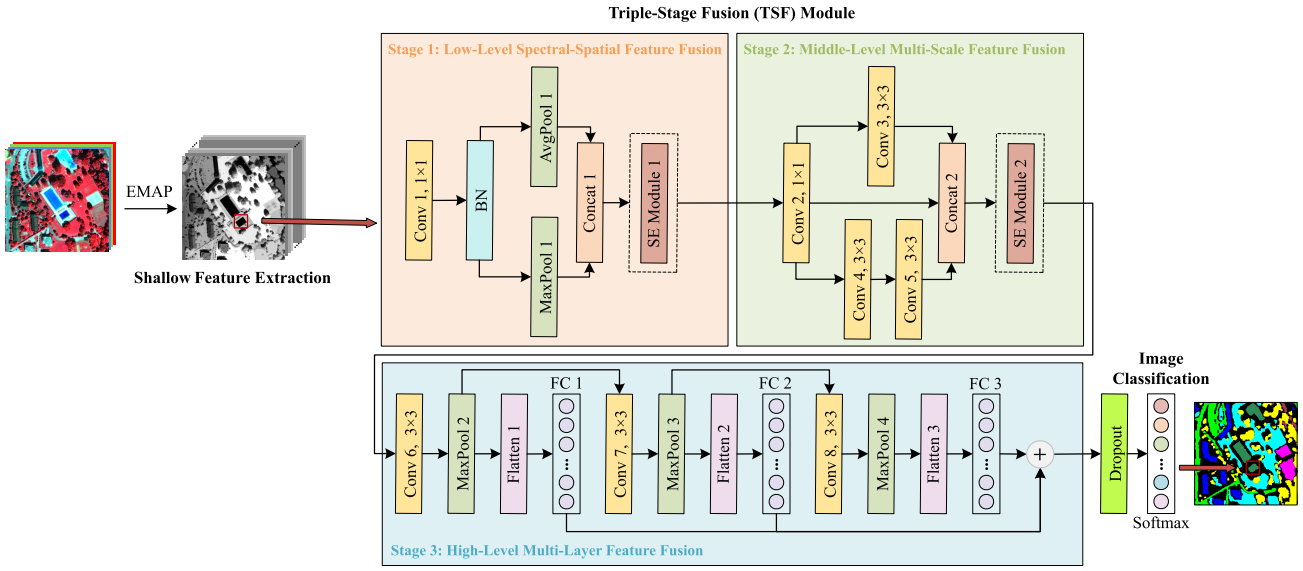


Fig. 2. Flowchart of the proposed SDF^{2}N for VHR image classification. It comprises shallow feature extraction, TSF, and image classification three modules, where the TSF module consists of three sequential feature fusion stages: low-level spectral–spatial feature fusion, middle-level multiscale feature fusion, and high-level multilayer feature fusion.

where $\boldsymbol{u}_m$ is the mth feature map of $\boldsymbol{U}$, i, and j are the elements of the feature map, and z_m is the output of the squeeze operation.

2) Excitation: The excitation transform F_{ex} performs a nonlinear transformation on the squeezed result based on a FC layer and compresses the weights of different features to 0–1 through the sigmoid function

$$\boldsymbol{s} = F_{\text{ex}}(z, W) = f_\sigma(W_2 f_\delta(W_1 z)) \tag{6}$$

where $W_1 \in \mathbb{R}^{(W/r)\times W}$ and $W_2 \in \mathbb{R}^{W\times(W/r)}$ are the FC layers for reducing and increasing dimension, respectively. r is a reduction ratio, $\boldsymbol{s}$ is the output of the excitation operation (which also can be seen as the weight vector), and f_σ represents the sigmoid function.

3) Scale: The scale transform F_{sc} is also called the reweight or feature recalibration. The previous output $\boldsymbol{s}$ is applied to weight the input feature set $\boldsymbol{U}$. So the final output $\widetilde{\boldsymbol{u}}_m$ can be obtained through the F_{sc} operation

$$\widetilde{\boldsymbol{u}}_m = F_{\text{sc}}(\boldsymbol{u}_m, \boldsymbol{s}_m) = \boldsymbol{s}_m \boldsymbol{u}_m. \tag{7}$$

III. Proposed SDF^{2}N

Fig. 2 illustrates the architecture of the proposed SDF^{2}N approach, which consists of three main parts: 1) shallow feature extraction; 2) TSF; and 3) image classification. In particular, a novel TSF strategy is designed to sequentially capture and fuse the shallow-to-deep features in VHR images at different levels. In particular, the rich shallow artificial spectral–spatial features are fused in stage 1 at the low level, the multiscale features are fused in stage 2 at the middle level, and the multilayer abstract and discriminative features are fused in stage 3 at the high level. More details are provided as follows.

A. Shallow Feature Extraction

Considering that VHR images usually contain several broad spectral bands, EMAP can effectively capture spectral–spatial features in an unsupervised fashion and then provide richer shallow features as the input of DL-based networks. Therefore, we first extracted the EMAP features from the original VHR image $\boldsymbol{X}$. Let $\boldsymbol{X}' \in \mathbb{R}^{H\times W\times B}$ be the new data, where B is the number of EMAP features (B is equal to $7M$ in this article). A patch with a size of $w \times w$ is created as the feature region (around each pixel). Therefore, the actual size of the input data is $\boldsymbol{I} \in \mathbb{R}^{w\times w\times B}$.

B. TSF Module

1) Stage 1—Low-Level Spectral–Spatial Feature Fusion: The handcrafted EMAP features generated in the previous step contain rich spectral–spatial information but with high redundancy. Therefore, the first stage of the TSF module is designed to overcome the drawback so as to effectively fuse the spectral–spatial information in a compound set of discriminate features (see Fig. 2).

Table I lists the structure parameter settings in this stage. Specifically, a 1×1 convolution layer with 128 kernels is first employed to transform $\boldsymbol{I}$ into $F(\boldsymbol{I}) \in \mathbb{R}^{w\times w\times 128}$ for capturing the complex and learnable interactions of cross-channel information. Next, the BN is connected behind to avoid the gradient vanishing phenomenon. It can be formalized as

$$\boldsymbol{I}^2 = F(\boldsymbol{I}^1) = f_\delta[\omega^1 * (\boldsymbol{I}^1) + b^1] \tag{8}$$

$$\boldsymbol{I}^3 = \text{BN}(\boldsymbol{I}^2) = \frac{\boldsymbol{I}^2 - \text{E}(\boldsymbol{I}^2)}{\sqrt{\text{Var}(\boldsymbol{I}^2) + \epsilon}} \tag{9}$$

where $\boldsymbol{I}^1$ is equal to $\boldsymbol{I}$, $\text{E}(\boldsymbol{I}^2)$, and $\text{Var}(\boldsymbol{I}^2)$ are the expectation and variance function of $\boldsymbol{I}^2$, respectively, and ϵ is a very small constant value (i.e., 1e^{-5}) that maintain stability.

Then, two kinds of pooling layers, that is, the global max pooling (MaxPool) and the AvgPool, are combined to obtain the texture detailed information and background information, respectively. This can also improve the representability of

Fig. 3. Illustration of an example of the AvgPool and MaxPool operations.

TABLE I
NETWORK AND PARAMETER SETTINGS IN THE STAGE 1 OF THE PROPOSED SDF²N

Layers	Filter Size	Activation	Strides	Padding	Output Shape
Conv 1	128×1×1	ReLU	1	same	32×32×128
BN	/	/	/	/	32×32×128
AvgPool 1	2×2	/	2	valid	16×16×128
MaxPool 1	2×2	/	2	valid	16×16×128
Concat 1	/	/	/	/	16×16×256
SE Module 1	/	/	/	/	16×16×256

geometric details of complex objects. The pooling operations are illustrated in Fig. 3. Let the input feature maps of two pooling layers be $\boldsymbol{I}^3 \in \mathbb{R}^{w\times w\times 128}$. For each feature map $\boldsymbol{I}_d^3 \in \mathbb{R}^{w\times w}$, the MaxPool selects the maximum value, while the AvgPool calculates the average value of a specific area $R_{k,k}^d$ as its representation

$$\boldsymbol{I}_t^4 = \text{MaxPool}(\boldsymbol{I}^3) = \max_{t\in R_{k,k}^d} \boldsymbol{I}_t^3 \tag{10}$$

$$\boldsymbol{I}_t^5 = \text{AvgPool}(\boldsymbol{I}^3) = \frac{1}{\left|R_{k,k}^d\right|} \sum_{t\in R_{k,k}^d} \boldsymbol{I}_t^3 \tag{11}$$

$$\boldsymbol{I}^6 = \text{Concat}(\boldsymbol{I}^4, \boldsymbol{I}^5) = \left\{\boldsymbol{I}^4, \boldsymbol{I}^5\right\} \tag{12}$$

where $1 \le d \le 128$, and $1 \le k \le w$.

Finally, in order to improve the feature representation by modeling the interdependencies between different channels [43], the SE attention module is adopted to realize the weighted recalibration for low-level features

$$\boldsymbol{I}^7 = \text{SE}(\boldsymbol{I}^6). \tag{13}$$

2) Stage 2—Middle-Level Multiscale Feature Fusion: After fusing the spectral–spatial information in the previous stage, stage 2 aims to generate and fuse the middle-level multiscale features. Table II lists the structure parameter settings in this stage. Filters are pivotal for the convolution operation in CNN [44]. The larger the scale filter, the larger the receptive field and stronger semantic representation (see Fig. 4). For complex image objects, the joint use of different scales can better retain the discriminant information. Therefore, in this stage, two types of filters (1×1 and 3×3) are used to construct a multireceptive field feature learning mechanism, that is, 1×1: $\boldsymbol{I}^8 = F(\boldsymbol{I}^7)$, 3×3: $\boldsymbol{I}^9 = F(\boldsymbol{I}^8)$, and 5×5: $\boldsymbol{I}^{10} = F(F(\boldsymbol{I}^8))$. Due to the fact that larger receptive results in weaker spatial geometric features but requires more parameters, two 3×3 convolution layers are used in this work instead of the 5×5 convolution. This reduces the number of parameters and increases the nonlinear expression ability. Finally, the same as in stage 1, the concatenation operation and SE module are used to further fuse the generated

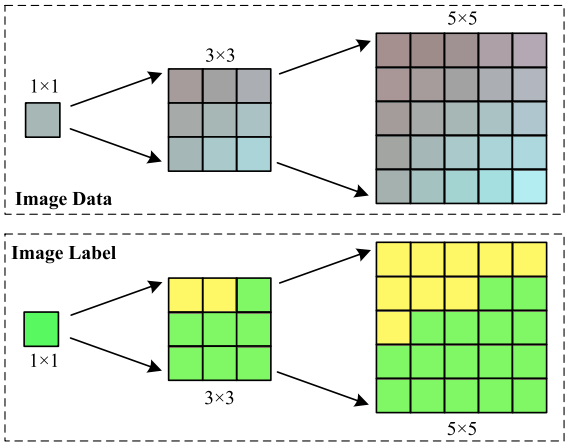


Fig. 4. Illustration of multireceptive fields.

TABLE II
NETWORK AND PARAMETERS SETTINGS IN THE STAGE 2 OF THE PROPOSED SDF²N

Layers	Filter Size	Activation	Strides	Padding	Output Shape
Conv 2	128×1×1	ReLU	1	same	16×16×128
Conv 3	128×3×3	ReLU	1	same	16×16×128
Conv 4	128×3×3	ReLU	1	same	16×16×128
Conv 5	128×3×3	ReLU	1	same	16×16×128
Concat 2	/	/	/	/	16×16×384
SE Module 2	/	/	/	/	16×16×384

TABLE III
NETWORK AND PARAMETERS SETTINGS IN THE STAGE 3 OF THE PROPOSED SDF²N

Layers	Filter Size	Activation	Strides	Padding	Output Shape
Conv 6	128×3×3	ReLU	1	same	16×16×128
MaxPool 2	2×2	/	2	valid	8×8×128
Flatten 1	/	/	/	/	8192
FC 1	128	ReLU	/	/	128
Conv 7	128×3×3	ReLU	1	same	8×8×128
MaxPool 3	2×2	/	2	valid	4×4×128
Flatten 2	/	/	/	/	2048
FC 2	128	ReLU	/	/	128
Conv 8	128×3×3	ReLU	1	same	4×4×128
MaxPool 4	2×2	/	2	valid	2×2×128
Flatten 3	/	/	/	/	512
FC 3	128	ReLU	/	/	128
Add	/	/	/	/	128

multiscale features as

$$\boldsymbol{I}^{11} = \text{Concat}(\boldsymbol{I}^8, \boldsymbol{I}^9, \boldsymbol{I}^{10}) \tag{14}$$

$$\boldsymbol{I}^{12} = \text{SE}(\boldsymbol{I}^{11}). \tag{15}$$

3) Stage 3—High-Level Multilayer Feature Fusion: In this stage, inspired by the classic VGG [45] and SSUN frameworks [33], middle-level features go through three pairs of convolution and pooling layers to extract the discriminative and abstract high-level features in a hierarchical manner. As shown in Table III, three high-level feature extraction architectures, that is, a 3×3 convolution layer with 128 kernels and the corresponding 2×2 MaxPool layer are first

stacked layer-by-layer as follows:

$$\boldsymbol{I}^{l+1} = F(\boldsymbol{I}^{l}) \tag{16}$$
$$\boldsymbol{I}^{l+2} = \text{MaxPool}(\boldsymbol{I}^{l+1}) \tag{17}$$
$$\boldsymbol{I}^{l+7} = \text{Flatten}(\boldsymbol{I}^{l+2}) \tag{18}$$
$$\boldsymbol{I}^{l+8} = \text{FC}(\boldsymbol{I}^{l+7}) \tag{19}$$

where $l \in \{12, 14, 16\}$. Then three pairs of convolution and MaxPool layers are followed by a flatten layer and an FC layer.

Finally, three high-level FC vectors are fused according to the add operation. Therefore, after the above sequential fusion operations, the high-level information at different layers is acquired to further enhance the semantic representation ability for land-objects in VHR images and thus improve the classification performance

$$\boldsymbol{I}^{25} = \text{Add}(\boldsymbol{I}^{20}, \boldsymbol{I}^{22}, \boldsymbol{I}^{24}). \tag{20}$$

C. Image Classification

In the final classification step, the obtained high-level features are first fed to the dropout layer to randomly discard half of the features, which can avoid over-fitting and enhance the stability and generalization ability of the model. After that, the representative and discriminative features are input into the FC layer with the Softmax classifier for classification, where cross-entropy is used as the loss function. For the given output feature vector $\boldsymbol{I}'$ and its category label $y \in \{1, 2, \ldots, C\}$, the probability distribution can be expressed as

$$P(y = c \mid \boldsymbol{I}') = \text{softmax}(\omega_c \boldsymbol{I}') = \frac{\exp(\omega_c \boldsymbol{I}')}{\sum_{c'=1}^{C} \exp(\omega_{c'} \boldsymbol{I}')} \tag{21}$$

where ω_c is the weight vector of the cth class. The final decision function of the Softmax can be formulated as follows:

$$\hat{y} = \arg\max_{c=1}^{C} P(y = c \mid \boldsymbol{I}') = \arg\max_{c=1}^{C} \omega_c \boldsymbol{I}' \tag{22}$$

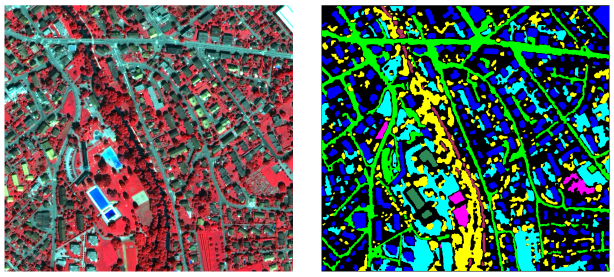
where $\hat{y}$ represents the predicted label of the feature vector $\boldsymbol{I}'$.

IV. Experimental Results and Analysis

A. Description of Datasets

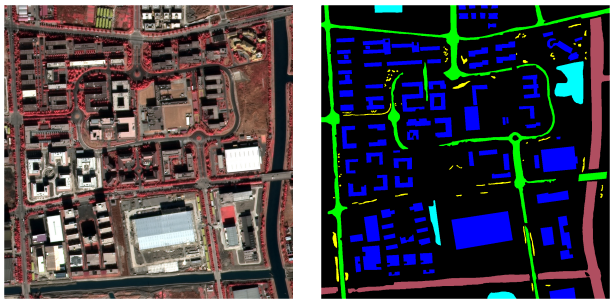
Experiments were conducted on three real VHR remote sensing datasets, including two satellite MS images, and one airborne HS image.

1) Zurich 17 (ZH17): The first dataset was acquired by the QuickBird satellite over the urban area of Zurich, Switzerland. The image contains 1025 × 1112 pixels and four spectral (blue, green, red, and near-infrared) bands with an approximate resolution of 0.62 m after the pansharpening operation. The false-color composite image and the corresponding ground reference map of the ZH17 dataset are visualized in Fig. 5. As shown in Table IV, in this scenario, there are seven land-cover classes including roads, buildings, trees, grass, bare soil, water, and swimming pools.



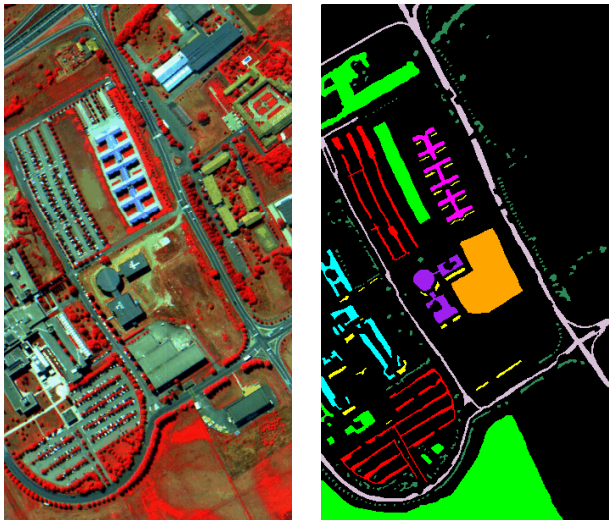
(a) (b)

Fig. 5. ZH17 dataset. (a) False color composite image (RGB: near-infrared, red, and green bands). (b) Ground reference map.



(a) (b)

Fig. 6. SH dataset. (a) False color composite image (RGB: near-infrared, red, and green bands). (b) Ground reference map.



(a) (b)

Fig. 7. UP dataset. (a) False color composite image (RGB: bands 90, 50, and 10). (b) Ground reference map.

TABLE IV
NUMBER OF SAMPLES OF THE ZH17 DATASET

No.	Class Name	Color	Samples (pixel)
1	Roads		154786
2	Buildings		150627
3	Trees		111072
4	Grass		129125
5	Bare Soil		10619
6	Water		9040
7	Swimming Pools		6052

2) Shanghai (SH): The second dataset was acquired by the Gaofen-2 satellite over the urban areas of

TABLE V
NUMBER OF SAMPLES OF THE SH DATASET

No.	Class Name	Color	Samples (pixel)
1	Buildings		195439
2	Roads		84444
3	Water		78043
4	Trees		11381
5	Grass		24868

TABLE VI
NUMBER OF SAMPLES OF THE UP DATASET

No.	Class Name	Color	Samples (pixel)
1	Asphalt		6631
2	Meadows		18649
3	Gravel		2099
4	Trees		3064
5	Painted metal sheets		1345
6	Bare soil		5029
7	Bitumen		1330
8	Self-blocking bricks		3682
9	Shadows		947

Shanghai, China. The image contains 1220 × 1200 pixels and four spectral (blue, green, red, and near-infrared) bands with a spatial resolution of 1 m after the pansharpening operation. Fig. 6(a) and (b) presents the false-color composite image and the ground reference map, respectively. There exist five land-cover classes in the study area (i.e., buildings, roads, water, trees, and grass). Detailed information on these classes is provided in Table V.

3) PaviaU (UP): The third dataset was acquired by the reflective optics systems imaging spectrometer (ROSIS) sensor at the Pavia University, northern Italy. This image consists of 103 spectral bands (wavelength from 0.43 to 0.86 μm) having a size of 610 × 340 pixels and a spatial resolution of 1.3 m. In order to remove the redundancy in spectral bands, the PCA transformation was performed on the original full bands, where the first four PCs that retain 99% information of the input bands were kept for classification. Fig. 7(a) and (b) presents the false-color composite image and the ground reference map, respectively. There are nine complex classes in this dataset, that is, asphalt, meadows, gravel, trees, painted metal sheets, bare soil, bitumen, self-blocking bricks, and shadows (more information on classes can be seen in Table VI).

B. Parameter Settings

To demonstrate the effectiveness of the proposed SDF^{2}N approach, six reference methods were compared on the three considered datasets, including two traditional ML-based classification methods, that is, support vector machines (SVM) and random forest (RF), and four state-of-the-art DL-based classification approaches, that is, FDSSC [32], SSUN [33], SSRN [34], and SpectralFormer [36]. For the SVM classifier, the radial basis function (RBF) was selected as the kernel function. For the RF classifier, the number of decision trees was set to 500. For three CNN-based reference networks FDSSC, SSUN, and SSRN, the spatial window size of the input was set as 32 × 32, the batch size was set to 128, and the value of epochs was defined as 100. For the SpectralFormer, the values of patches, band-epochs, and epochs were set to [7, 7, 480], [7, 3, 900], and [7, 7, 480] in ZH17, SH, and UP three datasets, respectively. For the proposed SDF^{2}N, the Adam optimizer with a learning rate of 0.001 was used for model training. In addition, the parameter settings of the input window, batch, and epoch were consistent with the aforementioned CNN-based reference methods. Finally, in order to keep consistency with the proposed method, the EMAP features were also used as input to the six reference methods.

The DL-based methods (i.e., FDSSC, SSUN, SSRN, SpectralFormer, and SDF^{2}N) were implemented by TensorFlow or PyTorch on an NVIDIA P40 GPU with 24-GB memory. The ML-based methods (i.e., SVM and RF) were implemented by MATLAB R2020b on a computer with Intel[1] Core[2] i5-7300 CPU, RAM 8 GB.

To quantitatively evaluate the classification performance among all compared methods, different indices such as the

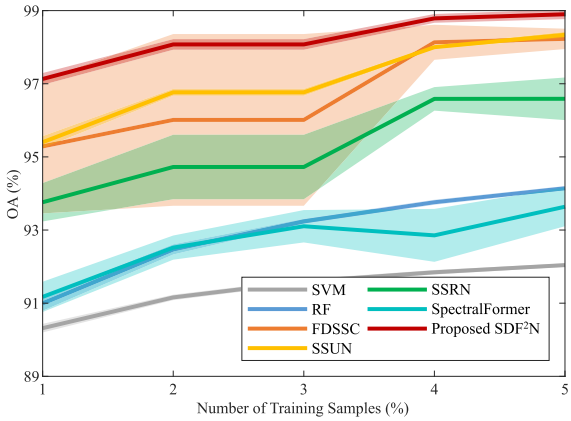


Fig. 8. OA results obtained by different methods with different numbers of randomly selected training samples (ZH17 dataset). Each curve represents the average OA after ten times of random sampling, and the shaded area represents the SD of ten OA values.

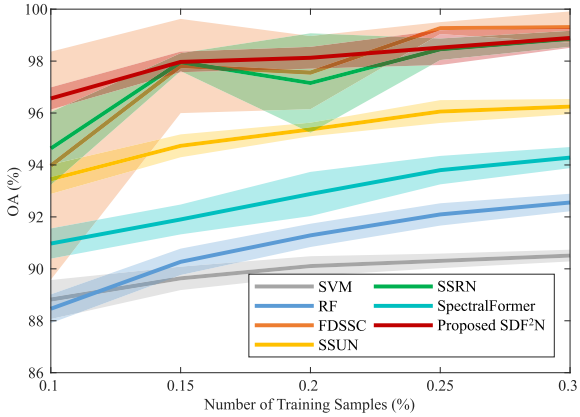


Fig. 9. OA results obtained by different methods with different numbers of randomly selected training samples (SH dataset). Each curve represents the average OA after ten times of random sampling, and the shaded area represents the SD of ten OA values.

[1]Registered trademark.
[2]Trademarked.

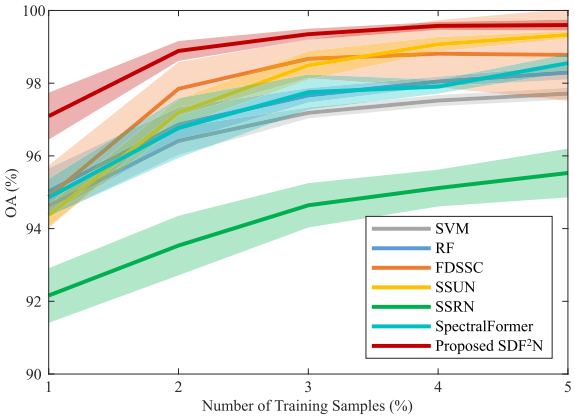


Fig. 10. OA results obtained by different methods with different numbers of randomly selected training samples (UP dataset). Each curve represents the average OA under ten times of random sampling, and the shaded area represents the SD of ten OA values.

overall accuracy (OA), the class accuracy (CA), the kappa coefficient (Kappa), and the computational time cost (T) were calculated. Final experimental results were obtained by repeat running ten times of each method with randomly generated training samples.

C. Experimental Results

In this section, we report the quantitative (see Figs. 8–10 and Tables VII–IX) and the qualitative (see Figs. 11–16) analysis of results obtained by the proposed approach and six reference methods on three VHR remote sensing datasets.

1) Results on the ZH17 Dataset: A detailed comparison of the classification performance of seven methods was done under different numbers of training samples, which were randomly selected as 1%, 2%, 3%, 4%, and 5% of samples for each class. Note that the randomization and classification were repeated ten times to assess average performance. The standard deviations (SD) of OA values are illustrated by the shaded areas in Fig. 8. In the figure, the red curve corresponds to the OA values obtained by the proposed SDF^{2}N method, which clearly shows higher values than the others with subtle fluctuation. Among four DL-based reference methods, FDSSC and SSUN outperformed SSRN and SpectralFormer. Unfortunately, the FDSSC performance fluctuated greatly under different random samplings. The SpectralFormer achieved the lowest accuracy but with stable performance. In addition, although the SVM and RF have the most stable performance, their accuracies are much lower than those of DL-based methods.

Table VII summarizes the classification accuracies obtained by different methods with 1% training samples. From the table, one can see that two ML-based methods obtained the worst average OA values among all considered methods (i.e., SVM: 90.32% and RF: 91.00%). Considering five DL-based methods, the proposed SDF^{2}N approach resulted in the highest classification accuracy (OA = 97.13%), significantly outperforming four reference deep networks, that is, FDSSC (95.29%), SSUN (95.40%), SSRN (93.76%), and SpectralFormer (91.17%). Moreover, the proposed SDF^{2}N approach produced the smallest SD value of OA ($\pm$0.16), indicating its stability.

The classification maps obtained by different methods under the first group of random samples are shown in Fig. 11. One subset highlighted in Fig. 11 is further compared in Fig. 12. We can see that the classification maps proposed by SVM and RF present much noise, which leads to a decrease in OA values. Classification maps obtained by FDSSC [see Fig. 12(c)], SSUN [see Fig. 12(d)], SSRN [see Fig. 12(e)], and SpectralFormer [see Fig. 12(f)] contain some confusion between roads (in green) and trees (in yellow). Compared with the ground reference map [see Fig. 5(b)], the best classification map is obtained by using the proposed SDF^{2}N approach that resulted in an OA = 97.14% [see Figs. 11(g) and 12(g)]. Most importantly, it is easy to observe that the proposed method alleviates the misclassification pixels in the edges and interiors of adjacent objects (e.g., roads and trees), thus confirming its superiority over the other six compared methods.

2) Results on the SH Dataset: Fig. 9 illustrates the OA values obtained by different methods with the number of training samples increasing from 0.1% to 0.3% of total samples. In Fig. 9, among seven curves, the proposed SDF^{2}N approach (represented by the red curve) achieved higher OA values with more stable performance. Due to the fact that on the SH dataset, the land-cover types are relatively simple, all DL-based methods except the SpectralFormer obtained high classification accuracies with a limited number of training samples. The SpectralFormer may be more suitable for HS image classification owing to its strong capability in spectral feature learning, rather than the large-scene VHR image especially the one with very few broad spectral bands. In the meantime, performance with a small number of training samples demonstrates the stability of the advanced network. Therefore, although the accuracies of the FDSSC method are slightly higher than those of the SDF^{2}N in some conditions, its overall fluctuation is more significant than that of SDF^{2}N. In addition, compared with other reference methods, SVM, RF, and SpectralFormer resulted in more stable performances while their accuracies are quite lower.

Table VIII reports the average classification accuracies and their SD values with 0.1% training samples. We can see that on this dataset, the obtained results are in line with the results of the previous dataset. In particular, despite the lower SD values and computational time costs, SVM, and RF resulted in the lowest average OA values which are equal to 88.82% and 88.46%, respectively. By taking advantage of the powerful capability of extracting high-level semantic features, all five DL-based methods obtained high classification accuracies even in the few-shot learning cases. Among them, the proposed SDF^{2}N achieves the highest accuracy (i.e., OA = 96.56%) with the smallest SD value (i.e., SD = 0.43).

Fig. 13 shows the classification maps obtained by different methods by using the first group samples. Fig. 14 further illustrates the local classification results of the areas highlighted in Fig. 13. From Fig. 14, one can see that the proposed SDF^{2}N method obtained the fewest misclassified pixels [see Fig. 14(g)]. Compared with the other six reference methods, the SDF^{2}N approach better models the object's external edges and internal homogeneity of similar classes, thus reducing the misclassification errors, such as buildings (in blue) and roads

TABLE VII
CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS ON THE ZH17 DATASET

Classes	SVM	RF	FDSSC	SSUN	SSRN	SpectralFormer	SDF^{2}N
Roads	89.62	89.54	97.11	95.29	95.97	87.76	97.48
Buildings	87.78	91.09	94.84	95.80	96.38	91.71	98.10
Trees	90.19	90.33	92.76	94.59	90.83	92.56	95.26
Grass	93.86	92.87	95.69	95.95	91.86	93.38	97.09
Bare soil	86.91	88.71	96.59	95.83	95.77	89.03	97.47
Railways	95.02	94.58	94.01	91.62	70.33	91.88	97.11
Swimming pools	97.04	96.59	96.88	96.35	97.72	95.41	98.80
OA	90.32 ±0.12	91.00 ±0.20	95.29 ±1.83	95.40 ±0.17	93.76 ±0.53	91.17 ±0.41	**97.13 ±0.16**
Kappa	87.38 ±0.16	88.26 ±0.27	93.85 ±2.40	94.01 ±0.69	91.86 ±0.69	88.50 ±0.54	96.26 ±0.21
T(s)	117.18 ±6.07	59.07 ±0.67	2606.99 ±41.62	300.80 ±3.28	1388.83 ±13.91	6545.93 ±122.48	485.59 ±11.93

TABLE VIII
CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS ON THE SH DATASET

Classes	SVM	RF	FDSSC	SSUN	SSRN	SpectralFormer	SDF^{2}N
Buildings	85.51	88.77	92.54	93.42	94.25	90.35	96.15
Roads	84.72	77.75	98.52	89.77	97.41	82.66	94.07
water	99.55	99.91	96.70	99.04	98.66	99.05	99.34
Trees	86.12	74.64	54.14	72.15	43.85	93.46	96.78
Grass	96.29	92.81	99.46	98.39	98.93	97.68	99.43
OA	88.82 ±0.75	88.46 ±0.55	93.97 ±4.39	93.45 ±0.57	94.64 ±1.40	90.97 ±0.58	**96.56 ±0.43**
Kappa	83.43 ±1.09	82.67 ±0.81	91.16 ±5.97	90.16 ±0.84	91.97 ±2.07	86.48 ±0.93	94.84 ±0.63
T(s)	14.37 ±0.90	26.86 ±1.06	490.15 ±22.10	129.87 ±2.80	325.17 ±25.52	6399.58 ±112.22	165.48 ±6.69

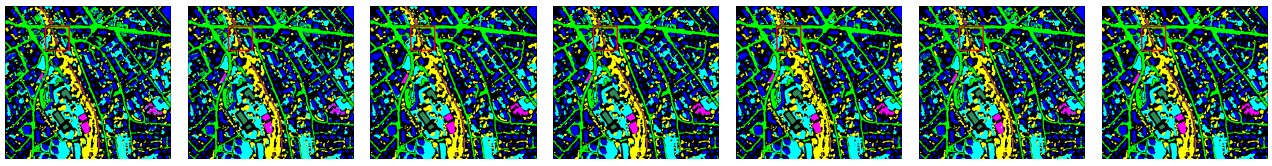
TABLE IX
CLASSIFICATION ACCURACIES (%) OBTAINED BY DIFFERENT METHODS ON THE UP DATASET

Classes	SVM	RF	FDSSC	SSUN	SSRN	SpectralFormer	SDF^{2}N
Asphalt	93.85	95.06	98.67	96.53	96.04	91.66	98.73
Meadows	99.53	99.25	99.36	99.81	99.61	99.93	99.93
Gravel	65.31	68.23	94.11	76.45	94.17	72.35	90.40
Trees	90.17	86.64	66.29	89.79	39.32	92.28	91.72
Painted metal sheets	98.85	98.34	99.29	94.00	100.00	99.95	99.11
Bare soil	98.22	99.25	99.87	99.60	100.00	98.23	99.97
Bitumen	75.96	87.43	97.11	67.15	92.58	78.33	81.91
Self-blocking bricks	91.24	90.35	97.29	90.49	96.88	90.60	95.47
Shadows	97.29	99.81	28.98	52.62	13.18	90.29	71.66
OA	94.63 ±0.31	95.02 ±0.63	94.88 ±0.86	94.36 ±0.29	92.16 ±0.75	94.86 ±0.52	**97.09 ±0.64**
Kappa	92.85 ±0.42	93.38 ±0.85	93.19 ±1.15	92.48 ±1.02	89.55 ±1.02	93.17 ±0.69	96.14 ±0.86
T(s)	13.92 ±2.51	14.35 ±0.25	211.12 ±21.82	26.47 ±0.74	115.03 ±8.63	451.74 ±11.66	39.31 ±1.21

(in green) classes, which are easy to be confused as shown in Figs. 13(g) and 14(g).

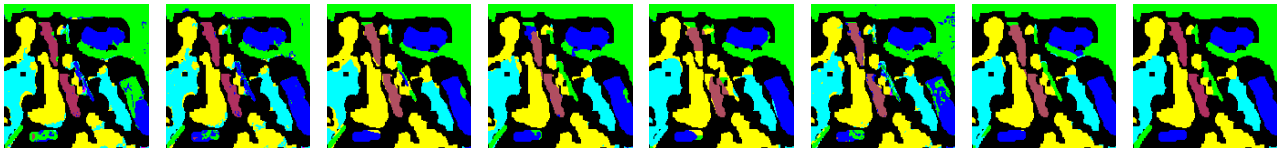
3) Results on the Up Dataset: Fig. 10 illustrates the accuracy obtained by different methods by using different numbers of samples. Compared with the previous two datasets, OA results better show the advantages of the proposed SDF^{2}N that provides the highest classification accuracy. On the contrary, the SSRN method achieved the lowest OA value. This may be due to the insufficient number of training samples in the model training process. In addition, the SpectralFormer obtained higher performance than in other datasets. This demonstrated it is more suitable for HS image classification with sufficient spectral information.

Table IX presents the average classification accuracies of the seven methods with 1% training samples, which are consistent with the quantitative results of the previous two datasets.



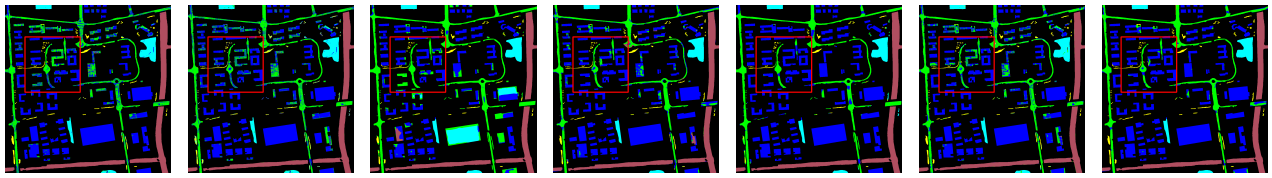
(a) (b) (c) (d) (e) (f) (g)

Fig. 11. Classification maps obtained by different methods on the ZH17 dataset. (a) SVM (90.23%). (b) RF (90.78%). (c) FDSSC (96.09%). (d) SSUN (95.67%). (e) SSRN (93.56%). (f) SpectralFormer (91.23%). (g) SDF^{2}N (97.14%).



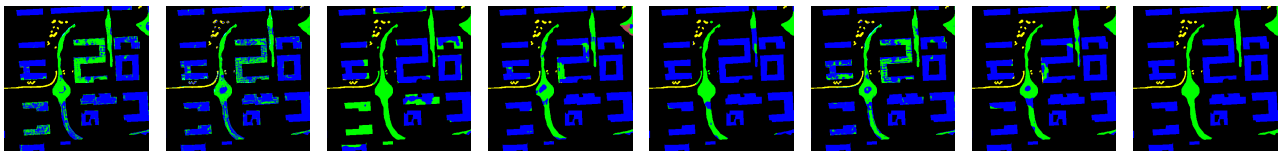
(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 12. Classification maps obtained by different methods at a local subset on the ZH17 dataset. (a) SVM. (b) RF. (c) FDSSC. (d) SSUN. (e) SSRN. (f) SpectralFormer. (g) SDF^{2}N. (h) Ground reference map.



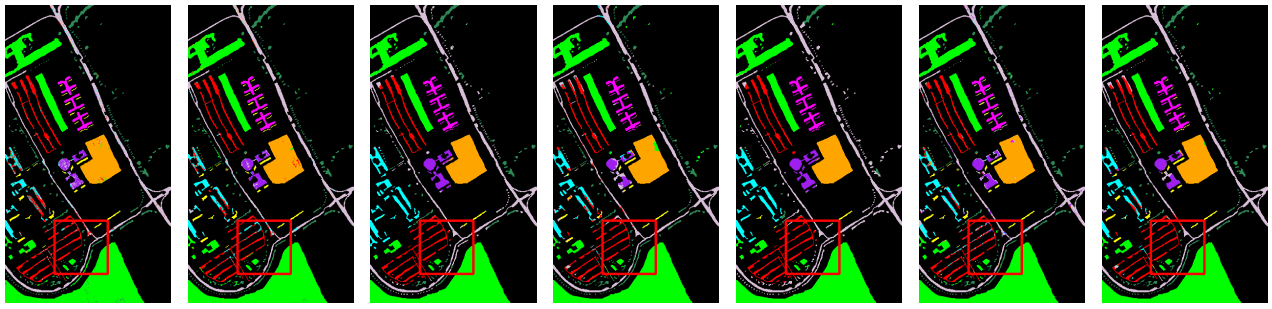
(a) (b) (c) (d) (e) (f) (g)

Fig. 13. Classification maps obtained by different methods on the SH dataset. (a) SVM (88.16%). (b) RF (88.57%). (c) FDSSC (82.02%). (d) SSUN (93.37%). (e) SSRN (96.19%). (f) SpectralFormer (90.82%). (g) SDF^{2}N (97.20%).



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 14. Classification maps obtained by different methods at a local subset on the SH dataset. (a) SVM. (b) RF. (c) FDSSC. (d) SSUN. (e) SSRN. (f) SpectralFormer. (g) SDF^{2}N. (h) Ground reference map.



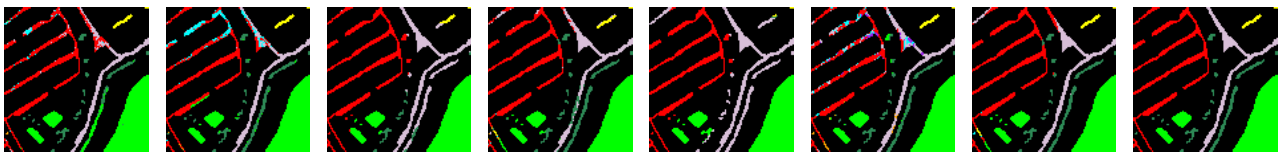
(a) (b) (c) (d) (e) (f) (g)

Fig. 15. Classification maps obtained by different methods on the UP dataset. (a) SVM (94.77%). (b) RF (95.67%). (c) FDSSC (94.61%). (d) SSUN (94.16%). (e) SSRN (91.55%). (f) SpectralFormer (95.51%). (g) SDF^{2}N (96.68%).

The proposed SDF^{2}N obtains the highest classification accuracy (OA = 97.09%) and relatively small standard deviation (SD = 0.64). In particular, the OA of the SDF^{2}N is higher by roughly 2%–5% than those of other methods and is obtained with a low computation cost (39.31 s).

Fig. 15 visualizes the classification maps of the first random sampling group obtained by different methods. In addition, subsets highlighted in the red rectangle in Fig. 15 are further compared in Fig. 16. It should be noted that the classification map obtained by the SDF^{2}N method presents more regular and correct classification results with fewer confusions among classes [see Figs. 15(g) and 16(g)]. It effectively reduces the misclassification especially for those complex objects with adjacent edges [e.g., asphalt (thistle) and trees (dark green)] and similar spectral characteristics [e.g., trees (dark green) and meadows (bright green)].

4) Ablation Study for the Proposed SDF^{2}N: To further validate the effectiveness of the proposed SDF^{2}N, a detailed ablation study was also made based on different combinations of the three fusion stages in the TSF module on the three



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 16. Classification maps obtained by different methods at a local subset on the UP dataset. (a) SVM. (b) RF. (c) FDSSC. (d) SSUN. (e) SSRN. (f) SpectralFormer. (g) SDF^{2}N. (h) Ground reference map.

TABLE X
ABLATION STUDIES OF DIFFERENT FUSION STAGES

Combination strategies	OA(%)		
	ZH17	SH	UP
Stage 1	95.44 ±0.19	94.28 ±0.89	95.65 ±0.76
Stage 2	95.81 ±0.23	94.75 ±0.40	95.74 ±0.62
Stage 3	96.37 ±0.16	94.57 ±0.65	95.54 ±0.57
Stage 1&2	96.46 ±0.25	95.55 ±0.80	96.85 ±0.57
Stage 1&3	96.98 ±0.16	95.25 ±1.37	96.68 ±0.46
Stage 2&3	96.26 ±0.41	95.35 ±0.64	96.38 ±0.42
Stage 1&2&3	**97.13 ±0.16**	**96.56 ±0.43**	**97.09 ±0.64**

datasets. As shown in Table X, when only a single stage is considered (see rows 1–3), relatively poor performance is obtained as the multilevel features are not fused and utilized. In addition, although the combinations of two fusion stages show better performance (see rows 4–6) than using a single stage, the proposed shallow-to-deep TSF structure (see row 7) resulted in the highest performance. This further demonstrates the effectiveness of the proposed SDF^{2}N that sequentially takes advantage of the three fusion stages to improve the classification performance in VHR images.

V. CONCLUSION

In this article, a novel shallow-to-deep feature fusion network (SDF^{2}N) has been proposed to hierarchically extract and fuse the saliency and discriminative features for VHR remote sensing image classification. Specifically, the SDF^{2}N contains three core feature fusion stages: 1) the low-level feature fusion stage, which is used to fuse the rich spectral–spatial features; 2) the middle-level feature fusion stage, which utilizes different size filters for integrating multiscale spatial context information; and 3) the high-level feature fusion stage, which includes three hierarchical layers for learning abstract and discriminative information. Compared with six popular and state-of-the-art reference methods, experimental results obtained on three real VHR remote sensing datasets confirmed the effectiveness of the proposed SDF^{2}N approach. It effectively alleviates the inaccurate identification problems of complex objects, especially in the high-detailed edges, and improves classification accuracy. In addition, the proposed SDF^{2}N approach has better model stability, especially in the small-sample cases, where it is superior to the other considered reference state-of-the-art methods.

For future developments, we will explore more efficient shallow-to-deep feature fusion modules for large complex scene classification in VHR satellite images.

ACKNOWLEDGMENT

The authors would like to thank Dr. Michele Volpi from the Swiss Federal Institute of Technology, Zürich, Switzerland, for providing the QB dataset; the China Centre for Resources Satellite Data and Application for providing the Gaofen-2 images; and Prof. Paolo Gamba of The University of Pavia, Pavia, Italy, for providing the University of Pavia hyperspectral dataset.

REFERENCES

[1] S. Liu, Y. Zheng, Q. Du, A. Samat, X. Tong, and M. Dalponte, "A novel feature fusion approach for VHR remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 464–473, 2021.

[2] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[3] S. Liu, Q. Du, X. Tong, A. Samat, and L. Bruzzone, "Unsupervised change detection in multispectral remote sensing images via spectral—spatial band expansion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3578–3587, Sep. 2019.

[4] F. Yang, W. Li, H. Hu, W. Li, and P. Wang, "Multi-scale feature integrated attention-based rotation network for object detection in VHR aerial images," *Sensors*, vol. 20, no. 6, p. 1686, Mar. 2020.

[5] J. R. Bergado, C. Persello, and A. Stein, "Recurrent multiresolution convolutional networks for VHR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6361–6374, Jun. 2018.

[6] Y. Tao, M. Xu, F. Zhang, B. Du, and L. Zhang, "Unsupervised-restricted deconvolutional neural network for very high resolution remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 6805–6823, Dec. 2017.

[7] S. Dong, Y. Zhuang, Z. Yang, L. Pang, H. Chen, and T. Long, "Land cover classification from VHR optical remote sensing images by feature ensemble deep learning network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1396–1400, Aug. 2020.

[8] A. Samat, C. Persello, S. Liu, E. Li, Z. Miao, and J. Abuduwaili, "Classification of VHR multispectral images using extratrees and maximally stable extremal region-guided morphological profile," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3179–3195, Sep. 2018.

[9] X. Kang, C. Li, S. Li, and H. Lin, "Classification of hyperspectral images by Gabor filtering based deep network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1166–1178, Apr. 2017.

[10] S. Liu *et al.*, "A multi-scale superpixel-guided filter feature extraction and selection approach for classification of very-high-resolution remotely sensed imagery," *Remote Sens.*, vol. 12, no. 5, p. 862, Mar. 2020.

[11] D. S. Maia, M.-T. Pham, E. Aptoula, F. Guiotte, and S. Lefevre, "Classification of remote sensing data with morphological attribute profiles: A decade of advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 43–71, Sep. 2021.

[12] E. Zhang, X. Zhang, H. Liu, and L. Jiao, "Fast multifeature joint sparse representation for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1397–1401, Jul. 2015.

[13] S. Niazmardi, A. Safari, and S. Homayouni, "A novel multiple kernel learning framework for multiple feature classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3734–3743, Aug. 2017.
[14] J. Wang, Y. Zheng, M. Wang, Q. Shen, and J. Huang, "Object-scale adaptive convolutional neural networks for high-spatial resolution remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 283–299, 2021.
[15] A. Ma *et al.*, "Fast sequential feature extraction for recurrent neural network-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5920–5937, Jul. 2021.
[16] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[17] 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.
[18] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and LiDAR data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.
[19] J. Li, H. Zhang, and L. Zhang, "Supervised segmentation of very high resolution images by the use of extended morphological attribute profiles and a sparse transform," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1409–1413, Aug. 2014.
[20] L. Y. Fang, S. T. Li, X. D. Kang, and J. A. Benediktsson, "Spectral—spatial classification of hyperspectral images with a superpixel-based discriminative sparse model," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4186–4201, Aug. 2015.
[21] C. Zhao, X. Gao, Y. Wang, and J. Li, "Efficient multiple-feature learning-based hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 4052–4062, Jul. 2016.
[22] S. Jia, J. Hu, Y. Xie, L. Shen, X. Jia, and Q. Li, "Gabor cube selection based multitask joint sparse representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3174–3187, Jun. 2016.
[23] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.
[24] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.
[25] L. Shu, K. McIsaac, and G. R. Osinski, "Learning spatial–spectral features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5138–5147, Sep. 2018.
[26] Y. Gu, K. Feng, and H. Wang, "Spatial—spectral multiple kernel learning for hyperspectral image classification," in *Proc. 5th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens. (WHISPERS)*, Jun. 2013, pp. 1–4.
[27] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Jun. 2020.
[28] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.
[29] Z. Zhong, J. Li, D. A. Clausi, and A. Wong, "Generative adversarial networks and conditional random fields for hyperspectral image classification," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3318–3329, Jul. 2020.
[30] Y. Zheng, S. Liu, Q. Du, H. Zhao, X. Tong, and M. Dalponte, "A novel multitemporal deep fusion network (MDFN) for short-term multitemporal HR images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10691–10704, 2021.
[31] S. Liu, H. Zhao, Q. Du, L. Bruzzone, A. Samat, and X. Tong, "Novel cross-resolution feature-level fusion for joint classification of multispectral and panchromatic remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.
[32] W. Wang, S. Dou, Z. Jiang, and L. Sun, "A fast dense spectral–spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, p. 1068, 2018.
[33] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.
[34] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[35] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
[36] D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[37] P. Ghamisi, J. Atli Benediktsson, G. Cavallaro, and A. Plaza, "Automatic framework for spectral–spatial classification based on supervised feature extraction and morphological attribute profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2147–2160, Jun. 2014.
[38] Z. Zhang, L. Yang, and Y. Zheng, "Multimodal medical volumes translation and segmentation with generative adversarial network," in *Handbook of Medical Image Computing and Computer Assisted Intervention* (The Elsevier and MICCAI Society Book Series), S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. New York, NY, USA: Academic, 2020, ch. 8, pp. 183–204.
[39] S. Shajun Nisha and M. N. Meeral, "Applications of deep learning in biomedical engineering," in *Handbook of Deep Learning in Biomedical Engineering*, V. E. Balas, B. K. Mishra, and R. Kumar, Eds. New York, NY, USA: Academic, 2021, ch. 9, pp. 245–270.
[40] Z. Ge, G. Cao, X. Li, and P. Fu, "Hyperspectral image classification method based on 2D–3D CNN and multibranch feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5776–5788, 2020.
[41] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.
[42] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Jun. 2020.
[43] A. Alshehri, Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan, "Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery," *IEEE Access*, vol. 7, pp. 119873–119880, 2019.
[44] M. Liang, Z. Ren, J. Yang, W. Feng, and B. Li, "Identification of colon cancer using multi-scale feature fusion convolutional neural network based on shearlet transform," *IEEE Access*, vol. 8, pp. 208969–208977, 2020.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556v6*.

Sicong Liu (Senior Member, IEEE) received the B.Sc. degree in geographical information system and the M.E. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2009 and 2011, respectively, and the Ph.D. degree in information and communication technology from the University of Trento, Trento, Italy, in 2015.

He is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University, Shanghai, China. His research interests include multitemporal data analysis, change detection, multi-spectral/hyperspectral remote sensing in earth observation, and planetary exploration.

Dr. Liu was a Winner (ranked in third place) of the Paper Contest of the 2014 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest. He is the Technical Co-Chair of the Tenth International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp 2019). He serves as the Program Committee Member for SPIE Remote Sensing Symposium: Image and Signal Processing for Remote Sensing XXVI–XXVIII from 2020 to 2022 and also served as the Session Chair for many international conferences such as International Geoscience and Remote Sensing Symposium from 2017 to 2019. He is/was a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) and *Remote Sensing*.

Yongjie Zheng received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2018, and the M.S. degree in photogrammetry and remote sensing from Tongji University, Shanghai, China, in 2021.

Her research interests include deep learning, feature extraction and fusion, and multispectral/hyperspectral image classification/change detection.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore, Baltimore, MD, USA, in 2000.

She is currently the Bobby Shackouls Professor of the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, and machine learning.

Dr. Du has been a member of the IEEE Periodicals Review and Advisory Committee and SPIE Publications Committee. She is a fellow of the SPIE-International Society for Optics and Photonics (SPIE). She was a recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society (GRSS). She was the Co-Chair of the Data Fusion Technical Committee of the IEEE GRSS from 2009 to 2013, the Chair of the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014, and the General Chair of the Fourth IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing held at Shanghai, China, in 2012. She has served on the editorial board for many journals, such as IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, *Journal of Applied Remote Sensing*, IEEE SIGNAL PROCESSING LETTERS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, and *Remote Sensing* (MDPI). From 2016 to 2020, she was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree *(summa cum laude)* in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is the Founder and the Director of the Remote Sensing Laboratory (https://rslab.disi.unitn.it/), Department of Information Engineering and Computer Science, University of Trento. He is the Principal Investigator of many research projects. Among the others, he is also the Principal Investigator of the Radar for icy Moon exploration (RIME) instrument in the framework of the JUpiter ICy moons Explorer (JUICE) mission of the European Space Agency (ESA) and the Science Lead of the High Resolution Land Cover Project in the framework of the Climate Change Initiative of ESA. He is the author (or a coauthor) of 294 scientific publications in referred international journals (221 in IEEE journals), more than 340 papers in conference proceedings, and 22 book chapters. He is the editor/coeditor of 18 books/conference proceedings and one scientific book. His papers are highly cited, as proved by the total number of citations (more than 41 000) and the value of the H-index (95) (source: Google Scholar). He was invited as a keynote speaker at more than 40 international conferences and workshops. His research interests are in the areas of remote sensing, radar and synthetic aperture radar (SAR), signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects.

Dr. Bruzzone has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) since 2009, where he has been the Vice-President of professional activities since 2019. He ranked First Place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, July 1998. He was a recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, and the 2019 WHISPER Outstanding Paper Award. He was a Guest Coeditor of many special issues of international journals. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) series and is currently a member of the Permanent Steering Committee of this series of workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the Founder of *IEEE Geoscience and Remote Sensing Magazine* for which he was the Editor-in-Chief from 2013 to 2017. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He was a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society from 2012 to 2016.

Alim Samat (Member, IEEE) received the B.S. degree in geographic information system from Nanjing University, Nanjing, China, in 2009, the M.S. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2012, and the Ph.D. degree in cartography and geographic information system from Nanjing University, in 2015.

He is currently an Associate Researcher with the State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Ürümqi, China. His research interests include PolSAR and optical remote sensing for land applications, image processing and pattern recognition, and machine learning.

Dr. Samat serves as a Reviewer for several international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and *Pattern Recognition*.

Xiaohua Tong (Senior Member, IEEE) received the Ph.D. degree from Tongji University, Shanghai, China, in 1999.

He has worked as a Post-Doctoral Researcher with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, from 2001 to 2003. He was a Research Fellow with The Hong Kong Polytechnic University, Hong Kong, in 2006, and a Visiting Scholar with the University of California at Santa Barbara, Santa Barbara, CA, USA, from 2008 to 2009. He is currently a Professor with the College of Surveying and Geoinformatics, Tongji University. His research interests include remote sensing, geographic information system (GIS), uncertainty and spatial data quality, image processing for high resolution, and hyperspectral images.

Dr. Tong serves as the Vice-Chair for the Commission on Spatial Data Quality of the International Cartographical Association and the Co-Chair for the ISPRS Working Group (WG II/4) on spatial statistics and uncertainty modeling.

Yanmin Jin received the B.S. degree from the Taiyuan University of Technology, Taiyuan, China, in 2007, and the Ph.D. degree from Tongji University, Shanghai, China, in 2015.

She is currently an Associate Professor with the College of Surveying and Geo-Informatics, Tongji University. Her research interests are in GIS spatial data processing and data quality control.

Chao Wang received the B.S. degree in remote sensing and geography information system from the China University of Mining and Technology, Xuzhou, China, in 2010, and the Ph.D. degree in cartography and geography information system from East China Normal University, Shanghai, China, in 2016.

From 2014 to 2016, he was with the Laboratoire de Météorologie Dynamique/IPSL, CNRS, Sorbonne Université, Paris, France, funded by the China Scholarship Council, Beijing, China, as a joint Ph.D. Student. He is currently an Assistant Professor with Tongji University, Shanghai. His research interests include hyperspectral remote sensing and planetary atmosphere.