Novel Cross-Resolution Feature-Level Fusion for Joint Classification of Multispectral and Panchromatic Remote Sensing Images

Sicong Liu, *Senior Member*, *IEEE*, Hui Zhao, *Student Member*, *IEEE*, Qian Du, *Fellow*, *IEEE*, Lorenzo Bruzzone, *Fellow*, *IEEE*, Alim Samat, *Member*, *IEEE* and Xiaohua Tong, *Senior Member*, *IEEE*

***Abstract*—With the increasing availability and resolution of satellite sensor data, multispectral (MS) and panchromatic (PAN) images are the most popular data that are used in remote sensing among applications. This paper proposes a novel cross-resolution hidden layer features fusion (CRHFF) approach for joint classification of multi-resolution MS and PAN images. In particular, shallow spectral and spatial features at a global scale are firstly extracted from a MS image. Then deep cross-resolution hidden layer features extracted from MS and PAN are fused from patches at a local scale according to an Autoencoder (AE) like deep network. Finally, the selected multi-resolution hidden layer features are classified in a supervised manner. By taking advantage of integrated shallow-to-deep and global-to-local features from the high-resolution MS and PAN images, the cross-resolution latent information can be extracted and fused in order to better model imaged objects from the multi-model representation, and finally increase the classification accuracy. Experimental results obtained on three real multiresolution data sets covering complex urban scenarios confirm the effectiveness of the proposed approach in terms of higher accuracy and robustness with respect to literature methods.**

***Index Terms*—Multi-resolution images, feature-level fusion, remote sensing, shallow and deep features, classification.**

I. Introduction

Nowadays, due to the increasing satellite sensor data availability and quality, Earth's land cover/use change detection and classification at a fine resolution have received more attentions [1-4]. Many emerging new applications in urban, agriculture, disaster, forestry fields require the full use and fusion of multi-source or multitemporal remote sensing images in order to exploit complementary information and promote identification accuracy [5-10]. In the current scenario, there are many Earth-Observation satellites that can simultaneously acquire multi-model images in the same scene, among which the multispectral (MS) and panchromatic (PAN) images are the most widely used data for fusion in the practical applications. Different from the traditional moderate-resolution satellites, such as Landsat or SPOT families, newly launched satellites can acquire high-resolution (HR) and even very-high-resolution (VHR) MS and PAN images. This provides a great opportunity as well as challenges to properly integrate the multiresolution data, promoting their applications at a fine level. Table 1 lists some examples of these satellites (with a spatial resolution of MS image > 5m) and their corresponding parameters.

Table 1. Examples of EO satellites that simultaneously acquire MS and PAN images.

Country	**Satellites**	**MS**	**PAN**	**Launch Time**
		(meter/pixel)		
USA	IKONOS	4	1	1999
	QuickBird	2.44	0.61	2001
	GeoEye-1	1.65	0.41	2008
	WorldView-2	1.84	0.46	2009
	WorldView-3, -4	1.24	0.31	2014, 2016
	Planet Labs	5	3	2014
Spain	DEIMOS-2	4	1	2013
France	Pleiades-1, -2	2	0.5	2011, 2012
UK	TripleSat	3.2	0.8	2015
South Korea	KOMPSAT-3	2.8	0.7	2012
	KOMPSAT-3A	2.2	0.55	2015
China	GaoFen 2	4	1	2014
	GaoFen 6	8	2	2018
	GaoFen 7	3.2	0.8	2019
	Super-View-1 01/02/03/04	2	0.5	2016, 2018
	JiLin-1	2.88	0.72	2015

Usually, MS sensors collect data in red, green, blue, and near-infrared four bands, with relatively lower spatial resolution compared to PAN sensors, due to physical

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0505000, in part by the Natural Science Foundation of China under Grants 42071324, 42071424 and by the Shanghai Rising-Star Program (21QA1409100) (*Corresponding authors: Qian Du, Hui Zhao*)

S. Liu, Hui Zhao and Xiaohua Tong are with the College of Surveying and Geoinformatics, Tongji University, Shanghai, 200092, China (e-mail: sicong.liu@tongji.edu.cn, zhaohui@tongji.edu.cn, xhtong@tongji.edu.cn).

Q. 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, I-38123, Italy (e-mail: lorenzo.bruzzone@unitn.it).

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

limitations and technical constraints of onboard storage and bandwidth transmission [11]. In contrast, PAN image with only a single broad band has a much higher spatial resolution [12]. Accordingly, MS image is usually used for identifying different types of land objects, while the PAN image can accurately describe the geometrical properties of objects, which is of great benefit to image interpretation at high-resolution. The fusion of these two images can utilize both spatial and spectral information to further increase identification capability [13]. However, with the unprecedentedly increased spatial resolution, especially for the HR and VHR PAN/MS images that reach a sub-meter/meter level, their effective fusion becomes a very important yet challenging task.

In general, methods for fusing MS and PAN images include pixel-level fusion which is known as pan-sharpening (PS) techniques, and feature-level fusion methods. For the former, traditional algorithms include component substitution (CS), multi-scale decomposition-based method, hybrid method, and model-based algorithms [14]. In [13], five PS algorithms (*i.e.*, Gram-Schmidt, Principal Component Analysis, High Pass Filter, Wavelet Transform, Generalized Intensity-Hue-Saturation) and decision fusion were designed, and their impacts on the performance of change detection were compared and analyzed. In [15], eight advanced PS methods including various state-of-the-art and advanced Deep Learning (DL) methods were studied through the task of anomaly detection. In recent years, DL methods have also been also used for PS [16-18]. The basic idea is to train a PS model between the fused image and the observations based on a DL architecture, then the model is used to construct the final fused MS image. However, the PS process inevitably introduces spectral and spatial distortions in the resultant fused MS image, which influence the final detection or classification results [19]. Despite a DL-based PS method may achieve desirable results with less spectral distortion, it requires more prior data to train a robust network [20]. For the latter, feature-level fusion methods first extract representative features from MS and PAN images, then integrate these features via a robust fusion model for further classification or detection. Accordingly, they are more straightforward for applications without producing a pan-sharpened image, and avoid the limitations of pixel-level fusion methods to some extent. In [21], texture features (*i.e.*, homogeneity, contrast and entropy) were extracted from PAN images using the co-occurrence matrix, and spectral features (*i.e.*, normalized band values) were calculated from MS images. Then, object-based classification using the standard nearest neighbor was applied as a fusion analysis for forest type classification. In [22], a graph cut method was combined with the linear mixture model, and MS and PAN data were integrated to generate a context classification map. In [23], a unified Bayesian framework was presented to iteratively discover semantic segments from a PAN image and inferring cluster labels for the segments from a MS image to obtain the classification maps.

The above feature-level fusion methods mainly focus on artificial features that require a domain expert's knowledge. On the contrary, DL-based techniques that can automatically learn abstract and robust deep features from the original data are becoming a very promising way for dealing with the fusion of MS and PAN images at feature level. Within this context, in [24], a super-pixel based multiple local network model was proposed to classify MS image; then a PAN image was used to fine tuning the classification results. In [25], a stacked autoencoder was used to extract the spectral features from a MS image, and a Convolutional Neural Network (CNN) was used to extract spatial features from a PAN image; then spectral and spatial features were concatenated to obtain final classification results. In [26], two CNN modules inspired from the VGG model were designed for MS and PAN images at their original resolution; then, they were combined to perform land-cover classification. In [27], a novel framework was proposed via 3-D and 2-D adaptive multi-scale convolutional networks and a perceptual loss function for MS and PAN images classification. In [28], based on a data-driven DL, a spatial attention module (SA-module) for PAN images and a channel attention module (CA-module) for MS images were designed to extract the features that were then fused. In [29], a local spatial attention module (LSA-module) for the PAN image and a global channel attention module (GCA-module) for the MS image were designed; then an interaction module effectively reduced the differences in the characteristics obtained by the PAN branch and the MS branch. Then the GCA-module was used to further enhance feature representation from the fused features for classification.

The above existing feature-level fusion methods are proven to potentially outperform pixel-level fusion methods. However, there are still some open issues that require further investigation: 1) In the current DL-based methods, the patch is fixed as a rectangle; thus the spatial integrity and connectivity of land objects may be mishandled. 2) In the two-branch deep network fusion methods (*e.g.*, in [25, 26, 28, 29]), high-level abstract features in the last layer of each branch are concatenated. Thus, the cross-resolution representation in the middle-layer features is ignored. How to properly extract and utilize the intermediate cross-resolution information has not been fully investigated. 3) In feature-level fusion methods (*e.g.*, in [25]), the multiresolution MS and PAN images require a resampling operation, which will inevitably introduce interpolation errors, and increase data processing burden.

By considering the aforementioned open issues, in this paper, we propose a cross-resolution hidden layer feature fusion (CRHFF) approach to HR/VHR MS and PAN image classification. To the best of our knowledge, there is no similar work in the literature that deals with the same task. Main contributions of this paper are highlighted as follows.

1) By taking advantage of shallow-to-deep integration and global-to-local features in the proposed CRHFF approach, the inconsistent feature representation problem of the local patches can be solved, where the objects can be modeled in a more comprehensive way, while increasing the classification accuracy. Moreover, shallow-to-deep feature extraction procedure is designed in an unsupervised and automatic fashion, which makes it very interesting for practical applications.

2) The spatial and spectral information in MS and PAN

images can be neutralized through the novel autoencoder-like deep network. Intermediate hidden layer features at different resolutions are fused using a multi-branch CNN. This leads to a more detailed and precise cross-resolution feature representation than the traditional pan-sharpening or features stacking, thus further enhancing the classification performance.

3) Different from the conventional operation where a MS image is first up-sampled, in this work the proposed architecture is built by considering the low-resolution shallow features from MS image as input and the high-resolution PAN image as output. The cross-resolution conversion is made during the process of network training, where the bias and computational burden introduced by the up-sampling process are significantly reduced.

The rest of this paper is organized as follows. The proposed CRHFF approach is described in details in Section II. Data sets used in experiments are introduced in Section III. Experimental results and the related analysis are presented in Section IV. Finally, Section V draws the conclusions.

II. Proposed Feature-Level Fusion Approach

The proposed CRHFF approach aims to extract and fuse the global-to-local and shallow-to-deep features hidden in multi-resolution MS and PAN images for classification. Fig. 1 shows its block diagram that mainly consists of three steps: 1) shallow spectral-spatial feature extraction at a global scale, 2) deep multi-hidden layer feature extraction at a local scale, and 3) cross-resolution feature fusion and classification.

Step 1: Shallow Spectral-spatial Feature Extraction

Features used for image classification can be shallow features (*i.e.*, artificial features extracted from the original image by some specific image processing operations) or deep features (*i.e.*, extract from a deep-network by DL approaches) [30]. In order to take fully advantages of the context information in the HR/VHR images, spectral-spatial shallow spatial features are usually considered, such as the extended multi-attribute profile (EMAP) [31, 32], extinction profile [33], edge-preserving filtering features [34], Gabor features [35], superpixel-guided filter features [36], etc. In the proposed CRHFF approach, we selected EMAP as an example of shallow spatial features, which are combined with the original MS bands as extended shallow spectral features. Note that such shallow spectral-spatial features focus on the global representation of image objects and consider their integrity and connectivity, which will benefit the deep local information extraction and fusion in the deep feature generation step. Other effective shallow features can also be integrated in the proposed framework.

As shown in Step 1 in Fig. 1, we define the size of input MS and PAN images as $H \times D \times c$ and $nH \times nD \times 1$, respectively, where H and D represent the height and weight of the MS image, respectively, c is the number of MS bands, n is the resolution ratio between PAN and MS images, and t represents the number of EMAP features based on all MS bands, which are described in detail as follows.

Attribute profiles (APs) are an extension of the widely used morphological profiles (MPs). The AP operation replaces the structural elements of traditional morphological operation with general attribute criteria, which can reflect the structural characteristics of objects more effectively. In particular, APs are obtained by processing a scalar grayscale image α, according to a criterion T, with m attribute thickening (ϕ^T) operators and m attribute thinning (γ^T) operators, instead of the conventional morphological filters by reconstruction [32]:

$$AP(\alpha) = \left\{ \phi_m^T(\alpha), \phi_{m-1}^T(\alpha), \cdots, \phi_1^T(\alpha), \alpha, \gamma_1^T(\alpha), \cdots, \gamma_{m-1}^T(\alpha), \gamma_m^T(\alpha) \right\} \qquad (1)$$

Extended attribute profiles (EAPs) are built based on APs, and EMAP is the combination of different EAPs [31, 32]. In this work, we compute the APs on each band of the MS image, so the corresponding EAP can be expressed as:

$$EAP = \left\{ AP(g_1), AP(g_2), \cdots, AP(g_c) \right\} \qquad (2)$$

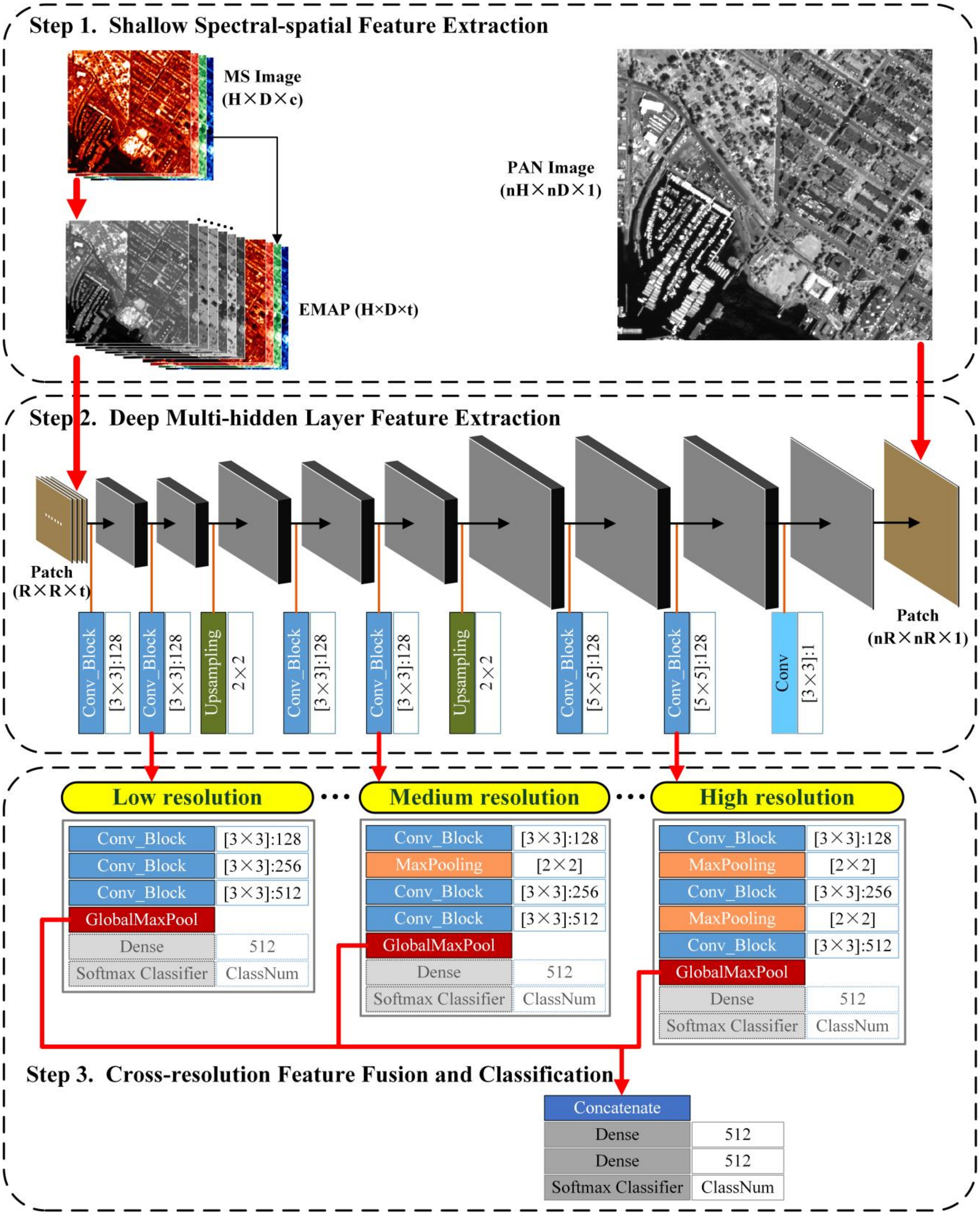
where $g_1, g_2, \cdots, g_c$ are the MS bands. In particular, in this work the following four attributes are selected in APs: 1) area of the regions a; 2) length of the diagonal of the box bounding the region d; 3) first moment invariant of Hu, moment of inertia i; 4) standard deviation of the gray-level values of the pixels in the regions s. For each individual attribute profile, EAP can be expressed as EAP_a, EAP_d, EAP_i, EAP_s, respectively. Then the final EMAP can be formulated as:

$$EMAP = \left\{ EAP_a, EAP_d, EAP_i, EAP_s \right\} \qquad (3)$$

Step 2: Deep Multi-hidden Layer Feature Extraction

Differently from the simple concatenation of MS and PAN features based on a two-branch structure deep network used in the literature, we extract cross-resolution latent features of MS and PAN images through an end-to-end deep network using an Autoencoder (AE) architecture. The AE network was first proposed to reduce data dimensionality [37]. The architecture of an AE involves an encoder and a decoder. The former converts the input into a hidden representation that only keeps the most representative information, and the latter recovers the input data from the hidden representation. Accordingly, the hidden representation can be viewed as the input features for the reconstruction. In recent years, the AE networks have been widely applied to image super-resolution [38, 39] and pan-sharpening [40].

In the original AE, the input and recovered data are exactly the same. In the proposed AE-like deep network, the input data are the patches derived from the EMAP, while the recovered data are the patches derived from PAN image. In the process of training AE-like deep network, low-resolution patches of EMAP (denoted as $x(Patches_{EMAP})$) are automatically aligned to the size of high-resolution patches of PAN image (denoted as $x(Patches_{PAN})$). In Step 2 of Fig. 1, the conversion between the multiresolution $x(Patches_{EMAP})$ and $x(Patches_{PAN})$ is illustrated in detail. Each pair of $x(Patches_{EMAP})$ ($R \times R \times t$) and $x(Patches_{PAN})$ ($nR \times nR \times 1$) is acquired over the same area to ensure the highly correlation between the two types of source data. Let us assume $\hat{x}(Patches_{EMAP})$ denotes the reconstructed data of $x(Patches_{EMAP})$ through convolutional layers and up-sampling layers. The energy function of the reconstruction error is defined as:



Conv_Block	[3×3]:128
Conv_Block	[3×3]:256
Conv_Block	[3×3]:512
GlobalMaxPool	
Dense	512
Softmax Classifier	ClassNum

Conv_Block	[3×3]:128
MaxPooling	[2×2]
Conv_Block	[3×3]:256
Conv_Block	[3×3]:512
GlobalMaxPool	
Dense	512
Softmax Classifier	ClassNum

Conv_Block	[3×3]:128
MaxPooling	[2×2]
Conv_Block	[3×3]:256
MaxPooling	[2×2]
Conv_Block	[3×3]:512
GlobalMaxPool	
Dense	512
Softmax Classifier	ClassNum

Concatenate	
Dense	512
Dense	512
Softmax Classifier	ClassNum

Fig. 1 Block diagram of the proposed CRHFF approach.

$$J = \frac{1}{k}\sum_{j=1}^{k}\left[\left(x\left(Patches_{PAN}\right)\right)^{j} - \left(\hat{x}\left(Patches_{EMAP}\right)\right)^{j}\right]^{2} \qquad (4)$$

where k is the number of $x(Patches_{EMAP})$.

Specifically, there are six conv_bolck, two up-sampling layers and one convolutional layer in Step 2, where conv_block contains the convolutional layer, the batch normalization and an activation function. Parameters of kernel and feature maps are [3×3]:128, which means that the kernel size of the convolution is 3×3, and 128 feature maps are generated. Note that resolutions are different between $x(Patches_{EMAP})$ and $x(Patches_{PAN})$, thus two up-sampling layers are used after conv_block with an up-sampling factor for rows and columns of 2×2. Thus, $x(Patches_{EMAP})$ are resampled into a unified spatial resolution of PAN image, whereas the hidden layers contain different resolution features.

There are two kinds of convolution kernels (*i.e.*, 3×3 and 5×5). In the low-resolution and medium-resolution feature maps, a 3×3 convolution kernel is used, and while a 5×5 convolution kernel is used in the last high-resolution feature maps. This is due to the fact that usually low-resolution feature maps have small size patches, while high-resolution feature maps have larger size patches. Since the dimensionality of feature maps from the last convolutional layer is different from the single band PAN image, the last convolutional layer with one feature map is used to make them consistent.

This deep multi-hidden layer feature extraction method is unsupervised. Here, and the number of training samples of the network is also the number of patches, rather than labeled samples of classes. In order to enlarge the number of training samples, each pixel in EMAP is used to generate training patches, and zero values are filled up for boundaries of EMAP.

Step 3: Cross-resolution Feature Fusion and Classification

The deep hidden layer features extracted between $x(Patches_{EMAP})$ and $x(Patches_{PAN})$ in Step 2 represent distinct latent features cross different scales in two data. Lower-resolution feature maps close to $x(Patches_{EMAP})$ contain more homogeneous spectral-spatial information, which is beneficial to identify pixels in the same object. Higher-resolution feature maps close to $x(Patches_{PAN})$ contain more detailed spatial features, which are useful for fine classification. In order to take advantage of multi-resolution deep hidden features, they are fused in Step 3 as shown in Fig. 1. In particular, we choose three hidden layers, *i.e.*, layers 2, 5, 8, of size $R\times R$, $2R\times 2R$, $4R\times 4R$, respectively. Three parallel CNN modules are designed to extract deep features at different resolutions. Then, the three deep feature sets are connected followed by two dense layers, and the SoftMax classifier is applied to perform the final classification.

The convolution process for the parallel CNN modules is described as follows. Let F_{hidden_u} (u = 1, 2, 3) be the input features of the three CNN modules. The output in the lth layer can be written as:

$$Z^l = \begin{cases} W^l \otimes F_{hidden\_u} + b^l, l = 1 \\ W^l \otimes Z^{l-1} + b^l, l = 2, \dots p \end{cases} \tag{5}$$

where W is the weight, b is the bias, and $\otimes$ denotes the convolution operation. Then a batch normalization (BN) layer to accelerate network convergence and mitigate gradient explosion or vanishing problem is added over the output Z^l; it can be denoted as $BN(Z^l)$. Before importing $BN(Z^l)$ into the next block, a Rectifier Linear Unit (ReLU) activation function is implemented:

$$\mathrm{Re}\,LU\left(BN\left(Z^l\right)\right) = Max\left(0, BN\left(Z^l\right)\right) \tag{6}$$

Output features in the last fully-connected layer are then transformed into a probability distribution for specific categories, where the cross entropy is used to measure the prediction loss of the network. To minimize the loss function, the stochastic gradient descent algorithm is adopted to update parameters and to optimize the model.

It is worth noting that unlike Step 2 that is unsupervised, Step 3 is a supervised process. As shown in Fig. 1, three parallel CNN modules are similar, and the only difference is the number of Maxpooling layers. Considering different patch sizes of hidden features, there are no Maxpooling layers in branch 1 (hidden layer 2), one Maxpooling layer in branch 2 (hidden layer 5), and two Maxpooling layers in branch 3 (hidden layer 8). Accordingly, after several convolutional layers, the size of the feature maps in different branches remain the same.

As in Step 2, conv_block contains a convolutional layer, a batch normalization and an activation function. Parameters of each layer are [3×3]:128, which means that the kernel size of the convolution is 3×3, and 128 feature maps are produced. All max pooling layers are implemented with a polling size of 2×2 with a stride equal to 2. There is a global max pooling layer followed by the last convolution layer of each branch. The global max pooling operation extracts one feature from each feature map. It acts as a high-pass filter and reduces the number of parameters.

Finally, a fine-tuning strategy is employed to avoid the difficulty of simultaneous parameters optimization in the three branches. A pre-trained model is required before fine-tuning. Therefore, as described in Step 3 (see Fig. 1), three branches are first trained with a large learning rate separately. Then, the layers after GlobalMaxPool (*i.e.*, Dense and Softmax layer with grey shading and dotted box marking in Step 2 of Fig. 1) of each pre-trained model are removed and the pre-trained parameters of remaining layers are fixed. We denote the GlobalMaxPool features of the three CNN branches as F_{branch_u} (u = 1, 2, 3), which are concatenated as:

$$F_{Merge} = f\left(W \otimes \left(F_{branch\_1} \| F_{branch\_2} \| F_{branch\_3}\right) + b\right) \tag{7}$$

where ‖ means concatenating the GlobalMaxPool features of the three CNN branches, and f is a nonlinear activation function. Then two dense layers and Softmax classifier layers are used to generate the final classification map.

III. Experimental Results

Experiments were conducted on three real multiresolution remote sensing data sets, which were acquired by three different satellite sensors (QuickBird, Deimos-2 and GaoFen-2). Ground reference maps are built according to careful image interpretation, where the spatial resolution of the reference maps is fixed as the same as that of the corresponding PAN image.

Algorithms were implemented by using Matlab and Python, and the DL networks were built using Tensorflow[1] with the high-level API Keras[2], which is a simplified interface to Tensorflow. In particular, experiments based on DL networks were carried out on the Ubuntu 18.04.5, with Intel(R) Xeon(R) Gold 6130 CPUs @ 2.10GHz, 159GB RAM, and GPU of NVIDIA GRID P40-24Q, 22GB.

[1] http://tensorflow.org/

[2] https://github.com/fchollet/keras

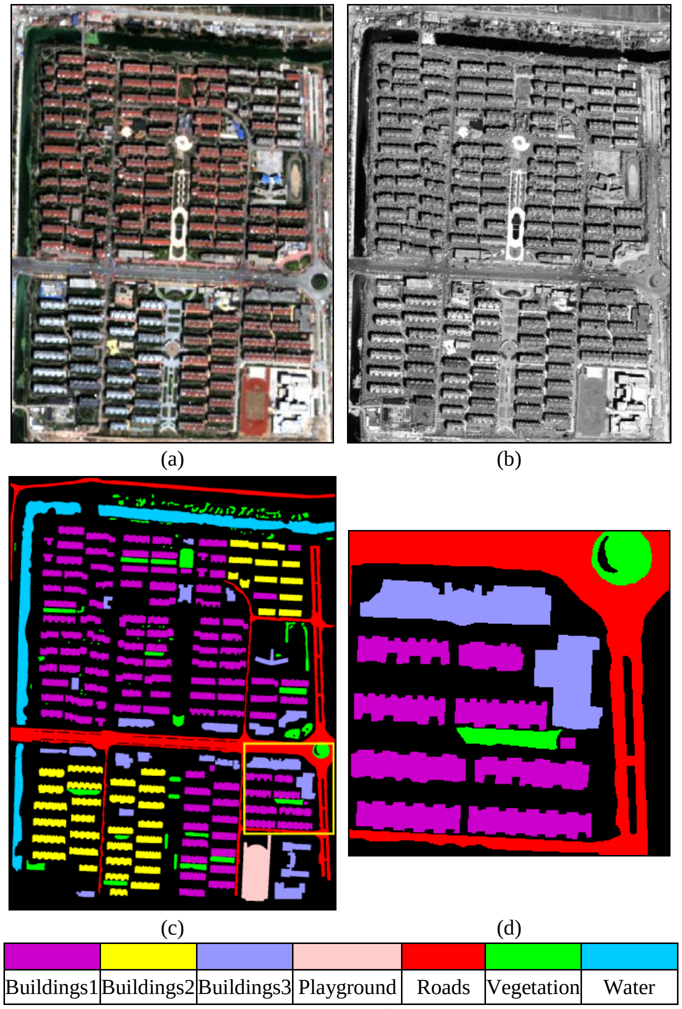


Fig. 2 XZ data set: (a) true color composite of the MS image, (b) PAN image, (c) ground reference map and (d) zoom of the portion of the image highlighted in the yellow box in (c).

Table 2. Number of training and test samples for the XZ data set.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings1	200	209250
2	Buildings2	200	80919
3	Buildings3	200	55554
4	Playground	200	18152
5	Roads	200	113484
6	Vegetation	200	41040
7	Water	200	72620

Table 3 Number of training and test samples for the VC data set.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings1	200	89867
2	Buildings2	200	13519
3	Roads	200	38206
4	Railways	200	9330
5	Trees	200	20703
6	Water	200	347554

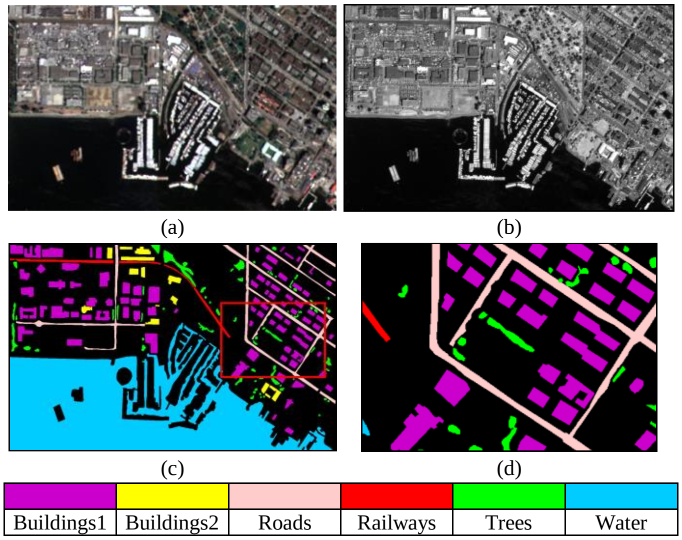


Fig. 3 VC data set: (a) true color composite of the MS image, (b) PAN image and (c) ground reference map and (d) zoom of the portion of the image highlighted in the yellow box in (c).

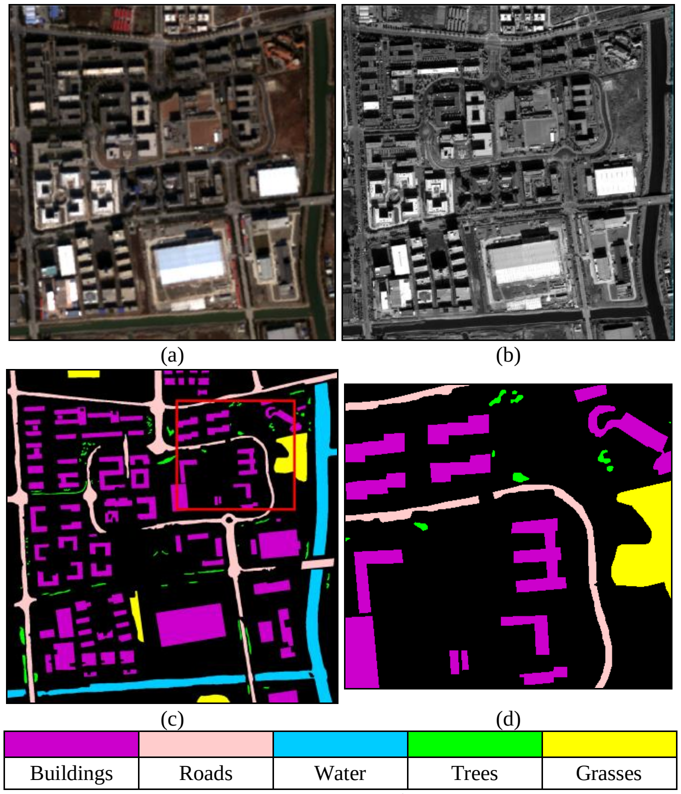


Fig. 4 SH data set: (a) true color composite of the MS image, (b) PAN image and (c) ground reference map and (d) zoom of the portion of the image highlighted in the yellow box in (c).

Table 4 Number of training and test samples for the SH data set.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings	200	195239
2	Roads	200	84244
3	Water	200	77843
4	Trees	200	11181
5	Grasses	200	24668

A. Description of Data Sets

1) *Xuzhou data set (XZ)*: This data set was acquired by the QuickBird satellite over urban area of Xuzhou city, China. The PAN image has a size of 1132×1516 pixels, with a spatial resolution of 0.6m, and the MS image has a size of 283×379 pixels with a spatial resolution of 2.4m. This scene contains seven land-cover classes, including buildings1 (with red roofs), buildings2 (with bluish roofs), buildings3 (with gray roofs), playground, roads, vegetation and water. 200 training samples for each class were selected, and the rest were used for testing. Fig. 2a shows the true color composite image of the MS image, Fig. 2b its corresponding PAN image, and Fig. 2c the ground reference map. Fig. 2d shows zoom of the portion of the image highlighted in the yellow box in Fig. 2c. Note that the reference map is made according to a careful manual image interpretation of VHR images and Google maps. Table 2 lists the number of training and test samples used in the experiments.

2) *Vancouver data set (VC)*: The 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest [41] offered MS and PAN images which were acquired on March 31 and May 30, 2015, over Vancouver city, Canada, from the DEIMOS-2 satellite. A subset of the whole image was selected for experiments. The spatial resolutions of the PAN and the MS images (with blue, green, red and near-infrared bands) are 1m and 4m, respectively. The size of the MS and PAN images are 345×219 and 1380×876 pixels, respectively. There are mainly six classes in the scene, including buildings1 (with brown roofs), buildings2 (with white roofs), roads, railways, trees and water. 200 training samples for each class were selected, and the others were considered as test samples. Fig. 3 presents the true color composite images of the MS and the PAN images, and their ground reference map. The training and testing samples are listed in Table 3.

3) *Shanghai data set (SH)*: This data set was made up of a pair of MS and PAN images acquired by the Chinese GaoFen-2 satellite over Shanghai, China, on January 2, 2015. The spatial resolution of MS (with blue, green, red and near-infrared bands) and PAN images are 4m and 1m, respectively. Corresponding image sizes are 300×305 and 1200×1220 pixels. There are five classes in this image scene, and the training and testing samples used in experiments are listed in Table 4. Fig. 4 shows the true color composite of the MS and the PAN images, and the corresponding ground reference map.

B. Parameter Tuning

The proposed feature-level fusion architecture represents a proper definition of the parameter values to enhance the classification performance. For shallow features, optimal parameters of EMAP were selected after multiple trials: parameters were set as a = 2000, d = 200, i = 0.5, and s = 10. In order to analyze and validate in details the proposed CRHFF approach, the obtained classification results are compared after parameter tuning according to different patch sizes and different layers.

1) *Multi-scale Comparison*: The performance of different patch sizes in the deep network is compared. Results obtained based on PAN image patch sizes of 20×20, 24×24, 28×28, 32×32, 36×36, 40×40 and 44×44 are provided in Fig. 5. We can see that the patch sizes resulting in the highest classification accuracies are 28×28 for both the VC and the SH data sets, and 36×36 for the XZ data set. Patches with different sizes contain spatial features at different scales. Since classification of large-scale objects may contain isolated noise, large patches are usually preferred. However, using large patches is always time-consuming and may increase misclassification of small objects. An optimal patch is determined to reach a compromise between image resolution and object size.

2) *Multi-resolution Layer Comparison*: As mentioned before, different hidden layers contain different distinct features. To further study the potential performance of hidden layers at different resolutions, the classification accuracy obtained on different branches are compared in Fig. 6. For single branch results, the branch 3 (*i.e.*, high-resolution layer) close to $Patches_{PAN}$ achieves poor results. The overall accuracy of branch 2 (*i.e.*, medium-resolution layer) is the highest, and it is slightly higher than that of branch 1 (*i.e.*, low-resolution layer). The dotted lines shown in Fig. 6 are the final classification accuracy obtained by merging three branches. One can clearly see that after combining the three branches, the classification accuracy is higher than for any single branch in all three data sets. This demonstrates the effectiveness of using multi-resolution latent information contained in different hidden layers.

3) *Learning Rate and Batch Size*: The step of gradient descent in the training process is determined by the learning rate, and also affects the learning behavior of the network. Based on multiple trials on the experimental data sets, the learning rate was set as 0.001 with Adam optimization [42], and the batch size was defined as 128.

C. Classification Performance Evaluation and Analysis

In order to validate the effectiveness of the proposed CRHFF approach, five reference methods were also considered for comparison purpose. We selected two state-of-the-art feature-level fusion methods, *i.e.*, Deep Multiple Instance Learning (DMIL) [25] and MultiResolution Land Cover Classification (MultiResoLCC) [26], and three popular methods, *i.e.*, Support Vector Machine (SVM), Random Forest (RF) and VGG-Like based on pan-sharpening and feature stacking fusion strategies. Specifically, SVM was implemented using the libsvm toolbox[3], where the radial basis function was selected as the kernel function. RF was defined with 200 trees. VGG-Like contains three fewer Maxpool layers compared to the original VGG16 [43] limited by the patch size of input data, and adds a fully connected layer with ClassNum (the number of classes) neurons to the last layer. It is important to note that final results were generated based on the average of 10 times running of each method in order to test the robustness of the methods.

For convenience of expression, the following abbreviations are defined: F_{PS} means the pan-sharpened results of MS and PAN images using the Gram-Schmidt (GS) pan-sharpening algorithm, F_{Stack} represents the concatenation of MS (after

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

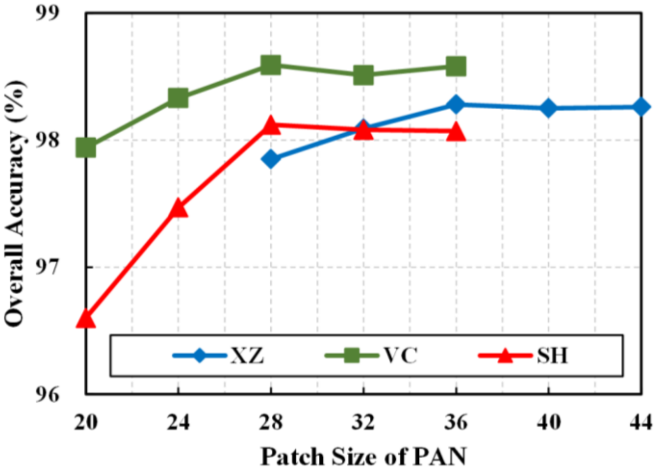


Fig. 5 Overall classification accuracy versus the patch sizes for the proposed CRHFF approach.

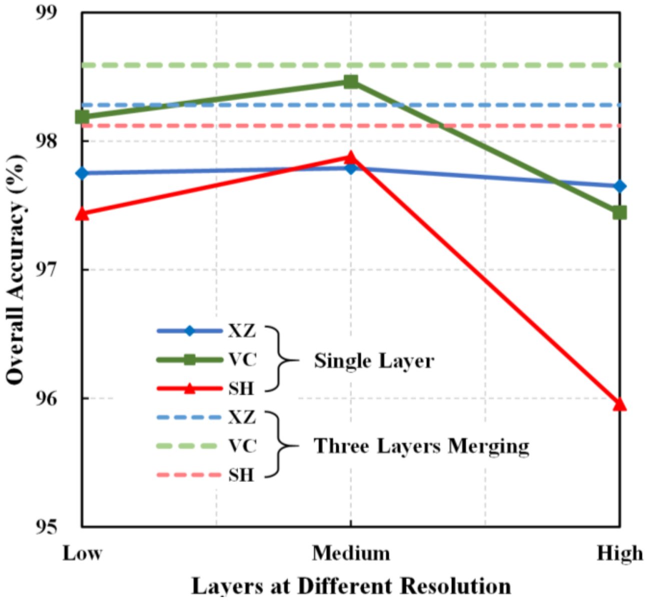


Fig. 6 Overall classification accuracy versus the number of layers in the proposed CRHFF approach on the three considered data sets.

up-sampling) and PAN images, F_{MS*PAN} and $F_{EMAP*PAN}$ represent the feature-level fusion based on the MS and PAN images, and on the EMAP and PAN images in the proposed CRHFF approach, respectively.

1) *Results on the XZ data set*. The overall accuracy (OA), the average accuracy (AA), the Kappa coefficient (Kappa) and the class-by-class accuracies are listed in Table 5. To further compare the stability of methods, standard deviation is also calculated. One can see that the proposed CRHFF ($F_{EMAP*PAN}$) approach is superior to other reference methods providing the highest OA=98.28%, with an improvement of 2.58%, 4.68%, and 1.83% with respect to VGG-Like (F_{Stack}), DMIL and MultiResoLCC methods, respectively. Moreover, it also outperforms the classical SVM (F_{Stack}) and RF (F_{Stack}) by sharply increasing the OA values of approximately 27.1% and 24.93%, respectively. The proposed CRHFF (F_{MS*PAN}) approach is also superior to other reference methods in terms of class accuracies on buildings2, buildings3, roads and water. By combination with EMAP, the proposed CRHFF ($F_{EMAP*PAN}$) greatly improves the classification performance especially for roads, buildings3 and vegetation compared to the CRHFF (F_{MS*PAN}). Furthermore, for the F_{Stack} fusion strategy, the obtained OA values of SVM, RF, and VGG-Like are 71.18%, 73.35%, 95.70%, respectively, with an improvement of 2.18%, 4.21% and 0.92% on the F_{PS} fusion strategy. This also demonstrates the advantage of feature-level fusion strategies compared with the pixel-level fusion ones.

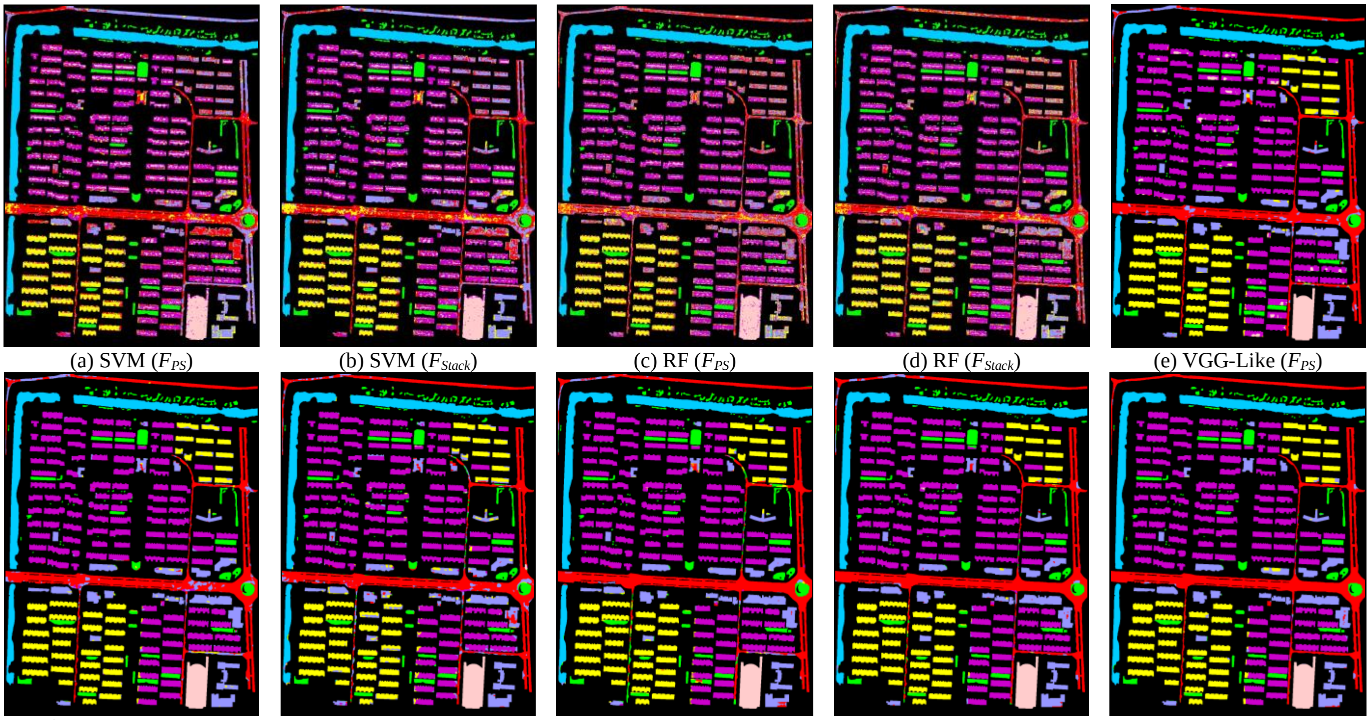
For a qualitative evaluation, classification maps associated with the average OA values among 10 runs are provided in Fig. 7 (which corresponds to the quantitative results in Table 5). In order to better evaluate the classification performance, Fig. 8 shows the classification results at a local scale. We can see that the proposed CRHFF (F_{MS*PAN}) provides more reliable classification maps than the reference methods. Besides, by joining EMAP information, CRHFF ($F_{EMAP*PAN}$) further enhanced the classification performance, particularly in preserving inner-class homogeneity for roads and buildings. Moreover, compared to the DL-based methods (*i.e.*, VGG-Like, DMIL, MultiResoLCC, CRHFF), classification maps obtained by the SVM and RF methods exhibit more salt-and-pepper noises, showing commission errors mainly on roads and buildings. The standard deviation of the OA values of different methods are illustrated in Fig. 11. It is clear that the standard deviation of the proposed CRHFF method is much lower than those of the other state-of-the-art DL-methods, which further demonstrates its robustness in feature fusion and classification.

2) *Results of VC data set*. Table 6 lists the obtained classification accuracy. We can see that the OA of the proposed CRHFF ($F_{EMAP*PAN}$) is 98.59%, which is 2.94%, 6.19% and 1.22% higher than the ones of the VGG-Like (F_{Stack}), the DMIL and the MultiResoLCC methods, respectively. We can also observe from the confusion matrix that buildings1 and roads, building2 and railways are more likely to be misclassified due to the similar spectral representations. However, in the proposed CRHFF (F_{MS*PAN}), accuracies of buildings1 and roads have improvements of 2.38% and 4.15% with respect to those of MultiResoLCC, which are the highest among all the reference methods. The proposed CRHFF ($F_{EMAP*PAN}$) further improves the accuracy of the above-mentioned misclassified classes, by extracting more spatial information with the EMAP at global scale.

Moreover, the proposed CRHFF approach outperformed the classical SVM (F_{Stack}) and RF (F_{Stack}) by approximately increasing of 9.79% and 8.22% OA values, respectively. Similar to the previous data set, for the SVM, the RF and the VGG-Like methods, feature-level fusion (F_{Stack}) strategy is superior to pixel-level fusion (F_{PS}) strategy. Fig. 9 provides the obtained classification maps for a detailed qualitative analysis at global and local scales. The proposed CRHFF (F_{MS*PAN}) preserves the inner-class homogeneity and achieves an accuracy on boundaries superior than the others, especially for roads and buildings that are difficult to be distinguished in the high-resolution images. The CRHFF ($F_{EMAP*PAN}$) integrating the EMAP information can better preserve inner-class homogeneity for roads and buildings, thus further enhancing the classification performance.

Table 5 Comparison of the classification accuracies (%) provided by different methods (XZ data set)

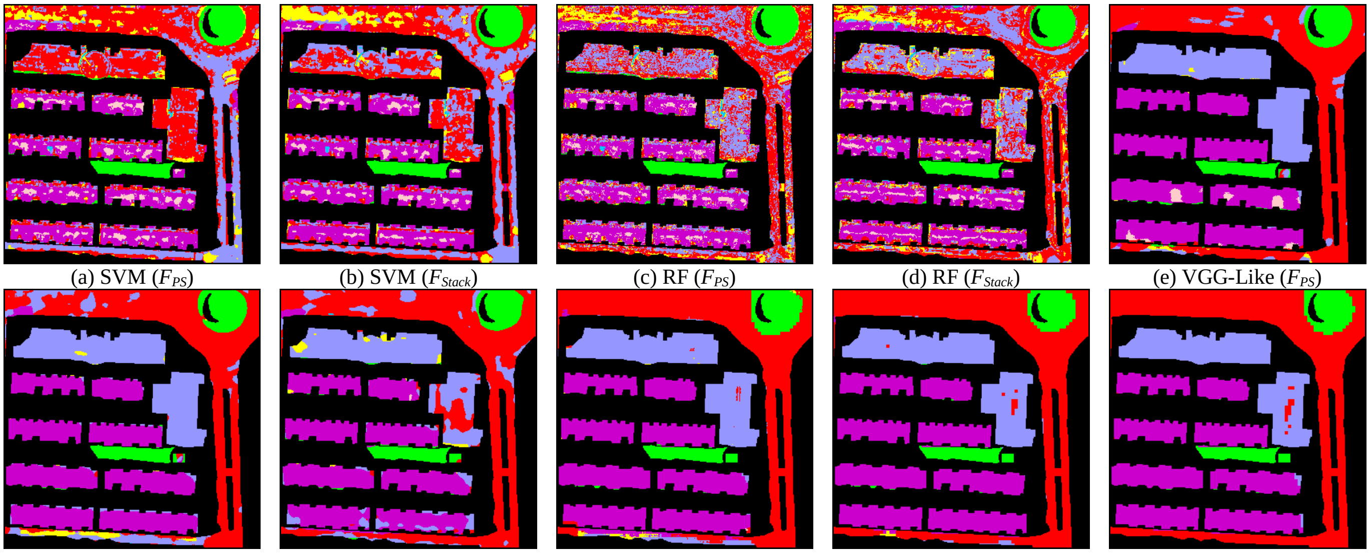
Class	SVM		RF		VGG-Like		DMIL	MultiResoLCC	CRHFF	
	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{MS*PAN}	F_{MS*PAN}	F_{MS*PAN}	$F_{EMAP*PAN}$
Buildings1	65.43	68.57	69.87	74.46	96.39±2.21	98.08±0.76	95.43±0.72	98.78±0.35	98.69±0.14	98.68±0.25
Buildings2	59.20	62.39	56.26	61.56	89.79±4.56	91.69±7.22	90.98±1.46	96.69±4.80	98.79±0.17	98.50±0.42
Buildings3	50.19	60.35	49.27	51.37	89.64±6.56	89.97±4.73	84.22±1.31	95.75±1.83	96.47±0.78	97.08±0.82
Playground	92.30	93.80	91.28	94.39	99.84±0.49	99.85±0.43	99.89±0.07	99.97±0.06	99.98±0.07	99.91±0.03
Roads	58.18	55.79	54.14	61.97	92.39±6.18	93.01±3.30	90.68±2.70	89.66±5.48	92.98±0.88	96.95±0.62
Vegetation	97.55	98.68	97.09	98.11	98.89±0.63	98.02±1.22	97.71±0.35	96.80±0.67	96.73±0.34	97.88±0.22
Water	99.54	99.62	98.72	98.61	99.78±0.16	99.55±0.91	99.07±0.26	99.58±0.30	99.55±0.04	99.73±0.14
OA	69.00	71.18	69.14	73.35	94.78±0.89	95.70±1.47	93.60±0.81	96.45±1.68	97.40±0.16	98.28±0.14
AA	74.63	77.03	73.80	77.21	95.25±4.58	95.74±4.07	94.00±5.65	96.71±3.50	97.57±2.38	98.39±1.23
Kappa	62.18	64.81	62.11	67.03	93.42±1.11	94.56±1.87	91.93±1.01	95.52±2.10	96.72±0.21	97.83±0.18



(a) SVM (F_{PS}) (b) SVM (F_{Stack}) (c) RF (F_{PS}) (d) RF (F_{Stack}) (e) VGG-Like (F_{PS})

(f) VGG-Like (F_{Stack}) (g) DMIL (h) MultiResoLCC (i) CRHFF (F_{MS*PAN}) (j) CRHFF ($F_{EMAP*PAN}$)

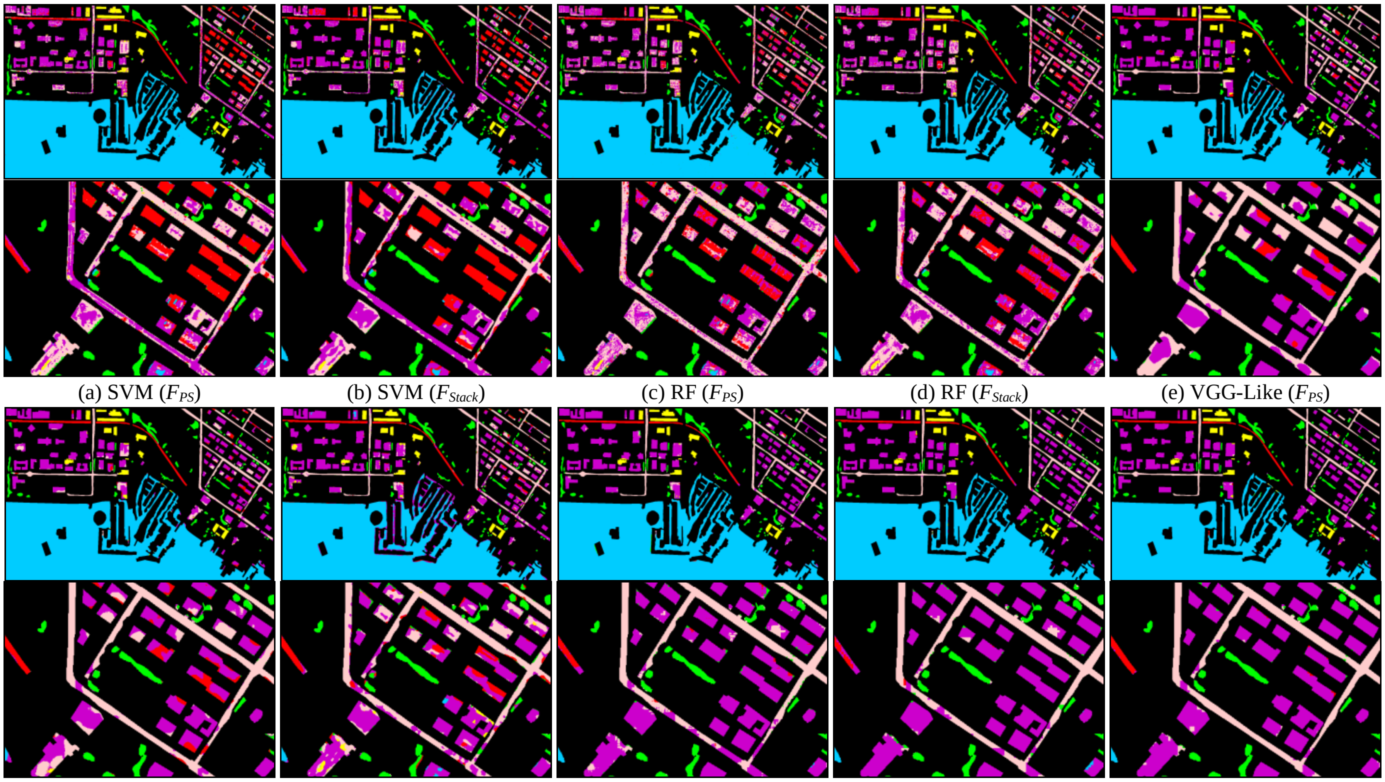
Fig. 7 Classification maps obtained by different methods on the XZ data set: (a) SVM (F_{PS}) (69.00%), (b) SVM (F_{Stack}) (71.18%), (c) RF (F_{PS}) (69.14%), (d) RF (F_{Stack}) (73.35%), (e) VGG-Like (F_{PS}) (94.98%), (f) VGG-Like (F_{Stack}) (95.72%), (g) DMIL (93.59%), (h) MultiResoLCC (96.48%), (i) proposed CRHFF (F_{MS*PAN}) (97.45%), (j) proposed CRHFF ($F_{EMAP*PAN}$) (98.38%).



(a) SVM (F_{PS}) (b) SVM (F_{Stack}) (c) RF (F_{PS}) (d) RF (F_{Stack}) (e) VGG-Like (F_{PS})

(f) VGG-Like (F_{Stack}) (g) DMIL (h) MultiResoLCC (i) CRHFF (F_{MS*PAN}) (j) CRHFF ($F_{EMAP*PAN}$)

Fig. 8 Classification maps obtained by different methods at local scale on the XZ data set: (a) SVM (F_{PS}), (b) SVM (F_{Stack}), (c) RF (F_{PS}), (d) RF (F_{Stack}), (e) VGG-Like (F_{PS}), (f) VGG-Like (F_{Stack}), (g) DMIL, (h) MultiResoLCC, (i) proposed CRHFF (F_{MS*PAN}), (j) proposed CRHFF ($F_{EMAP*PAN}$).



(a) SVM (F_{PS}) (b) SVM (F_{Stack}) (c) RF (F_{PS}) (d) RF (F_{Stack}) (e) VGG-Like (F_{PS})

(f) VGG-Like (F_{Stack}) (g) DMIL (h) MultiResoLCC (i) CRHFF (F_{MS*PAN}) (j) CRHFF ($F_{EMAP*PAN}$)

Fig. 9 Classification maps obtained by different methods on the VC data set: (a) SVM (F_{PS}) (88.20%), (b) SVM (F_{Stack}) (88.80%), (c) RF (F_{PS}) (89.36%), (d) RF (F_{Stack}) (90.37%), (e) VGG-Like (F_{PS}) (94.95%), (f) VGG-Like (F_{Stack}) (95.85%), (g) DMIL (92.43%), (h) MultiResoLCC (97.42%), (i) proposed CRHFF (F_{MS*PAN}) (98.20%), (j) proposed CRHFF ($F_{EMAP*PAN}$) (98.58%). The first and third rows represent the whole classification maps at global scale, whereas the second and fourth rows represent the subsets at local scale.

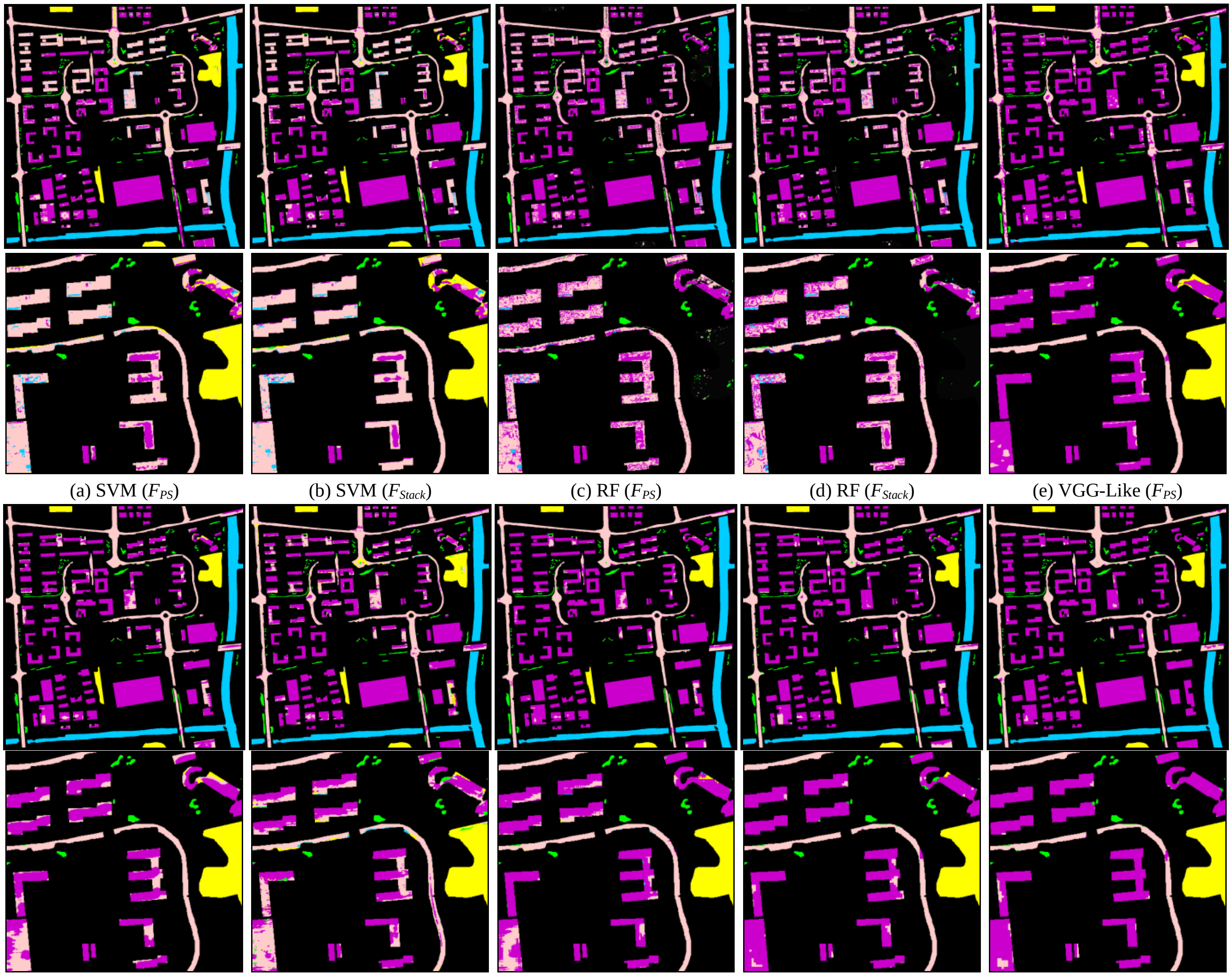
Table 6 Comparison of the classification accuracies (%) provided by different methods (VC data set)

Class	**SVM**		**RF**		**VGG-Like**		**DMIL**	**MultiResoLCC**	**CRHFF**	
	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{MS*PAN}	F_{MS*PAN}	F_{MS*PAN}	$F_{EMAP*PAN}$
Buildings1	51.45	57.47	54.61	57.10	82.56±6.95	82.75±6.96	81.23±2.00	93.08±1.63	95.46±0.59	95.79±0.71
Buildings2	94.34	94.41	95.10	95.16	98.40±0.95	94.95±9.57	96.34±1.34	99.53±0.22	99.26±0.23	99.70±0.16
Roads	61.48	59.01	75.10	78.43	93.01±1.85	91.84±3.00	82.94±2.44	86.06±2.44	90.21±0.90	93.99±0.58
Railways	85.23	91.32	82.44	89.66	97.17±2.88	98.83±0.87	94.36±0.94	95.41±1.55	94.63±1.49	99.35±0.28
Trees	97.75	97.72	97.95	97.70	96.69±3.77	96.24±2.95	96.53±0.55	97.15±0.42	97.53±0.48	97.72±0.31
Water	99.91	99.35	99.36	99.69	97.49±5.33	99.31±1.21	95.88±0.68	99.70±0.16	99.87±0.02	99.80±0.03
OA	88.20	88.80	89.36	90.37	94.56±4.78	95.65±1.51	92.40±0.36	97.37±0.23	98.19±0.07	98.59±0.09
AA	81.69	83.21	84.09	86.29	94.22±6.01	93.99±6.14	91.21±7.13	95.16±5.11	96.16±3.56	97.72±2.39
Kappa	77.37	78.56	79.69	81.56	89.94±8.04	91.65±2.84	85.66±0.63	94.89±0.44	96.48±0.14	97.26±0.18

3) *Results of SH data set*. Table 7 provides the qualitative classification results. Similar conclusions can be drawn as in the previous two data sets. The highest OA value was achieved by the proposed CRHFF (*i.e.*, 98.12%), which is 4.53%, 9.33% and 2.52% higher than the ones of the VGG-Like (F_{Stack}), DMIL and MultiResoLCC methods, respectively, and also significantly outperforms SVM and RF. Class accuracies of buildings and roads are also relatively low in the reference methods. The proposed CRHFF (F_{MS*PAN}) and CRHFF ($F_{EMAP*PAN}$) methods significantly improve the accuracies of these two difficult classes. Fig. 10 shows the classification maps obtained on the SH data set at global and local scales. The maps point out many commission errors in the SVM and RF methods. DL-based methods offer a great improvement by considering more spatial information, while the proposed CRHFF outperforms all the other with more regular and complete classification results.

IV. Conclusion

In this paper a cross-resolution hidden layer feature fusion approach (CRHFF) is proposed for joint classification of MS and PAN images. It fills the gap in the traditional ways that fusion of MS and PAN images mainly relies on the pan-sharpening or individual feature extraction followed by stacking. In particular, to alleviate the degradation of spatial integrity and connectivity of land objects by local patches, we first extract spatial features from MS image at a global scale, then deep hidden layer features are extracted from MS and PAN images and fused from patches at a local scale with an AE-like deep network. Moreover, different scale information of land objects is taken into account by means of cross-resolution latent features, without implementing up-/down-sampling operations. Experimental results obtained on three real multi-



(a) SVM (F_{PS}) (b) SVM (F_{Stack}) (c) RF (F_{PS}) (d) RF (F_{Stack}) (e) VGG-Like (F_{PS})

(f) VGG-Like (F_{Stack}) (g) DMIL (h) MultiResoLCC (i) CRHFF (F_{MS*PAN}) (j) CRHFF ($F_{EMAP*PAN}$)

Fig. 10 Classification maps obtained by different methods on the SH data set: (a) SVM (F_{PS}) (81.14%), (b) SVM (F_{Stack}) (82.29%), (c) RF (F_{PS}) (83.43%), (d) RF (F_{Stack}) (85.27%), (e) VGG-Like (F_{PS}) (92.42%), (f) VGG-Like (F_{Stack}) (94.01%), (g) DMIL (88.91%), (h) MultiResoLCC (95.71%), (i) proposed CRHFF (F_{MS*PAN}) (96.49%), (j) proposed CRHFF ($F_{EMAP*PAN}$) (98.13%). The first and third rows represent the whole classification maps at global scale, whereas the second and fourth rows represent the subsets at local scale.

Table 7 Comparison of the classification accuracies (%) provided by different methods (SH data set)

Class	**SVM**		**RF**		**VGG-Like**		**DMIL**	**MultiResoLCC**	**CRHFF**	
	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{PS}	F_{Stack}	F_{MS*PAN}	F_{MS*PAN}	F_{MS*PAN}	$F_{EMAP*PAN}$
Buildings	66.93	66.80	74.40	76.07	89.77±4.71	89.48±4.41	83.26±1.82	93.96±0.50	94.99±0.36	98.00±0.09
Roads	89.64	94.85	83.58	88.33	91.86±5.87	94.63±3.43	87.73±3.77	93.60±0.61	95.29±0.34	95.92±0.35
Water	99.95	99.96	99.87	99.92	99.58±0.73	99.96±0.08	99.79±0.15	99.99±0.03	100.00±0.00	99.94±0.04
Trees	98.43	99.03	97.16	97.40	99.06±0.45	99.17±0.95	98.33±0.86	99.72±0.21	99.90±0.04	99.91±0.04
Grasses	97.35	98.54	96.30	95.99	99.88±0.27	99.99±0.03	97.23±1.01	99.73±0.41	99.98±0.04	99.99±0.03
OA	81.14	82.29	83.43	85.27	93.06±2.09	93.59±1.80	88.79±1.61	95.60±0.21	96.50±0.13	98.12±0.08
AA	90.46	91.84	90.26	91.54	96.03±4.82	96.64±4.58	93.27±7.33	97.40±3.31	98.03±2.64	98.75±1.79
Kappa	73.31	75.01	76.11	78.70	89.72±2.99	90.55±1.54	83.64±2.27	93.43±0.31	94.77±0.19	97.17±0.13

resolution data sets acquired by QuickBird, Deimos-2 and GaoFen-2 satellites confirmed the effectiveness of the proposed approach, compared with the state-of-the-art methods in terms of higher classification accuracy and robustness.

In future work, we will explore other cross-resolution fusion network that could further improve the multi-resolution data classification efficiency and accuracy.

ACKNOWLEDGMENTS

The authors would like to thank Deimos Imaging for acquiring and providing the VC data set, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. We also thank the China Centre for Resources Satellite Data and Application for providing the Gaofen-2 images.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2021.3127710

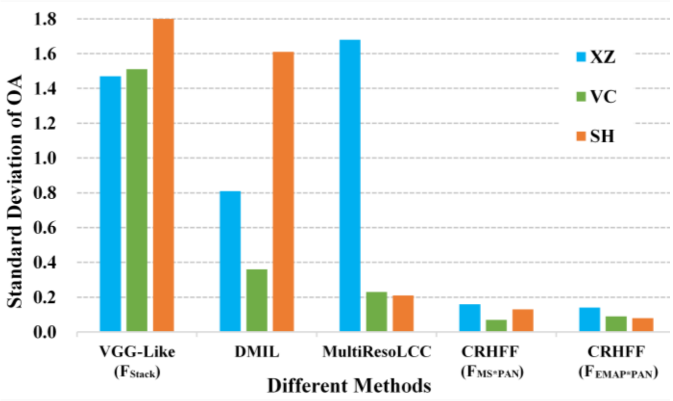


Fig. 11 Comparison of the standard deviation of OA values for the different considered methods.

REFERENCES

[1] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340-4354, 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 Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140-158, 2019.

[3] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966-5978, 2021.

[4] D. Hong, L. Gao, J. Yao, N. Yokoya, J. Chanussot, U. Heiden, B. Zhang, "Endmember-Guided Unmixing Network (EGU-Net): A General Deep Learning Framework for Self-Supervised Hyperspectral Unmixing," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[5] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 464-473, 2021.

[6] J. Kang, R. Fernandez-Beltran, P. Duan, S. Liu, and A. J. Plaza, "Deep Unsupervised Embedding for Remotely Sensed Images Based on Spatially Augmented Momentum Contrast," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 2598-2610, 2021.

[7] P. Du, S. Liu, P. Gamba, K. Tan, and J. Xia, "Fusion of Difference Images for Change Detection Over Urban Areas," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 4, pp. 1076-1086, 2012.

[8] X. Ma, X. Tong, S. Liu, C. Li, and Z. Ma, "A Multisource Remotely Sensed Data Oriented Method for “Ghost City” Phenomenon Identification," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 7, pp. 2310-2319, 2018.

[9] S. Liu, Y. Zheng, M. Dalponte, and X. Tong, "A novel fire index-based burned area change detection approach using Landsat-8 OLI data," European Journal of Remote Sensing, vol. 53, no. 1, pp. 104-112, 2020.

[10] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10691-10704, 2021.

[11] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of Multispectral Images to High Spatial Resolution: A Critical Review of Fusion Methods Based on Remote Sensing Physics," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1301-1312, 2008.

[12] H. Ghassemian, "A review of remote sensing image fusion methods," *Information Fusion*, vol. 32, pp. 75-89, 2016.

[13] P. Du, S. Liu, J. Xia, and Y. Zhao, "Information fusion techniques for change detection from multi-temporal remote sensing images," *Information Fusion*, vol. 14, no. 1, pp. 19-27, 2013.

[14] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110-120, 2020.

[15] Y. Qu, H. R. Qi, B. Ayhan, C. Kwan, and R. Kidd, "Does Multispectral / Hyperspectral Pansharpening Improve the Performance of Anomaly Detection ?," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 6130-6133, 2017.

[16] H. Wei, X. Liang, W. Zhihui, L. Hongyi, and T. Songze, "A New Pan-Sharpening Method With Deep Neural Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1037-1041, 2015.

[17] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978-989, 2018.

[18] Q. Liu, L. Han, R. Tan, H. Fan, W. Li, H. Zhu, B. Du, and S. Liu, "Hybrid Attention Based Residual Network for Pansharpening," *Remote Sensing*, vol. 13, no. 10, 2021.

[19] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of Pansharpened Urban Satellite Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 281-297, 2012.

[20] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, G. Yang, Q. Yuan, R. Fu, and H. Zhang, "A Large-Scale Benchmark Data Set for Evaluating Pansharpening Performance: Overview and Implementation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 18-52, 2021.

[21] N. Kosaka, T. Akiyama, Bien Tsai, and T. Kojima, "Forest type classification using data fusion of multispectral and panchromatic high-resolution satellite imageries," *in Proceedings. IEEE International Geoscience and Remote Sensing Symposium*, Jul. 25-29, 2005, vol. 4, pp. 2980-2983.

[22] G. Moser, A. De Giorgi, and S. B. Serpico, "Multiresolution Supervised Classification of Panchromatic and Multispectral Images by Markov Random Fields and Graph Cuts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5054-5070, 2016.

[23] T. Mao, H. Tang, J. Wu, W. Jiang, S. He, and Y. Shu, "A Generalized Metaphor of Chinese Restaurant Franchise to Fusing Both Panchromatic and Multispectral Images for Unsupervised Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4594-4604, 2016.

[24] W. Zhao, L. Jiao, W. Ma, J. Zhao, J. Zhao, H. Liu, X. Cao, and S. Yang, "Superpixel-Based Multiple Local CNN for Panchromatic and Multispectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4141-4156, 2017.

[25] X. Liu, L. Jiao, J. Zhao, J. Zhao, D. Zhang, F. Liu, S. Yang, and X. Tang, "Deep Multiple Instance Learning-Based Spatial-Spectral Classification for PAN and MS Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 461-473, Jan 2018.

[26] R. Gaetano, D. Ienco, K. Ose, and R. Cresson, "A Two-Branch CNN Architecture for Land Cover Classification of PAN and MS Imagery," *Remote Sensing*, vol. 10, no. 11, 2018.

[27] C. Shi and C.-M. Pun, "Adaptive multi-scale deep neural networks with perceptual loss for panchromatic and multispectral images classification," *Information Sciences*, vol. 490, pp. 1-17, 2019.

[28] H. Zhu, W. P. Ma, L. L. Li, L. C. Jiao, S. Y. Yang, and B. Hou, "A Dual-Branch Attention fusion deep network for multiresolution remote-Sensing image classification," *Information Fusion*, vol. 58, pp. 116-131, Jun 2020.

[29] W. Ma, J. Zhao, H, Zhu, J. Shen, L. Jiao, Y, Wu, and B. Hou, "A Spatial-Channel Collaborative Attention Network for Enhancement of Multiresolution Classification," *Remote Sensing*, vol. 13, no. 1, 2020.

[30] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. Benediktsson, "Feature Extraction for Hyperspectral Imagery: The Evolution From Shallow to Deep: Overview and Toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 60-88, Dec 2020.

[31] M. Dalla Mura, J. Atli Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975-5991, 2010.

[32] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of Hyperspectral Images by Using Extended Morphological Attribute Profiles and Independent Component Analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 542-546, 2011.

[33] P. Ghamisi, R. Souza, J. A. Benediktsson, X. X. Zhu, L. Rittner, and R. A. Lotufo, "Extinction Profiles for the Classification of Remote Sensing

Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5631-5645, 2016.

[34] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–Spatial Hyperspectral Image Classification With Edge-Preserving Filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2666-2677, 2014.

[35] X. Kang, C. Li, S. Li, and H. Lin, "Classification of Hyperspectral Images by Gabor Filtering Based Deep Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1166-1178, 2018.

[36] S. Liu, Q. Hu, X. Tong, J. Xia, Q. Du, A. Samat, and X. Ma, "A Multi-Scale Superpixel-Guided Filter Feature Extraction and Selection Approach for Classification of Very-High-Resolution Remotely Sensed Imagery," *Remote Sensing*, vol. 12, no. 5, 2020.

[37] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-7, Jul 28 2006.

[38] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," presented at the 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A New Deep Generative Network for Unsupervised Remote Sensing Single-Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6792-6810, 2018.

[40] C. Liu, Y. Zhang, S. Wang, M. Sun, Y, Ou, Y. Wan, and X. Liu, "Band-Independent Encoder–Decoder Network for Pan-Sharpening of Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 5208-5223, 2020.

[41] 2016 IEEE GRSS Data fusion contest. online: http://www.grss–ieee.org/community/technical–committees/data–fusion.

[42] D. P. Kingma and J. L. Ba,"Adam: A method for stochastic optimization," *in Proceedings. 3rd International conference on learning representations*, San Diego, CA, USA, May 2015.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In Proc. International Conference on Learning Representations* http://arxiv.org/abs/1409.1556 (2014).