



Cross-Scene Deep Transfer Learning With Spectral Feature Adaptation for Hyperspectral Image Classification

Chongxiao Zhong, *Student Member, IEEE*, Junping Zhang , *Member, IEEE*, Sifan Wu, and Ye Zhang , *Member, IEEE*

Abstract—The small size of labeled samples has always been one of the great challenges in hyperspectral image (HSI) classification. Recently, cross-scene transfer learning has been developed to solve this problem by utilizing auxiliary samples of a relevant scene. However, the disparity between hyperspectral datasets acquired by different sensors is a tricky problem which is hard to overcome. In this article, we put forward a cross-scene deep transfer learning method with spectral feature adaptation for HSI classification, which transfers the effective contents from source scene to target scene. The proposed framework contains two parts. First, the distribution differences of spectral dimension between source domain and target domain are reduced through a joint probability distribution adaptation approach. Then, a multiscale spectral-spatial unified network with two-branch architecture and a multiscale bank is designed to extract discriminating features of HSI adequately. Finally, classification of the target image is achieved by applying a model-based deep transfer learning strategy. Experiments conducted on several real hyperspectral datasets demonstrate that the proposed approach can explicitly narrow the disparity between HSIs captured by different sensors and yield ideal classification results of the target HSI.

Index Terms—Cross-scene deep transfer learning, hyperspectral image (HSI) classification, multiscale spectral-spatial unified network (MSSN), spectral feature adaptation (SFA).

I. INTRODUCTION

HYPERSPECTRAL imagery can provide a wealth of information about an imaged scene due to the combination of spatial and spectral information [1]. As an important application among numerous domains, hyperspectral image (HSI) classification has been attracted extensive attentions for years. However, the size of labeled samples is usually small, which becomes a challenge of HSI classification. In order to tackle this problem, cross-scene transfer learning strategy is developed, which classifies the target image with only few labeled samples by introducing auxiliary information acquired from a relevant scene.

Manuscript received February 27, 2020; revised May 9, 2020; accepted May 23, 2020. Date of publication June 2, 2020; date of current version June 16, 2020. The work was supported by the National Natural Science Foundation of China under Grant 61871150. (*Corresponding author: Junping Zhang.*)

The authors are with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: zhongcx_hit@163.com; zhangjp@hit.edu.cn; woo_sifan@163.com; zhye@hit.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2999386

The most straightforward way for cross-scene transfer learning is to train a classifier to classify the target image by utilizing labeled samples of the source image directly. However, the classification results can be seriously deteriorated as the distribution disparity between images is completely neglected. In fact, spectrums of the same ground object can vary in distributions considerably due to many affecting factors during the imaging process [2]. Therefore, the significant issue of cross-scene transfer learning is to reduce the distribution differences between source and target datasets, i.e., the domain adaptation problem.

Inspired by the typical methods in machine learning, several domain adaptation approaches have been proposed for HSI classification [3]–[7]. Based on the rotation-based ensemble and the transfer component analysis [8], Xia *et al.* [9] generated a transfer learning model for HSI classification. In [10], a dictionary learning based feature-level domain adaptation technique is proposed to overcome the spectral shift phenomenon that appears in different hyperspectral image (HSIs). Shen *et al.* [11] developed a feature adaptation and augmentation method for cross-scene knowledge transfer, which learned a common subspace by introducing two different projection matrices to extract the transferable knowledge from the source HSI to the target HSI. Recent years, deep learning model was introduced to transfer learning and gained great progress [12]. The deep networks can availably exploit high-level nonlinear feature representation of images compared with the traditional shallow models [13]–[16]. Based on this concept, Yang *et al.* [17] proposed a deep transfer learning model, which is able to extract hierarchical spectral-spatial features of HSI and achieve decent classification accuracy for the target scene with scarce available samples via a transfer-learning-based training strategy.

Although the classification results obtained by above methods are encouraging, they are limited to the cross-scene tasks where source and target datasets are imaged by same hyperspectral sensor. In order to bridge the gap of images captured by different sensors, arduous efforts have been made. Kemker and Kanan [18] introduced the self-taught learning strategy to HSI classification, during which a large number of samples from various hyperspectral datasets are input to a stacked convolutional auto-encoder to learn fairly similar features, and the obtained encoder can be utilized to classify the target scene through a fine-tuning process [19]. However, this model failed

to fully consider the distribution disparity triggered by spectral shift between different hyperspectral sensors. In [20], Lin *et al.* presented a deep transfer learning method for HSI classification, which effectively established a correlation of the trained top-level network between the source domain and the target domain. The framework can be applicable for HSIs acquired from different sensors, yet it is only available to the binary classification task. In [21], a dual dictionary nonnegative matrix factorization algorithm was proposed. By learning individual domain-specific dictionaries for each image and utilizing graph regularization strategy, the two datasets are bridged in a unified feature space.

In this article, we propose a framework composed of a spectral feature adaptation (SFA) process and a deep transfer learning model for HSI classification. In the first part, the distribution disparity of co-occurrence classes in HSI cross-scene transfer learning is mainly focused. For the spectral vectors obtained by different hyperspectral sensors, a novel domain adaption method is proposed to project the original samples into a new feature space where the spectral differences of marginal distribution and conditional distribution between the source scene and the target scene can be significantly reduced. In the second part, a cross-scene training strategy based on a deep model that incorporates the multiscale strategy into a two branch convolutional neural network (CNN) [22] is generated to transfer effective knowledge for the target domain. During the training process, bottom layers of the pretrained network is transferred to the target image directly while the parameters of top layer are initialized randomly. With a fine-tuning operation, the final network is obtained. Contributions of this article can be summarized as follows.

- 1) A domain adaptation method based on the probability distribution of statistics is introduced to spectral dimension of HSI to effectively reduce the distribution distance between the source image and target image captured by different sensors, thus the effectiveness of knowledge transfer can be ensured.
- 2) A multiscale spectral-spatial unified network (MSSN) model with cross-scene training strategy is designed for HSI classification. The network consists of a two-branch architecture and a multiscale bank, which extract multiscale spatial features and exploit spectral information simultaneously.

The rest part of this article is organized as follows. In Section II, previous works on domain adaptation and spectral-spatial classification for HSI are briefly reviewed. Then, the framework of the proposed method is detailed in Section III. In Section IV, experimental datasets, analysis and discussion of the experiment results are presented. Finally, summaries and conclusions of our work are drawn in Section V.

II. RELATED WORK

A. Domain Adaptation

Domain adaptation is the most attended research subject in transfer learning field [23]–[25], which aims at overcoming the cross-domain disparity between the source domain and target domain.

Take X as the input space, and Y as the output space for clear mathematical expression. The traditional supervised learning method aims at learning a model H from the labeled samples to predict the unlabeled samples of domain D_S . As for domain adaptation in transfer learning, the main goal is to learn a proper projection H from the source domain D_S to make predictions on the samples of the target domain D_T as accurately as possible. Depending on the sample size of target domain, the adaptation problems can be grouped into three situations depending on the sample size of target domain.

- 1) The labeled samples are sufficient in target domain.
- 2) The sample size of target domain is small.
- 3) There are no labeled samples in the target domain.

In this article, we focus on the case that the target HSI contains very limited labeled samples.

B. Spectral-Spatial Classification for HSI

With the combination of spatial and spectral information, spectral-spatial classification methods have been proved to achieve better accuracy than traditional pixel-level algorithms [26]–[29]. For example, Huo and Tang [27] introduced the Gabor filter to learn spatial information of HSI and combine it with the spectral feature to achieve a better classification result, and Jia *et al.* [30] proposed a spectral-spatial Gabor surface feature fusion approach by generating a group of predefined 2-D Gabor filters to extract spectral-spatial features of HSI. Considering the shading component of hyperspectral data is uncorrelated with the material of the imaged object, an intrinsic image decomposition approach is presented in [31] to separate the reflectance part from hyperspectral data. In the latest studies, Wang *et al.* [32] generated a locality and structure-regularized low-rank representation model which combines the spectral and spatial features into a unified distance metric and achieves decent classification accuracies without any complex classifiers, and Zhou *et al.* [33] employed the extracted spectral-spatial features as the input of the proposed compact and discriminative stacked autoencoder due to the effectiveness and simplicity.

However, extracting features on one scale may not meet the requirement of classification task since different ground objects are of different scales. Therefore, several multiscale spectral-spatial classification methods have been generated to use the spatial information more sufficiently in recent years. For example, the 3-D wavelet transform [34] and the 3-D Gabor filter [35] designed for multiscale spatial feature extraction have both achieved decent classification results. Furthermore, Dundar and Ince [36] applied the multiscale superpixels to obtain local information from different region scales so that the small and large local regions are formed to acquire spatial information well, and Li *et al.* [37] proposed a multiscale spectral-spatial classification method by decomposing the dimension-reduced image into several Gaussian pyramids to extract the multiscale features. The multiscale strategy in these methods certainly improved the classification accuracy, yet they are reliant on manual interventions, and the extracted low-level features are not adequate to describe the characteristics of different ground objects discriminately.

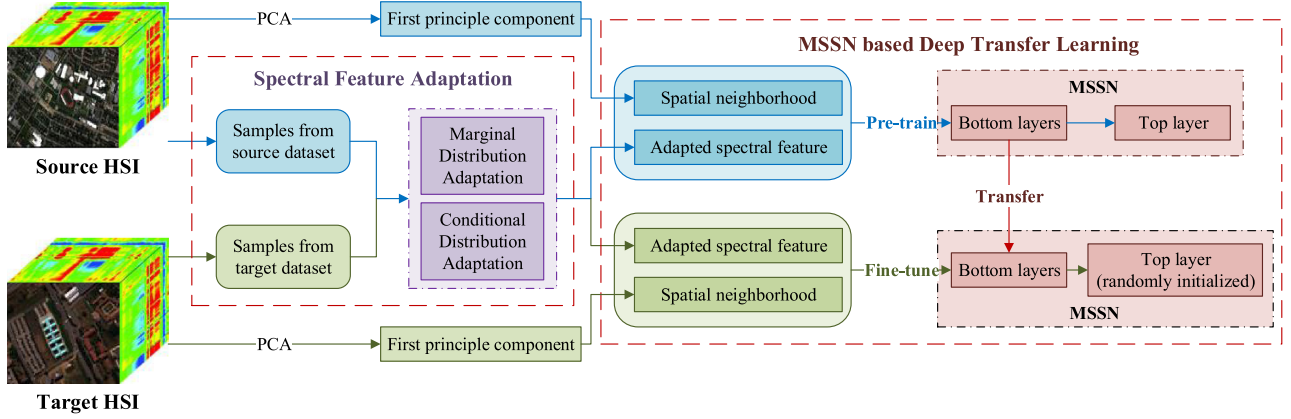


Fig. 1. Framework of the proposed method.

In this article, we generate a deep CNN model, which is well suited for hyperspectral data to automatically learn the multi-scale spectral-spatial features sufficiently, for our cross-scene transfer learning framework.

III. PROPOSED FRAMEWORK

The distribution disparity between HSIs captured by different sensors has been a tough problem which severely limited the effectiveness of knowledge transfer in HSI classification task. In this article, we present a novel cross-scene transfer learning method by jointly applying a distribution adaptation for spectral dimension and an MSSN-based deep learning model. The overall flowchart of the proposed approach is shown in Fig. 1. During the SFA process, a proper projection for spectral vectors is learned to explicitly reduce the distribution distance of co-occurrence categories between the source and target datasets. As for the deep transfer learning part, a two branch CNN architecture with a multiscale bank is designed to extract spectral and spatial features simultaneously, and the final classifier network for target HSI is obtained by a cross-scene training approach. Details of our method are described in the following subsections.

A. Spectral Feature Adaptation

Spectral vectors acquired by different hyperspectral sensors have different data distributions, which brings great challenges to cross-scene transfer learning. Therefore, the significant task is to exploit internal associations between the source and target domains and reduce the distribution disparity efficiently, which is summarized as the domain adaptation problem.

In domain adaptation, marginal probability distribution and conditional probability distribution are two important data features need to be matched between different domains [38]. Taking D as an example that consists of a sample set $X = \{x_1, \dots, x_n\}$ and the corresponding label set $Y = \{y_1, \dots, y_n\}$, the marginal probability of X is described as $P(X)$, while the conditional probability is formulated by $P(y|x)$. Due to the different acquisition conditions, the underlying probability distributions of target domain are generally different than that of the source domain. Based on this, in this article, we present SFA for cross-scene

transfer learning by introducing a joint probability distribution adaptation model [39].

Take D_s and D_t as two pixel-sets of the source HSI and target HSI, where D_s is composed of a sample set $X_s = \{x_1, \dots, x_{n_s}\}$ and the corresponding label set $Y_s = \{y_1, \dots, y_{n_s}\}$, and D_t is a sample set $X_t = \{x_1, \dots, x_{n_t}\}$ with a large amount of unlabeled samples. Assuming that the samples in both source and target domains are classified in C categories, the goal of the adaptation is to seek a proper projection H for $x_s \in X_s$ and $x_t \in X_t$. With the projection, spectral vectors are transformed to a new feature space where the marginal distribution distance and the conditional distribution distance between the two domains are explicitly reduced. Considering the labeled samples of target domain are limited to fully describe the probability distributions, unlabeled samples are also introduced during SFA. The adaptation process is detailed as follows.

1) *Marginal Distribution Adaptation (MDA)*: The MDA is to reduce the difference between $P(H^T x_s)$ and $P(H^T x_t)$, and the key of which is to choose a fine function to measure the distance. Due to the effective measure of the differences in probability distributions, the maximum mean discrepancy (MMD), defined as the mean distance between source domain and target domain in the infinite-dimensional kernel space, is adopted [40]. The expression is as follows:

$$D_{\text{MDA}}(D_s, D_t) = \left\| \frac{1}{n} \sum_{i=1}^n H^T x_{s_i} - \frac{1}{m} \sum_{i=1}^m H^T x_{t_i} \right\|^2. \quad (1)$$

Considering the complex solving procedure, this function can be transformed to a kernel learning problem [41] and become

$$D_{\text{MDA}}(D_s, D_t) = \text{tr}(H^T X M_0 X^T H) \quad (2)$$

where X is the combined dataset of X_s and X_t , and $M_0 = [(M_0)_{ij}]$ is a matrix computed as follows:

$$(M_0)_{ij} = \begin{cases} 1/n_s^2, & x_i, x_j \in D_s \\ 1/n_t^2, & x_i, x_j \in D_t \\ -1/n_s n_t, & \text{otherwise} \end{cases} \quad (3)$$

where n and m represent the number of samples in the source domain and target domain, respectively. With the kernel strategy, function (1) can be simplified and convenient to solve.

2) *Conditional Distribution Adaptation (CDA)*: In order to further narrow the gap between source domain and target domain, the unification of conditional distributions is also considered. However, it is tricky to minimize the distance between $P(y_s|H^T x_s)$ and $P(y_t|H^T x_t)$, since the two distributions can be hardly obtained due to the uncertain classification model. To overcome this difficulty, we introduce the sufficient statistic concept, which is commonly used in studying statistical problems, is introduced here. According to theory, some other sample statistics can be selected to approximate the distribution when there are too many unknown attributes in the large sample set and the samples are of good quantity. Based on this, the involved conditional distribution $P(y|x)$ of sample set X can be approximately instead by class-conditional distribution $P(x|y)$. Thus, $P(y_s|H^T x_s)$ and $P(y_t|H^T x_t)$ can be replaced by $P(H^T x_s|y_s)$ and $P(H^T x_t|y_t)$ respectively. Moreover, considering the posterior probabilities are incalculable since most samples in the target domain are unlabeled, the pseudo-labeling strategy is applied. By simply using the labeled samples to train a classifier, the unlabeled ones are preclassified. Similar to the case of MDA, MMD distance is utilized to measure the conditional distribution differences

$$D_{\text{CDA}}(D_s, D_t) = \sum_{c=1}^C \left\| \frac{1}{n_c} \sum_{x_{s_i} \in D_s^{(c)}} H^T x_{s_i} - \frac{1}{m_c} \sum_{x_{t_i} \in D_t^{(c)}} H^T x_{t_i} \right\|^2 \quad (4)$$

where $n_{s(c)}$ is the number of the samples of class c in the source domain, while $n_{t(c)}$ is of the target domain. Similar to MDA, the kernel method is applied here to simplify

$$D_{\text{CDA}}(D_s, D_t) = \sum_{c=1}^C \text{tr}(H^T X M_c X^T H) \quad (5)$$

where matrix $M_c = [(M_c)_{ij}]$ is computed by

$$(M_c)_{ij} = \begin{cases} 1/n_{s(c)}^2, & x_i, x_j \in D_s^{(c)} \\ 1/n_{t(c)}^2, & x_i, x_j \in D_t^{(c)} \\ -1/n_{s(c)}n_{t(c)}, & \begin{cases} x_i \in D_s^{(c)}, & x_j \in D_t^{(c)} \\ x_i \in D_t^{(c)}, & x_j \in D_s^{(c)} \end{cases} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

3) *Optimization of the Adaptation Model*: Since the adaptation goal is to reduce the distances of marginal distribution and conditional distribution between source and target datasets simultaneously, the mathematical model can be established by integrating (2) and (5) into a new function as follows:

$$\min \frac{\sum_{c=0}^C \text{tr}(H^T X M_c X^T H) + \lambda \|H\|_F^2}{H^T X A X^T H} \quad (7)$$

where the regularization term $\lambda \|H\|_F^2$ is added to make the function robust, and the denominator $H^T X A X^T H$ is to maintain the variance of the dataset. Note that this function represents the MMD distance of MDA and CDA when $c = 0$ and $c = 1, \dots, C$, respectively, thus the two distributions is incorporated. For solving purpose, function (7) can be further optimized as follows:

$$\begin{aligned} \min \sum_{c=0}^C \text{tr}(H^T X M_c X^T H) + \lambda \|H\|_F^2 \\ \text{s.t. } H^T X A X^T H = I. \end{aligned} \quad (8)$$

So far, the optimal function can be solved through the Lagrange multiplier algorithm, and the obtained result is expressed as

$$\left(X \sum_{c=0}^C M_c X^T + \lambda I \right) A = X A X^T H \Phi \quad (9)$$

where $\Phi = \text{diag}(\phi_1, \dots, \phi_k)$ is the Lagrange multiplier. Note that the pseudo-labeling strategy applied to the CDA might lead to a comparatively low accuracy for classification at first, yet this can be improved during the solution process. Specifically, the prediction results obtained by the previous iteration are utilized as the new pseudolabels for the current iteration, and thus the accuracy of pseudo-labels increases steadily with each iteration.

B. Multiscale Spectral-Spatial Network (MSSN)

In order to better exploit efficient spectral-spatial features of HSI, the MSSN is designed. As indicated in the graphical illustration in Fig. 2, the network is a two-branch CNN with a multiscale bank. Detail descriptions of MSSN are explained in the following sections.

1) *Two-Branch Architecture*: As shown in Fig. 2, two branches of the network are designed for spectral and spatial feature extraction separately. For the pixel x_n of hyperspectral dataset, the spectral branch takes the adapted spectral feature $\text{spec}_n = H^T x_n$ of the corresponding pixel as input data. After l layers of convolutional and max-pooling operations, the output of the spectral branch is represented as $F^l(\text{spec}_n)$, which can be regarded as the extracted spectral features. Since the input spec_n is 1-D signal, the convolutional and pooling operations correspond to 1-D computation. As for the spatial branch, considering the high dimensionality characteristic of HSI, principal component analysis (PCA) is performed on the spectral dimension, and the spatial neighboring patch $\text{spat}_n \in R^{r \times r}$ of the first principal component is utilized as the input data source of the branch. After l layers of convolutional and max-pooling operations, the output spatial features $F^l(\text{spat}_n)$ can be obtained. It is clearly that all convolutional and pooling operations in spatial branch belong to 2-D computations.

2) *Multiscale Bank*: Deep neural networks can learn multiple hierarchical nonlinear representation, which means that the extracted features in different convolution layer are corresponding to different scale characteristics. Specifically, the extracted features are usually some edges and textures during the first convolution layer, while the features obtained at the higher layer are likely to be the parts of the ground object. This concept

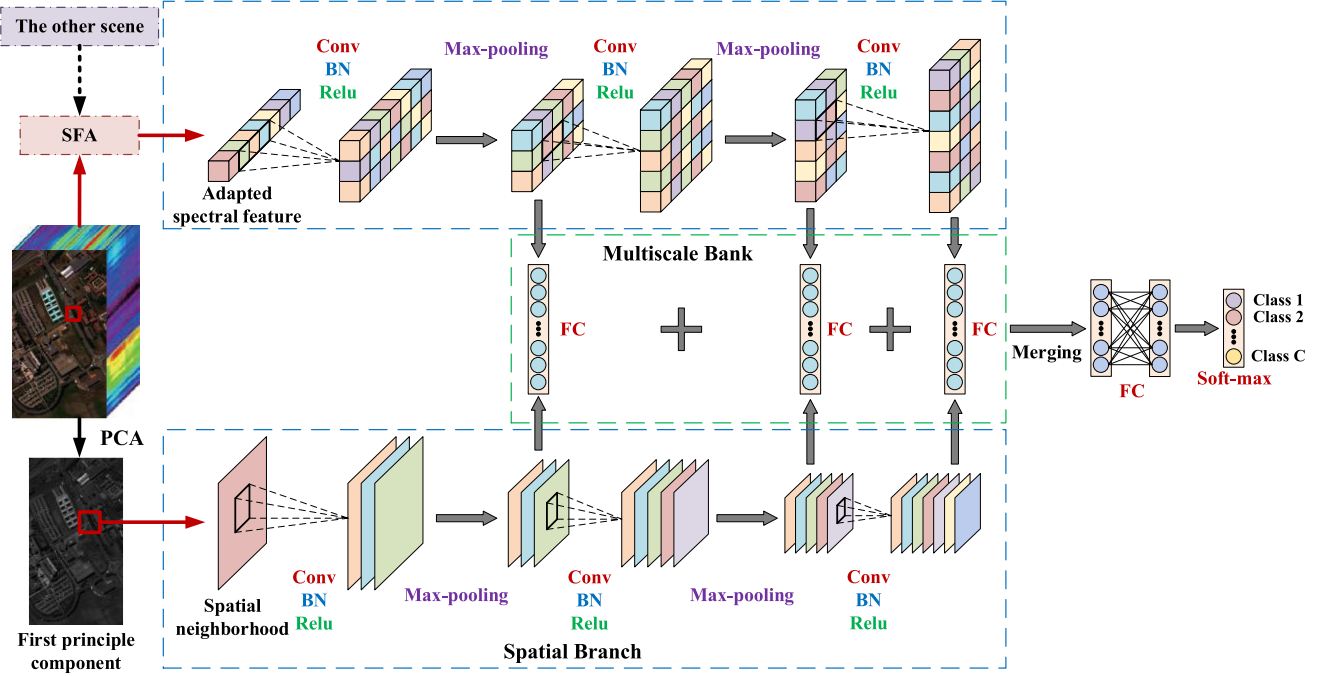


Fig. 2. Detailed architecture of MSSN.

has been applied to optimally exploit diverse local structures of HSI [42], whereas multiscale convolutional filter bank is used as an initial component of the CNN pipeline to provide higher classification accuracy.

Based on the elaboration above, a strategy that effectively combine the multiscale spectral and spatial features simultaneously is introduced to the spectral-spatial unified network. The multiscale bank is shown in the green dashed box in Fig. 2, where a fully-connected (FC) layer is added to each convolutional layer in both branches with the same size, and the outputs of these FC layers are merged into a new FC layer. Take the l th FC layer as an example, the corresponding output can be formulated as

$$F_n^l = f [W^l(spe^l + spa^l) + b^l] \quad (10)$$

where $f(*)$ is the activation function, W^l is the weight matrix and b^l is the bias term. spe^l and spa^l represent the flattened outputs of the l th max-pooling layer in spectral and spatial branches, respectively. The obtained F^l can be considered as the spectral-spatial feature extracted by the l th convolution layer. Finally, all outputs of the FC layers in the multiscale bank are merged to be a feature vector in a concatenate way

$$F_{\text{multi}} = \text{concat}(F^1, \dots, F^L) \quad (11)$$

where L is the number of the total FC layers of the multiscale bank. As the output of the proposed bank, F_{multi} is input to two FC layers to achieve the final multiscale feature F_{final} , and a soft-max regression layer is adopted to accomplish the classification task. Denote $\{\text{spec}_n, \text{spat}_n\}$ as the input data, and $F_{\text{final}}(n)$ as the corresponding feature learned by MSSN, the normalized probability that the input data belongs to i th

category is calculated by

$$p_i(n) = \frac{e^{W_i F_{\text{final}}(n)}}{\sum_{c=1}^C e^{W_c F_{\text{final}}(n)}}, \quad i = 1, \dots, C \quad (12)$$

where $W_c, c = 1, \dots, C$ is the c th row of the soft-max layer. For a labeled dataset contains N training samples, the loss function of the soft-max regression layer can be formulated as

$$\mathcal{J} = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C 1\{c = y_n\} \log p_c(n) \quad (13)$$

where y_n is the class label of the n th sample of the dataset, and $1\{*\}$ is an indicative function. The adaptive moment estimation (Adam) [43] is applied to optimize the loss function.

C. MSSN-Based Cross-Scene Deep Transfer Learning

In order to transfer the efficient contents from source domain to target domain, we train the MSSN with a cross-scene strategy that inspired by the transfer learning method reported in [44]. It is shown in the paper that the bottom layers of the deep network principally extract features that are universality for different datasets and well suited for cross-domain transfer learning. As for the top layers, the extracted features tend to be more abstract and carry with discriminant information of different datasets to be classified. In addition, experiments of the paper has proved that the classification accuracy can be enhanced effectively by continually applying the lower layers of the network trained by the source domain and fine-tuning the network with samples of the target domain. Based on these facts, we develop a cross-scene network training method for the proposed MSSN, which is summarized as follows.

- Step 1:* By inputting sufficient labeled samples of the source HSI to MSSN, the deep network is fully trained and thus achieves good feature extraction performance.
- Step 2:* The well-trained MSSN model is transferred to the target domain, with the bottom layers are remained the same as that of the source domain, while the parameters of the top layers are initialized randomly.
- Step 3:* The labeled samples of the target HSI are input to the network model for fine-tune operation, and thus the adaptive MSSN for target HSI classification is obtained.

IV. EXPERIMENTS AND ANALYSIS

A. Dataset Descriptions and Evaluation Indexes

Four real-world HSI datasets are applied during the experiments to validate the efficacy of the proposed cross-scene transfer learning architecture. Details of these datasets are described as follows.

The first dataset includes two scenes: Pavia Center and Pavia University. They were acquired during a flight over an urban area of Pavia in Northern Italy by ROSIS sensor, which contains spatial resolution of 1.3 m and high-resolution spectral information in the range from 430 to 860 nm. The Pavia center data has a spatial dimension of 1096×490 and 102 spectral bands, while the Pavia University consists of 103 bands of size 610×340 . Three-band false color images are shown in Fig. 3(a) and (b).

The second hyperspectral dataset Salinas was collected by AVIRIS sensor, it contains 16 labeled land cover classes and consists of 224 spectral bands over the 400–2500 nm wavelength range. The spatial dimension of Salinas is 512×217 pixels, and three-band false color image is shown in Fig. 3(c).

The third real dataset is the University of Houston campus and the neighboring urban area, which was captured by ITRES-CASI 1500 hyperspectral imager. It has 144 spectral channels that cover a wavelength range of 380–1050 nm. The spatial dimension of Houston data consists of 349×1905 pixels, and the resolution is 2.5 m. Fig. 3 shows the three-band false color image of Houston data.

The fourth dataset is Chikusei [45] captured by Headwall Hyperspec-VNIR-C imaging sensor in 2014. The ground sampling distance is 2.5 m, and the original image includes 2517×2235 pixels with 128 bands that covers a spectral range of 363–1018 nm. The three-band false color image of Chikusei is shown in Fig. 3(e). We choose a 2100×2200 pixel-size image in the experiment.

To evaluate the performance of the competing methods comprehensively, commonly used indexes, such as overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) are applied to record and assess the performance of different classification methods in our experiments. Among these metrics, OA measures the percentage of correctly classified pixels, AA indicates the average value of the percentage of pixels that are classified correctly for each class, and Kappa estimates the percentage of pixels which are classified correctly through a series of agreements [46]. All the classification results reported in this article are averaged values of over ten trials.

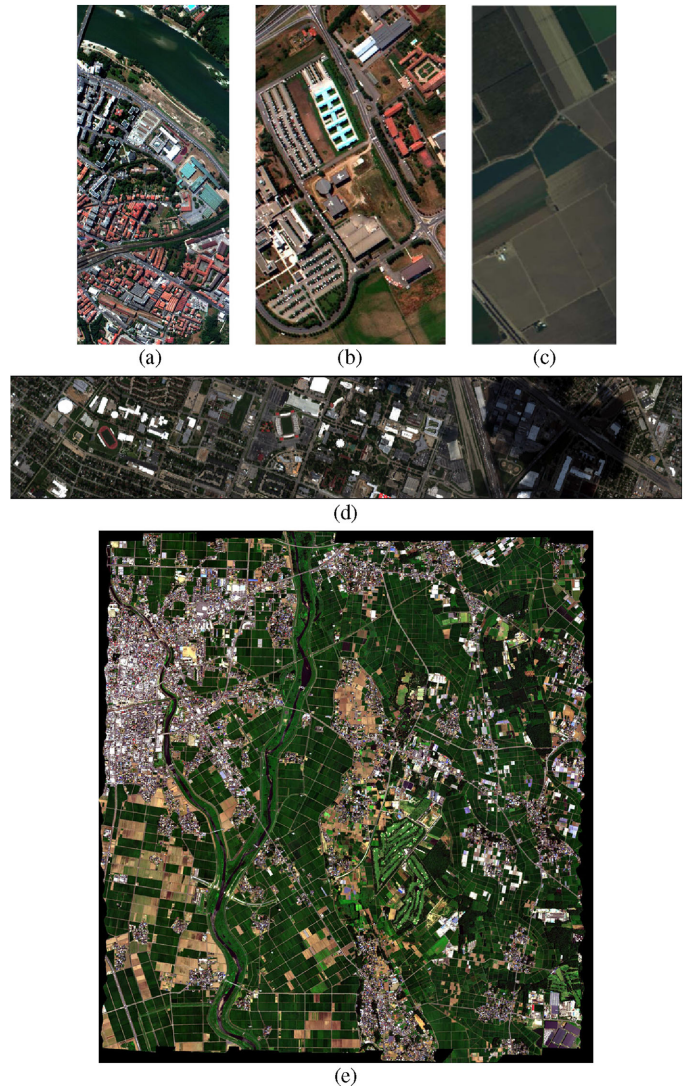


Fig. 3. False color images of the applied hyperspectral datasets. (a) Pavia Center. (b) Pavia University. (c) Salinas. (d) University of Houston campus. (e) Chikusei.

B. Rationality Analysis of the Proposed Model

In order to verify the rationality and effectiveness of the proposed method, some experiments are conducted on SFA part and MSSN part. Experimental results are specifically analyzed as follows.

1) *Analysis of SFA:* Two HSIs are chosen for the adaptation experiment: the Houston data (source domain) and University of Pavia (target domain). Since the adaptation process is developed for class-conditional distributions, only the co-occurrence classes of the two images are considered, including road, grass, trees and soil. Sample sizes of different classes are given in Table I. Fig. 4 shows the original spectral curves of the source and target HSI. It can be obviously seen from Fig. 4(a) and (b) that there are large differences between the original spectral curves of co-occurrence classes of the two images, and directly utilizing these spectral vectors will definitely trigger negative transfer phenomenon in classification. Note that only the first 103 bands

TABLE I
SAMPLE SIZE OF THE CO-OCCURRENCE CLASSES IN UNIVERSITY OF PAVIA AND HOUSTON DATASETS

Class	Sample size	
	University of Pavia	Houston
Road	6631	1059
Grass	18649	1053
Trees	3064	1056
Soil	5029	1056

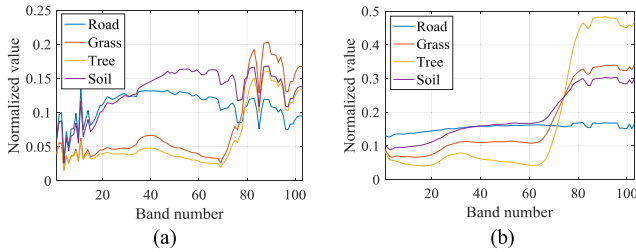


Fig. 4. (a) Original spectral curves of Houston data. (b) Original spectral curves of University of Pavia data.

of the source image that roughly equivalent to the band range of the target image are applied for adaptation since Houston data covers a wider spectral range compared with University of Pavia data.

In order to make quantitative analysis, we choose 200 samples and 500 samples of each class randomly from the source and target domains. During the domain adaptation process, only ten samples of each class are assumed to be labeled in target domain, while the rest ones are assumed to unlabeled. By simply adopting the support vector machine (SVM) classifier, we conduct cross-domain classification experiment on the data pair to measure the effectiveness of the proposed SFA method. In the experiment, we train SVM classifier with the samples of source domain and test the classification performance by samples of target domain. Two cases are compared, one is using the chosen original samples directly, the other is using the transformed features obtained by SFA. With the cross-domain classification experiment, we test the distribution adaptation model with regularization parameter λ of different magnitude. The results indicate that the optimal range is $[0.5, 1.0]$. In this experiment, λ is selected to be 0.56 with the highest classification accuracy.

Results of the cross-domain classification experiment are given in Table II. Compared with the results predicted by classifier that trained with original samples, the transformed spectral features applied for cross-scene classification can improve the accuracy by over 20% on average. To further illustrate the adaptation performance intuitively, the first two components of the transformed features are visualized in Fig. 5 (second row). Since domain adaptation of spectral vectors can be regarded as a feature transformation process essentially, we compare it with the traditional PCA approach (displayed in the first row). It can be clearly seen that in the new feature space transformed by SFA, the disparity between source and target domains is

TABLE II
CLASSIFICATION ACCURACY RESULTS ON UNIVERSITY OF PAVIA OBTAINED BY SVM CLASSIFIER

Class	Accuracy	
	Original spectrums	Transformed features
Road (%)	31.44	96.11
Grass (%)	31.44	39.44
Trees (%)	73.22	93.11
Soil (%)	40.67	49.33
OA (%)	44.19	69.90
AA (%)	44.19	69.90
Kappa (%)	25.59	59.87

TABLE III
PARAMETERS SETTING OF MSSN

Layer	Size	
	Spectral branch	Spatial branch
Conv1	$11 \times 11 \times 64$	$3 \times 3 \times 32$
Max-pooling1	2×1	2×2
Conv2	$11 \times 11 \times 128$	$3 \times 3 \times 128$
Max-pooling2	2×1	2×2
Conv3	$11 \times 11 \times 256$	$3 \times 3 \times 512$

comparatively reduced, and components of different samples from the same class are distributed in similar regions.

Although the experimental results in this subsection demonstrate that the proposed SFA can certainly strengthens the correlation between co-occurrence classes of source domain and target domain while extracting effectively spectral features, it is obvious that the clustering maps of grass and soil are not satisfactory enough and the corresponding classification accuracies are lower than 50%. This is due to the high similarity of original spectral curves between the two categories. Besides, the low-level spectral features obtained by linear transformation original spectral curves between the two categories. Besides, the low-level spectral features obtained by linear transformation are inadequate to accomplish accurate classification of the two categories. Therefore, we further complete the cross-scene transfer learning model by combining SFA with the proposed MSSN, which is analyzed in the following section.

2) *Settings and Rationality Analysis of MSSN*: The spectral branch and the spatial branch of MSSN have similar architectures. As depicted in Fig. 2, there are three convolutional layers and two max-pooling layers in each branch. The parameter setting of the two branches is given in Table III. Besides, a batch-normalization layer is added after each convolutional operation to avoid the vanishing gradient problem and accelerate the training process. All spatial neighboring patches input to the spatial branch are fixed to the size of 21×21 during the experiments. The number of neurons in each FC layer of the multiscale bank is set to 1024, and the two FC layers before soft-max classifier both contains 400 neurons. The MSSN model is implemented using the TensorFlow open source library. During the network training process, all convolutional kernels and weight matrix of the FC layers are initialized through the initialize function of the library, while the bias values are initialized as 0. The learning rate is fixed as 0.001 and remains unchanged during the whole

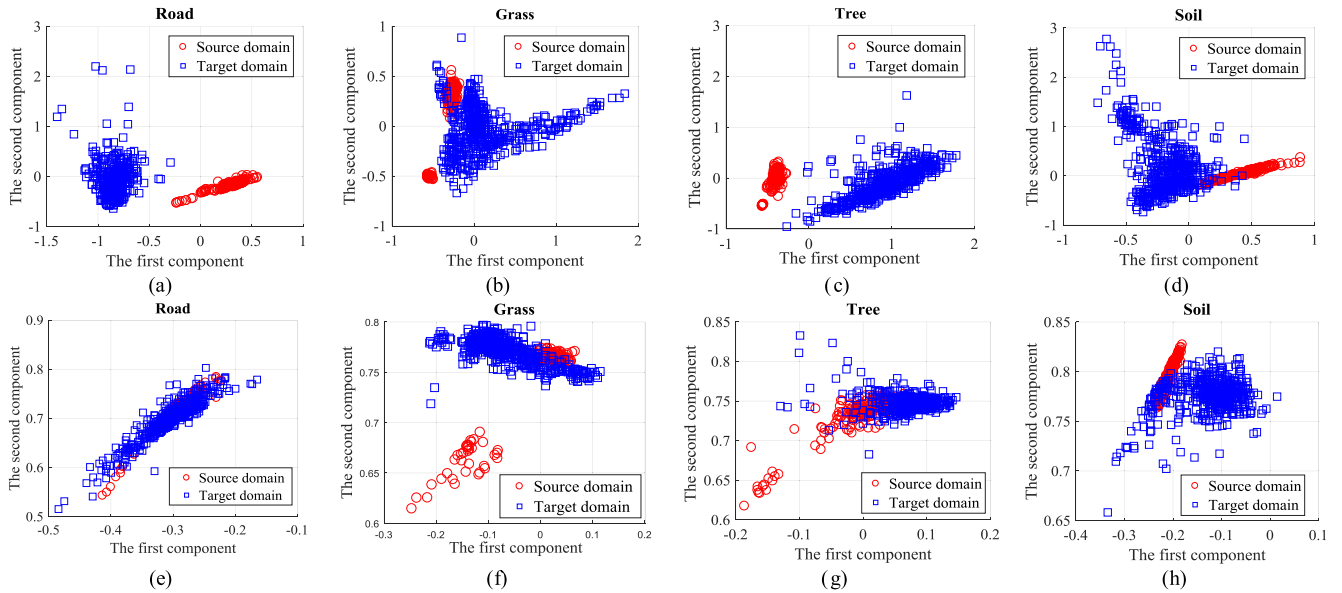


Fig. 5. Visualization of the first two spectral components obtained by PCA and distribution adaptation. (a) First two components of road obtained by PCA. (b) First two components of grass obtained by PCA. (c) First two components of tree obtained by PCA. (d) First two components of soil obtained by PCA. (e) First two components of road obtained by feature adaptation. (f) First two components of grass obtained by feature adaptation. (g) First two components of tree obtained by feature adaptation. (h) First two components of soil obtained by feature adaptation.

TABLE IV
CLASSIFICATION RESULTS FOR RATIONALITY ANALYSIS OF MSSN

Dataset	Indexes	MSSN-spec	MSSN-spat	MSSN-nomul	MSSN
University of Pavia	OA (%)	84.85	86.03	92.36	93.20
	AA (%)	87.92	87.14	92.60	93.46
	Kappa (%)	81.25	82.64	90.71	91.29
Salinas	OA (%)	88.13	89.78	92.27	92.44
	AA (%)	95.83	95.34	97.05	97.17
	Kappa (%)	89.52	90.37	91.78	92.83

procedure, and the batch size is set to 128. β_1 , β_2 and ε of Adam are all set to the default values. Maximal number of the training iterations is 1×10^3 .

In order to illustrate the significance of combining the spectral-spatial information with the multiscale strategy, we remove different part of the proposed MSSN to conduct a comparison research on the classification ability. First, the multiscale bank and one of the two branches from MSSN are removed, respectively, to testify the effectiveness of spectral-spatial feature extraction method, which means that only spectral or spatial features is exploited in the network. The network that simply contains spectral branch is represented as MSSN-Spec, while the spatial branch structured network is denoted by MSSN-Spat. Then, the multiscale bank is removed from the MSSN to test the validity of the multiscale bank. Since only the two-branch architecture is left, the network is denoted as MSSN-nomul.

Two datasets (University of Pavia and Salinas) are used for classification during the experiment, and 50 labeled samples of each class are randomly selected for network training. Classification results of the two images are given in Table IV. It can be clearly seen that the proposed MSSN outperforms other cases in both datasets, which demonstrates the dominant position of multiscale spectral-spatial feature extraction. Furthermore,

the MSSN-nomul achieves higher classification accuracy than MSSN-spec and MSSN-spat, which proves the superiority of the two-branch architecture.

C. Experimental Results

In this section, cross-scene transfer learning experiments are carried out on different data pairs, and the classification results of different methods are reported and discussed.

Three data pairs are tested during the experiments: the first is the Center of Pavia data (source domain) and the University of Pavia data (target domain), captured by same hyperspectral sensor. The second and the third are scenes acquired from different imagers, the data pairs are Salinas data (source domain) and University of Pavia data (target domain), and Houston data (source domain) and University of Pavia data (target domain).

1) *Experiment 1: Center of Pavia Data and University of Pavia Data:* In the first experiment, we conduct cross-scene transfer learning on the Center of Pavia (source domain) and University of Pavia (target domain). Co-occurrence categories and the corresponding sample sizes of the data pair are given in Table V. Since the two images are acquired by ROSIS sensor during a flight campaign, they share a strong correlation in

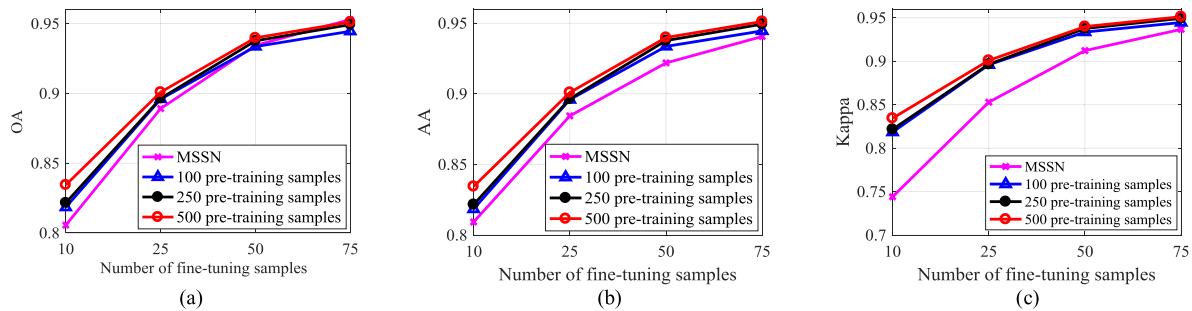


Fig. 6. Line charts of classification accuracies on University of Pavia dataset. (a) Overall accuracies. (b) Average accuracies. (c) Kappa coefficients.

TABLE V
SAMPLE SIZE OF THE CO-OCCURRENCE CLASSES IN CENTER OF PAVIA AND UNIVERSITY OF PAVIA

Class	Sample size	
	Center of Pavia	University of Pavia
Road	801	6631
Grass	6515	18649
Trees	20516	3064
Soil	2859	5029
Bitumen	7287	1330
Self-blocking bricks	2140	3682
Shadow	2165	947

both spectral and spatial dimensions. Therefore, the proposed transfer learning model is applied without the SFA process in this experiment. For the settings of network training, the bottom two layers of the pre-trained model are transferred, while the parameters of the third layer are initialized randomly.

In order to fully test the performance of the proposed MSSN model, we randomly choose 100, 250, and 500 samples of each class from the source dataset for network pre-training, and the corresponding fine-tune samples of the target domain are randomly chosen with the number of {10, 25, 50, 75}. We compare the proposed cross-scene transfer learning method with the case that only the samples of target domain are used for network training.

The classification results are given in Table VI. It can be observed that the classification accuracy of the target dataset can be effectively improved by applying our cross-scene transfer learning method, especially when the samples size of the target domain are relatively small. For instance, the transfer learning strategy brings increase of 1.79%, 1.83%, and 3.45% accuracy compared with the case that only ten labeled samples of the target domain are involved during the network training process.

Furthermore, we draw polylines to illustrate how the number of training samples influence the classification accuracy in Fig. 6. It is worth-while to note that the gaps between these polylines gradually close as the size of fine-tuning samples getting larger, which actually confirms the fact that transfer learning strategy is better suited to the case when labeled samples of target domain are scarce.

2) *Experiment 2: Houston data and University of Pavia Data:* The second experiment is conducted on Houston data (source domain) and University of Pavia data (target domain),

which are campus scenarios of two different cities. The spectral distributions of the two datasets are quite different since they were captured by different hyperspectral sensors, and the spatial resolution of the source image is lower than that of the target image. Common classes and the corresponding sample sizes of the data pair are already given in Table I.

Considering the different wavelength ranges, we chose the first 103 bands from the source image to make sure that the two datasets are of the same spectral dimension. Since our method is presented for the case where labeled samples of the target image are limited, we randomly chose {10, 25, 50, 75} labeled samples of each class from University of Pavia. Besides, 200 labeled samples are randomly chosen from Houston data to ensure the effectiveness of transferred knowledge and the sufficiency of network pretraining. During domain adaptation part, regularization parameter λ is set to be 0.56 according to the previous analysis. As for the cross-scene network training process, the first and second layers of the pre-trained MSSN are transferred to the target image, while the top convolutional layer is randomly initialized before fine-tuning.

In order to better illustrate the performance of the proposed method, three strategies are compared during the experiment: the traditional SVM, the MSSN trained by original labeled samples of the target domain (MSSN), and MSSN trained by spectral feature adapted labeled samples of the target domain (SFA-MSSNtar).

The obtained classification accuracies are given in Table VII, and it can be noted that the results yielded by the four methods are in a progressive relation. The MSSN achieves higher accuracies than SVM overall due to the reasonable network structure, yet the improvements are not obvious since the network cannot be fully trained under the situation of limited labeled samples. In contrast, SFA-MSSNtar performs better than MSSN by inputting SFA projected vectors to the spectral branch, which demonstrates the effectiveness of integrating SFA with MSSN model. Finally, by transferring lower layers of the network pretrained by samples of the source domain, the classification results can be further improved. For instance, the OA and AA can reach over 92% when only ten labeled samples of the target domain are applied for fine-tuning, and the classification accuracies of grass and soil are both up to 95% as the fine-tuning sample size is enlarged to 75 per class. The results indicate that the proposed method can overcome the distribution disparity problem in cross-scene transfer learning and thus effectively

TABLE VI
CLASSIFICATION RESULTS FOR THE FIRST DATA PAIR

Number of pre-training samples in source dataset	Indexes	Number of fine-tuning samples in target dataset			
		10 (per class)	25 (per class)	50 (per class)	75 (per class)
--	OA (%)	80.56	88.91	93.44	95.30
	AA (%)	80.93	88.44	92.19	94.06
	Kappa (%)	74.41	85.29	91.22	93.67
100 (per class)	OA (%)	82.35	90.25	94.12	95.49
	AA (%)	81.85	89.59	93.36	94.46
	Kappa (%)	76.69	87.04	92.12	93.93
250 (per class)	OA (%)	82.39	90.08	94.51	95.81
	AA (%)	82.16	89.65	93.77	94.94
	Kappa (%)	76.74	86.85	92.64	94.36
500 (per class)	OA (%)	84.01	90.58	94.70	95.97
	AA (%)	83.44	90.09	93.99	95.12
	Kappa (%)	78.81	87.48	92.89	94.58

TABLE VII
CLASSIFICATION RESULTS FOR UNIVERSITY OF PAVIA IN SECOND DATA PAIR

Number of labeled samples in target domain	Method	Accuracy						
		Road (%)	Grass (%)	Trees (%)	Soil (%)	OA (%)	AA (%)	Kappa (%)
10 (per class)	SVM	97.81	65.10	85.85	57.60	72.61	76.59	60.07
	MSSN	97.74	63.87	88.30	66.35	78.50	79.07	71.33
	SFA-MSSNtar	93.52	85.91	99.41	80.73	89.70	89.70	86.26
	Proposed	96.27	89.00	99.43	84.72	92.28	92.28	89.71
25 (per class)	SVM	98.07	77.05	91.95	74.32	82.39	85.35	74.08
	MSSN	98.23	69.20	95.46	80.00	85.08	85.72	80.10
	SFA-MSSNtar	96.69	84.09	99.35	83.27	90.56	90.85	87.41
	Proposed	97.43	90.22	99.79	89.31	94.15	94.19	92.20
50 (per class)	SVM	98.33	81.39	94.01	81.51	86.14	88.81	79.26
	MSSN	95.82	84.84	91.20	85.00	89.30	89.21	85.73
	SFA-MSSNtar	98.40	90.37	99.07	91.54	94.80	94.85	93.07
	Proposed	97.64	92.47	99.52	93.04	95.58	95.67	94.11
75 (per class)	SVM	98.58	87.19	95.06	88.37	90.51	92.30	85.65
	MSSN	99.20	85.08	94.74	88.09	91.75	91.78	89.00
	SFA-MSSNtar	98.85	92.24	99.72	93.61	96.07	96.10	94.76
	Proposed	98.61	95.27	99.80	96.14	97.43	97.45	96.60

solve the problem of limited samples in HSI classification. Fig. 7 displays the ground truth and classification maps obtained by different methods when the number of labeled samples is ten in Pavia University. It can be observed that the proposed transfer learning framework achieves much better classification performance than the other three with a result map that closest to the ground truth.

3) *Experiment 3: Chikusei Data and University of Pavia Data*: In the third experiment, cross-scene transfer learning study is carried on Chikusei data (source domain) and University of Pavia data (target domain) to further test the robustness of our method. According to the dataset descriptions, the two scenes are of different spectral ranges and spatial resolutions as they were captured by different imagers. Besides, the Chikusei data is in a larger scale that contains both agricultural and urban areas compared with University of Pavia, which means that the correlations between source and target images are even lower than that of the data pair in experiment 2.

Since the spectral dimensions of the two images are different, we choose 103 bands of the Chikusei data (band 13 to band 115) to match with the bands of University of Pavia data according to the wavelength range. Sample sizes of the co-occurrence classes

TABLE VIII
SAMPLE SIZE OF THE CO-OCCURRENCE CLASSES IN CHIKUSEI AND UNIVERSITY OF PAVIA

Class	Sample size	
	Chikusei	University of Pavia
Road	801	6631
Grass	6515	18649
Trees	20516	3064
Soil	2859	5029

in two datasets are given in Table VIII. Same with experiment 2 200 labeled samples and {10, 25, 50, 75} labeled samples of each class are randomly chosen from the source and target domain.

During the SFA process, spectral distribution differences between the two datasets are reduced by performing the domain adaptation method, and the regularization parameter λ of the objective function is selected as 0.58. The network training settings of MSSN model based transfer learning are the same as experiment 2: the pretrained bottom layers (first and second layers) are transferred to the target domain directly, while the parameters of top layer (third layer) is randomly initialized.

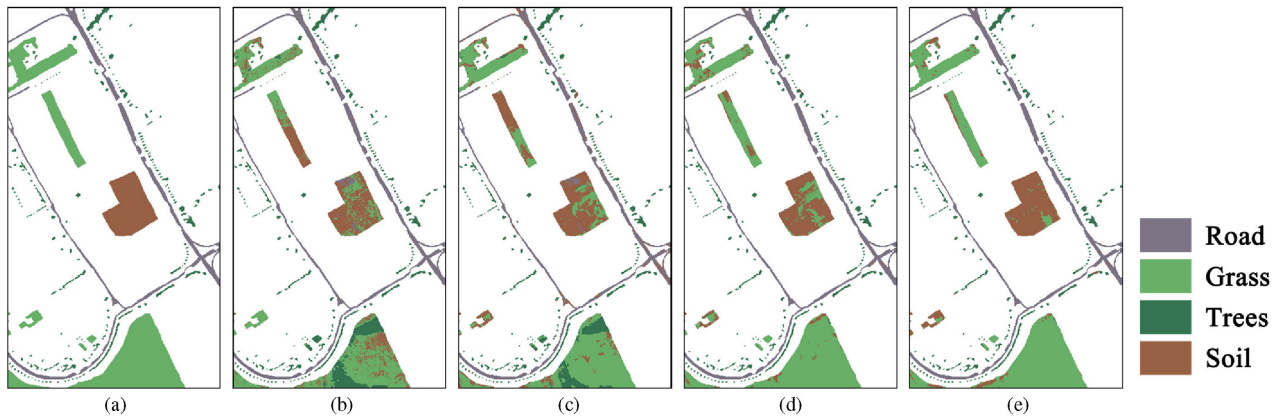


Fig. 7. Classification maps for the University of Pavia dataset of the second data pair. (a) Ground truth map. (b) SVM. (c) MSSN. (d) SFA-MSSNtar. (e) Proposed.

TABLE IX
CLASSIFICATION RESULTS FOR UNIVERSITY OF PAVIA IN THE THIRD DATA PAIR

Number of labeled samples in target domain	Method	Accuracy						
		Road (%)	Grass (%)	Trees (%)	Soil (%)	OA (%)	AA (%)	Kappa (%)
10 (per class)	SVM	97.81	65.10	85.85	57.60	72.61	76.59	60.07
	MSSN	97.74	63.87	88.30	66.35	78.50	79.07	71.33
	SFA-MSSNtar	97.86	74.43	98.32	80.99	87.21	87.90	82.95
	Proposed	98.96	83.27	99.19	84.57	91.30	91.50	88.40
25 (per class)	SVM	98.07	77.05	91.95	74.32	82.39	85.35	74.08
	MSSN	98.23	69.20	95.46	80.00	85.08	85.72	80.10
	SFA-MSSNtar	96.72	81.35	99.62	85.76	90.51	90.87	87.36
	Proposed	97.64	88.07	99.64	87.35	93.05	93.18	90.73
50 (per class)	SVM	98.33	81.39	94.01	81.51	86.14	88.81	79.26
	MSSN	95.82	84.84	91.20	85.00	89.30	89.21	85.73
	SFA-MSSNtar	98.16	88.80	99.90	91.41	94.43	94.57	92.57
	Proposed	98.50	92.54	99.04	93.34	95.82	95.85	94.43
75 (per class)	SVM	98.58	87.19	95.06	88.37	90.51	92.30	85.65
	MSSN	99.20	85.08	94.74	88.09	91.75	91.78	89.00
	SFA-MSSNtar	99.50	90.24	99.70	94.95	95.95	96.10	94.60
	Proposed	98.88	94.36	99.40	96.03	97.15	97.17	96.20

TABLE X
RUNTIME (SECONDS) OF THE PROPOSED METHOD

Data pair	SFA	Pre-train			Fine-tune			
		100 (per class)	250 (per class)	500 (per class)	10 (per class)	25 (per class)	50 (per class)	75 (per class)
First (7 common classes)	--	70.74	143.24	268.75	42.13	48.11	55.28	62.98
			200 (per class)		10 (per class)	25 (per class)	50 (per class)	75 (per class)
Second (4 common classes)	9.43	79.66			26.36	27.85	31.99	36.55
			200 (per class)		10 (per class)	25 (per class)	50 (per class)	75 (per class)
Third (4 common classes)	9.87	80.23			25.94	27.55	32.36	36.78

Finally, the classifier network is obtained by performing fine-tune operation on the transferred MSSN.

Table IX shows the classification results of the experiment. Comparing with the traditional SVM, MSSN trained by original samples of the target domain, and MSSN trained by spectral feature adapted labeled samples of the target domain (SFA-MSSNtar), the proposed method yields the highest accuracies. Similar to the case in experiment 2, SFA-MSSNtar can improve the results to some extent when comparing with MSSN, and the proposed method further enhances the classification accuracies. Specifically, compared with applying SVM to the raw data, the proposed method increases OA from 72.61% to 91% by using ten labeled samples in the considered image. In addition, the

experimental data shows that the classification accuracies of grass and soil can be greatly enhanced by transferring relevant knowledge from the source image. As for road and trees, the improvements are likely to be limited since the classification accuracies acquired by traditional SVM are already up to 97% even only ten samples are applied to train the classifier. It should be noted that the margin between these methods is narrowed as the sample size of the target domain enlarged, which gains coincident conclusion with experiment 1 that transfer learning can be more efficient to the situation of small sample size.

Finally, runtime of the proposed method conducted on these data pairs is reported in Table X. The computation cost of the proposed approach mainly contains two parts: the SFA and the

MSSN based deep transfer learning which contains pretraining and fine-tuning processes. The SFA is performed by using MATLAB, and the training process of MSSN is implemented on Pycharm platform with an NVIDIA GTX 1070 graphic card. Each training time recorded in the table is computed in the case where training iteration is taken as the maximum value. Obviously, it takes longer to train the network with the increase of the number of training samples. Note that testing on the whole dataset costs less than one second, since only the feed-forward propagation is performed.

V. CONCLUSION

In this article, we presented a novel cross-scene deep transfer learning model for HSI classification. The model contains two parts. With the proposed SFA in the first part, the distribution differences between spectral vectors acquired by different sensors can be explicitly narrowed. In the second part, the MSSN is designed to exploit hierarchical spectral-spatial features of hyperspectral data. Through a cross-scene network training strategy based on MSSN, relevant knowledge of the source dataset can be efficiently transferred to the target dataset to help with the classification task. Experimental results have shown that the method can be very meaningful in dealing with the small-sample problem in HSI classification, as it can significantly overcome the cross-domain disparity and achieve comparatively ideal HSI classification accuracies.

It is worth noting that the research in this article is still at an early stage, since the co-occurrence categories appeared in available hyperspectral datasets captured by different sensors are relatively limited. Therefore, cross-scene transfer learning experiments should be conducted on more hyperspectral data pairs with diverse co-occurrence categories to further verify the effectiveness of the proposed method.

ACKNOWLEDGEMENT

The authors would like to thank Prof. Gamba for providing the Pavia scenes, and the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) for providing the Houston dataset. We also gratefully acknowledge the Space Application Laboratory, Department of Advanced Interdisciplinary Studies, the University of Tokyo for providing the Chikusei dataset.

REFERENCES

- [1] A. Zare, J. Bolton, J. Chanussot, "Foreword to the special issue on hyperspectral image and signal processing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1841–1843, Jun. 2014.
- [2] W. Kim and M. M. Crawford, "Adaptive classification for hyperspectral image data using manifold regularization kernel machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4110–4121, Nov. 2010.
- [3] S. Rajan, J. Ghosh, and M. M. Crawford, "Exploiting class hierarchies for knowledge transfer in hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3408–3417, Nov. 2006.
- [4] D. Tuia, E. Pasolli, and W. J. Emery, "Using active learning to adapt remote sensing image classifiers," *Remote Sens. Environ.*, vol. 115, no. 9, pp. 2232–2242, Sep. 2011.
- [5] Z. Sun, C. Wang, P. Li, H. Wang, and J. Li, "Hyperspectral image classification with SVM-based domain adaption classifiers," in *Proc. Int. Conf. Comput. Vis. Remote Sens.*, Dec. 2012, pp. 268–272.
- [6] Z. Sun, C. Wang, H. Wang, and J. Li, "Learn multiple-kernel SVMs for domain adaptation in hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1224–1228, Sep. 2013.
- [7] D. Tuia, C. Persello, and L. Bruzzone, "Domain adaptation for the classification of remote sensing data: An overview of recent advances," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 41–57, Jun. 2016.
- [8] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
- [9] J. Xia, N. Yokoya, and A. Iwasaki, "Ensemble of transfer component analysis for domain adaptation in hyperspectral remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2017, pp. 4762–4765.
- [10] M. Ye, Y. Qian, J. Zhou, and Y. Y. Tang, "Dictionary learning-based feature-level domain adaptation for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1544–1562, Mar. 2017.
- [11] J. Shen, X. Cao, Y. Li, and D. Xu, "Feature adaptation and augmentation for cross-scene hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 622–626, Apr. 2018.
- [12] H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep-learning systems for domain adaptation in computer vision: Learning transferable feature representations," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 117–129, Nov. 2017.
- [13] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
- [14] J. Li, H. Zhang, Y. Huang, and L. Zhang, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 33, no. 3, pp. 53–69, May 2015.
- [15] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.
- [16] C. Yu *et al.*, "Hyperspectral image classification method based on CNN architecture embedding with hashing semantic feature," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1866–1881, Jun. 2019.
- [17] J. Yang, Y. Zhao, and J. C. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
- [18] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. Int. Conf. Proc. Ser.*, Jun. 2007, pp. 759–766.
- [19] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2693–2705, May 2017.
- [20] J. Lin, R. Ward and Z. J. Wang, "Deep transfer learning for hyperspectral image classification," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process.*, Aug. 2018, pp. 1–5.
- [21] H. Chen, M. Ye, H. Lu, L. Lei, and Y. Qian, "Dual dictionary learning for mining a unified feature subspace between different hyperspectral image scenes," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 1096–1099.
- [22] S. Wu, J. Zhang, and C. Zhong, "Multiscale spectral-spatial unified networks for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 2706–2709.
- [23] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.
- [24] H. Wang, F. Nie, H. Huang, and C. Ding, "Dyadic transfer learning for cross-domain image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 551–556.
- [25] B. Banerjee *et al.*, "A novel graph-matching-based approach for domain adaptation in classification of remote sensing image pair," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 4045–4062, Jul. 2015.
- [26] Z. Xue, J. Li, L. Cheng, and P. Du, "Spectral-spatial classification of hyperspectral data via morphological component analysis-based image separation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 70–84, Jan. 2015.
- [27] L. Huo and P. Tang, "Spectral and spatial classification of hyperspectral data using SVMs and Gabor textures," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 1708–1711.

- [28] T. C. Bau, S. Sarkar, and G. Healey, "Hyperspectral region classification using a three-dimensional Gabor filterbank," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3457–3464, Sep. 2010.
- [29] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.
- [30] S. Jia, K. Wu, J. Zhu, and X. Jia, "Spectral-spatial gabor surface feature fusion approach for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1142–1154, Feb. 2019.
- [31] X. Kang, S. Li, L. Fang, and J. A. Benediktsson, "Intrinsic image decomposition for feature extraction of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2241–2253, Apr. 2015.
- [32] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 911–923, Feb. 2019.
- [33] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.
- [34] Y. Qian, M. Ye, and J. Zhou, "Hyperspectral image classification based on structured sparse logistic regression and three-dimensional wavelet texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2276–2291, Apr. 2013.
- [35] L. Wang, X. Cao, Y. Zheng, and Q. Dai, "Multiscale feature extraction of hyperspectral image with guided filtering," *J. Remote Sens.*, vol. 22, no. 2, pp. 293–303, Mar. 2018.
- [36] T. Dundar and T. Ince, "Sparse representation-based hyperspectral image classification using multiscale superpixels and guided filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 246–250, Feb. 2019.
- [37] S. Li, Q. Hao, X. Kang, and J. A. Benediktsson, "Gaussian pyramid based multiscale feature fusion for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3312–3324, Sep. 2018.
- [38] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, 2015.
- [39] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2200–2207.
- [40] K. M. Borgwardt *et al.*, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, pp. 49–57, 2006.
- [41] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. Nat. Conf. Artif. Intell.*, Jul. 2008, pp. 677–682.
- [42] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
- [43] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, May 2015.
- [44] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 3320–3328.
- [45] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," *Space Appl. Lab.*, Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, May 2016.
- [46] R. Nishii and S. Tanaka, "Accuracy and inaccuracy assessments in land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 491–498, Jan. 1999.



Chongxiao Zhong (Student Member, IEEE) received the B.S. degree in communication engineering from the Northeastern University, Shenyang, China, in 2014, and the M.S. degree in electronics and communication engineering in 2017 from the Harbin Institute of Technology, Harbin, China, where she is currently working toward the Ph.D. degree.

Her current research interests include image processing, hyperspectral image classification, transfer learning, and application.



Junping Zhang (Member, IEEE) received the B.S. degree in biomedical engineering and instrument from Harbin Engineering University and Harbin Medical University, Harbin, China, in 1993, and the M.S. and Ph.D. degrees in signal and information processing from the Harbin Institute of Technology (HIT), Harbin, China, in 1998 and 2002, respectively.

She is currently a Professor with the Department of Information Engineering, School of Electronics and Information Engineering, HIT. Her research interests include hyper-spectral data analysis and image processing, multisource information fusion, pattern recognition, and classification.



Sifan Wu received the B.S. degree in electronic information engineering from the Harbin Institute of Technology, Weihai, China, in 2017, and the M.S. degree in electronics and communication engineering from the Harbin Institute of Technology, Harbin, China, in 2019.

He is currently a Software Engineer with Xiamen Airlines. His current research interests include image processing and computer vision.



Ye Zhang (Member, IEEE) received the B.S. degree in communication engineering, and the M.S. and Ph.D. degrees in communication and electronic system from the Harbin Institute of Technology (HIT), Harbin, China, in 1982, 1985, and 1996, respectively.

Since 1985, he has been a Teacher with HIT. Between 1998 and 1999, he was a Visiting Scholar with the University of Texas at San Antonio. He is currently a Professor and a Doctoral Supervisor in information and communication engineering. He is the Director of Institute of Image and Information

Technology with the School of Electronic and Information Engineering, HIT. His research interests are remote sensing hyperspectral image analysis and processing, image video compression and transmission, as well as multisource information collaboration processing and applications.