Joint and Progressive Subspace Analysis (JPSA) With Spatial–Spectral Manifold Alignment for Semisupervised Hyperspectral Dimensionality Reduction

Danfeng Hong, *Member, IEEE*, Naoto Yokoya, *Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, Jian Xu, *Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*—Conventional nonlinear subspace learning techniques (e.g., manifold learning) usually introduce some drawbacks in explainability (explicit mapping) and cost effectiveness (linearization), generalization capability (out-of-sample), and representability (spatial–spectral discrimination). To overcome these shortcomings, a novel linearized subspace analysis technique with spatial–spectral manifold alignment is developed for a semisupervised hyperspectral dimensionality reduction (HDR), called joint and progressive subspace analysis (JPSA). The JPSA learns a high-level, semantically meaningful, joint spatial–spectral feature representation from hyperspectral (HS) data by: 1) jointly learning latent subspaces and a linear classifier to find an effective projection direction favorable for classification; 2) progressively searching several intermediate states of subspaces to approach an optimal mapping from the original space to a potential more discriminative subspace; and 3) spatially and spectrally aligning a manifold structure in each learned latent subspace in order to preserve the same or similar topological property between the compressed data and the original data. A simple but effective classifier, that is, nearest neighbor (NN), is explored as a potential application for validating the algorithm performance of different HDR approaches. Extensive experiments are conducted to demonstrate the superiority and effectiveness of the proposed JPSA on two widely used HS datasets: 1) Indian Pines (92.98%) and 2) the University of Houston (86.09%) in comparison with previous state-of-the-art HDR methods. The demo of this basic work (i.e., ECCV2018) is openly available at https://github.com/danfenghong/ECCV2018_J-Play.**

***Index Terms*—Dimensionality reduction (DR), hyperspectral (HS) data, joint learning, manifold alignment, progressive learning, semisupervised, spatial–spectral, subspace learning (SL).**

Manuscript received July 15, 2019; revised August 12, 2020; accepted September 21, 2020. Date of publication November 11, 2020; date of current version June 23, 2021. The work of Danfeng Hong was supported in part by the German Research Foundation (DFG) under Grant ZH 498/7-2, and in part by the Helmholtz Association through the Framework of Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Local Unit "Munich Unit @ Aeronautics, Space and Transport (MASTr)." The work of Naoto Yokoya was supported by the Japan Society for the Promotion of Science (KAKENHI) under Grant 18K18067. The work of Jocelyn Chanussot was supported by the AXA Research Fund. The work of Xiao Xiang Zhu was supported in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (*So2Sat*) under Grant ERC-2016-StG-714087; in part by the HAICU—MASTr and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research," and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond." This article is an extended version in algorithm of [1] published in ECCV2018. This article was recommended by Associate Editor S. Ventura. *(Corresponding author: Xiao Xiang Zhu.)*

Danfeng Hong is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany, and also with the GIPSA-lab, Grenoble INP, CNRS, Université Grenoble Alpes, 38000 Grenoble, France (e-mail: hongdanfeng1989@gmail.com).

Naoto Yokoya is with the Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-8561, Japan, and also with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: naoto.yokoya@riken.jp).

Jocelyn Chanussot is with the INRIA, CNRS, Grenoble INP, LJK, Université Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn@hi.is).

Jian Xu is with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: jian.xu@dlr.de).

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.3028931.

Digital Object Identifier 10.1109/TCYB.2020.3028931

I. Introduction

HYPERSPECTRAL (HS) data are often characterized by rich and diverse spectral information, which enables us to locate or identify targets more easily. However, their high dimensionality also raises some crucial issues that need to be carefully considered, including information redundancy, complex noise effects, the need for large storage capacities and high-performance computing, and the curse of dimensionality. A general way to address this problem is to compress the original data to a low-dimensional and highly discriminative subspace with the preservation of the topological structure. In general, it is also referred to as dimensionality reduction (DR) or subspace learning (SL).

Over the past decade, SL techniques have been widely used in remote sensing data processing and analysis [2]–[11], particularly hyperspectral DR (HDR) [12]. Li *et al.* [13] carried out the HDR and classification by locally preserving neighborhood relations. In [14], spectral-spatial noise estimation can largely enhance the performance of DR, and the proposed method not only can extract high-quality features

but also can well deal with nonlinear problems in hyperspectral (HS) image classification. Huang *et al.* [15] introduced the sparseness property [16] into the to-be-estimated subspace in order to better structure the low-dimensional embedding space. Rasti *et al.* [17] extracted the HS features in an unsupervised fashion using the orthogonal total Variation component analysis (OTVCA), yielding a smooth spatial–spectral HSI feature extraction. In [18], spatial–spectral manifold (SSM) embedding was developed to compress the HS data into a more robust and discriminative subspace. Wang *et al.* [19] proposed selecting representative features hierarchically by the means of random projection in an end-to-end neural network, which has shown the effectiveness in the large-scale data. Very recently, Huang *et al.* [20] followed the trail of drawbacks of spatial–spectral techniques, and fixed them by designing a new spatial–spectral-combined distance to select spatial–spectral neighbors of each HS pixel more effectively. In the combined distance, the pixel-to-pixel distance measurement between two spectral signatures is converted to the weighted summation distance between spatially adjacent spaces of the two target pixels.

Despite the good performance of these methods in HDR, yet most of them only adhere to either the unsupervised or the supervised strategy, and fail to jointly consider the labeled and unlabeled information in the process of HDR. Some recent works for semisupervised HDR have been proposed by the attempt to preserve the potentially global data structure that lies in the entire high-dimensional space. For example, Liao *et al.* [21] simultaneously exploited labeled and unlabeled data to extract the feature representation from the HSI in a semisupervised fashion, called semisupervised local discriminant analysis (SELD). Different from [21] that utilizes the similarity measurement to construct the graph structure, in [22], the performance of local discriminant analysis (LDA) is enhanced with the joint use of the labels and "soft-labels" predicted by label propagation, yielding a soft-label LDA (SLLDA) for semisupervised HDR. A similar semisupervised strategy was presented in [23] to reduce the spectral dimension of HSI by embedding pseudolabels obtained using the pretrained classifier into local Fisher discriminant analysis (LFDA), called semisupervised LFDA (SSLFDA). The use of "soft-labels" or "pseudolabels" is effective for the process of low-dimensional embedding. Since more pixels considered can help us better capture the global manifold of the data, even though these soft or pseudolabels could be noisy and inaccurate. It should be noted that these techniques are commonly applied as a disjunct feature learning step before classification, whose limitation mainly lies in a weak connection between features by SL and label space (see the top panel of Fig. 1). It is unknown which learned features can accurately improve the classification. In [24], the features can adequately be exploited by using the *t*-distributed stochastic neighbor embedding and a multiscale scheme, and the proposed neural network shows outstanding and reliable performance in HS image classification.

A feasible solution to this problem can be generalized into a joint learning framework [26] that simultaneously learns a linearized subspace projection and a classifier, as illustrated in the middle panel of Fig. 1. Inspired by it, a large amount of work has been proposed for various applications, such as cross-modality learning and retrieval [27], and heterogeneous joint features learning [28]. Although these works have tried to make a connection between the learned subspaces and label information using regression techniques (e.g., linear regression) to adaptively find a latent subspace in favor of classification, they fail to find an optimal subspace. It is that the representative ability only using a single linear projection remains limited for the complex transformation from the original data space to the potential optimal subspace. Similar to the joint learning model, deep neural networks (DNNs) have attracted increasing attention due to its powerful ability in HS feature extraction. Chen *et al.* [29] designed a stacked autoencoder (SAE) for feature extraction and classification of HSI. Kemker and Kanan [30] investigated the performance of self-taught feature learning [e.g., convolutional autoencoder (CAE)] by jointly considering the spatial–spectral information embedding with the application to HSI classification.

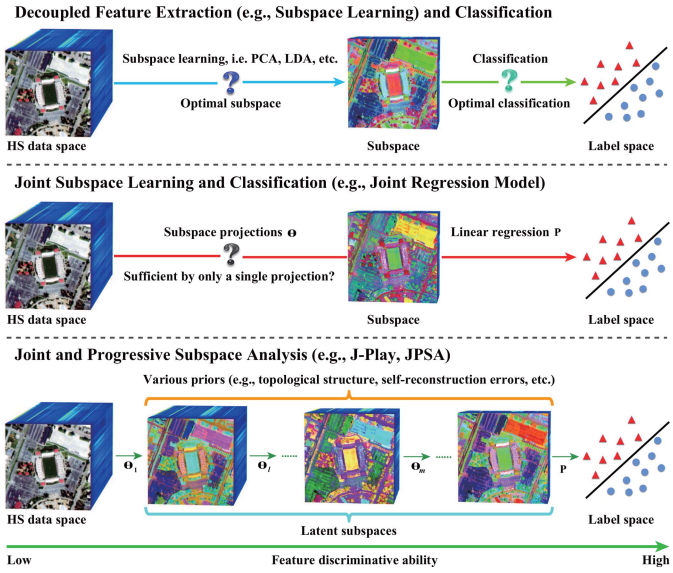


Fig. 1. Motivation interpolation from separately learning subspaces and training classifier [25], to jointly learning subspaces and classifier [26], to joint and progressive learning multicoupled subspaces and classifier again [1]. The green bottom line from left to right indicates a gradual improvement in feature discriminative ability. Ideally, the features (subspaces) learned by our model are expected to have a higher discrimination ability, which benefits from the proposed joint and progressive learning strategy.

A. Motivation and Objectives

To sum up, these aforementioned methods can be approximately categorized into linear HDR and nonlinear HDR techniques. Consequently, the strengths and weaknesses of the two methods can be summarized as follows.

1) Theoretically, nonlinear HDR strategies, such as manifold learning [31] and DNN-based DR methods (e.g., SAE and CAE) [32], can overfit the data perfectly, owing to their powerful model learning capability. However, this type of method is relatively sensitive to complex spectral variability inevitably caused by complex noise, atmospheric effects, and various physical and chemical

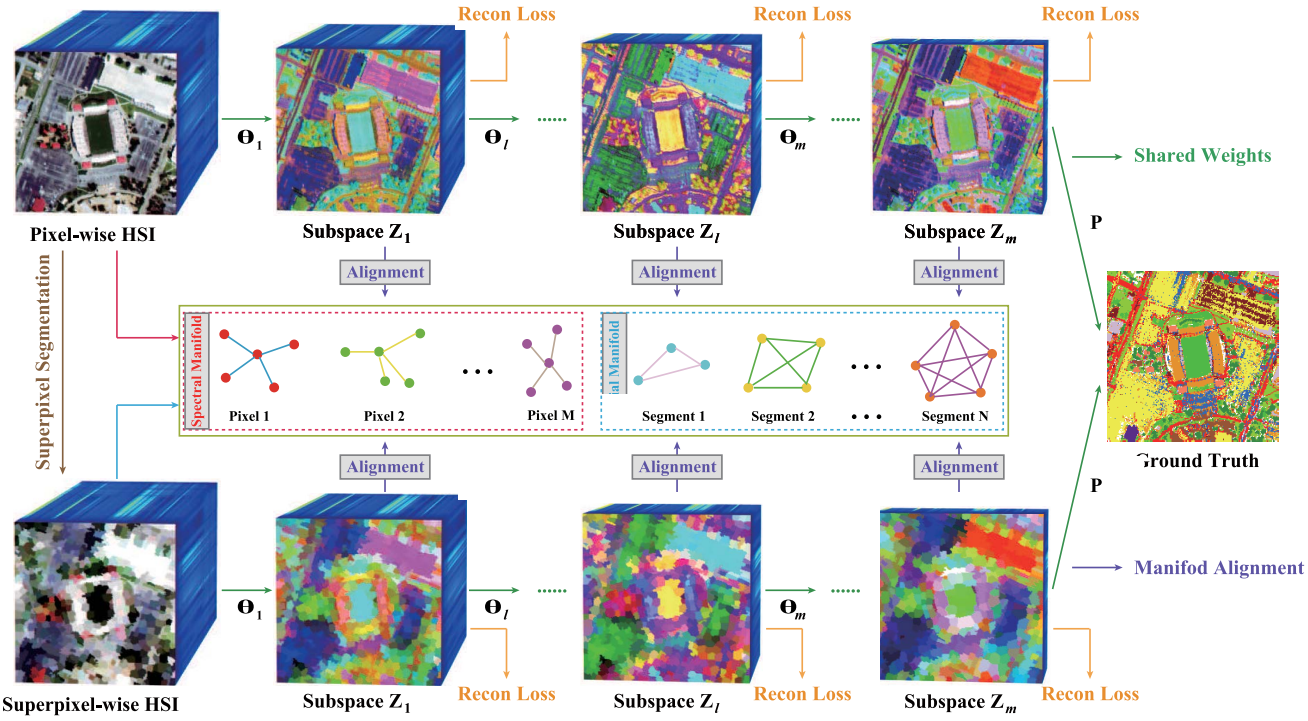


Fig. 2. Illustration of the proposed JPSA framework.

factors in HS imaging. Because the spectral variability tends to be absorbed by the DNN-based methods [33], the discriminative ability of the dimension-reduced feature gets possibly hurt.

2) In turn, the linearized SL methods, such as principal component analysis (PCA) [34], linearized manifold learning (e.g., locality preserving projection (LPP) [35], LDA [25], and LFDA [36]) can well address the above drawbacks, yet they usually provide limited performance due to the defects of the model itself, that is, the single linearized model is lack of data representation ability.

The above tradeoff motivates us to develop a multilayered linearized SL technique for HDR with more discriminative and robust data representation and to preserve the structural consistency between the compressed data and the original data.

B. Method Overview and Contributions

To effectively pursue high spectral discrimination and preservation of the spatial–spectral topological structure in compressing the HS data, we propose a novel joint and progressive subspace analysis (JPSA) to linearly find an optimal subspace for the low-dimensional data representation, as shown in the bottom panel of Fig. 1. A promising idea of simultaneous SL and classification is used to form the basic skeleton of the proposed JPSA model. In the framework, we learn a series of subspaces instead of a single subspace, making the original data space being progressively converted to a potentially optimal subspace through multicoupled intermediate transformations. To avoid trivial solutions, a self-reconstruction (SR) strategy in the form of regularization is applied in each latent subspace. Furthermore, we not only consider structure consistency (topology) between the compressed data and the original data in both spatial and spectral domains but also align the two (spatial and spectral) manifolds in each latent subspace, yielding the SSM embedding in the process of HDR.

Beyond previous existing works, that is, [1] and [37], the main contributions of our work can be summarized as follows.

1) We develop a novel semisupervised HDR framework (JPSA) for better learning the spatial–spectral low-dimensional embedding by modeling relations between superpixels and pixels in a joint and progressive fashion.
2) With the SR term simultaneously performed on superpixels and pixels, the linearized JPSA shows its robustness and effectiveness in handling the spectral variability over many nonlinear HDR approaches, which will be well demonstrated in the following experiment section.
3) SSMs are preserved in each latent subspace and are further aligned for spatial–spectral structure consistency between the compressed data and the original data, where the manifold structure in spectral space is computed by Gaussian kernel function, and the spatial manifold structure is determined by superpixels, e.g., simple linear iterative clustering (SLIC) [38].
4) To avoid falling into bad local optimums, a pretraining model, called autoreconstructing unsupervised learning (AutoRULe), is proposed as an initialization of JPSA to jointly initialize the branches of pixels and superpixels.
5) An iterative optimization algorithm based on the alternating direction method of multipliers (ADMMs) is designed to solve the newly proposed model.

II. JPSA: Joint and Progressive Subspace Analysis

Fig. 2 illustrates the workflow of the proposed JPSA. Intuitively, the JPSA is a two-stream multilayered regression model involving the two input sources: 1) pixelwise and

2) superpixelwise spectral signatures and the same output (ground truth). In the learning process of the two-stream model, the to-be-estimated parameters (projections) are shared with a spatial–spectral alignment constraint in each latent subspace. Moreover, each learned subspace is expected to be capable of projecting back to its former high-dimensional product, which is measured by the reconstruction loss.

A. Review of Joint Regression

Before introducing our JPSA, we first briefly introduce the basis of developing our method: a joint regression model [26], in which SL and classification are simultaneously performed to reduce the gap between the estimated subspace and labels. This model has been proven to be effective in extracting the discriminative low-dimensional representation [39]. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_N] \in \mathbb{R}^{d_0 \times N}$ be an HS data matrix with d_0 bands by N pixels, and $\mathbf{Y} \in \{0,1\}^{L \times N}$ be the one-hot encoded class matrix corresponding to labels, whose kth column is defined as $\mathbf{y}_k = [\mathbf{y}_{k1}, \ldots, \mathbf{y}_{kt}, \ldots, \mathbf{y}_{kL}]^T \in \mathbb{R}^{L \times 1}$, we then have

$$\min_{\mathbf{P},\boldsymbol{\Theta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{P}\boldsymbol{\Theta}\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \ \text{ s.t. } \ \boldsymbol{\Theta}\boldsymbol{\Theta}^T = \mathbf{I} \tag{1}$$

where $\|\bullet\|_{\mathrm{F}}$ represents a Frobenius norm, $\mathbf{P} \in \mathbb{R}^{L \times d_m}$ (d_m denotes the dimension of the latent subspace) is regression matrix to explicitly bridge the learnt latent subspace and labels, and the projection $\boldsymbol{\Theta} \in \mathbb{R}^{d_m \times d_0}$ is usually called as intermediate transformation and the corresponding subspace $\boldsymbol{\Theta}\mathbf{X}$ is called the latent subspace. It has been proven in [40] that the feature is prone to be structurally learned and represented in such a latent subspace.

Furthermore, by considering the graph structure measured by an adjacency matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ as a regularizor [41], the joint regression model in (1) can be extended to the following improved version [37]:

$$\begin{aligned} \min_{\mathbf{P},\boldsymbol{\Theta}} \ & \frac{1}{2}\|\mathbf{Y} - \mathbf{P}\boldsymbol{\Theta}\mathbf{X}\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\operatorname{tr}\big(\boldsymbol{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{\Theta}^T\big) \\ \text{s.t.} \ & \boldsymbol{\Theta}\boldsymbol{\Theta}^T = \mathbf{I} \end{aligned} \tag{2}$$

where $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$ is defined as a degree matrix and the Laplacian matrix $\mathbf{L}$ can be computed by $\mathbf{L} = \mathbf{D} - \mathbf{W}$ [42]. The third term of (2), that is, graph regularization, can provide additional prior knowledge by modeling relations between samples, thereby improving the regression performance.

B. Problem Formulation

A single linear transformation is hardly capable of describing the complex mapping relationship between the data and labels well, particularly for HS data suffering from a variety of spectral variabilities. On the other hand, although the nonlinear techniques (e.g., manifold learning or DL) hold a powerful representation ability for the HS data, yet they are usually vulnerable to the attack of spectral variability, inevitably degrading the quality of dimension-reduced features. As a tradeoff, we propose to progressively learn multicoupled linear projections on the basis of the joint regression framework. Thus, the resulting JPSA with necessary priors can be formulated as the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{P},\{\boldsymbol{\Theta}_l\}_{l=1}^m} \ & \frac{1}{2}\boldsymbol{\Upsilon}\big(\{\boldsymbol{\Theta}_l\}_{l=1}^m\big) + \frac{\alpha}{2}\mathbf{E}\big(\mathbf{P}, \{\boldsymbol{\Theta}_l\}_{l=1}^m\big) + \frac{\beta}{2}\boldsymbol{\Phi}\big(\{\boldsymbol{\Theta}_l\}_{l=1}^m\big) \\ & + \frac{\gamma}{2}\boldsymbol{\Psi}(\mathbf{P}) \\ \text{s.t.} \ & \mathbf{X}_l \succeq 0, \ \|\mathbf{x}_{lk}\|_2 \preceq 1, \ \mathbf{X}_l^{sp} \succeq 0, \ \big\|\mathbf{x}_{lk}^{sp}\big\|_2 \preceq 1 \end{aligned} \tag{3}$$

where $\{\boldsymbol{\Theta}_l\}_{l=1}^m \in \mathbb{R}^{d_l \times d_{l-1}}$ are defined as a set of intermediate transformations, m is the number of subspace projections, and $\{d_l\}_{l=1}^m$ stand for as the dimensions of those latent subspaces. Moreover, $\mathbf{X}_l$ denotes the lth layer subspace features, where $\mathbf{X}_0$ represents original data ($\mathbf{X}$), while $\mathbf{X}_l^{sp}$ denotes the superpixel representation of $\mathbf{X}_l$. To effectively solve the two-stream joint regression model in (3), several key terms are featured in the following.

1) SR Loss Term $\boldsymbol{\Upsilon}(\{\boldsymbol{\Theta}_l\}_{l=1}^m)$: Without any constraints or prior, jointly estimating multiple successive variables in JPSA can hardly be implemented, especially when the number of estimated variables gradually increases. This can be well explained by gradient missing between the two neighboring variables estimated in the process of optimization. In other words, the variations between two neighboring variables approach to a tiny value or even 0. When the number of estimated projections accumulates to a certain extent, most of the valid values could only gather a few projections, making other projections being close to the identity matrix and become meaningless. To address the above-mentioned issue, a kind of autoencoder-like scheme is adopted to reduce the information loss in the process of propagation between two neighboring spaces. The benefits of the scheme are two-fold. On the one hand, this term can prevent overfitting of the data to a great extent, especially avoiding all kinds of spectral variabilities from being considered, since we found that those variabilities are difficult to be linearly reconstructed. On the other hand, it can also establish an effective link between the original space and the subspace, enabling the learned subspace to project back to the former one as much as possible. Such a strategy can be formulated by simultaneously considering pixels and superpixels of HSI

$$\boldsymbol{\Upsilon}\big(\{\boldsymbol{\Theta}_l\}_{l=1}^m\big) = \sum_{l=1}^{m}\big\|\big[\mathbf{X}_{l-1} \ \ \mathbf{X}_{l-1}^{sp}\big] - \boldsymbol{\Theta}_l^T\boldsymbol{\Theta}_l\big[\mathbf{X}_{l-1} \ \ \mathbf{X}_{l-1}^{sp}\big]\big\|_{\mathrm{F}}^2. \tag{4}$$

Note that we propose to utilize (4) in each latent subspace to maximize the advantages of this term.

2) Prediction Loss Term $\mathbf{E}(\mathbf{P}, \{\boldsymbol{\Theta}_l\}_{l=1}^m)$: This term is to minimize the empirical risk between the original data and the label matrix through a set of subspace projections and a linear regression coefficient, which can be written as

$$\mathbf{E}\big(\mathbf{P}, \{\boldsymbol{\Theta}_l\}_{l=1}^m\big) = \big\|[\mathbf{Y} \ \ \mathbf{Y}] - \mathbf{P}\boldsymbol{\Theta}_m \cdots \boldsymbol{\Theta}_l \cdots \boldsymbol{\Theta}_1\big[\mathbf{X} \ \ \mathbf{X}^{sp}\big]\big\|_{\mathrm{F}}^2. \tag{5}$$

Theoretically, with the increase of the number of estimated subspaces, the variations between neighboring subspaces are gradually narrowed down to a very small range. In this case, such small variations can be approximately represented via a

linear transformation. This allows us to find a good solution in a simple way, especially for the nonconvex model.

3) Alignment-Based SSM Regularization $\Phi(\{\Theta_l\}_{l=1}^m)$: As introduced in [43], the manifold structure is an important prior for compressing high-dimensional data, which can effectively capture the intrinsic structure between samples. For this reason, we not only embed the locally spectral manifold structure computed between the pixels but also model the nonlocal-like spatial manifolds constructed by superpixels. Therefore, the two graph structures can be formulated as

$$\mathbf{W}_{i,j}^{p} = \begin{cases} \exp\frac{-\|\mathbf{X}_i-\mathbf{X}_j\|_2^2}{2\sigma^2}, & \text{if } \mathbf{X}_j \in \phi_k(\mathbf{X}_i); \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

$$\mathbf{W}_{i,j}^{sp} = \begin{cases} \exp\frac{-\left\|\mathbf{X}_i^{sp}-\mathbf{X}_j^{sp}\right\|_2^2}{2\sigma^2}, & \text{if } \mathbf{X}_j^{sp} \in \phi_k\left(\mathbf{X}_i^{sp}\right); \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\phi_k(\mathbf{X}_i)$ and $\phi_k(\mathbf{X}_i^{sp})$ are the k neighbors of the pixel $\mathbf{X}_i$ and the superpixel $\mathbf{X}_i^{sp}$, respectively.

In addition, we also align the SSMs in each learned subspace to enhance the model's ability to distinguish and generalize, further yielding the structure consistency of the two-stream joint regression model. The alignment operator can be expressed by the form of a graph

$$\mathbf{W}_{i,j}^{a} = \begin{cases} 1, & \text{if } \mathbf{X}_i \in \phi\left(\mathbf{X}_j^{sp}\right); \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\phi(\mathbf{X}_j^{sp})$ denotes the pixel set in the jth superpixel.

By collecting the above subgraphs, we have the final graph structure ($\mathbf{W}^f$) by considering spatial and spectral neighbors of each pixel as well as their alignment information

$$\mathbf{W}^f = \begin{bmatrix} \mathbf{W}^p & \mathbf{W}^a \\ \mathbf{W}^a & \mathbf{W}^{sp} \end{bmatrix}. \tag{9}$$

Thus, the resulting manifold alignment-based spatial–spectral regularization can be written as

$$\mathbf{\Phi}\left(\{\mathbf{\Theta}_l\}_{l=1}^m\right) = \sum_{l=1}^{m} \operatorname{tr}\left(\mathbf{\Theta}_l\left[\mathbf{X}_{l-1}\ \mathbf{X}_{l-1}^{sp}\right]\mathbf{L}^f\left[\mathbf{X}_{l-1}\ \mathbf{X}_{l-1}^{sp}\right]^T\mathbf{\Theta}_l^T\right) \tag{10}$$

where $\mathbf{L}^f$ can be computed by $\mathbf{D}^f - \mathbf{W}^f$. In this study, each pixel's spatial neighbors are other pixels in the same segment obtained by SLIC, while its k spectral neighbors are selected with the Euclidean measurement on a kernel-induced space. Fig. 3 illustrates the spatial–spectral graph structure.

4) Regression Coefficient Regularization $\Psi(\mathbf{P})$: This regularization term ensures a reliable solution and improves the generalization ability of the model, which is

$$\mathbf{\Psi}(\mathbf{P}) = \|\mathbf{P}\|_{\mathrm{F}}^2. \tag{11}$$

HS data are non-negative either in radiance or reflectance. To inherit this physical nature, we expect to learn non-negative features with respect to each learned low-dimensional feature (e.g., $\{\mathbf{X}_l\}_{l=1}^m \succeq 0$). The hard orthogonal constraint with respect to the variable $\mathbf{\Theta}$ could lead to nonconvergence of the model or reach a bad solution. To provide a proper search space of the solution, we, therefore, relax the constraint by imposing a sample-based norm constraint [44] on each latent subspace as $\|\mathbf{x}_{lk}\|_2 \preceq 1\ \forall k = 1, \ldots, N$ and $l = 1, \ldots, m$. Note that these constraints are similarly applicable to the superpixel-guided optimization problem.

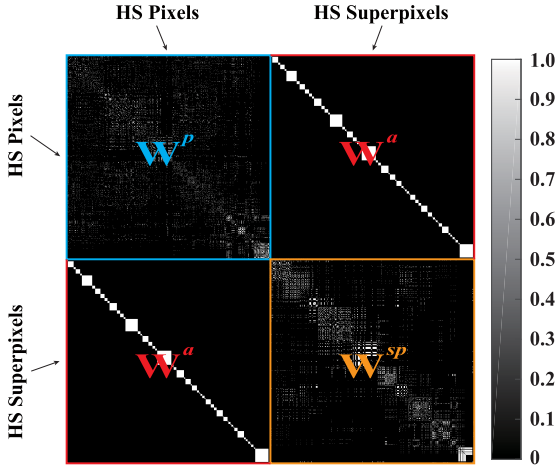


Fig. 3. Showcase to illustrate the graph structure used in the alignment-based SSM regularization term.

C. Model Learning

Considering the fact that we need to successively estimate multicoupled variables in JPSA, which obviously results in the increasing complexity and the nonconvexity of our model, a group of good initial approximations of subspace projections $\{\mathbf{\Theta}_l\}_{l=1}^m$ would greatly reduce training time and help finding a better local optimal solution. This is a common tactic that has been widely used to address this issue [45]. Inspired by this trick, we pretrain our model by simplifying (3) as

$$\min_{\mathbf{\Theta}_l} \frac{1}{2}\mathbf{\Upsilon}(\mathbf{\Theta}_l) + \frac{\eta}{2}\mathbf{\Phi}(\mathbf{\Theta}_l) \ \text{ s.t. } \ \tilde{\mathbf{X}}_l \succeq 0, \ \left\|\tilde{\mathbf{x}}_{lk}\right\|_2 \preceq 1 \tag{12}$$

where $[\mathbf{X}_l\ \ \mathbf{X}_l^{sp}]$ is collectively rewritten as $\tilde{\mathbf{X}}_l$ for convenience of writing and model optimization.

We call (12) as AutoRULe. Given the outputs of AutoRULe to the problem of (3) as the initialization, $\{\mathbf{\Theta}_l\}_{l=1}^m$ and $\mathbf{P}$ tend to obtain the better estimations. In detail, Algorithm 1 summarizes the global algorithm of JPSA, where AutoRULe is initialized by LPP.

We propose to use the ADMM-based optimization method to solve the pretraining method (AutoRULe), hence an equivalent form of (12) is considered by introducing multiple auxiliary variables $\mathbf{H}$, $\mathbf{G}$, $\mathbf{Q}$, and $\mathbf{S}$ to replace $\tilde{\mathbf{X}}_l$, $\mathbf{\Theta}_l$, $\tilde{\mathbf{X}}_l^+$, and $\tilde{\mathbf{X}}_l^\sim$, respectively, where $()^+$ denotes an operator for converting each component of the matrix to its absolute value and $()^\sim$ is a proximal operator that solves the constraint of $\left\|\tilde{\mathbf{x}}_{lk}\right\|_2 \preceq 1$ [46]. Therefore, the resulting expression is

$$\begin{aligned} \min_{\mathbf{\Theta}_l,\mathbf{H},\mathbf{G},\mathbf{Q},\mathbf{S}} \ & \frac{1}{2}\left\|\tilde{\mathbf{X}}_{l-1} - \mathbf{G}^T\mathbf{H}\right\|_{\mathrm{F}}^2 + \frac{\eta}{2}\operatorname{tr}\left(\mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}\mathbf{L}^f\tilde{\mathbf{X}}_{l-1}^T\mathbf{\Theta}_l^T\right) \\ \text{s.t. } & \tilde{\mathbf{X}}_l = \mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}, \ \mathbf{Q} \succeq 0, \ \|\mathbf{s}_k\|_2 \preceq 1. \\ & \tilde{\mathbf{X}}_l = \mathbf{H} = \mathbf{Q} = \mathbf{S}, \ \mathbf{\Theta}_l = \mathbf{G}. \end{aligned} \tag{13}$$

The constrained optimization problem in (13) can be converted to its augmented Lagrangian version by introducing the

Algorithm 1: JPSA: Global Algorithm

Input: $\mathbf{Y}, \tilde{\mathbf{X}}, \mathbf{L}^f$, and parameters α, β, γ and *maxIter*.
Output: $\{\mathbf{\Theta}_l\}_{l=1}^m$.

Initialization Step:
Greedily initialize $\mathbf{\Theta}_l$ corresponding to each latent subspace:
for $l = 1{:}m$ **do**
  $\mathbf{\Theta}_l^0 \leftarrow LPP(\tilde{\mathbf{X}}_{l-1})$
  $\mathbf{\Theta}_l \leftarrow AutoRULe(\tilde{\mathbf{X}}_{l-1}, \mathbf{\Theta}_l^0, \mathbf{L}^f)$
  $\tilde{\mathbf{X}}_l \leftarrow \mathbf{\Theta}_l \tilde{\mathbf{X}}_{l-1}$
end
Fine-tuning Step:
$t = 0, \zeta = 1e-4$;
while $t > maxIter$ **do**
  Update $\mathbf{P}$ by solving a subproblem in Eq. (16).
  for $i = 1{:}m$ **do**
    Update $\mathbf{\Theta}_l^{t+1}$ by solving a subproblem in Eq. (18).
  end
  Compute the objective function value Obj^{t+1} and check the convergence condition:
  if $|\frac{Obj^{t+1}-Obj^t}{Obj^t}| < \zeta$ **then**
    Stop iteration;
  else
    $t \leftarrow t+1$;
  end
end

Algorithm 2: AutoRULe: Initialization Step for JPSA

Input: $\tilde{\mathbf{X}}_{l-1}, \mathbf{\Theta}_l^0, \mathbf{L}^f$, and parameters η and *maxIter*.
Output: $\mathbf{\Theta}_l$.

Initialization:
$\mathbf{H}^0 = \mathbf{\Theta}_l^0 \tilde{\mathbf{X}}_{l-1}, \mathbf{G}^0 = \mathbf{0}, \mathbf{Q}^0 = \mathbf{P}^0 = \mathbf{0}, \mathbf{\Lambda}_2^0 = \mathbf{0}, \mathbf{\Lambda}_1^0 = \mathbf{\Lambda}_3^0 = \mathbf{\Lambda}_4^0 = \mathbf{0}, \mu^0 = 1e-3, \mu_{max} = 1e6, \rho = 2, \varepsilon = 1e-6, t = 0.$
while $t > maxIter$ **do**
  Fix $\mathbf{H}^t, \mathbf{G}^t, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{\Theta}_l^{t+1}$ by Eq. (26).
  Fix $\mathbf{\Theta}_l^{t+1}, \mathbf{G}^t, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{H}^{t+1}$ by Eq. (28).
  Fix $\mathbf{H}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{Q}^t, \mathbf{P}^t$ to update $\mathbf{G}^{t+1}$ by Eq. (30).
  Fix $\mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{P}^t$ to update $\mathbf{Q}^{t+1}$ by Eq. (32).
  Fix $\mathbf{H}^{t+1}, \mathbf{G}^{t+1}, \mathbf{\Theta}_l^{t+1}, \mathbf{Q}^{t+1}$ to update $\mathbf{P}^{t+1}$ by Eq. (34).
  Update Lagrange multipliers using Eq. (35).
  Update penalty parameter using $\mu^{t+1} = min(\rho\mu^t, \mu_{max})$.
  Check the convergence conditions:
  if $\left\|\mathbf{H}^{t+1} - \mathbf{\Theta}_l^{t+1}\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}} < \varepsilon$ *and* $\left\|\mathbf{G}^{t+1} - \mathbf{\Theta}_l^{t+1}\right\|_{\mathrm{F}} < \varepsilon$ *and* $\left\|\mathbf{Q}^{t+1} - \mathbf{\Theta}_l^{t+1}\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}} < \varepsilon$ *and* $\left\|\mathbf{P}^{t+1} - \mathbf{\Theta}_l^{t+1}\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}} < \varepsilon$ **then**
    Stop iteration;
  else
    $t \leftarrow t+1$;
  end
end

Lagrange multipliers $\{\mathbf{\Lambda}_n\}_{n=1}^4$ and the penalty parameter μ, where the non-negativity and norm constraint can be relaxed by defining two kinds of proximal projection operators $l_R^+(\bullet)$ and $l_R^{\sim}(\bullet)$. More specifically, $l_R^+(\bullet)$ can be expressed as

$$\max(\bullet) = \begin{cases} \bullet, & \bullet \succ 0 \\ 0, & \bullet \preceq 0 \end{cases} \tag{14}$$

while $l_R^{\sim}(\bullet_k)$ is a sample-based normalization operator

$$\mathrm{prox}_f(\bullet_k) = \begin{cases} \frac{\bullet_k}{\|\bullet_k\|_2}, & \|\bullet_k\|_2 \succ 1 \\ \bullet_k, & \|\bullet_k\|_2 \preceq 1 \end{cases} \tag{15}$$

where $\bullet_k$ is the kth column of matrix $\bullet$ in our case.

Algorithm 2 lists the optimization procedures of AutoRULe, and the solution to each subproblem is detailed in the Appendix.

After running the AutoRULe, its outputs can be fed into JPSA for the model initialization, and then the two subproblems (solve $\mathbf{P}$ and $\{\mathbf{\Theta}_l\}_{l=1}^m$) in (3) can be optimized alternatively as follows.

Optimization With Respect to $\mathbf{P}$ *Subproblem:* Typically, this is a Tikhonov-regularized least square regression problem, which can be formulated as

$$\min_{\mathbf{P}} \frac{\alpha}{2}\left\|\tilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}_m \cdots \mathbf{\Theta}_l \cdots \mathbf{\Theta}_1 \tilde{\mathbf{X}}\right\|_{\mathrm{F}}^2 + \frac{\gamma}{2}\|\mathbf{P}\|_{\mathrm{F}}^2 \tag{16}$$

where the variable $\tilde{\mathbf{Y}}$ is a collection of $[\mathbf{Y} \ \ \mathbf{Y}]$ similar to the variable $\tilde{\mathbf{X}}$. Intuitively, the analytical solution of (16) can be directly derived as

$$\mathbf{P} \leftarrow \left(\alpha\tilde{\mathbf{Y}}\mathbf{V}^T\right)\left(\alpha\mathbf{V}\mathbf{V}^T + \gamma\mathbf{I}\right)^{-1} \tag{17}$$

where $\mathbf{V}$ is assigned to $\mathbf{\Theta}_m \cdots \mathbf{\Theta}_l \cdots \mathbf{\Theta}_1 \tilde{\mathbf{X}}$ $\forall l = 1, \ldots, m$.

Optimization With Respect to $\{\mathbf{\Theta}_l\}_{l=1}^m$*:* When other variables are fixed, the variable $\mathbf{\Theta}_l$ can be individually solved, hence the optimization problem for any $\mathbf{\Theta}_l$ can be written in the following general form:

$$\begin{aligned} \min_{\mathbf{\Theta}_l} \ & \frac{1}{2}\left\|\tilde{\mathbf{X}}_{l-1} - \mathbf{\Theta}_l^T\mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\tilde{\mathbf{Y}} - \mathbf{P}\mathbf{\Theta}_m \cdots \mathbf{\Theta}_1\tilde{\mathbf{X}}\right\|_{\mathrm{F}}^2 \\ & + \frac{\beta}{2}\mathrm{tr}\left(\mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}\mathbf{L}^f\tilde{\mathbf{X}}_{l-1}^T\mathbf{\Theta}_l^T\right) \\ \text{s.t.} \ & \tilde{\mathbf{X}}_l = \mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}, \ \tilde{\mathbf{X}}_l \succeq 0, \ \left\|\tilde{\mathbf{x}}_{lk}\right\|_2 \preceq 1. \end{aligned} \tag{18}$$

Likewise, the problem of (18) can basically be solved by following the framework of Algorithm 2. (More details regarding the variable optimization can be found in the Appendix.) The only difference lies in the optimization of the subproblem with respect to $\mathbf{H}$. Herein, we supplement the optimization problem of the variable $\mathbf{H}$ as follows:

$$\begin{aligned} \min_{\mathbf{H}} \ & \frac{1}{2}\left\|\tilde{\mathbf{X}}_{l-1} - \mathbf{G}^T\mathbf{H}\right\|_{\mathrm{F}}^2 + \frac{\alpha}{2}\left\|\tilde{\mathbf{Y}} - \mathbf{P}_l\mathbf{H}\right\|_{\mathrm{F}}^2 \\ & + \mathbf{\Lambda}_1^T\left(\mathbf{H} - \mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right) + \frac{\mu}{2}\left\|\mathbf{H} - \mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2 \\ \text{s.t.} \ & \mathbf{P}_l = \mathbf{P}_{l-1}\mathbf{\Theta}_{l+1}, \ \mathbf{P}_0 = \mathbf{P} \end{aligned} \tag{19}$$

whose analytical solution is given by

$$\begin{aligned} \mathbf{H} \leftarrow & \left(\alpha\mathbf{P}_l^T\mathbf{P}_l + \mathbf{G}\mathbf{G}^T + \mu\mathbf{I}\right)^{-1} \\ & \times \left(\alpha\mathbf{P}_l^T\tilde{\mathbf{Y}} + \mathbf{G}\tilde{\mathbf{X}}_{l-1} + \mu\mathbf{\Theta}_l\tilde{\mathbf{X}}_{l-1} - \mathbf{\Lambda}_1\right). \end{aligned} \tag{20}$$

Finally, the aforementioned optimization procedures are repeated until a stopping criterion is satisfied.

D. Convergence Analysis

The iterative alternating strategy used in Algorithm 1 is nothing but a block coordinate descent, whose convergence

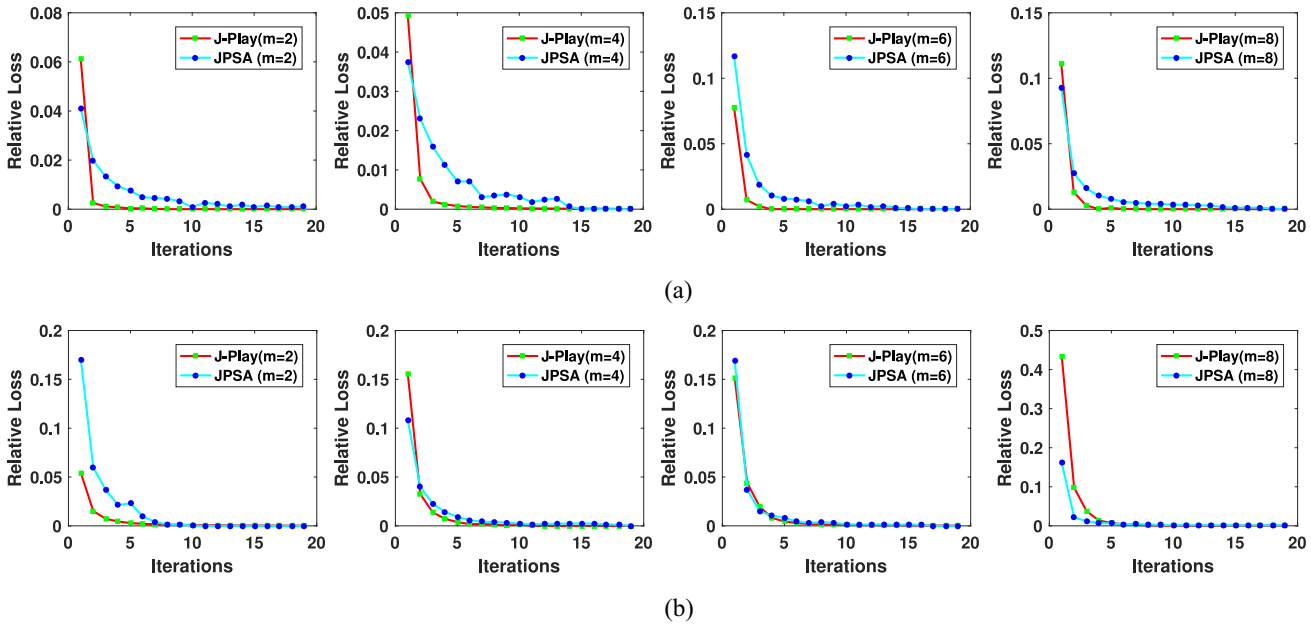


Fig. 4. Convergence analysis of J-Play and JPSA with different m values of $2, 4, 6, 8$ (left to right) was experimentally performed on the two HS datasets. (a) Indian Pines dataset. (b) University of Houston dataset.

is theoretically guaranteed as long as each subproblem of (12) is exactly minimized [47]. Each subproblem optimized in Algorithm 2 is strongly convex, and thus the ADMM-based optimization strategy can converge to a unique minimum when the parameters are updated in finite steps [48], [49]. Moreover, we experimentally illustrate to clarify the convergences of J-Play and the proposed JPSA on the two HS datasets, where the relative errors of the objective function value are recorded in each iteration (see Fig. 4).

III. Experiments

A. Description of the Data

The experiments are performed on two different standard HS datasets, corresponding to different contexts, different sensors, and different resolutions.

1) Indian Pines AVIRIS Image: The first HS cube was acquired by the AVIRIS sensor with 16 classes of vegetation. It consists of 145×145 with the spectral 220 bands covering the wavelength range from 400 to 2500 nm in a 10-nm spectral resolution. A set of widely used training and test sets [1] with the specific categories is listed in Table I. A false-color image of the data is given in Fig. 5.

2) University of Houston Image: The second HSI was provided for the 2013 IEEE GRSS data fusion contest. It was acquired by an ITRES-CASI-1500 sensor over the campus of the University of Houston, Houston, USA, with a size of $349 \times 1905 \times 144$ in the wavelength from 364 to 1046 nm. The information regarding classes as well as training and test samples can be also found in Table I. The first image of Fig. 6 shows a false color image of the study scene.

B. Experimental Setup and Preparation

We learn the subspaces for the different methods on the training set and then the test set can be simply projected to the subspace where training and test samples can be further classified by the nearest neighbor (NN). The reason for selecting the simple but effective classifier in our case is that the NN classifier tends to avoid the confusing evaluation, as it is unknown whether the performance improvement originates from either the classifiers or the features themselves if more advanced classifiers are used.

Moreover, the original spectral features (OSFs) without DR and ten popular and advanced methods are compared with our JPSA, including:

1) *Unsupervised HDR:* PCA [34], OTVCA [17];
2) *Supervised HDR:* LDA [25], LFDA [36], and J-Play [1];
3) *Semisupervised HDR:* SELD [21], SLLDA [22], and SSLFDA [23];
4) *DNN-Based HDR:* SAE [29], CAE [30].

TABLE I
Scene Categories, the Number of Training (TR), and Test (TE) Samples for Each Class on the Two Datasets: Indian Pines and University of Houston

No.	Indian Pines Dataset			University of Houston Dataset		
	Class Name	TR	TE	Class Name	TR	TE
1	CornNotill	50	1384	HealthyGrass	198	1053
2	CornMintill	50	784	StressedGrass	190	1064
3	Corn	50	184	Synthetic Grass	192	505
4	GrassPasture	50	447	Tree	188	1056
5	GrassTrees	50	697	Soil	186	1056
6	HayWindrowed	50	439	Water	182	143
7	SoybeanNotill	50	918	Residential	196	1072
8	SoybeanMintill	50	2418	Commercial	191	1053
9	SoybeanClean	50	564	Road	193	1059
10	Wheat	50	162	Highway	191	1036
11	Woods	50	1244	Railway	181	1054
12	BuildingsGrassTrees	50	330	Parking Lot1	192	1041
13	StoneSteelTowers	50	45	Parking Lot2	184	285
14	Alfalfa	15	39	Tennis Court	181	247
15	GrassPastureMowed	15	11	Running Track	187	473
16	Oats	15	5	–	–	–
	Total	695	9671	Total	2832	12197

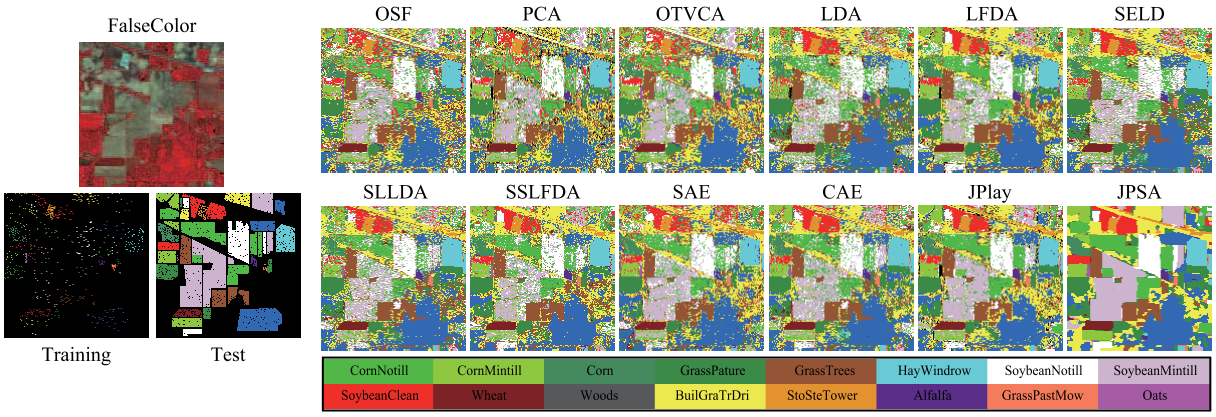


Fig. 5. False color image, the distribution of training and test sets with category names, and classification maps of the different algorithms obtained using the NN classifier on the Indian Pines dataset.

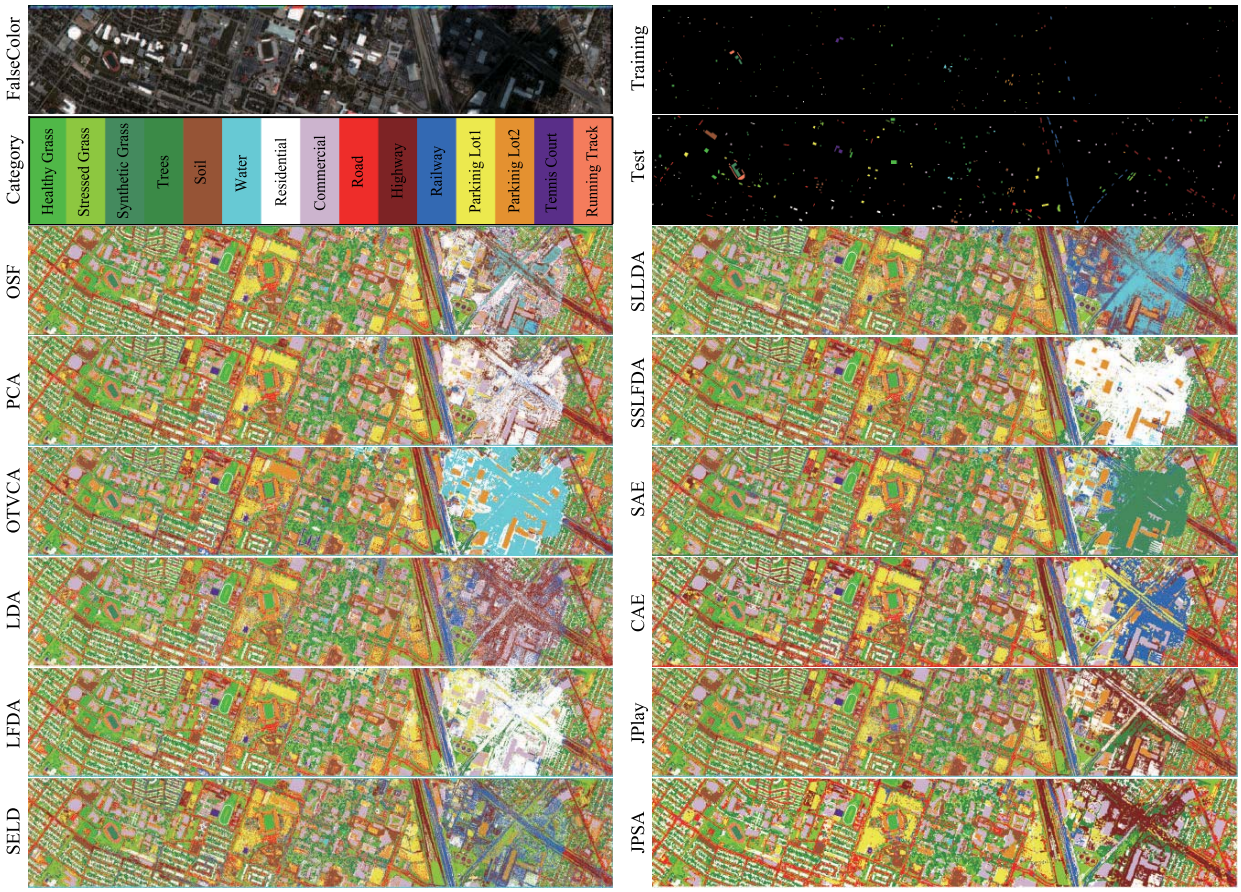


Fig. 6. False color image, the distribution of training and test sets with category names, and classification maps of the different algorithms obtained using the NN classifier on the University of Houston dataset.

Furthermore, we maximize the performances of the different algorithms by tuning their parameters, such as dimension (d), regularization parameters (α, β, γ), etc., using ten-fold cross-validation on training data. Regarding the dimensions ($\{d_l\}$) which are common parameters for all algorithms, they can be selected ranging from 10 to 50 at an interval of 10. For the number of NNs (k) and the standard deviation of the Gaussian kernel function (σ) in those algorithms that need to construct the graph structure (e.g., LFDA, SELD, SSLFDA, J-Play, and JPSA), we select them in the range of $\{10, 20, \ldots, 50\}$ and $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$, respectively, and three regularization parameters (α, β, γ) in J-Play or JPSA are all chosen from $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$. For the OTVCA algorithm, we directly used the parameter setting suggested in [17]: that is, d is equal to the number of classes, and λ can be automatically determined by 1% of the maximum intensity range of the datasets.

We adopt three criteria to quantitatively assess the algorithm performance, including overall accuracy (OA), average accuracy (AA), and Kappa Coefficient (κ). They can be formulated

TABLE II
QUANTITATIVE PERFORMANCE COMPARISONS OF DIFFERENT ALGORITHMS ON THE INDIAN PINES DATASET WITH THE OPTIMAL PARAMETER COMBINATION IN TERMS OF OA, AA, AND κ, AS WELL AS THE ACCURACY FOR EACH CLASS. THE BEST ONE IS SHOWN IN BOLD. JPLAY_4 MEANS A FOUR-LAYERED J-PLAY MODEL ($m = 4$), WHILE JPSA_4 DENOTES A FOUR-LAYERED JPSA MODEL ($m = 4$)

Method	OSF	PCA	OTVCA	LDA	LFDA	SELD	SLLDA	SSLFDA	SAE	CAE	JPlay_4	JPSA_4
d	220	20	16	15	15	15	15	15	20	20	20	20
k	–	–	–	–	10	10	–	5	–	–	10	10
σ	–	–	–	–	0.1	0.01	–	0.1	–	–	0.1	0.1
α	–	–	–	–	–	–	–	–	–	–	1	1
β	–	–	–	–	–	–	–	–	–	–	0.1	0.1
γ	–	–	–	–	–	–	–	–	–	–	0.1	0.1
OA	65.89	65.40	68.87	64.14	73.86	75.81	70.93	75.26	71.39	76.89	83.92	**92.98**
AA	75.71	75.43	79.04	74.52	85.59	83.37	82.20	85.91	78.88	84.94	89.35	**95.40**
κ	0.6148	0.6097	0.6490	0.5964	0.7042	0.7265	0.6713	0.7200	0.6765	0.7379	0.8169	**0.9197**
Class1	51.66	50.79	54.55	51.45	67.77	72.40	57.73	70.23	60.62	66.47	79.05	**91.04**
Class2	57.40	57.14	59.69	48.47	65.05	65.69	59.69	67.35	56.51	72.19	80.74	**90.18**
Class3	70.65	69.02	69.57	69.57	83.15	83.15	71.74	87.50	82.07	86.96	85.87	**99.46**
Class4	88.14	87.92	90.60	90.60	**95.30**	**95.30**	94.63	94.85	90.38	94.63	94.63	95.08
Class5	81.78	81.64	84.22	86.80	**94.55**	91.68	88.52	93.54	88.95	90.10	90.24	91.25
Class6	95.90	95.67	95.67	97.95	97.95	98.63	98.41	98.41	94.99	99.32	96.58	**99.77**
Class7	66.56	67.32	77.89	58.06	70.81	74.95	73.20	75.16	73.09	73.31	81.37	**97.39**
Class8	55.21	54.18	55.29	42.97	52.94	57.49	54.43	55.21	57.78	63.52	76.51	**87.80**
Class9	53.01	52.30	54.96	71.45	79.61	82.27	68.44	78.01	72.34	81.56	84.40	**93.26**
Class10	98.15	98.15	98.15	**99.38**	**99.38**	**99.38**	**99.38**	**99.38**	96.30	**99.38**	**99.38**	**99.38**
Class11	82.88	82.40	86.90	85.53	89.79	88.18	87.94	89.87	86.58	89.31	93.41	**98.63**
Class12	50.91	51.21	61.21	77.88	83.03	82.73	81.21	81.52	73.03	82.12	79.09	**96.06**
Class13	97.78	97.78	97.78	97.78	97.78	**100.00**	97.78	97.78	97.78	95.56	**100.00**	**100.00**
Class14	79.49	79.49	87.18	74.36	92.31	82.05	82.05	94.87	58.97	84.62	**97.44**	87.18
Class15	81.82	81.82	90.91	**100.00**	**100.00**	**100.00**	**100.00**	90.91	72.73	**100.00**	90.91	**100.00**
Class16	**100.00**	**100.00**	**100.00**	40.00	**100.00**	60.00	**100.00**	**100.00**	**100.00**	80.00	**100.00**	**100.00**

by using the following equations:

$$\text{OA} = \frac{N_c}{N_a} \tag{21}$$

$$\text{AA} = \frac{1}{C}\sum_{i=1}^{C}\frac{N_c^i}{N_a^i} \tag{22}$$

and

$$\kappa = \frac{\text{OA} - P_e}{1 - P_e} \tag{23}$$

where N_c and N_a denote the number of samples classified correctly and the number of total samples, respectively, while N_c^i and N_a^i correspond to the N_c and N_a of each class, respectively. P_e in κ is defined as the hypothetical probability of chance agreement [50], which can be computed by

$$P_e = \frac{N_r^1 \times N_p^1 + \cdots N_r^i \times N_p^i + \cdots + N_r^C \times N_p^C}{N_a \times N_a} \tag{24}$$

where N_r^i and N_p^i denote the number of real samples for each class and the number of predicted samples for each class, respectively.

C. Results Analysis and Discussion

1) Indian Pines Dataset: Table II presents the classification performances of the different methods with the optimal parameter setting tuned by cross-validation on the training set using the NN classifier. Correspondingly, the classification maps are given in Fig. 5 for visual assessment.

Overall, PCA provides similar performances with the baseline (OSF), as the PCA more focuses on maximizing the information but could fail to excavate the potential data structure that lies in reality. By smoothing the spatial structure of HSI, OTVCA enables better identification of the materials than OSF and PCA. For the supervised HDR methods, the classification performances of classic LDA are even lower than those previously mentioned, due to the limited amount of training samples. Holding a more powerful intraclass homogeneity and interclass separation, LFDA obtains more competitive results by locally focusing on discriminative information, which is obviously better than those of the baseline, PCA, and LDA around 8%. However, the features learned by LFDA are relatively difficult to be generalized, due to the small-size labeled samples. Comparatively, SELD learns a robust low-dimensional feature representation with a higher generalization ability, since unlabeled samples are involved in the process of model training. In SELD, the unlabeled information is embedded by computing the similarities between samples, which is more effective than that using the pseudolabels (e.g., SLLDA and SSLFDA) to some extent. However, these semisupervised methods are still bad at handling noisy data. A direct proof can be shown in Fig. 5 that there exist obvious salt-and-pepper-like noises in classification maps of SELD, SLLDA, and SSLFDA. Likewise, although the SAE holds a strong nonlinear learning ability in data representation, its performance is still limited by complex spectral variability and pixelwise feature embedding. Thanks to the spatial information modeling, CAE locally extracts the spatial information and thus obtains a relatively smooth classification result. With the benefit of a multilinear regression system, the J-Play algorithm performs much better (at least 7% OAs) than DNN-based nonlinear HDR (SAE and CAE). Such a strategy makes the learned features more robust

TABLE III
QUANTITATIVE PERFORMANCE COMPARISONS OF DIFFERENT ALGORITHMS ON THE UNIVERSITY OF HOUSTON DATASET WITH THE OPTIMAL PARAMETER COMBINATION IN TERMS OF OA, AA, AND κ, AS WELL AS THE ACCURACY FOR EACH CLASS. THE BEST ONE IS SHOWN IN BOLD. JPLAY$_3$ MEANS A THREE-LAYERED J-PLAY MODEL ($m = 3$), WHILE JPSA$_3$ DENOTES A THREE-LAYERED JPSA MODEL ($m = 3$)

Method	OSF	PCA	OTVCA	LDA	LFDA	SELD	SLLDA	SSLFDA	SAE	CAE	JPlay$_3$	JPSA$_3$
d	144	20	15	14	14	14	14	14	30	30	30	30
k	–	–	–	–	20	20	–	30	–	–	10	10
σ	–	–	–	–	0.1	0.1	–	0.1	–	–	0.1	0.1
α	–	–	–	–	–	–	–	–	–	–	1	1
β	–	–	–	–	–	–	–	–	–	–	0.1	0.1
γ	–	–	–	–	–	–	–	–	–	–	0.1	0.1
OA	72.83	72.75	74.18	74.18	75.52	77.45	77.18	78.94	79.52	80.68	80.13	**86.09**
AA	76.16	76.09	77.61	79.04	79.10	80.40	79.59	82.09	82.45	83.23	82.99	**87.90**
κ	0.7079	0.7071	0.7228	0.7374	0.7355	0.7555	0.7537	0.7716	0.7789	0.7905	0.7845	**0.8490**
Class1	82.15	82.15	82.24	81.67	81.96	81.29	81.96	82.43	82.53	82.15	**82.72**	81.10
Class2	81.86	81.86	82.05	82.14	82.99	83.46	83.36	82.14	83.27	83.74	82.61	**84.68**
Class3	99.60	99.60	99.60	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.80	99.80	**100.00**	99.41
Class4	91.76	91.76	91.86	90.44	91.00	93.75	91.57	92.42	88.83	91.48	**96.78**	94.89
Class5	97.06	97.06	97.73	97.25	97.82	99.34	98.01	99.05	98.11	**99.91**	99.62	99.72
Class6	95.10	95.10	94.41	99.30	99.30	**100.00**	94.41	99.30	95.10	97.90	96.50	96.50
Class7	73.60	73.60	79.10	72.57	81.72	81.34	72.76	**85.73**	79.10	78.64	80.50	78.17
Class8	36.37	36.37	36.56	59.92	40.65	42.17	50.71	42.55	43.59	64.20	52.23	**76.35**
Class9	66.19	66.29	67.14	57.60	74.13	72.33	63.74	73.84	74.41	74.22	77.15	**79.70**
Class10	49.23	49.03	48.26	57.53	42.95	45.08	57.92	49.52	58.30	51.25	53.28	**81.95**
Class11	67.74	67.55	70.68	74.76	74.19	82.64	81.02	81.50	79.22	**85.77**	75.81	74.00
Class12	54.27	53.60	57.64	58.79	59.46	69.36	69.55	75.60	84.53	75.70	77.52	**97.31**
Class13	51.93	51.93	58.95	56.49	62.81	58.60	51.93	69.82	71.58	68.77	72.63	**80.00**
Class14	97.57	97.57	99.19	99.19	99.19	99.19	99.19	99.19	**100.00**	**100.00**	98.79	**100.00**
Class15	97.89	97.89	**98.73**	97.89	98.31	97.46	97.67	98.31	98.31	94.93	**98.73**	94.71

against various spectral deformation and degradation, in spite of without accounting for the spatial information.

The performances of the proposed JPSA are superior to the other methods, which indicates that JPSA can learn a more discriminative and robust spectral embedding. The alignment-based SSM embedding enables us to identify the materials at a more accurate level on a small-scale training set. As shown in Fig. 5, the classification map obtained by JPSA is smoother than others, demonstrating that our method is capable of effectively aggregating the spatially contextual information in the process of HDR by means of superpixels. It is worth noting that the JPSA not only outperforms others from the whole perspective but also obtains highly competitive results for each class, particularly for *Corn*, *Soybean-Notill*, *Soybean-Mintill*, *Soybean-Clean*, and *Building-Grass-Trees* that have a dramatic improvement of about 10% in classification accuracy.

2) University of Houston Dataset: Fig. 6 shows a visual comparison among the different algorithms, and the specific classification accuracies for various compared methods, which were optimally parameterized by cross-validation as listed in Table III.

Generally, there is a basically consistent trend in classification performance between OSF and PCA: around 72% OA as a baseline. For another unsupervised HDR method, OTVCA approximately yields a 2% improvement on the basis of OSF and PCA. Owing to the use of total variation operator in OTVCA (see the smooth classification map in Fig. 6), it shares similar performances with discriminant analysis-based approaches, such as LDA and LFDA. This reason why the unsupervised OTVCA is comparable to the supervised HDR methods could be, to some extent, two-fold. On the one hand, the local smoothing strategy is a good fit for HS feature extraction and HDR tasks; on the other hand, the small-size training set hinders the supervised LDA and LFDA finding a generalized or transferable discriminative subspace. Nevertheless, LFDA is capable of steadily performing better than OTVCA owing to the consideration of local manifold structure. This might be seen as indirect evidence to show the effectiveness of the manifold embedding in HDR. More intuitively, the performance of semisupervised methods is superior to that of those only considering the labeled samples, where the SSLFDA achieves the best classification results. This demonstrates the effectiveness of embedding unlabeled samples in improving the generalization ability of the learned model. Although these semisupervised methods show the discriminative power between different classes, yet there is still room for improvement in spatial information modeling and model learning ability. As a member of deep learning, SAE is capable of better reducing the gap between the original data and compressed data, thus yielding better classification performance. Another DL-based technique for HDR is CAE, which can extract a low-dimensional spectral representation with the attention of spatial contextual information. As a result, CAE performs better than the pixelwise SAE with an about 1% slight increase of OA. Due to the lack of modeling spectral variability, SAE or CAE fails to transfer the trained model to out-of-sample (i.e., test set) effectively, even though there is a powerful learning ability in SAE and CAE. Unlike them, J-Play adopts a multilinear modeling strategy with the SE constraint in order to remove the spectral variabilities effectively and maintain the learned features as discriminative as possible, which results in basically the same results with CAE and slightly higher than SAE.

JPSA outperforms other HDR algorithms significantly, which indicates that the proposed method is capable of effectively approximating an optimal mapping from the

TABLE IV
CLASSIFICATION PERFORMANCE (OA, AA, AND κ) WITH THE DIFFERENT NUMBER OF LEARNT PROJECTIONS (m) ON THE TWO DATASETS

Method	Indian Pines			University of Houston		
	OA	AA	κ	OA	AA	κ
$JPSA_1$	87.41	93.13	0.8565	81.75	83.82	0.8019
$JPSA_2$	90.74	94.58	0.8942	82.27	84.35	0.8074
$JPSA_3$	92.28	94.97	0.9116	**86.09**	**87.90**	**0.8490**
$JPSA_4$	**92.98**	**95.40**	**0.9197**	84.82	86.88	0.8353
$JPSA_5$	91.35	95.02	0.9012	84.20	86.10	0.8285
$JPSA_6$	92.76	95.23	0.9173	82.89	85.19	0.8143
$JPSA_7$	89.74	94.21	0.8831	82.44	84.56	0.8094
$JPSA_8$	90.79	94.43	0.8948	81.54	82.97	0.7997

original space to the label space by fully considering a tradeoff between spectral discrimination and subspace robustness, thus providing a robust and discriminative low-dimensional feature representation. Further, the embedding of spatial–spectral information enables semantically meaningful object-based HS classification results. Notably, JPSA is able to more effectively eliminate the effects of shadow covered by clouds in image acquisition, compared to other methods as shown in Fig. 6. Accordingly, JPSA also shows the superiority in identifying different materials, as quantified in Table III, especially for those challenging classes, such as *Commercial*, *Highway*, and *Parking Lot1*.

D. Parameter Sensitivity Analysis of JPSA

The quality of low-dimensional feature embedding, to some extent, depends on the parameter selection, it is, therefore, indispensable to investigate the sensitivity of parameter setting in JPSA. Five main parameters involved in the JPSA, which need to be emphatically analyzed and discussed, would result in a significant effect on dimension-reduced features and even final classification results. They include three regularization parameters (α, β, and γ) in (3), subspace dimension (d), and the number of layers (m).

Significantly, we start to analyze the effects of different m for JPSA. With the different number of learnt projections, we successively specify the proposed model as $JPSA_1, \ldots, JPSA_l, \ldots, JPSA_m$ $\forall l = 1, \ldots, m$. To investigate the trend of OAs, m is uniformly set up to 8 on the two datasets. We experimentally set the number of clusters in SLIC as 10% of the total samples. As listed in Table IV, with the increase of m, the performances of JPSA with SSM embedding steadily increase to the best with around 3 layers for both datasets and then gradually decrease with a slight perturbation. This might be explained by overfitting and error accumulation of the model in the multilayered regression process, since our model is only trained on a limited number of samples. Note that more results about the JPlay in terms of the parameter m can be found in [1], and the code is openly available from the link: https://github.com/danfenghong/ECCV2018_J-Play.

Apart from the parameter m, the regularization parameters and subspace dimension also play a crucial role in improving the model's performance. More specifically, the resulting quantitative analysis of the two datasets is given in Fig. 7, where the parameter combinations of ($\alpha = 1,\ \beta = 0.1,\ \gamma = 0.1,\ d = 20$) and ($\alpha = 1,\ \beta = 0.1,\ \gamma = 0.1,\ d = 30$) achieve the best classification performance on the test sets for the first and second datasets, respectively. The resulting parameter selection for the two sets of datasets is basically consistent with that determined by ten-fold cross-validation on the training set (see Section III-B for more details). The cross-validation is, therefore, an effective strategy to automatically determine the model's parameters so that other researchers are able to produce the results for their own tasks. More specifically, the optimal parameters can be determined by testing all of the parameter combinations. Furthermore, we only show the 2-D figures (see Fig. 7) for the convenience of visualization, where other variables are set to be the optimal ones except for the current investigated variable.

Moreover, we can observe from Fig. 7 that with the increase of d, the JPSA's performance increases to the optimal value with the dimension of 20 for the Indian Pines dataset and 30 for the University of Houston dataset, respectively, then starts to reach a relatively stable state, and finally decreases with a slight perturbation when the subspace dimension is approaching to that of original spectral signature. For the variable α that mainly controls the prediction errors between the input data and labels, it is a very important factor that needs to be carefully considered in the model learning, since the setting of α is sensitive to the feature embedding and even to the final classification results. Similarly, the terms of SR and SSM alignment also have great effects on the classification performance, which indicates the importance of the two terms. What is more, the subspace dimension is a noteworthy factor as well, although the OAs with different dimensions are relatively stable when the variable d reaches a larger value (e.g., 10).

E. Ablation Studies of JPSA

In addition, we analyze the performance gain of JPSA by stepwise adding the different components, that is, SR term, SSM alignment term, etc. Table V details the increasing performance when different terms are fused. As it turns out successively embedding each component into the JPSA would lead to a progressive enhancement in feature representation ability. This demonstrates the advancement and effectiveness of the proposed JPSA model for HDR.

IV. CONCLUSION

In this article, we proposed a JPSA technique to learn an optimal mapping for effective HS data compression along the spectral dimension. JPSA is expected to find a discriminative subspace where the samples can be semantically (label information) and structurally (SSM or topology perseveration and alignment) represented and thereby be better classified. Oriented by assessing pixelwise HS classification performances, we conduct extensive experiments using JPSA in comparison with some previous state-of-the-art HDR methods. The desirable results using JPSA demonstrate its superiority and effectiveness, particularly in handling various complex spectral variabilities compared to other nonlinear DR techniques (e.g., DL-based methods). In the future, we will further develop and apply the JPSA framework to the multimodality learning.

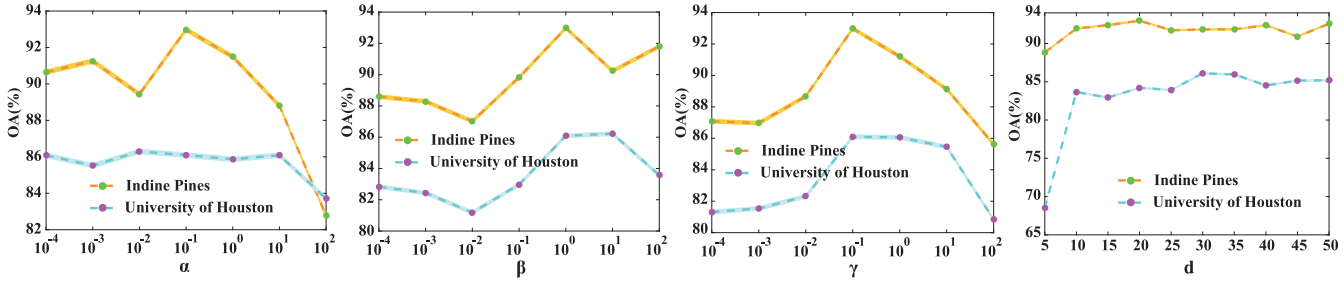


Fig. 7. Parameter sensitivity analysis of JPSA for three regularization parameters (α, β, and γ) and the subspace dimension (d) on the two datasets.

TABLE V
ABLATION ANALYSIS OF JPSA WITH A PROGRESSIVE COMBINATION OF DIFFERENT TERMS ON THE TWO DATASETS

Terms	Indian Pines			University of Houston		
	OA	AA	κ	OA	AA	κ
None	87.92	93.16	0.8623	82.22	84.74	0.8068
SR	89.40	93.81	0.8791	84.44	86.60	0.8310
SR+SSM	**92.98**	**95.40**	**0.9197**	**86.09**	**87.90**	**0.8490**

APPENDIX
SOLUTION TO AUTORULE

The solution to problem (12) can be transferred to equivalently solve the problem (13) with ADMM. Considering the fact that the object function in (13) is not convex with respect to all variables simultaneously, but it is a convex problem regarding the separate variable when other variables are fixed, therefore we successively minimize $\mathscr{L}_\mu$ (13) with respect to $\boldsymbol{\Theta}_l, \mathbf{H}, \mathbf{G}, \mathbf{Q}, \mathbf{S}, \{\boldsymbol{\Lambda}_n\}_{n=1}^4$ as follows.

$\boldsymbol{\Theta}_l$ *Problem:* The optimization problem for $\boldsymbol{\Theta}$ is

$$\begin{aligned}\min_{\boldsymbol{\Theta}_l}\ & \frac{\eta}{2}\operatorname{tr}\left(\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\mathbf{L}^f\tilde{\mathbf{X}}_{l-1}^T\boldsymbol{\Theta}_l^T\right)+\frac{\mu}{2}\left\|\mathbf{H}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2\\ &+\boldsymbol{\Lambda}_1^T\left(\mathbf{H}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)+\frac{\mu}{2}\|\mathbf{G}-\boldsymbol{\Theta}_l\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_2^T(\mathbf{G}-\boldsymbol{\Theta}_l)\\ &+\frac{\mu}{2}\left\|\mathbf{Q}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_3^T\left(\mathbf{Q}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)+l_R^+(\mathbf{Q})\\ &+\frac{\mu}{2}\left\|\mathbf{S}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_4^T\left(\mathbf{S}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)+l_R^{\sim}(\mathbf{S})\end{aligned}\tag{25}$$

which has a closed-form solution

$$\begin{aligned}\boldsymbol{\Theta}_l\leftarrow&\begin{pmatrix}\mu\mathbf{H}\tilde{\mathbf{X}}_{l-1}^T+\mu\mathbf{G}+\mu\mathbf{Q}\tilde{\mathbf{X}}_{l-1}^T+\mu\mathbf{P}\tilde{\mathbf{X}}_{l-1}^T\\ +\boldsymbol{\Lambda}_1\tilde{\mathbf{X}}_{l-1}^T+\boldsymbol{\Lambda}_2+\boldsymbol{\Lambda}_3\tilde{\mathbf{X}}_{l-1}^T+\boldsymbol{\Lambda}_4\tilde{\mathbf{X}}_{l-1}^T\end{pmatrix}\\ &\times\left(\eta\left(\tilde{\mathbf{X}}_{l-1}\mathbf{L}^f\tilde{\mathbf{X}}_{l-1}^T\right)+3\mu\left(\tilde{\mathbf{X}}_{l-1}\tilde{\mathbf{X}}_{l-1}^T\right)+\mu\mathbf{I}\right)^{-1}.\end{aligned}\tag{26}$$

$\mathbf{H}$ *Problem:* The variable $\mathbf{H}$ can be estimated by solving the following problem:

$$\begin{aligned}\min_{\mathbf{H}}\ & \frac{1}{2}\left\|\tilde{\mathbf{X}}_{l-1}-\mathbf{G}^T\mathbf{H}\right\|_{\mathrm{F}}^2+\frac{\mu}{2}\left\|\mathbf{H}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2\\ &+\boldsymbol{\Lambda}_1^T\left(\mathbf{H}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)\end{aligned}\tag{27}$$

its analytical solution is given by

$$\mathbf{H}\leftarrow\left(\mathbf{G}\mathbf{G}^T+\mu\mathbf{I}\right)^{-1}\left(\mathbf{G}\tilde{\mathbf{X}}_{l-1}+\mu\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}-\boldsymbol{\Lambda}_1\right).\tag{28}$$

$\mathbf{G}$ *Problem:* The optimization problem can be written as

$$\min_{\mathbf{G}}\frac{\mu}{2}\|\mathbf{G}-\boldsymbol{\Theta}_l\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_2^T(\mathbf{G}-\boldsymbol{\Theta}_l)\tag{29}$$

which can be effectively solved as

$$\mathbf{G}\leftarrow\left(\mathbf{H}\mathbf{H}^T+\mu\mathbf{I}\right)^{-1}\left(\mathbf{H}\tilde{\mathbf{X}}_i+\mu\boldsymbol{\Theta}_l-\boldsymbol{\Lambda}_2\right).\tag{30}$$

$\mathbf{Q}$ *Problem:* The optimization problem of $\mathbf{Q}$ is

$$\min_{\mathbf{Q}}\frac{\mu}{2}\left\|\mathbf{Q}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_3^T\left(\mathbf{Q}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)+l_R^+(\mathbf{Q}).\tag{31}$$

Here, the update rule for $\mathbf{Q}$ can be expressed as

$$\mathbf{Q}\leftarrow\max\left(\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}-\boldsymbol{\Lambda}_3/\mu,\mathbf{0}\right).\tag{32}$$

$\mathbf{S}$ *Problem:* The variable $\mathbf{S}$ is estimated by solving

$$\min_{\mathbf{S}}\frac{\mu}{2}\left\|\mathbf{S}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right\|_{\mathrm{F}}^2+\boldsymbol{\Lambda}_4^T\left(\mathbf{S}-\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}\right)+l_R^{\sim}(\mathbf{S})\tag{33}$$

whose solution can be updated in each iteration by the vector-based projection operator of (15)

$$\mathbf{S}\leftarrow\operatorname{prox}_f\left(\boldsymbol{\Theta}_l\tilde{\mathbf{X}}_{l-1}-\boldsymbol{\Lambda}_4/\mu\right).\tag{34}$$

Lagrange Multipliers ($\{\boldsymbol{\Lambda}_i\}_{i=1}^4$) Update: Before stepping into the next iteration, the Lagrange multipliers are updated by

$$\begin{aligned}&\boldsymbol{\Lambda}_1=\boldsymbol{\Lambda}_1+\mu\left(\mathbf{H}-\boldsymbol{\Theta}_i\tilde{\mathbf{X}}_{l-1}\right),\ \boldsymbol{\Lambda}_2=\boldsymbol{\Lambda}_2+\mu(\mathbf{G}-\boldsymbol{\Theta}_i)\\ &\boldsymbol{\Lambda}_3=\boldsymbol{\Lambda}_3+\mu\left(\mathbf{Q}-\boldsymbol{\Theta}_i\tilde{\mathbf{X}}_{l-1}\right),\ \boldsymbol{\Lambda}_4=\boldsymbol{\Lambda}_4+\mu\left(\mathbf{P}-\boldsymbol{\Theta}_i\tilde{\mathbf{X}}_{l-1}\right).\end{aligned}\tag{35}$$

ACKNOWLEDGMENT

The authors would like to thank the HS Image Analysis group and the NSF Funded Center for Airborne Laser Mapping (NCALM) at the University of Houston for providing the CASI University of Houston dataset. The authors would like to express their appreciation to Prof. D. Cai and Dr. C. Wang for providing MATLAB codes for LPP and manifold alignment algorithms.

REFERENCES

[1] D. Hong, N. Yokoya, J. Xu, and X. Zhu, "Joint & progressive learning from high-dimensional data for multi-label classification," in *Proc. ECCV*, 2018, pp. 469–484.
[2] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise mrf optimization," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2966–2977, Dec. 2016.
[3] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Mar. 2016.

[4] X. Lu, W. Zhang, and X. Li, "A hybrid sparsity and distance-based discrimination detector for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1704–1717, Dec. 2017.

[5] Q. Wang, F. Zhang, and X. Li, "Optimal clustering framework for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5910–5922, Oct. 2018.

[6] X. Wu, D. Hong, J. Tian, J. Chanussot, W. Li, and R. Tao, "Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5146–5158, Jan. 2019.

[7] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 24, 2020, doi: 10.1109/TGRS.2020.3016820.

[8] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2018.

[9] X. Wu, D. Hong, J. Chanussot, Y. Xu, R. Tao, and Y. Wang, "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 302–306, May 2020.

[10] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, early access, Jun. 18, 2020, doi: 10.1109/TGRS.2020.3000684.

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

[12] W. Li, F. Feng, H. Li, and Q. Du, "Discriminant analysis-based dimension reduction for hyperspectral image classification: A survey of the most recent advances and an experimental comparison of different techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 1, pp. 15–34, Mar. 2018.

[13] W. Li, S. Prasad, J. E. Fowler, and M. Cui, "Locality-preserving non-negative matrix factorization for hyperspectral image classification," in *Proc. IEEE IGARSS*, 2012, pp. 1405–1408.

[14] L. Gao, B. Zhao, X. Jia, W. Liao, and B. Zhang, "Optimized kernel minimum noise fraction transformation for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 6, p. 548, 2017.

[15] H. Huang, F. Luo, J. Liu, and Y. Yang, "Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 106, pp. 42–54, Apr. 2015.

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

[17] B. Rasti, M. Ulfarsson, and J. R. Sveinsson, "Hyperspectral feature extraction using total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 6976–6985, Aug. 2016.

[18] D. Hong, N. Yokoya, and X. Zhu, "Learning a robust local manifold representation for hyperspectral dimensionality reduction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2960–2975, Jul. 2017.

[19] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1581–1586, Sep. 2019.

[20] H. Huang, G. Shi, H. He, Y. Duan, and F. Luo, "Dimensionality reduction of hyperspectral imagery based on spatial-spectral manifold learning," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2604–2616, Jun. 2020, doi: 10.1109/TCYB.2019.2905793.

[21] W. Liao, A. Pizurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 184–198, Jan. 2013.

[22] M. Zhao, Z. Zhang, T. Chow, and B. Li, "A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction," *Neural Netw.*, vol. 55, pp. 83–97, May 2014.

[23] H. Wu and S. Prasad, "Semi-supervised dimensionality reduction of hyperspectral imagery using pseudo-labels," *Pattern Recognit.*, vol. 74, pp. 212–224, Feb. 2018.

[24] L. Gao, D. Gu, L. Zhuang, J. Ren, D. Yang, and B. Zhang, "Combining *t*-distributed stochastic neighbor embedding with convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1368–1372, Jun. 2020.

[25] A. M. Martínez and C. K. Avinash, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Oct. 2001.

[26] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *Proc. IJCAI*, vol. 9, 2009, pp. 1077–1082.

[27] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1470–1474, Mar. 2020.

[28] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, no. 2, pp. 12–23, 2020.

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

[30] 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, Mar. 2017.

[31] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, no. 8, pp. 35–49, 2019.

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

[33] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[34] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometr. Intell. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987.

[35] X. He and P. Niyogi, "Locality preserving projections," in *Proc. NIPS*, 2004, pp. 153–160.

[36] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[37] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Trans. Geos. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Apr. 2019.

[38] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, May 2012.

[39] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. Zhu, "Learnable manifold alignment (LEMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, Jan. 2019.

[40] J. Hu, W. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov. 2017.

[41] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Aug. 2007.

[42] F. R. K. Chung, *Spectral Graph Theory*, Amer. Math. Soc., Providence, RI, USA, 1997.

[43] D. Hong, L. Gao, J. Yao, B. Zhang, P. Antonio, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Aug. 18, 2020, doi: 10.1109/TGRS.2020.3015157.

[44] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. NIPS*, 2007, pp. 801–808.

[45] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[46] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *Proc. CVPR*, 2015, pp. 5135–5143.

[47] D. P. Bertsekas, *Nonlinear Programming*. Belmont, CA, USA: Athena Sci., 1999.

[48] S. Boyd *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[49] J. Kang, D. Hong, J. Liu, G. Baier, N. Yokoya, and B. Demir, "Learning convolutional sparse coding on complex domain for interferometric phase restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 9, 2020, doi: 10.1109/TNNLS.2020.2979546.

[50] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

Danfeng Hong (Member, IEEE) received the M.Sc. degree (*summa cum laude*) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, and the Dr.-Ing degree (*summa cum laude*) in signal processing in Earth observation, Technical University of Munich, Munich, Germany, in 2019.

Since 2015, he has been a Research Associate with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany. He is a Research Scientist and leads a Spectral Vision Working Group with IMF, DLR, and an Adjunct Scientist with the GIPSA-lab, Grenoble INP, CNRS, Université Grenoble Alpes, Grenoble, France. His research interests include signal/image processing and analysis, hyperspectral remote sensing, machine/deep learning, and artificial intelligence and their applications in Earth vision.

Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees in aeronautics and astronautics from the University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

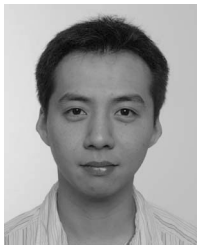
He is currently a Lecturer with the University of Tokyo and a Unit Leader with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Unit. He was an Assistant Professor with the University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow, working with the German Aerospace Center (DLR), Wessling, Germany, and Technical University of Munich (TUM), Munich, Germany. His research is focused on the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management.

Dr. Yokoya won the first place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He is the Chair from 2019 to 2021 and was the Co-chair from 2017 to 2019 of IEEE GRSS IADF TC and has been the Secretary of the IEEE GRSS All Japan Joint Chapter since 2018. He has been an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2018. He is/was a Guest Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2015–2016, *Remote Sensing* in 2016–2020, and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2018 and 2019.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor with the INRIA, CNRS, Grenoble INP, LJK, Université Grenoble Alpes. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA; the KTH Royal Institute of Technology, Stockholm, Sweden; and the National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavík, Iceland. In 2015–2017, he was a Visiting Professor with the University of California at Los Angeles, Los Angeles, CA, USA. He holds the AXA Chair of Remote Sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Prof. Chanussot is the Founding President of IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010 which received the 2010 IEEE GRS-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia from 2017 to 2019. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. He was the Chair from 2009 to 2011 and the Co-chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS and *Remote Sensing* from 2011 to 2015. In 2014, he served as a Guest Editor for the *IEEE Signal Processing Magazine*. He is a member of the Institut Universitaire de France from 2012 to 2017 and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, 2018–2019).



Jian Xu (Member, IEEE) received the B.E. degree in geographic information systems from Hohai University, Nanjing, China, in 2004, and the M.S. degree in earth-oriented space science and technology and the Ph.D. degree in atmospheric remote sensing from the Technical University of Munich, Munich, Germany, in 2009 and 2015, respectively.

Since 2010, he has been with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany. His research interests include remote sensing of atmospheric temperature and trace gases, radiative transfer modeling, and ill-posed inverse problems.

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

She is currently the Professor for Data Science in Earth Observation, TUM and the Head of the Department "EO Data Science" with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Wessling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School, Munich. Since 2019, she has been the Head of the Helmholtz Artificial Intelligence Cooperation Unit—Research Field "Aeronautics, Space and Transport." Since 2020, she has been the Director of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; the University of Tokyo, Tokyo, Japan, in 2015; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science with a special application focus on global urban mapping.

Prof. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin–Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.