UCSL: Towards Unsupervised Common Subspace Learning for Cross-Modal Image Classification

Jing Yao, *Member, IEEE,* Danfeng Hong, *Senior Member, IEEE,* Haipeng Wang, Hao Liu, Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—The emerging research line of cross-modal learning focuses on the issue of transferring feature representation manner learned from limited multimodal data with labelings to the testing phase with partial modalities. This is essentially common and practical in the remote sensing community when only modal-incomplete data are in users' hands due to inevitable imaging or access restrictions under large-scale observation scenarios. However, most of the existing cross-modal learning methods have been designed with exclusive reliance on labeling, which can be either limited or noisy due to their costly production. To address this issue, we explore in this paper the possibility to learn cross-modal feature representation in an unsupervised fashion. By integrating the multimodal data into a fully recombined matrix form, we propose 1) the use of common subspace representation as the regression target instead of conventionally adopted binary labels, and 2) the orthogonality and manifold alignment regularization terms to shrink the solution space whilst preserving the pairwise manifold correlations. Through this manner, the modality-specific and mutual latent representations in this common subspace as well as their corresponding projections can be learned simultaneously and their optimums can be efficiently reached through a nearly one-step computation with the help of Eigen decomposition. Finally, we show the superiority of our method through extensive image classification experiments on three multimodal datasets with four remotely sensed modalities involved (i.e., hyperspectral, multispectral, synthetic aperture radar, and light detection and ranging data). The code and dataset will be made freely available at https://github.com/jingyao16/UCSL after a possible publication to encourage the reproduction of our method and further use.**

***Index Terms*—Cross-modal, unsupervised, multimodal, remote sensing, common subspace, image classification, manifold alignment.**

I. Introduction and Contribution

THE rapid advancement of machine learning methodology in remote sensing (RS) has successfully enabled humans with the ability to understand what they see and predict the unseen on the Earth. Among the multitude of applications it has found in the past decades, the long-standing research problem of pixel-level classification using RS images has been gaining a surge of interest for its fundamental and considerable potential in a wide range of downstream use cases, such as object detection [1], [2], mineral exploration [3]–[5], environmental monitoring [6], urban mapping [7], disaster detection [8], [9], and so forth. Specifically, the aim of RS image classification is to assign a physically meaningful attribute to each pixel of a remotely sensed image according to its ground object distributions. To achieve this goal, four prominent procedures can be roughly summarized: front-side data preparation/pre-processing, feature extraction/selection, back-end classifier training/validating, and lastly the optional post-processing on the inference result, the mid-range of which becomes our main focus in this article.

As a common choice of the baseline along this line, one can naturally feed mature classifiers, such as k-nearest neighbors (k-NN), support vector machines [10], and deep neural networks [11], with ready-made input data–after Earth observations underwent necessary pre-processing–to obtain the mapping result. In this context, the advent of hyperspectral (HS) imaging has significantly expanded the boundaries of RS capability in recognizing targets of interest at a material level, owing to its simultaneous and dense spectral sampling of radiance at contiguous ranges of wavelengths [12]. Nevertheless, on one side, *Spectral Variability* and the *Curse of Dimensionality* are two non-negligible issues that accompany and bother the extensive use of HS images. For years, researchers have devoted themselves to developing more advanced machine learning tools by incorporating intrinsically structural and statistical prior knowledge to alleviate the redundancy underlying such high-dimensional formed data. Concrete ideas include, but are not limited to, manifold learning that seeks lower dimensional but meaningful representation while maintaining higher-order topological structures [13], representation learning that purses underlying sparsity via multiple spatial-spectral features aggregation [14], kernel learning by implicitly introducing nonlinearity representation power [15], ensemble learning based on ensembling a set of alternative classifiers that can guide the final decision towards a more favorable trade-off between the bias and variance [16], spatial information enhancement through edge preservation and texture smoothing [17], [18], and multi-view learning [19].

On the other side, despite the technical superiority achieved in sensing from HS data, it is still inevitable to meet an interpretation performance bottleneck in satisfying the ever-

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401. (*Corresponding author: Haipeng Wang*)

Jing Yao and Danfeng Hong are with the Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China. (e-mail: yaojing@aircas.ac.cn; hongdf@aircas.ac.cn)

Haipeng Wang is with the Harbin Institute of Technology (Shenzhen), 518055 Shenzhen, China, and also with the Naval Aviation University, 264001 Yantai, China. (e-mail: whp5691@163.com)

Hao Liu is with Shanghai Jiaotong University, 200240 Shanghai, China, and also with the Wuhan Digital Engineering Institute, 430074 Wuhan, China. (e-mail: liuhao2020@sjtu.edu.cn)

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

demanding industrial requirements for more precise, delicate, and customized applications. As one of the possible solutions, integrating multimodal RS data acquired from collaborated observations has become an exceedingly useful adjunct in pushing such a limit. Apart from the various modeling frameworks, either conventional optimization or deep learning-based ones [20], the efforts along this research line are generally thought to be of two kinds, i.e., concatenation-based and alignment-based models [21]. In specific, the first group of methods usually focuses on either multimodal feature learning based on input-level concatenation or modality-specific feature learning for output-level concatenation. For example, Liao *et al.* earlier proposed to generalize the handcrafted graph embedding by fusing the spectral, spatial, and elevation information contained in HS and height in light detection and ranging (LiDAR) data [22]. Yokoya *et al.* made the first attempt at fusing the spectral reflectance, spectral indexes, and morphological profiles for ensemble learning-based local climate zone classification [23]. More recently, Hang *et al.* proposed to append the coupled deep learning with a simultaneous feature and decision-level fusion, which has also received considerable attention [24], while Wu *et al.* introduced more advanced cross-channel reconstruction module in coupled networks [25]. Admittedly, a common prerequisite behind the success of these methods lies in the availability of multimodal observations for both training and testing use, which substantially limits their *cross-modal learning* ability in practical inference cases with modal-incomplete data [26].

Although there have been noteworthy efforts paid to adjust concatenation-based frameworks to fit the cross-modal feature learning scenario, such as zero-padding [26] and generative adversarial training [27], while both of their interpretability and ability in explicitly characterizing underlying multimodal structural information remains restricted. In comparison to those strategies, the alignment-based methods typically assume the alignability between multiple modalities and aim to learn their corresponding projection manners that can separately map multimodal input into a shared space. One popular way to encode such property is through manifold representation techniques. Tuia *et al.* first introduced semi-supervised manifold alignment for multitemporal, multisource, and multiangular very high resolution RS image classification without co-registration requirement [28] and further developed a nonlinear version with a guarantee for better domain generalization capability by applying the kernel tricks [29]. The series work by Hong *et al.* ably introduced the concept of subspace learning and achieved remarkable performance improvement in extensive multimodal and cross-modal learning applications, the details of which will be presented in the next section [21], [30]–[32]. It has also found places in other prominent RS applications, various examples include visualization of multi-band RS images [33], semi-supervised charting-based multimodal RS object recognition and semantic segmentation [34], and so on.

Throughout the above review, the vast majority of existing multimodal or cross-modal feature learning approaches so far, from either supervised or semi-supervised modeling perspectives, has been designed with indispensable dependence on label information. However, as is the case in our investigated RS image classification task, the pixel-wise labeling itself can be not only labor-costly for its field inspection and in-lab annotation by specialists, but also vulnerable to noise corruption caused by complex environmental factors [35]. What's more, we found that these supervised feature extraction methods mostly resort to a separately trained classifier to obtain the final classification results instead of using those learned representations that have been directly projected into the target space. This could probably be explained by that supervision in those commonly adopted one-hot encoded labels inevitably miss the relation information besides the correct category. To avoid such an over-constrained situation, we therefore strive to develop an unsupervised cross-modal feature extraction method that can get rid of using labeling as the regression target. With more relaxed constraints, we propose to implicitly set a latent representation that keeps the same multimodal manifold structure as the original recombined features, hoping to better reveal the underlying correlations among samples that were drawn from multimodal RS images. Our threefold contributions can be highlighted as follows:

1) A novel Unsupervised Common Subspace Learning method, abbreviated as UCSL, is proposed to effectively learn a discriminative representation without the need for annotations that can still generalize well for the task of cross-modal RS image classification.
2) With the guidance of our modified supervised graph, the proposed UCSL model can be readily upgraded into its supervised version, termed SCSL, which has shown its further superior capability of cross-modal feature learning on the basis of USCL in extensively conducted RS image classification experiments.
3) We design an efficient algorithm that can resolve the proposed models with nearly one-step computation by resorting to the well-studied spectral decomposition, which shows faster speed than conventional ones under practical situations with limited training samples.

The remainder of this article is structured as follows. Section II first introduces the background of related work and then describes our proposed method. Section III presents the description of investigated datasets, implementation details about training and testing, and experimental results and analysis. Section IV gives a conclusion of this article and plausible future outlooks.

II. Background and Methodology

In this section, we first briefly recall the basic preliminaries of common subspace learning-based works for cross-modal RS image classification. Then, we present our methods from mathematical problem formulation to its optimization, and lastly some necessary analysis in regards to detailed computations.

As for the mathematical notations, we follow the commonly adopted standard in signal processing as follows. Unless otherwise stated, we use non-bold case letters to denote scalars and bold upper case letters for matrices. Parenthesized superscript denotes a specific modality. Subscripts using single and two letters denote the indexes of a certain column and entry,

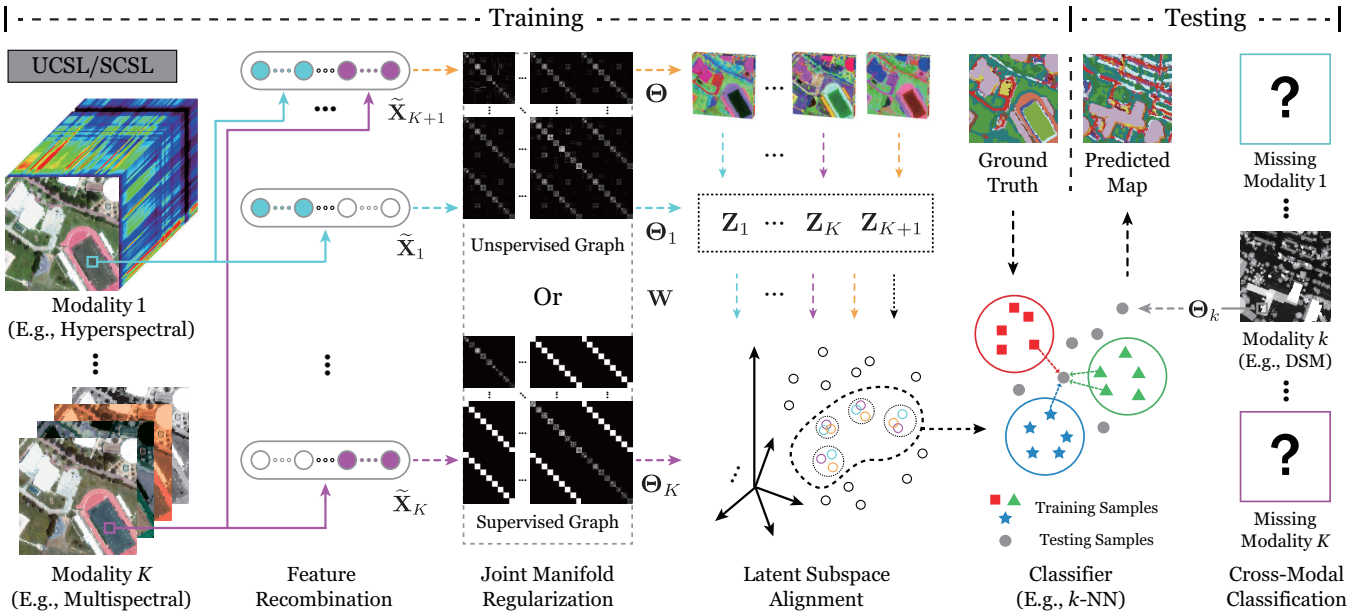


Fig. 1. An illustration to clarify the workflow of our proposed UCSL/SCSL model.

respectively. $\|\cdot\|_2$ is the vector ℓ_2-norm, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and $\mathrm{tr}(\cdot)$ computes the trace of a matrix.

A. Recalling Cross-Modal Common Subspace Learning

Let $\mathbf{Y} \in \mathbb{R}^{C\times N}$ denote the one-hot encoded labels in matrix form, where C and N are the numbers of ground object categories of interest and pixels within the investigated scene, respectively. For the k-th modality RS image ($k = 1, \ldots, K$) out of K considered modalities in total, let $\mathbf{X}^{(k)} \in \mathbb{R}^{d_k\times N}$ denotes the unfolded matrix with d_k channels. The task of cross-modal image classification typically assumes that $\{\mathbf{X}^{(k)}, \mathbf{Y}\}_{k=1}^{K}$ are available in the training phase whereas only a subset of $\{\mathbf{X}^{(k)}\}_{k=1}^{K}$ can be used to predict labels in the testing phase. The key point to solving this task, therefore, lies in how to extract features by exploring common information underlying multimodal training data.

The idea of subspace learning has been widely proved effective in the past decades, especially for the processing of high-dimensional RS data. Motivated by the joint framework of subspace learning and classification for single-modal data [36], one can readily derive its multimodal version as follows,

$$\min_{\mathbf{P},\boldsymbol{\Theta}\boldsymbol{\Theta}^\top=\mathbf{I}} \frac{1}{2}\|\tilde{\mathbf{Y}} - \mathbf{P}\boldsymbol{\Theta}\tilde{\mathbf{X}}\|_{\mathrm{F}}^2 + \Omega(\mathbf{P}), \tag{1}$$

where $\tilde{\mathbf{X}}$ horizontally collects K-length zero-padded single-modal observations $[\mathbf{0}; \ldots; \mathbf{X}^{(k)}; \ldots; \mathbf{0}]$, $\tilde{\mathbf{Y}} = [\mathbf{Y}; \ldots; \mathbf{Y}] \in \mathbb{R}^{C\times KN}$, $\mathbf{P}$ and $\boldsymbol{\Theta}$ are to-be-estimated projection matrices that bridging the data space, latent common subspace, and label space, respectively. The orthogonality assumption on $\boldsymbol{\Theta}$ and the regularization term on $\mathbf{P}$ which usually takes the form of its Frobenius norm, i.e., $\Omega(\mathbf{P}) = \lambda\|\mathbf{P}\|_{\mathrm{F}}^2$, are basic constraints that help to narrow the solution space.

To further introduce prior knowledge in the context of RS image classification, enormous efforts have been made on the basis of the baseline model in Eq. (1). Among them, the graph structural information encoded by the manifold regularization is undoubtedly one kind of panacea, which has been frequently adopted by Hong *et al.* in their works as

$$\frac{1}{2}\sum_{i,j=1}^{N}\|(\boldsymbol{\Theta}\tilde{\mathbf{X}})_i - (\boldsymbol{\Theta}\tilde{\mathbf{X}})_j\|_2^2\mathbf{W}_{ij} = \mathrm{tr}(\boldsymbol{\Theta}\tilde{\mathbf{X}}\mathbf{L}\tilde{\mathbf{X}}^\top\boldsymbol{\Theta}^\top), \tag{2}$$

where $\mathbf{L}$ is called the graph Laplacian [37] derived from the adjacency matrix $\mathbf{W}$ on graph. By doing so, the neighborhood relationships on the manifold can be enforced as well in the latent common subspace. Extensions to this idea include substituting $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{X}}$ with their superpixel representations [32], decomposing $\boldsymbol{\Theta}$ by considering modality-shared and modality-specific characteristics [21], and absorbing massive unlabeled data in a semi-supervised fashion to exploit more global information [31].

B. Problem Formulation

The above-mentioned series of works are all without exception built on a similar regression framework that tries to model the relationship between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$. However, it is worth noting that, the target space spanned by limited one-hot encodings may not be a good choice since those replicated binary labels are not only sparsely distributed but also vulnerable to inevitable disturbance. On the other hand, an interesting observation drawn from our reproductions shows that retraining mature classifiers with extracted latent features always outperforms those direct regression results (like the term $\mathbf{P}\boldsymbol{\Theta}\tilde{\mathbf{X}}$ in Eq. (1)). This strongly motivates us to rethink the possibility of developing a novel cross-modal feature extraction model without using labels.

Without relying on $\tilde{\mathbf{Y}}$, our main purpose now becomes to learn $\boldsymbol{\Theta}$ better only using $\{\mathbf{X}^{(k)}\}_{k=1}^{K}$. Unlike conventional set-

tings in CoSpace and its derivations, we propose to recombine the K-modality training samples as

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} & \dots & \mathbf{0} & \mathbf{X}^{(1)} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{X}^{(K)} & \mathbf{X}^{(K)} \end{bmatrix} \in \mathbb{R}^{\sum_{k=1}^{K} d_k \times (K+1)N}, \tag{3}$$

which includes both multimodal and single-modal representations so as to enrich the data distribution in the original high-dimensional space[1]. By introducing the modality-specific projections $\mathbf{\Theta}_k \in \mathbb{R}^{d \times d_k}$, we can then readily define their concatenation $\mathbf{\Theta} = [\mathbf{\Theta}_1, \dots, \mathbf{\Theta}_K]$ as the joint transformation matrix that projects $\widetilde{\mathbf{X}}$ into a d-dimensional latent subspace. Similarly, we assume the orthogonality of the target representation $\mathbf{Z} \in \mathbb{R}^{d \times (K+1)N}$ corresponding to those projected features to reduce the ill-posedness of our unsupervised regression formulation as follows,

$$\min_{\mathbf{\Theta}, \mathbf{Z}\mathbf{Z}^\top = \mathbf{I}} \frac{1}{2} \|\mathbf{\Theta}\widetilde{\mathbf{X}} - \mathbf{Z}\|_{\mathrm{F}}^2 + \Phi(\mathbf{\Theta}) + \Psi(\mathbf{Z}), \tag{4}$$

in which we use $1/2$ to simplify the following differentiation deductions. Our main motivation in constructing the regularization terms $\Phi(\mathbf{\Theta})$ and $\Psi(\mathbf{Z})$ are as follows. We first propose to regularize the latent representation $\mathbf{Z}$, or say, the regression target, using the joint graph Laplacian $\mathbf{L}$, to maintain the intrinsic geometry structure of multimodal input $\widetilde{\mathbf{X}}$,

$$\Psi(\mathbf{Z}) = \frac{\gamma}{2} \operatorname{tr}(\mathbf{Z}\mathbf{L}\mathbf{Z}^\top), \tag{5}$$

where the detailed computation of $\mathbf{L}$ will be discussed in the subsequent section. As for the projection matrix $\mathbf{\Theta}$, we on one hand adopt the Frobenius norm to pursue small-valued projection that has better generalization ability[2]. On the other hand, as Fig. 1 demonstrates, we enforce the joint manifold regularization on those projected features using the same graph Laplacian matrix as above, hoping to double the the alignment in latent subspace via coupling with the first regression term,

$$\Phi(\mathbf{\Theta}) = \frac{\alpha}{2} \|\mathbf{\Theta}\|_{\mathrm{F}}^2 + \frac{\beta}{2} \operatorname{tr}(\mathbf{\Theta}\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^\top\mathbf{\Theta}^\top), \tag{6}$$

where α, β, and γ are positive parameters that weight the relative importance of different priors. More experimental results and analysis on the necessity of these regularization terms can be found in the experiments section.

C. Optimization

The optimization problem of the proposed model in Eq. (4) is generally nonconvex with respect to the to-be-estimated variables. A direct solution is to resort to the alternating direction method of multipliers (ADMM) algorithm by introducing auxiliary variables to split the objective function and the corresponding constraints as shown in [39]. However, in this work, we propose to solve it more efficiently by the following deductions.

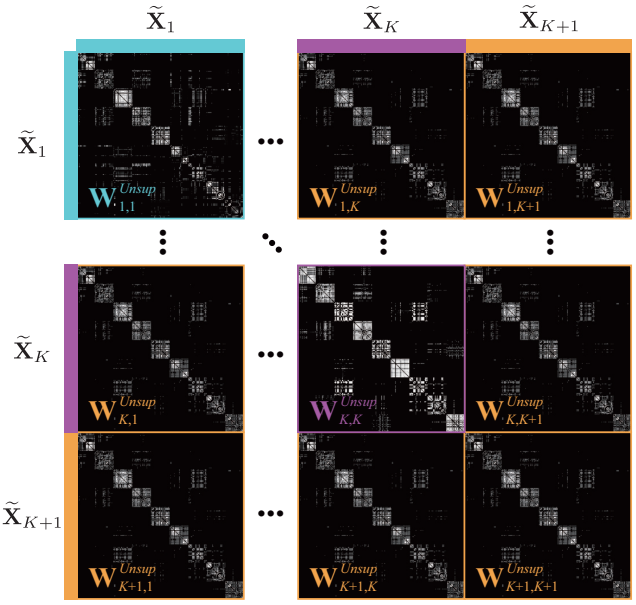


Fig. 2. An example to illustrate the adopted unsupervised adjacency matrix ($\mathbf{W}^{Unsup}$).

We first consider the subproblem with respect to $\mathbf{\Theta}$ by fixing $\mathbf{Z}$, the objective function of which takes the form of

$$J_{\mathbf{\Theta}} = \|\mathbf{\Theta}\widetilde{\mathbf{X}} - \mathbf{Z}\|_{\mathrm{F}}^2 + \alpha\|\mathbf{\Theta}\|_{\mathrm{F}}^2 + \beta \operatorname{tr}(\mathbf{\Theta}\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^\top\mathbf{\Theta}^\top). \tag{7}$$

It is convex and thus we can readily set its first-order derivative to zero, and get its closed-form solution as $\hat{\mathbf{\Theta}} = \mathbf{Z}\widetilde{\mathbf{X}}^\top\mathbf{H}^{-1}$, where $\mathbf{H} = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top + \alpha\mathbf{I} + \beta\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^\top$.

Then we can substitute the above optimal $\hat{\mathbf{\Theta}}$ into Eq. (4) to rewrite our overall objective as a function with respect to $\mathbf{Z}$,

$$\begin{aligned} J_{\mathbf{Z}} &= \operatorname{tr}\big(\hat{\mathbf{\Theta}}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top\hat{\mathbf{\Theta}}^\top - 2\hat{\mathbf{\Theta}}\widetilde{\mathbf{X}}\mathbf{Z}^\top + \mathbf{Z}\mathbf{Z}^\top + \alpha\hat{\mathbf{\Theta}}\hat{\mathbf{\Theta}}^\top \\ &\quad + \beta\hat{\mathbf{\Theta}}\widetilde{\mathbf{X}}\mathbf{L}\widetilde{\mathbf{X}}^\top\mathbf{\Theta}^\top + \gamma\mathbf{Z}\mathbf{L}\mathbf{Z}^\top\big) \\ &= \operatorname{tr}\big(\hat{\mathbf{\Theta}}\mathbf{H}\hat{\mathbf{\Theta}}^\top - 2\hat{\mathbf{\Theta}}\widetilde{\mathbf{X}}\mathbf{Z}^\top + \mathbf{Z}(\mathbf{I} + \gamma\mathbf{L})\mathbf{Z}^\top\big) \\ &= \operatorname{tr}\big(\mathbf{Z}(-\widetilde{\mathbf{X}}^\top\mathbf{H}^{-1}\widetilde{\mathbf{X}} + \mathbf{I} + \gamma\mathbf{L})\mathbf{Z}^\top\big). \end{aligned} \tag{8}$$

According to [40], the above optimization problem under the orthogonality constraint follows the form of spectral clustering, the solution $\hat{\mathbf{Z}}$ to which can be readily given by Eigen decomposition. Note that our algorithm can still provide comparable one-step solutions by eliminating any of the regularization terms.

D. Method Analysis

1) Unsupervised/Supervised Graph Construction: Like related works do, the classification performance largely depends on the construction of the $\mathbf{L}$. Ideally, an accurate estimation of the graph adjacency matrix $\mathbf{W}$ is beneficial for capturing global manifold structure and thus expected to result in better feature extraction. However, to avoid the burdensome whole graph computation and its vulnerability to complex noise, many recent efforts have verified the effectiveness of using locally computed alternatives [41], where only a certain size of neighbors to each sample, denoted as $\mathcal{N}(\cdot)$, are taken into

[1]Note that one can readily generalize such a recombination manner to superpixel case by following [38], which can better exploit prior knowledge on spatial structures and thus expect better classification performance.

[2]Instead, other kinds of prior can be encoded here, such as the sum of the column- or row-wise ℓ_2-norm that introduces feature selection function. However, the resulting optimization loses the efficiency caused by its iterative updating of auxiliary variable, which is not suggested in practice.

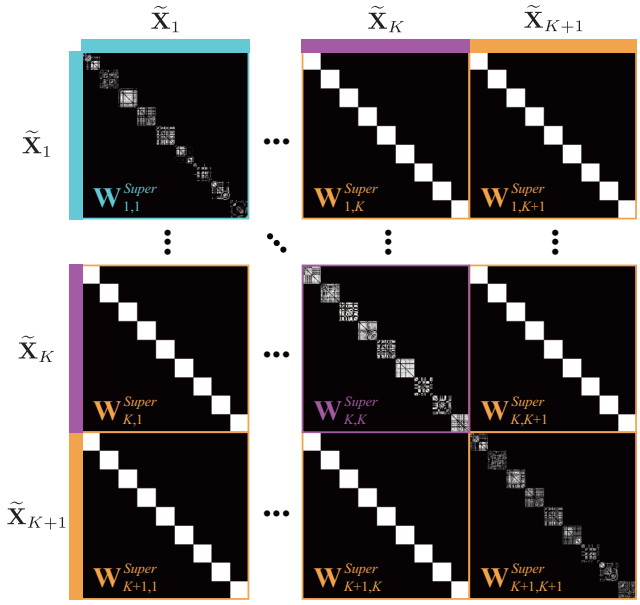


Fig. 3. An example to illustrate the adopted supervised adjacency matrix ($\mathbf{W}^{Super}$).

consideration. By following this rule, we first propose our unsupervised version as follows,

$$(\mathbf{W}^{Unsup}_{k,k})_{ij} = \begin{cases} r_\sigma((\widetilde{\mathbf{X}}_k)_i, (\widetilde{\mathbf{X}}_k)_j), & \text{if } (\widetilde{\mathbf{X}}_k)_i \in \mathcal{N}((\widetilde{\mathbf{X}}_k)_j), \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

in which $r_\sigma(a,b) = \exp(-\|a-b\|_2^2/2\sigma^2)$ is the so-called radial basis function kernel with parameter σ controlling the width. Considering that more information is contained in the recombined multimodal representations, we can therefore set the intra-modality graph as that computed by $\widetilde{\mathbf{X}}_{K+1}$, i.e., $\mathbf{W}^{Unsup}_{k_1,k_2} = \mathbf{W}^{Unsup}_{K+1,K+1}$, when $k_1 \neq k_2$, hoping to provide a more reliable guidance for the joint manifold regularization.

Note that although our regression model is in no need of the hard labeled information, we can still integrate it into the construction of a supervised version, i.e., $(\mathbf{W}^{Super}_{k,k})_{ij} = r_\sigma((\widetilde{\mathbf{X}}_k)_i, (\widetilde{\mathbf{X}}_k)_j)/N_c$, if $(\widetilde{\mathbf{X}}_k)_i \in \mathcal{N}((\widetilde{\mathbf{X}}_k)_j)$, and they belong to the same c-th class, which has N_c samples in total. In this case, we can also set $(\mathbf{W}^{Super}_{k_1,k_2})_{ij} = 1/N_c$, if $(\widetilde{\mathbf{X}}_{k_1})_i$ and $(\widetilde{\mathbf{X}}_{k_2})_j$ shares the same attribute, and 0 otherwise, for $k_1 \neq k_2$, by following the discriminative graph structure setting in [21]. Fig. 2 and Fig. 3 show examples of the two cases respectively.

With the above-mentioned adjacency matrix $\mathbf{W}$ in hand, constructed either in an unsupervised or supervised fashion, we can then compute the degree matrix $\mathbf{D}$ by $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, and obtain the normalized graph Laplacian matrix as $\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-\frac{1}{2}}$ for the final use, which is expected to be well-equipped to handle multimodal features with both regular and irregular graph structures [37].

2) Spectral Decomposition: Generally, the solution that minimizes Eq. (8) can be provided by either the left or right eigenvectors corresponding to the d smallest eigenvalues in Eigen decomposition of $\mathbf{M} = -\widetilde{\mathbf{X}}^\top \mathbf{H}^{-1}\widetilde{\mathbf{X}} + \mathbf{I} + \gamma\mathbf{L}$. Based on the above definitions, we can easily determine the symmetry of $\mathbf{M}$ by the fact that $\mathbf{H}^{-1} = (\mathbf{H}^{-1})^\top$. However, to avoid possible numerical instabilities caused by computing the inversion of $\mathbf{H}$ in practice, we propose to use $\texttt{eig}((\mathbf{M}+\mathbf{M}^\top)/2)$ instead. Note that the Eigen decomposition of such a real symmetric matrix is also termed as *spectral decomposition*, by which the orthogonal eigenvectors can be generated more stably and efficiently than $\texttt{eig}(\mathbf{M})$.

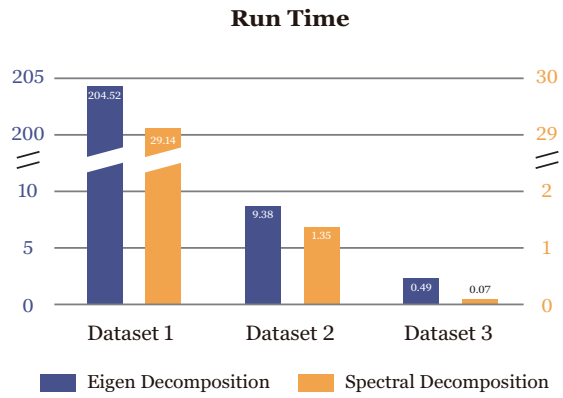


Fig. 4. Run time comparison of the conventional Eigen decomposition and the adopted spectral decomposition on the three investigated datasets.

3) Cross-Modal Inference: As aforementioned, our model can be resolved by nearly one-step Eigen decomposition, and obtain the optimal projection by $\hat{\mathbf{\Theta}} = \widetilde{\mathbf{Z}}\widetilde{\mathbf{X}}^\top\mathbf{H}^{-1}$. To guarantee a fair comparison without augmenting the training set, we feed the off-the-shelf classifier such as k-NN with $\hat{\mathbf{\Theta}}_k\mathbf{X}^{(k)}$ and finally use it for inference on the k-th modality testing data. Fig. 1 gives a whole picture of the proposed cross-modal feature learning framework for RS image classification.

4) Computational Efficiency: Besides the normal matrix multiplications in solving our proposed model, it essentially involves the Eigen decomposition of a $(K+1)N$-by-$(K+1)N$ matrix, which commonly takes a complexity of $\mathcal{O}(N^3)$. However, the efficiency of our algorithm can be guaranteed from the perspectives of numerical computation and application scenario. First, adopting the spectral decomposition as stated in subsection 2 can evidently speed up the run time in our practice (nearly 7x when compared with $\texttt{eig}(\mathbf{M})$ as shown in Fig. 4) by virtue of the spectral theorem [42]. Second, the N here only depicts the size of train set, which is commonly much more less than that of test set. Considering the fact that the inference computational cost of our model behaves no differences with those of the same type, it is reasonable to believe that our model can provide a better alternative than conventional common subspace learning methods for addressing the classification of large-scaled image.

III. Experiments and Discussion

In this section, we first provide a brief description of the three investigated multimodal RS image datasets, which are all publicly available from the websites. Then we describe the compared methods and the necessary implementation details in our experiments. At last, we show the parameter sensitivity analysis on the first dataset and analyze the extensive comparison results on all three datasets.

A. Dataset Description

Fig. 5 gives a whole picture of all three datasets with illustrations of their train and test set split. More specifically,

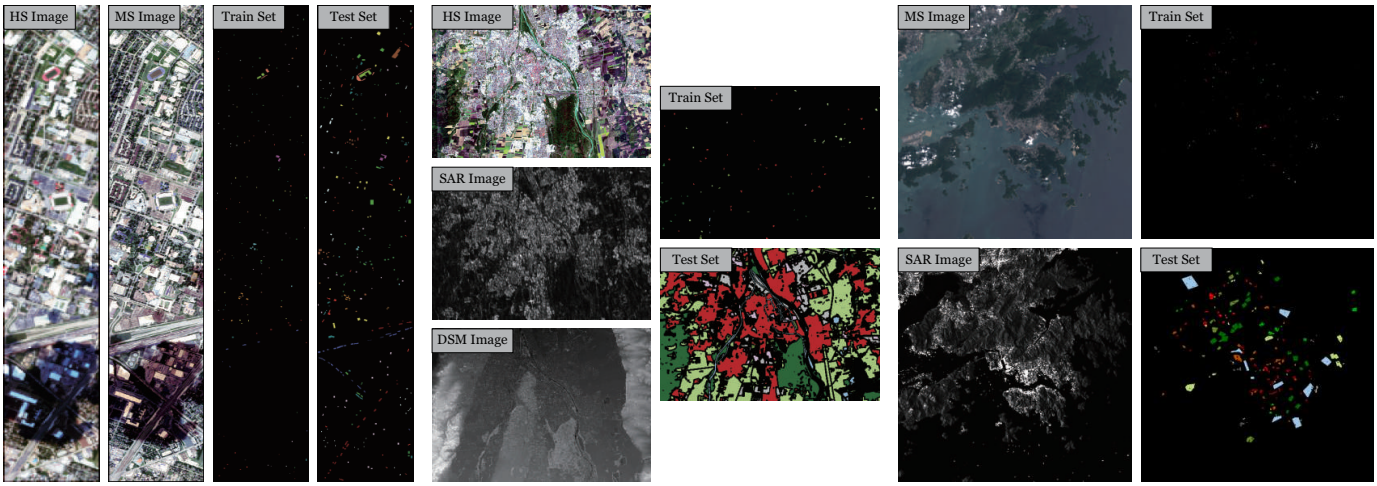


Fig. 5. Illustration of the three investigated multimodal RS image datasets and the train/test set split.

TABLE I
SUMMARY OF THE INVESTIGATED HOUSTON2013 HS-MS DATASET, INCLUDING THE CLASS NAMES AND THE CORRESPONDING NUMBERS OF TRAIN AND TEST SAMPLES.

Class No.	Class Name	Train	Test
1	Healthy Grass	198	1053
2	Stressed Grass	190	1064
3	Synthetic Grass	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot1	192	1041
13	Parking Lot2	184	285
14	Tennis Court	181	247
15	Running Track	187	473
–	Total	2832	12197

1) Houston2013 HS-MS Dataset: The HS data in the first dataset is available from the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest (DFC)[3]. It consists of 349×1905 pixels captured by an ITRES CASI-1500 imaging sensor over the University of Houston campus, Texas, USA, and its surrounding urban area. We follow the experimental settings in [21] to generate a homogeneous HS-MS dataset. Specifically, the spectral downsampling using the spectral response functions of the Sentinel-2 sensor was performed to obtain the MS image with 8 spectral bands and a ground sampling distance (GSD) of 2.5 m. The spatial downsampling using the bilinear interpolation was performed to obtain the HS image with 144 bands, covering the spectral range from 0.38 to 1.05 μm, and a GSD of 10 m. There are 15 ground objects of interest in this scene as the detailed statistics of class-wise ground truth reference are provided in Table I.

[3]http://www.classic.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/

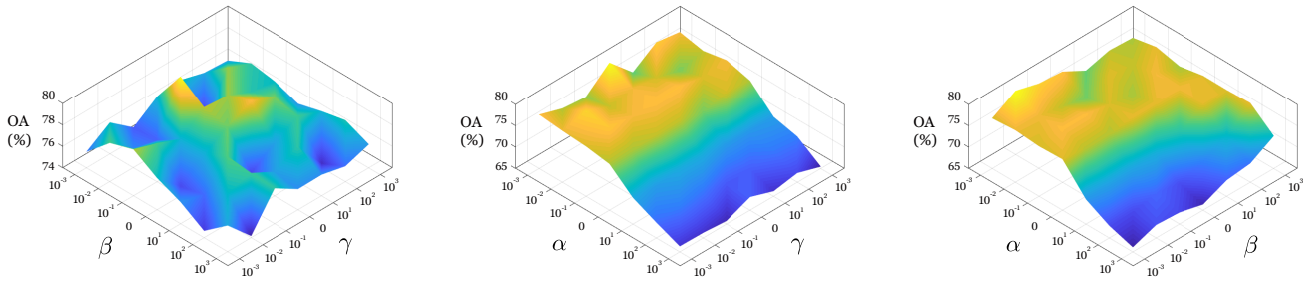
TABLE II
SUMMARY OF THE INVESTIGATED AUGSBURG HS-SAR-LIDAR DATASET, INCLUDING THE CLASS NAMES AND THE CORRESPONDING NUMBERS OF TRAIN AND TEST SAMPLES.

Class No.	Class Name	Train	Test
1	Forest	146	13361
2	Residential Area	264	30065
3	Industrial Area	21	3830
4	Low Plants	248	26609
5	Allotment	52	523
6	Commercial Area	7	1638
7	Water	23	1507
–	Total	761	77533

2) Augsburg HS-SAR-LiDAR Dataset: The second dataset comprises three-modality RS data, i.e., a spaceborne HS image, a spaceborne PolSAR image, and an airborne LiDAR image [43]. They were acquired by the HySpex sensor, the Sentinel-1 sensor, and the DLR 3K camera system [44] respectively in May 2018, covering the same area of the city of Augsburg, Germany. The spatial resolution of them was downsampled to the same 30 m by applying bilinear interpolation, resulting in 332×485 pixels with 180 bands ranging from 0.4 to 2.5 μm HS image, 1-channel LiDAR image, and 4-channel SAR image (VV intensity, VH intensity, the real part and the imaginary part of the off-diagonal element of the 2×2 polarimetric SAR covariance matrix [45]), respectively. The details of the ground truth reference generated by the OpenStreetMap are summarized in Table II.

3) Local Climate Zones (LCZs) Hong Kong MS-SAR Dataset: The MS data of this dataset was provided by the 2017 IEEE GRSS DFC[4]. It consists of 10 spectral bands at a GSD of 100 m. By following Hong *et al.* [26], we select the city of Hong Kong, which comprises MS and SAR images that were acquired by the Sentinel-2 and Sentinel-1 sensors respectively. The SAR image pre-processed in the same way as the Augsburg dataset was further downsampled to the same size as the MS image, i.e., 529×528 pixels. In Table III, we

[4]http://www.classic.grss-ieee.org/2017-ieee-grss-data-fusion-contest/



(a) (b) (c)

Fig. 6. Parameter sensitivity analysis in terms of the OA by fixing (a) α, (b) β, and (c) γ, respectively.

TABLE III
SUMMARY OF THE INVESTIGATED LCZ HONG KONG MS-SAR DATASET, INCLUDING THE CLASS NAMES AND THE CORRESPONDING NUMBERS OF TRAIN AND TEST SAMPLES.

Class No.	Class Name	Train	Test
1	Compact High-Rise	32	599
2	Compact Mid-Rise	32	147
3	Compact Low-Rise	30	296
4	Open High-Rise	34	639
5	Open Mid-Rise	31	95
6	Open Low-Rise	30	90
7	Large Low-Rise	30	107
8	Heavy Industry	31	188
9	Dense Trees	30	1586
10	Scattered Trees	30	510
11	Bush and Scrub	30	661
12	Low Plants	30	955
13	Water	32	2571
–	Total	402	8444

provide details for its ground truth information as well as the train and test set.

B. Implementation Details

1) Training and Testing Set Split: To split datasets without benchmark training and testing sets, like the LCZs Hong Kong MS-SAR dataset used in this work, we design a clear working flow as follows. We first conduct superpixel segmentation on the investigated scene. By setting the number of superpixels relatively large, these segments tend to share the same attribute within each. Then, the training set can be iteratively augmented by randomly selecting superpixels until a preset threshold of size is reached, which is 30 for each class in our case. The remaining ground truth is naturally for test or validation use.

2) Evaluation Metrics: Besides four metrics that are widely used for the quantitative classification assessment, i.e., class-wise accuracy (CA), overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ), we introduce the standard deviation of CAs (σ) in our comparison, hoping to reflect how CAs are spread out around AA, particularly for datasets with an imbalanced class distribution. Also, the run time of major computations are provided to offer more instructive guideline.

3) Compared Methods: To validate the effectiveness and generalization ability of our proposed algorithm, we select the following methods by ensuring diversity for comparison.

TABLE IV
QUANTITATIVE RESULTS OF ABLATION STUDY ON THREE ADOPTED REGULARIZATIONS, I.E., FROBENIUS NORM ON $\mathbf{\Theta}$ (Φ_F), MANIFOLD REGULARIZATIONS ON $\mathbf{\Theta}$ (Φ_{MR}) AND $\mathbf{Z}$ (Ψ_{MR}), RESPECTIVELY.

Method	UCSL-α	UCSL-β	UCSL-γ	UCSL
Φ_F	✗	✓	✓	✓
Φ_{MR}	✓	✗	✓	✓
Ψ_{MR}	✓	✓	✗	✓
OA (%)	77.01	78.01	78.23	79.91
AA (%)	80.10	81.14	80.98	82.29
κ	0.7507	0.7617	0.7641	0.7821

Besides the baseline that directly trains and tests using single-modality data, we reproduce the results of joint dimensionality reduction methods based on principal components analysis [46], unsupervised manifold alignment [33], and supervised manifold alignment [47], which will be abbreviated as JPCA, USMA, and SMA, respectively, basic common subspace learning framework ℓ_2-CoSpace[5] [30] and that with feature selection ℓ_1-CoSpace [48], shared and specific feature learning model S2FL[6] [21], multimodal deep learning framework for remote sensing images MDLRS[7] that adopting the fully connected architecture, and lastly ours using unsupervised and supervised adjacency matrix as UCSL and SCSL. Note that all the compared methods are pixel-based ones without explicit spatial information mining. Most of our experiments were carried out on the Windows platform using MATLAB R2019a on CPU expect for the MDLRS, which was reproduced under the PyTorch framework on GPU.

C. Results and Analysis on Houston2013 HS-MS Dataset

The first dataset has been widely used. Considering that HS data tend to generate more favorable classification results, we conduct cross-modal inference with MS data.

1) Parameter Sensitivity Analysis: Once the graph is constructed, the quality of the learned features and the performance of our model mainly depend on the parameter settings. According to Eqs. (4) to (5), three regularization parameters α, β, and γ need to be analyzed. Taking UCSL model as an example, we roughly tuned these parameters via grid search on interval of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$.

[5]https://github.com/danfenghong/IEEE_TGRS_CoSpace
[6]https://github.com/danfenghong/ISPRS_S2FL
[7]thttps://github.com/danfenghong/IEEE_TGRS_MDLRS

TABLE V
QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON HOUSTON2013 HS-MS DATASET. THE VALUES OF ACCURACY AND TRAIN TIME ARE SHOWN IN PERCENTAGES AND SECONDS, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD.

Method	Baseline	JPCA	USMA	SMA	ℓ_2-CoSpace	ℓ_1-CoSpace	S2FL	MDLRS	UCSL	SCSL
Class 1	82.53	82.43	81.29	79.87	82.53	80.53	**82.81**	82.81	81.96	80.91
Class 2	82.05	83.08	82.71	**83.46**	82.89	82.89	82.61	81.86	81.95	81.58
Class 3	98.61	99.41	99.80	99.60	99.80	99.80	99.80	99.80	99.80	**100.00**
Class 4	91.29	90.81	88.35	83.90	86.08	90.91	87.78	**98.77**	92.99	90.34
Class 5	96.50	96.69	97.44	97.44	**98.67**	98.01	98.48	98.01	98.30	98.30
Class 6	98.60	99.30	99.30	**100.00**	99.30	98.60	95.10	95.10	99.30	95.10
Class 7	71.55	75.75	81.16	71.27	**88.81**	82.84	86.66	88.43	74.25	76.77
Class 8	34.00	35.23	39.41	41.41	41.50	61.25	42.83	**93.22**	56.13	66.67
Class 9	63.64	66.01	63.93	66.67	76.77	68.56	78.66	**78.94**	66.67	69.31
Class 10	43.92	42.57	42.47	59.75	52.90	51.74	50.58	57.34	**84.85**	83.11
Class 11	63.85	67.36	65.65	65.56	**73.24**	69.26	73.15	71.73	72.96	68.88
Class 12	46.40	53.60	49.47	67.24	75.60	69.93	76.56	74.54	70.70	**76.85**
Class 13	52.28	56.84	56.14	55.44	64.91	59.30	65.61	**69.82**	59.30	57.54
Class 14	96.76	97.17	96.36	97.98	97.98	98.38	98.38	97.98	98.79	**99.19**
Class 15	**98.52**	**98.52**	96.41	96.83	97.46	96.19	97.46	97.25	96.41	96.83
OA↑	70.74	72.44	72.21	74.31	78.26	77.81	78.33	79.40	79.91	**80.89**
AA↑	74.70	76.32	75.99	77.76	81.23	80.55	81.10	82.11	82.29	**82.76**
κ↑	0.6857	0.7039	0.6993	0.7212	0.7644	0.7593	0.7580	0.7770	0.7821	**0.7926**
σ↓	22.71	21.72	21.49	18.47	17.66	16.30	17.36	**13.10**	15.09	13.52
Time↓	**0.13**	0.28	0.77	0.62	29.81	30.73	31.08	282.69	29.64	29.57

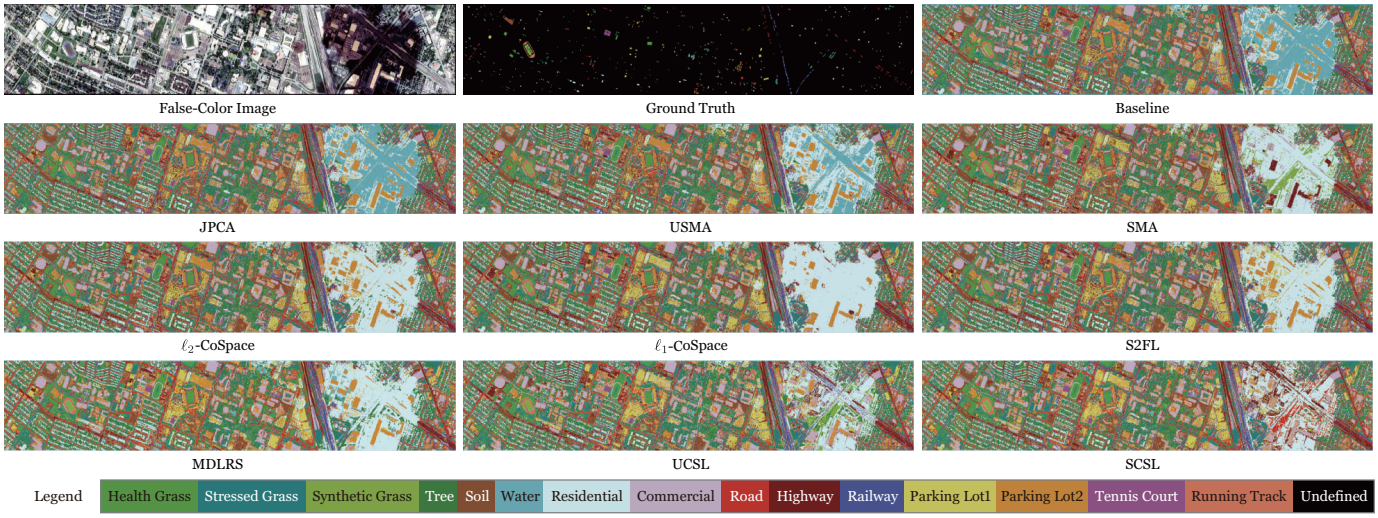


Fig. 7. Illustration of the false-color image, label ground truth, and classification maps obtained by compared methods on Houston2013 HS-MS dataset.

To better visualize the performance sensitivity (in terms of OA) to these parameters, we draw surface plots in Fig. 6 by fixing α, β, γ to the optimal combination as $(10^{-3}, 10^{-2}, 10^{0})$ while varying the other two variables, respectively. From the subfigures, we can basically conclude that too-small values of these parameters, corresponding to inactive regularizations, could even deteriorate the discriminative ability of the original data. In general, although the classification performance can be affected to some extent, it can still be kept at an acceptable level that is greater than 75% with a moderately set α.

2) Ablation Study on Regularizations: By following the parameter searching strategy mentioned above, we then conduct ablation experiments to validate the effectiveness of the three adopted regularizations. Although substituting the first Frobenius norm of projection matrix with more complicate norms like group sparsity term may bring further performance gain, we find it inevitably introduce more variables and iterative time-consuming optimization, thus limiting its feasibility in practical application. Hence we consider our UCSL models's variants by removing these terms weighted by α, β, and γ, respectively. Through the results summarized in Table IV, the first Frobenius norm term affects the performance the most, while the single use of other two manifold regularizations can result in comparably OA and AA that around 78% and 80%, respectively. The best result is achieved by employing all of these three terms.

3) Comparison Experiments: We conduct more comparison experiments to verify the effectiveness of the proposed method. In the following, we report the performance of all compared methods by exhaustive parameter tuning or following suggestions in related literature. The quantitative results are summarized in Table V with the best ones emphasized in bold. From the table, we can first observe that a large proportion of classes in this scene can be well separated, which is probably due to the high quality of airborne imaging products and labeling. The baseline using single MS data can achieve the best accuracy on classifying running track as JPCA does, while the latter further improves the OA by

TABLE VI
QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON AUGSBURG HS-SAR-LIDAR DATASET. THE VALUES OF ACCURACY AND TRAIN TIME ARE SHOWN IN PERCENTAGES AND SECONDS, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD.

Method	Baseline	JPCA	USMA	SMA	ℓ_2-CoSpace	ℓ_1-CoSpace	S2FL	MDLRS	UCSL	SCSL
Class 1	63.81	74.04	77.22	75.61	**77.29**	76.87	76.91	64.12	76.42	76.30
Class 2	72.88	81.79	81.96	82.18	83.81	83.52	84.11	82.79	**84.44**	84.41
Class 3	13.08	16.58	16.66	14.36	14.28	15.74	13.05	**19.53**	15.90	16.27
Class 4	71.57	80.69	82.58	83.90	83.10	82.91	83.39	65.44	84.55	**84.72**
Class 5	11.28	14.91	16.44	24.67	25.05	24.86	26.00	18.55	31.17	**35.18**
Class 6	7.51	10.93	**12.45**	6.84	8.00	7.81	5.01	4.27	7.26	6.29
Class 7	7.56	9.22	12.08	13.14	15.79	16.26	16.99	12.48	**22.96**	22.50
OA↑	64.84	73.50	74.86	74.97	75.69	75.51	75.74	67.12	76.52	**76.57**
AA↑	35.38	41.17	42.77	42.96	43.90	43.99	43.64	38.24	46.10	**46.52**
κ↑	0.5078	0.6240	0.6446	0.6454	0.6548	0.6531	0.6551	0.5264	0.6660	**0.6665**
σ↓	32.02	35.41	35.46	35.65	35.49	35.12	36.00	**31.56**	34.27	34.21
Time↓	**0.12**	0.24	0.39	0.35	7.83	7.97	8.45	134.30	1.62	1.49

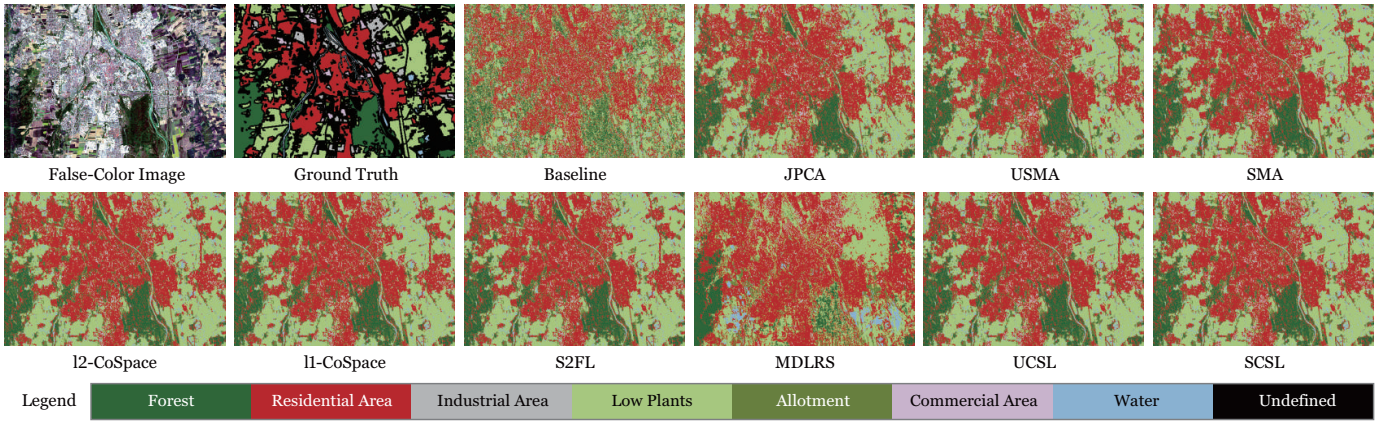


Fig. 8. Illustration of the false-color image, label ground truth, and classification maps obtained by compared methods on Augsburg HS-SAR-LiDAR dataset.

1.7%. An evident performance gain appears from the USMA to SMA (about 2% in OA), showing that supervised graph construction is more suitable in this dataset, which is consistent with our cases. CoSpace-based models largely increase the performance, reaching around 78% OA with 6 best CAs, in which the basic ℓ_2-regularized model and S2FL tend to behave with a strong generalization ability and share similar performances that are second to ours. It is noteworthy that our UCSL and SCSL can not only lead in all classic metrics as OA, AA, and κ, but also obtain the lowest σ, demonstrating their superior and unbiased feature extraction ability. As for the run time, those primitive methods (JPCA, USMA, SMA) are much scalable than more advanced ones. Our models still can be more efficient than CoSpace-families, while MDLRS costs the most because both the size and channel number of training data are much larger than the other two datasets.

Fig. 7 shows more visual evidence for qualitative assessment, especially for those compared methods that are hard to distinguish their differences quantitatively. The first thing that catches our eye is an irregularly shaped area at the bottom part. The classification mapping in this area exhibits various patterns among the investigated methods. According to the legend, the first three methods simultaneously detect the extensive existence of *Water* class, which does not quite match with actuality. While SMA mainly differs from CoSpace-based methods in discovering green spaces beside the road, despite their common failure in recognizing the rectangular buildings that belong to the *Commercial* class. However, the patterns of the road, green area, and buildings in our results are faithful to the real distributions with rich content.

D. Results and Analysis on Augsburg HS-SAR-LiDAR Dataset

In this dataset, we use SAR data for cross-modal testing since LiDAR has only one channel. We summarize the quantitative results and visualized classification maps in Table VI and Fig. 8, respectively. The number of categories is less than that of the former dataset, while using single SAR data for inference is more challenging, as the baseline only gives CAs around 10% for four classes (*Industrial Area*, *Allotment*, *Commercial Area*, and *Water*) according to the table. However, all the other joint feature learning methods can bring significant performance improvements of around 10% in terms of OA. In specific, JPCA and USMA dramatically increase the OA to around 74%, which is basically at the same level as that of SMA. ℓ_2-CoSpace, ℓ_1-CoSpace, and S2FL perform closely to each other in this dataset, as their OAs and AAs can reach the lines of 75% and 43%, respectively. Without any doubt, the best results in OA, AA, and κ are achieved by our SCSL, which is barely ahead of our UCSL. Although the baseline owns the lowest σ, the AA improvements brought by our methods deserve more attention since they indeed raise the performance on hard classes (*Allotment* and *Water*) to a great extent. From the figure, we can also verify the fact that our methods can give better estimates of the three major classes,

TABLE VII
QUANTITATIVE PERFORMANCE COMPARISON WITH THE DIFFERENT ALGORITHMS ON LCZ HONG KONG MS-SAR DATASET. THE VALUES OF ACCURACY AND TRAIN TIME ARE SHOWN IN PERCENTAGES AND SECONDS, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD.

Method	Baseline	JPCA	USMA	SMA	ℓ_2-CoSpace	ℓ_1-CoSpace	S2FL	MDLRS	UCSL	SCSL
Class 1	13.19	22.54	19.37	13.86	18.03	15.19	19.53	15.86	18.86	**24.21**
Class 2	29.25	32.65	31.29	31.29	34.01	34.01	34.69	13.61	**36.73**	34.69
Class 3	20.95	21.62	**22.97**	20.27	19.26	18.58	22.30	8.11	21.62	22.64
Class 4	12.21	11.42	14.08	11.58	13.46	12.99	11.74	**32.86**	14.40	13.30
Class 5	4.21	8.42	**9.47**	6.32	7.37	7.37	7.37	7.37	**9.47**	8.42
Class 6	14.44	14.44	21.11	17.78	15.56	16.67	16.67	3.33	20.00	**22.22**
Class 7	10.28	9.35	10.28	14.02	10.28	9.35	11.21	**16.82**	14.02	12.15
Class 8	20.74	19.68	17.55	22.34	24.47	18.09	23.40	8.51	**26.06**	23.94
Class 9	14.50	13.11	13.81	14.56	15.70	15.83	15.57	0.32	**17.09**	16.08
Class 10	20.78	22.75	19.02	21.37	24.71	21.76	22.16	**26.47**	21.57	22.75
Class 11	26.02	23.90	24.96	26.48	25.87	23.00	25.42	**65.20**	24.05	25.42
Class 12	19.58	20.84	19.48	20.84	**21.57**	20.63	21.26	0.00	20.73	20.42
Class 13	94.05	93.66	94.36	93.93	93.47	94.05	93.54	**96.03**	94.40	94.28
OA↑	40.76	41.18	41.27	41.11	41.85	41.14	41.72	40.66	42.39	**42.54**
AA↑	23.09	24.18	24.44	24.20	24.90	23.65	24.99	22.65	26.08	**26.19**
κ↑	0.3142	0.3188	0.3199	0.3175	0.3259	0.3178	0.3250	0.3137	0.3318	**0.3336**
σ↓	22.35	21.97	21.84	21.95	21.79	22.17	21.77	28.14	21.57	**21.55**
Time↓	**0.10**	0.21	0.30	0.26	3.67	0.54	0.65	43.52	0.25	0.21

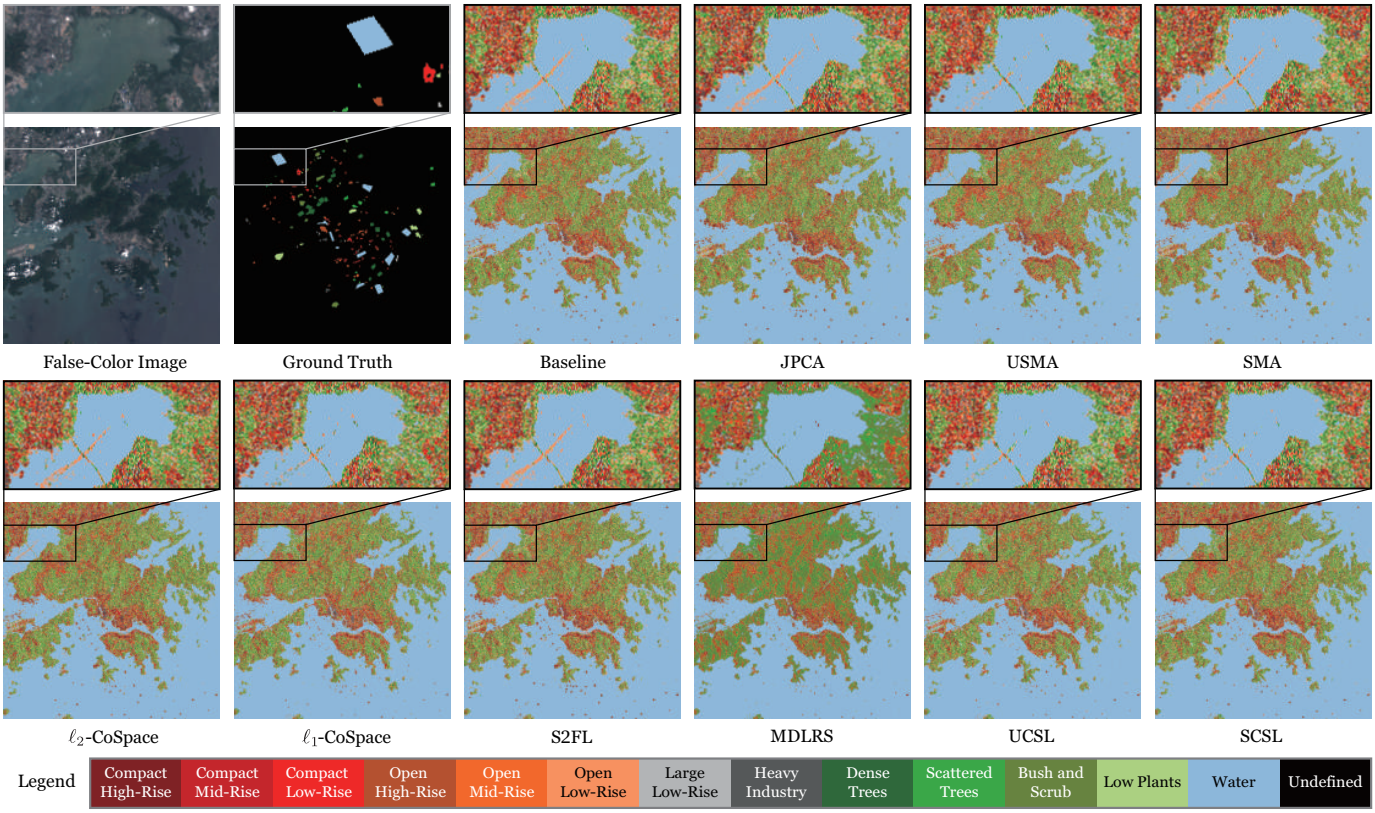


Fig. 9. Illustration of the false-color image, label ground truth, and classification maps obtained by compared methods on LCZ Hong Kong MS-SAR dataset.

i.e., *Forest*, *Residential Area*, and *Low Plants*, which confirms the good generalization ability of our model for heterogeneous cross-modal learning.

E. Results and Analysis on LCZ Hong Kong MS-SAR Dataset

The last dataset is much more challenging. Nearly all classes except for the *Water* class can be well recognized. We also select SAR as the testing modality to investigate how much it can benefit from joint feature learning with MS-SAR training data. Based on the results shown in Table VII and Fig. 9, a similar tendency can be observed in this dataset as that appeared in the former ones. All the OAs of existing cross-modal learning methods can not surpass the line of 42% whereas our SCSL and USCL are able to win the first and second places respectively, with a relatively large margin. It is also worth noting that our UCSL obtains the best CAs in more than half the classes (7 out of 13). Among the competitors, it is the method of ℓ_2-CoSpace behaves the best, as can be shown from the figure. Another interesting observation lies in that there does exist the *Shenzhen Bay Bridge* in the zoomed-in window though we can hardly draw any clues from the false-color image. However, the investigated methods can still detect

it more or less, which also illustrates the importance of pushing the limits of cross-modal learning from partial modality. The run time of our models is much more competitive in the last two datasets, showing its promising potential in dealing with limited training samples under practical situations.

IV. Conclusion and Outlook

In this paper we aim at alleviate the need of costly pixel-wise labeling and propose a fully unsupervised cross-modal feature learning method, UCSL, and its supervised version SCSL, for the task of RS image classification. By implicitly introducing the to-be-learned representations in the latent common subspace, we can directly learn the cross-modal projections from the compactly recombined multimodal data without the need of label target. Under the regularizations on orthogonality and manifold structural priors, the proposed models can be ably transformed into a trace norm optimization problem that can be efficiently resolved by Eigen decomposition. Extensive experiments conducted on parameter sensitivity and comparisons with relevant methods show the stability and constant superiority of proposed methods.

We will put our future research interest along the following two lines. First, there still leaves room to further improve the algorithm efficiency by resorting to mathematical and numerical approximation manners [49]. Second, we will also endeavor to integrate the unsupervised feature learning and deep learning techniques for the RS image classification.

References

[1] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 364–376, 2022.

[2] X. Wu, D. Hong, Z. Huang, and J. Chanussot, "Infrared small object detection using deep interactive u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[3] S. Lorenz, P. Ghamisi, M. Kirsch, R. Jackisch, B. Rasti, and R. Gloaguen, "Feature extraction for hyperspectral mineral domain mapping: A test of conventional and innovative methods," *Remote Sensing of Environment*, vol. 252, p. 112129, 2021.

[4] J. Yao, D. Meng, Q. Zhao, W. Cao, and Z. Xu, "Nonconvex-sparsity and nonlocal-smoothness-based blind hyperspectral unmixing," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2991–3006, 2019.

[5] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, "Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021. DOI: 10.1109/TGRS.2021.3069845.

[6] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.

[7] H. Li, J. Zech, D. Hong, P. Ghamisi, M. Schultz, and A. Zipf, "Leveraging openstreetmap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection," *International Journal of Applied Earth Observation and Geoinformation*, vol. 110, p. 102804, 2022.

[8] Y. Li, S. Martinis, and M. Wieland, "Urban flood mapping with an active self-learning convolutional neural network based on terrasar-x intensity and interferometric coherence," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 178–191, 2019.

[9] N. Yokoya, K. Yamanoi, W. He, G. Baier, B. Adriano, H. Miura, and S. Oishi, "Breaking limits of remote sensing by deep learning from simulated data for flood and debris-flow mapping," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2020.

[10] X. Huang and L. Zhang, "An svm ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 257–272, 2012.

[11] 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 Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 12–23, 2020.

[12] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. ECCV*, pp. 208–224, Springer, 2020.

[13] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (jpsa) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3602–3615, 2021.

[14] L. Fang, C. Wang, S. Li, and J. A. Benediktsson, "Hyperspectral image classification via multiple-feature-based adaptive sparse representation," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 7, pp. 1646–1657, 2017.

[15] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6547–6565, 2017.

[16] A. Merentitis, C. Debes, and R. Heremans, "Ensemble learning in hyperspectral image classification: Toward selecting a favorable bias-variance tradeoff," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1089–1102, 2014.

[17] P. Duan, X. Kang, S. Li, P. Ghamisi, and J. A. Benediktsson, "Fusion of multiple edge-preserving operations for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 10336–10349, 2019.

[18] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7726–7738, 2020.

[19] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multi-view satellite imagery," *Remote sensing of environment*, vol. 196, pp. 56–75, 2017.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[21] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 178, pp. 68–80, 2021.

[22] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and lidar data using morphological features," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 3, pp. 552–556, 2015.

[23] N. Yokoya, P. Ghamisi, and J. Xia, "Multimodal, multitemporal, and multisource global data fusion for local climate zones classification based on ensemble learning," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1197–1200, IEEE, 2017.

[24] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and lidar data using coupled cnns," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4939–4950, 2020.

[25] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[26] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 4340–4354, May 2021.

[27] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal gans: Toward crossmodal hyperspectral–multispectral image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5103–5113, 2021.

[28] D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls, "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 12, pp. 7708–7720, 2014.

[29] D. Tuia and G. Camps-Valls, "Kernel manifold alignment for domain adaptation," *PloS one*, vol. 11, no. 2, p. e0148655, 2016.

[30] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4349–4359, 2019.

[31] 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 Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 193–205, 2019.
[32] D. Hong, X. Wu, J. Yao, and X. Zhu, "Beyond pixels: Learning from multimodal hyperspectral superpixels for land cover classification," *Science China Technological Sciences*, vol. 65, no. 4, pp. 802–808, 2022.
[33] D. Liao, Y. Qian, J. Zhou, and Y. Y. Tang, "A manifold alignment approach for hyperspectral image visualization with natural color," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3151–3162, 2016.
[34] A. Pournemat, P. Adibi, and J. Chanussot, "Semisupervised charting for spectral multimodal manifold learning and alignment," *Pattern Recognition*, vol. 111, p. 107645, 2021.
[35] J. Yao, X. Cao, D. Hong, X. Wu, D. Meng, J. Chanussot, and Z. Xu, "Semi-active convolutional neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022. DOI: 10.1109/TGRS.2022.3206208.
[36] S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in *International Joint Conference on Artificial Intelligence*, Citeseer, 2009.
[37] F. R. Chung, *Spectral graph theory*, vol. 92. American Mathematical Society, 1997.
[38] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Joint and progressive subspace analysis (jpsa) with spatial–spectral manifold alignment for semisupervised hyperspectral dimensionality reduction," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3602–3615, 2020.
[39] J. Liu, Y. Chen, J. Zhang, and Z. Xu, "Enhancing low-rank subspace clustering by manifold regularization," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 4022–4030, 2014.
[40] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," *Advances in Neural Information Processing Systems*, vol. 24, 2011.
[41] X. He and P. Niyogi, "Locality preserving projections," *Advances in Neural Information Processing Systems*, vol. 16, 2003.
[42] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.
[43] J. Hu, R. Liu, D. Hong, A. Camero, J. Yao, M. Schneider, F. Kurz, K. Segl, and X. X. Zhu, "Mdas: A new multimodal benchmark dataset for remote sensing," *Earth System Science Data Discussions*, pp. 1–26, 2022.
[44] F. Kurz, D. Rosenbaum, J. Leitloff, O. Meynberg, and P. Reinartz, "Real time camera system for disaster and traffic monitoring," in *International Conference on Sensors and Models in Photogrammetry and Remote Sensing*, pp. 1–6, 2011.
[45] Y. Yamaguchi, T. Moriyama, M. Ishido, and H. Yamada, "Four-component scattering model for polarimetric sar image decomposition," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 8, pp. 1699–1706, 2005.
[46] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
[47] C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *International Joint Conference on Artificial Intelligence*, 2011.
[48] D. Hong, J. Chanussot, N. Yokoya, J. Kang, and X. X. Zhu, "Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 8, pp. 1470–1474, 2020.
[49] J. W. Demmel, *Applied numerical linear algebra*. SIAM, 1997.

Jing Yao (M'22) received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 2021.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. From 2019 to 2020, he was a visiting student at Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, and at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He put his recent research interests on hyperspectral and multimodal remote sensing image analysis, mainly including optimization and deep learning-based methods for image processing and interpretation tasks.

He was the recipient of the Jose Bioucas Dias Award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2021. He also serves as a Guest Editor of Remote Sensing.

Danfeng Hong (S'16–M'19–SM'21) received the M.Sc. degree (summa cum laude) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, the Dr. -Ing degree (summa cum laude) from the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019.

He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences (CAS). Before joining CAS, he has been a Research Scientist and led a Spectral Vision Working Group at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He was also an Adjunct Scientist at GIPSA-lab, Grenoble INP, CNRS, Univ. Grenoble Alpes, Grenoble, France. His research interests include artificial intelligence, remote sensing big data analysis, multimodal interpretation, and their applications in Earth Vision.

Dr. Hong is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing (TGRS), an Editorial Board Member of Remote Sensing, an Editorial Advisory Board Member of the ISPRS Journal of Photogrammetry and Remote Sensing. He was a recipient of the Best Reviewer Award of the IEEE TGRS in 2021 and 2022, and the Best Reviewer Award of the IEEE JSTARS in 2022, the Jose Bioucas Dias Award for recognizing the outstanding paper at WHISPERS in 2021, the Remote Sensing Young Investigator Award in 2022, the IEEE GRSS Early Career Award in 2022, and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) in 2022.

Haipeng Wang received the Ph.D. degree from Naval Aviation University, Yantai, China, in 2012. He is currently a Professor at Naval Aviation University, Yantai, China.

His research interests include the general area of intelligent perception and fusion, and big data technology and application. He also serves as a Reviewer for several distinguished journals, including IET Radar, Sonar & Navigation and IEEE Transactions on Aerospace and Electronic Systems. He also serves as a Guest Editor of IEEE JSTARS.

Hao Liu is currently a professor in Wuhan digital engineer institute, CSSC youth talent support program, member of the youth committee of CICC, member of the information fusion branch of CSAA.

His main research interests are information fusion and situation cognition. He won one 1st prize and one 2nd prize of the group's scientific and technological progress, one 1st prize of Shandong provincial scientific and technological progress, one 2nd prize of military scientific and technological progress, 10 authorized national invention patents, and published 9 SCI, EI and Chinese core papers.

Jocelyn Chanussot (M'04–SM'04–F'12) 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 of signal and image processing. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence. He has been a visiting scholar at Stanford University (USA), KTH (Sweden), and NUS (Singapore). Since 2013, he is an Adjunct Professor at the University of Iceland. In 2015-2017, he was a visiting professor at the University of California, Los Angeles (UCLA). He holds the AXA chair in remote sensing and is an Adjunct professor at the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing.

Dr. Chanussot is the founding President of the IEEE Geoscience and Remote Sensing French chapter (2007-2010) 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 (2017-2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair (2009-2011) and Cochair of the GRS Data Fusion Technical Committee (2005-2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006-2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing 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 (2011-2015) and an Associate Editor for IEEE Transactions on Image Processing. In 2014 he served as a Guest Editor for the IEEE Signal Processing Magazine. He is a Fellow of the IEEE, a member of the Institut Universitaire de France (2012-2017), and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters).