

1                    Learnable Manifold Alignment (LeMA) : A  
2                    Semi-supervised Cross-modality Learning Framework for  
3                    Land Cover and Land Use Classification

4                    Danfeng Hong<sup>a,b</sup>, Naoto Yokoya<sup>c</sup>, Nan Ge<sup>a</sup>, Jocelyn Chanussot<sup>d</sup>, Xiao Xiang Zhu<sup>a,b,\*</sup>

5                    <sup>a</sup>Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany

6                    <sup>b</sup>Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich,  
7                    Germany

8                    <sup>c</sup>Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo, Japan

9                    <sup>d</sup>Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France

---

10                    **Abstract**

In this paper, we aim at tackling a general but interesting cross-modality feature learning question in remote sensing community — *can a limited amount of highly-discriminative (e.g., hyperspectral) training data improve the performance of a classification task using a large amount of poorly-discriminative (e.g., multispectral) data?* Traditional semi-supervised manifold alignment methods do not perform sufficiently well for such problems, since the hyperspectral data is very expensive to be largely collected in a trade-off between time and efficiency, compared to the multispectral data. To this end, we propose a novel semi-supervised cross-modality learning framework, called learnable manifold alignment (LeMA). LeMA learns a joint graph structure directly from the data instead of using a given fixed graph defined by a Gaussian kernel function. With the learned graph, we can further capture the data distribution by graph-based label propagation, which enables finding a more accurate decision boundary. Additionally, an optimization strategy based on the alternating direction method of multipliers (ADMM) is designed to solve the proposed model. Extensive experiments on two hyperspectral-multispectral datasets demonstrate the superiority and effectiveness of the proposed method in comparison with several state-of-the-art methods.

11                    *Keywords:*

12                    Cross-modality, graph learning, hyperspectral, manifold alignment, multispectral,  
13                    remote sensing, semi-supervised learning.

---

\*Corresponding author  
Preprint submitted to ISPRS

---

14 **1. Introduction**

15 Multispectral (MS) imagery has been receiving an increasing interest in the urban  
16 area (e.g. a large-scale land-cover mapping [1] [2], building localization [3]), agri-  
17 culture [4], and mineral products [5], as operational optical broadband (multispectral)  
18 satellites (e.g. Sentinel-2 and Landsat-8 [6]) enable the multispectral imagery openly  
19 available on a global scale. In general, a reliable classifier needs to be trained on a  
20 large amount of labeled, discriminative, and high-quality samples. Unfortunately, la-  
21 beling data, in particular large-scale data, is very gruelling and time-consuming. A  
22 natural alternative way to this issue is to consider tons of unlabeled data, yielding a  
23 semi-supervised learning. On the other hand, MS data fails to spectrally discriminate  
24 similar classes due to its broad spectral bandwidth. A simple way is to improve the data  
25 quality by fusing high-discriminative hyperspectral (HS) data [6]. Although such data  
26 is expensive to collect, we may be able to expect a small amount of such data available.  
27 The aforementioned two points motivate us to raise a question related to transfer learn-  
28 ing and cross-modality learning: *Can a limited amount of HS training data partially*  
29 *overlapping MS data improve the performance of a classification task using a large*  
30 *coverage of MS testing data?*

31 Over the past decades, land-cover and land-use classification tasks of optical re-  
32 mote sensing imagery has received increasing attention in the unsupervised [7] [8] [9],  
33 supervised [10] [11], and semi-supervised ways [12] [13]. To our best knowledge,  
34 the classifying ability in unsupervised learning (or dimensionality reduction) still re-  
35 mains limited, due to missing label information. By fully considering the variability of  
36 intra-class and inter-class from labels, supervised learning is able to perform the clas-  
37 sification task better. In reality, a limited number of labeled samples usually hinders  
38 the trained classier towards a high classification performance, further leading to a pos-  
39 sible failure in some challenging classification or transferring tasks owing to the lack  
40 of generalization and representability. Alternatively, semi-supervised learning draws  
41 into plenty of unlabeled data in learning process. This is capable of better capturing  
42 the distribution of different categories in order to find an accurate decision boundary.

43 On the other hand, considerable work related to transfer learning (TL) or domain  
44 adaptation (DA) has been successfully developed and applied in the remote sensing  
45 community [14, 15, 16, 17, 18, 19]. According to the different transferred objects, the  
46 TL or DA approaches can be roughly categorized into three groups, including parame-  
47 ter adaptation, instance-based transfer, and feature-based alignment or representation.

48 The seminal work dealing with parameter adaptation was presented in [20] and  
49 [21], aiming at transferring an existing classifier (or parameters) trained or learned  
50 from the source domain to the target domain. Differently, the instance-based trans-  
51 ferring technique transfers the knowledge by reweighting [22] or resampling [23] the  
52 samples of the source domain to those of the target domain. A similar idea based on  
53 active learning [24] has also been proposed to address this issue, by selecting the most  
54 informative samples in the target domain to replace with those samples of the source  
55 domain that do not match the data distribution of the target domain [25].

56 For the final group of feature-based alignment or representation, manifold align-  
57 ment (MA) is one of the most popular semi-supervised learning framework [26] that  
58 facilitates transfer learning. MA has been successfully applied to various tasks in  
59 remote sensing community, e.g. classification [27], data visualization [28], multi-  
60 modality data analysis [13], etc. The key idea of MA can be generalized as learning a  
61 common (or shared) subspace where different data can be aligned to learn a joint fea-  
62 ture representation. Generally, existing MA methods can be approximately categorized  
63 into unsupervised, supervised, and semi-supervised approaches. The unsupervised ap-  
64 proach usually fails to align multimodal data sufficiently well, as their corresponding  
65 low-dimensional embeddings may be quite diverse [29]. In the supervised case, only  
66 aligning the limited number of training samples to learn a common subspace leads to  
67 weak transferability. While preserving a joint manifold structure created by both la-  
68 beled and unlabeled data, semi-supervised alignment allows different data sources to  
69 be better transformed into the common subspace [30].

70 Although the joint manifold structure used in conventional semi-supervised MA  
71 approaches can relate features or instances, poor connections between the common  
72 subspace and label information still hinder the low-dimensional feature representa-  
73 tion from being more discriminative. More importantly, in most graph-based semi-

74 supervised learning algorithms (e.g. graph-based label propagation (GLP) [31], semi-  
75 supervised manifold alignment (S-SMA [13]) [30]), the topology of unlabeled samples  
76 is merely given by a fixed Gaussian kernel function, which is computed in the original  
77 space rather than in the common space. This makes it difficult to adaptively transfer  
78 unlabeled samples into the learned common subspace, particularly when applied to  
79 multimodal data due to different numbers of dimensions. To address these issues, we  
80 propose a learnable manifold alignment (LeMA) by a data-driven graph learning di-  
81 rectly from a common subspace so as to make the multimodal data comparable as well  
82 as improve the explainability of the learned common subspace, which further results  
83 in a better transferability. More specifically, our contributions can be summarized as  
84 follows:

- 85 • We propose a novel semi-supervised cross-modality learning framework called  
86 learnable manifold alignment (LeMA) for a large-scale land-cover classification  
87 task. One spectrally-poor MS and one spectrally rich HS data are considered as  
88 two different modalities and applied for this task, where the spatial extent of the  
89 former is a true superset of that of the latter.
- 90 • Unlike jointly feature learning in which the model is both trained and tested from  
91 completed HS-MS correspondences, LeMA learns an aligned feature subspace  
92 from the labeled HS-MS correspondences and partially unlabeled MS data, and  
93 allows to identify out-of-samples using either MS data or HS data; Such the  
94 learnt subspace is a good fit for our case of cross-modality learning <sup>1</sup>.
- 95 • Instead of directly computing graph structure with a Gaussian kernel function, a  
96 data-driven graph learning method is exploited behind LeMA in order to strengthen  
97 the abilities of transferring and generalization;
- 98 • An optimization framework based on the alternating direction method of multi-  
99 pliers (ADMM) is designed to fast and effectively solve the proposed model.

---

<sup>1</sup>In contrast to multi-modal learning (bi-modality for example), cross-modal learning trains on single modality and tests on bi-modality, or *vice versa* (train on bi-modality and test on single modality).

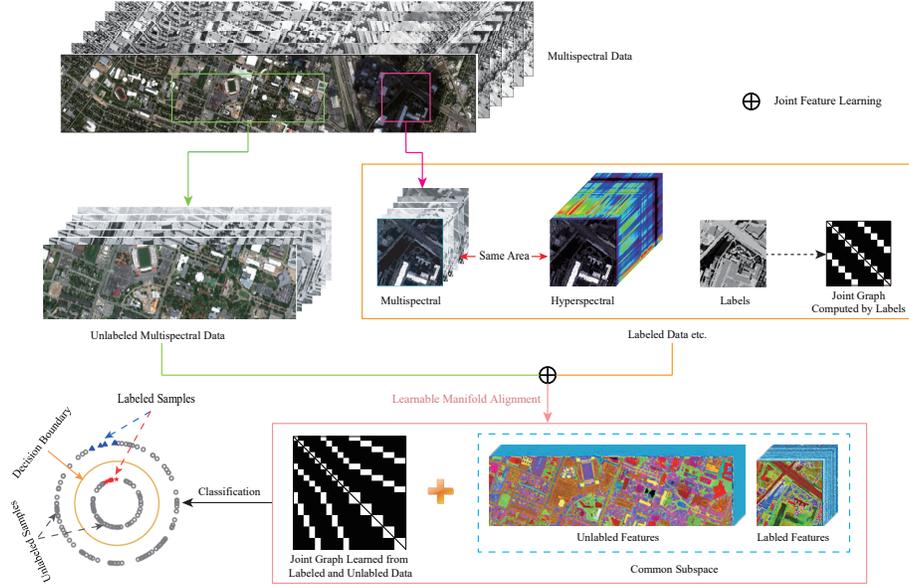


Figure 1: An illustration of the proposed LeMA method.

100 The remainder of this paper is organized as follows. Section II elaborates on our  
 101 motivation and proposes the methodology for the LeMA and the corresponding opti-  
 102 mization algorithm. In Section III, we present the experimental results on two HS-MS  
 103 datasets over the areas of the University of Houston and Chikusei, respectively, and  
 104 meanwhile discuss the qualitative and quantitative analysis. Section IV concludes with  
 105 a summary.

106 **2. Learnable Manifold Alignment (LeMA)**

107 In this section, a cross-modality learning problem is firstly casted and the moti-  
 108 vation is stated in the following. Accordingly, we formulate the methodology of our  
 109 proposed and then elucidate an ADMM-based optimization algorithm to solve it.

110 *2.1. Problem Statement and Motivation*

111 For many high-level data analysis tasks in remote sensing community, such as  
 112 land-cover classification, data collection plays an important role, since informa-  
 113 tion-rich training samples enable us to easily find an optimal decision boundary.

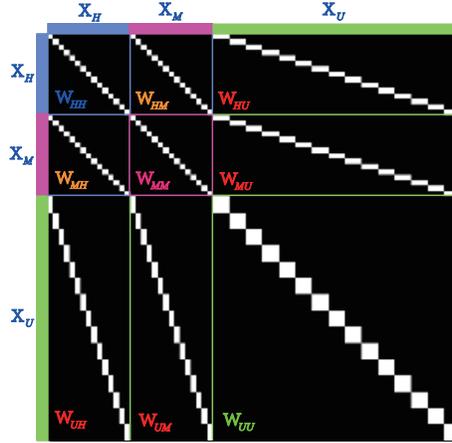


Figure 2: An example for the joint adjacency matrix  $\widetilde{W}$ .

114 There is, however, a typical bottleneck in collecting a large amount of labeled and  
 115 discriminative data. Despite the MS data available at a global scale from the satel-  
 116 lites of Sentinel-2 and Landsat-8, the identification and discrimination of materials are  
 117 unattainable at an accuracy level by MS data, resulting from its poorly spectral infor-  
 118 mation. On the contrary, HS data is characterized by rich spectral information, but only  
 119 can be acquired in very small areas, due to the limitations of imaging sensors. This is-  
 120 sue naturally guides us to jointly utilize the HS and MS bi-modal data, specifically  
 121 leading to the following interesting and challenging question *can a limited number of*  
 122 *HS training data contribute to the classification task of a large-scale MS data?*

123 A feasible solution to the issue can be unfolded to two parts: 1) *cross-modality*  
 124 *learning*: learning a common subspace where the features are expected to absorb the  
 125 different properties from the HS-MS modalities and meanwhile the HS and MS data  
 126 can be transferred each other; 2) *semi-supervised learning*: Embedding massive unlabeled  
 127 MS samples which are relatively in large quantities and easy to be collected, so  
 128 as to learn a more discriminative feature representation. Fig. 1 illustrates the workflow  
 129 of LeMA.

130 *2.2. Problem Formulation*

131 To effectively model the aforementioned issue, we intend to develop a joint learning  
 132 framework which better learns a discriminative common subspace from high-quality  
 133 HS data and low-quality MS data. Intuitively, such a common subspace can be shaped  
 134 by selectively absorbing the benefits of both high-quality data with more details and  
 135 low-quality data with more structural information. Therefore, following a popular joint  
 136 learning framework [32], we formulate the common subspace learning problem as

$$\min_{\mathbf{P}, \Theta} \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P}\Theta\tilde{\mathbf{X}}\|_{\mathbb{F}}^2 + \frac{\alpha}{2} \|\mathbf{P}\|_{\mathbb{F}}^2 + \frac{\beta}{2} \text{tr}(\mathbf{E}\mathbf{L}\mathbf{E}^{\text{T}}) \text{ s.t. } \mathbf{E} = \Theta\tilde{\mathbf{X}}, \Theta\Theta^{\text{T}} = \mathbf{I}, \quad (1)$$

137 where  $\tilde{\mathbf{Y}} = [\mathbf{Y}, \mathbf{Y}] \in \mathbb{R}^{d \times 2N}$  and  $\mathbf{Y} \in \mathbb{R}^{d \times N}$  is the label matrix represented by  
 138 one-hot encoding,  $\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}_H & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_M \end{bmatrix} \in \mathbb{R}^{(d_H+d_M) \times 2N}$  and  $\mathbf{X}_H$  and  $\mathbf{X}_M$  stand re-  
 139 spectively for the data from hyperspectral and multispectral domains,  $\Theta = [\Theta_H, \Theta_M]$   
 140 and  $\mathbf{P}$  are respectively the common subspace projection and the linear projection to  
 141 bridge the common subspace and label information.  $\mathbf{L} = \mathbf{D} - \mathbf{W} \in \mathbb{R}^{2N \times 2N}$  stands  
 142 for a joint Laplacian matrix,  $\mathbf{W}$  is an adjacency matrix and  $\mathbf{D}_{ii} = \sum_{i \neq j} \mathbf{W}_{i,j}$ .  $\mathbf{W}$  is  
 143 generally used to measure the similarity between samples. With the orthogonal con-  
 144 straint ( $\Theta\Theta^{\text{T}} = \mathbf{I}$ ), the global optimal solutions with respect to the variables  $\Theta$  and  $\mathbf{P}$   
 145 can be theoretically guaranteed [32].

146 The first term of Eq. (1) is a fidelity term, and the regularization term  $\frac{\alpha}{2} \|\mathbf{P}\|_{\mathbb{F}}^2$   
 147 parameterized by  $\alpha$  aims to achieve a reliable generalization of the proposed model.  
 148 The third term acts as supervised manifold alignment (SMA) [26]. We refer to the  
 149 proposed framework for joint common subspace learning as CoSpace.

150 To further exploit the information of unlabeled samples, we extend the CoSpace  
 151 in Eq. (1) to LeMA by learning a joint Laplacian matrix, which can be formulated as  
 152 follows with extra constraints related to necessary conditions of  $\tilde{\mathbf{L}}$ :

$$\min_{\mathbf{P}, \Theta, \tilde{\mathbf{L}}} \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P}\Theta\tilde{\mathbf{X}}\|_{\mathbb{F}}^2 + \frac{\alpha}{2} \|\mathbf{P}\|_{\mathbb{F}}^2 + \frac{\beta}{2} \text{tr}(\mathbf{H}\tilde{\mathbf{L}}\mathbf{H}^{\text{T}}) \quad (2)$$

s.t.  $\mathbf{H} = \Theta\tilde{\mathbf{X}}'$ ,  $\Theta\Theta^{\text{T}} = \mathbf{I}$ ,  $\tilde{\mathbf{L}} = \tilde{\mathbf{L}}^{\text{T}}$ ,  $\tilde{\mathbf{L}}_{i,j,i \neq j} \preceq 0$ ,  $\tilde{\mathbf{L}}_{i,j,i=j} \succeq 0$ ,  $\text{tr}(\tilde{\mathbf{L}}) = s$ ,

---

**Algorithm 1: Learnable Manifold Alignment (LeMA)**


---

**Input:**  $\tilde{\mathbf{Y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{X}}', \tilde{\mathbf{L}}, \alpha, \beta, \maxIter$ .  
**Output:**  $\mathbf{P}, \Theta, \tilde{\mathbf{L}}$   
1  $t = 1, \zeta = 1e - 4$ ;  
2 **Initializing**  $\mathbf{P}$  and  $\Theta$   
3 **while** *not converged* or  $t > \maxIter$  **do**  
4     Fix other variables to update  $\mathbf{P}$  by Eq. (6)  
5     Fix other variables to update  $\Theta$  by **Algorithm 2**  
6     Fix other variables to update  $\tilde{\mathbf{L}}$  by equivalently optimizing  $\tilde{\mathbf{W}}$  in a distributed fashion:  
7         1. update  $\tilde{\mathbf{W}}_{HU}$  by **Algorithm 3**;  
8         2. update  $\tilde{\mathbf{W}}_{MU}$  by **Algorithm 3**;  
9         3. align  $\tilde{\mathbf{W}}_{HU}$  and  $\tilde{\mathbf{W}}_{MU}$  by  $\max(\tilde{\mathbf{W}}_{HU}, \tilde{\mathbf{W}}_{MU})$ ;  
10        4. update  $\tilde{\mathbf{W}}_{UU}$  by **Algorithm 4**  
11        5. compute  $\tilde{\mathbf{L}} = \tilde{\mathbf{D}} - \tilde{\mathbf{W}}, \tilde{\mathbf{D}}_{ii} = \sum_{i \neq j} \tilde{\mathbf{W}}_{ij}$   
12     Compute the objective function value  $E^{t+1}$  and check the convergence condition: **if**  
13          $\left| \frac{E^{t+1} - E^t}{E^t} \right| < \zeta$  **then**  
14             Stop iteration;  
15         **else**  
16              $t \leftarrow t + 1$ ;  
17         **end**  
18     **end**

---

153 where  $\tilde{\mathbf{X}}' = \begin{bmatrix} \mathbf{X}_H & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_M & \mathbf{X}_U \end{bmatrix} \in \mathbb{R}^{(d_H + d_M) \times (2N + N_U)}, \tilde{\mathbf{L}} \in \mathbb{R}^{(2N + N_U) \times (2N + N_U)}$ ,  
154 and  $\mathbf{X}_U \in \mathbb{R}^{d_M \times N_U}$  represents the unlabeled MS samples and  $s > 0$  controls the  
155 scale. Note that a feasible and effective approach to choose the unlabeled data with  
156 respect to the variable  $\tilde{\mathbf{X}}'$  is to group total samples besides the training samples into  
157 some landmarks (cluster centers). These landmarks are used as the unlabeled data,  
158 which can fully take into account the available information and meanwhile effectively  
159 reduce the computational cost. Due to the use of clustering technique in unlabeled  
160 data, we experimentally and empirically set the ratio of labeled and unlabeled data to  
161 approximately be 1:1.

162 The model in Eq. (2) can be simplified by optimizing the adjacency matrix ( $\tilde{\mathbf{W}}$ )  
163 instead of directly solving a hard optimization problem of  $\tilde{\mathbf{L}}$ , then we have

$$\text{tr}(\mathbf{H}\tilde{\mathbf{L}}\mathbf{H}^T) = \frac{1}{2} \text{tr}(\tilde{\mathbf{W}}\mathbf{Z}) = \frac{1}{2} \|\tilde{\mathbf{W}} \odot \mathbf{Z}\|_{1,1}, \quad (3)$$

164 where  $\tilde{\mathbf{W}} \in \mathbb{R}^{(2N + N_U) \times (2N + N_U)}, \mathbf{Z} \in \mathbb{R}^{(2N + N_U) \times (2N + N_U)}$  is defined as a *pairwise*  
165 *Euclidean distance matrix* :  $\mathbf{Z}_{i,j} = \|\mathbf{H}_i - \mathbf{H}_j\|^2$ .  $\odot$  denotes the Schur-Hadamard

---

**Algorithm 2:** Solving the subproblem for  $\Theta$ 


---

**Input:**  $\tilde{\mathbf{Y}}, \mathbf{P}, \mathbf{J}, \tilde{\mathbf{X}}, \tilde{\mathbf{X}}', \tilde{\mathbf{L}}, \beta, \text{maxIter}$ .  
**Output:**  $\Theta$ .  
1 **Initialization:**  $\Theta = \mathbf{0}, \mathbf{G} = \mathbf{0}, \Lambda_1 = \mathbf{0}, \Lambda_2 = \mathbf{0}, \mu = 10^{-3}, \mu_{\max} = 10^6, \rho = 1.5, \varepsilon = 10^{-6}$ ,  
 $t = 1$ .  
2 **while** *not converged* or  $t > \text{maxIter}$  **do**  
3     Fix other variables to update  $\mathbf{J}$  by  $\mathbf{J} = (\mathbf{P}^T \mathbf{P} + \mu \mathbf{I})^{-1} (\mathbf{P}^T \tilde{\mathbf{Y}} + \mu \Theta \tilde{\mathbf{X}} - \Lambda_1)$ .  
4     Fix other variables to update  $\Theta$  by  

$$\Theta = (\mu \mathbf{J} \tilde{\mathbf{X}}^T + \Lambda_1 \tilde{\mathbf{X}}^T + \mu \mathbf{G} + \Lambda_2) \times (\mu \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T + \mu \mathbf{I} + \beta \tilde{\mathbf{X}}' \tilde{\mathbf{L}} \tilde{\mathbf{X}}'^T)^{-1}$$
.  
5     Fix other variables to update  $\mathbf{G}$  by  

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\Theta - \Lambda_2 / \mu), \quad \mathbf{G} = \mathbf{U} \mathbf{I}_{n \times m} \mathbf{V}$$
.  
6     Update Lagrange multipliers by  

$$\Lambda_1 \leftarrow \Lambda_1 + \mu(\mathbf{J} - \Theta \tilde{\mathbf{X}}), \quad \Lambda_2 \leftarrow \Lambda_2 + \mu(\mathbf{G} - \Theta)$$
.  
7     Update penalty parameter by  $\mu = \min(\rho \mu, \mu_{\max})$ .  
8     Check the convergence conditions: **if**  $\|\mathbf{J} - \Theta \tilde{\mathbf{X}}\|_F < \varepsilon$  **and**  $\|\mathbf{G} - \Theta\|_F < \varepsilon$  **then**  
9         | Stop iteration;  
10     **else**  
11         |  $t \leftarrow t + 1$ ;  
12     **end**  
13 **end**

---

166 (termwise) product.

167 Using Eq. (3), we can equivalently convert the optimization problem of smooth  
168 manifold in (2) to that of graph sparsity

$$\begin{aligned}
& \min_{\mathbf{P}, \Theta, \tilde{\mathbf{W}}} \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P} \Theta \tilde{\mathbf{X}}\|_F^2 + \frac{\alpha}{2} \|\mathbf{P}\|_F^2 + \frac{\beta}{4} \|\tilde{\mathbf{W}} \odot \mathbf{Z}\|_{1,1} \\
& \text{s.t. } \mathbf{H} = \Theta \tilde{\mathbf{X}}', \quad \Theta \Theta^T = \mathbf{I}, \quad \tilde{\mathbf{W}} = \tilde{\mathbf{W}}^T, \quad \tilde{\mathbf{W}}_{i,j} \succeq 0, \quad \|\tilde{\mathbf{W}}\|_{1,1} = s,
\end{aligned} \tag{4}$$

169 where  $\|\tilde{\mathbf{W}} \odot \mathbf{Z}\|_{1,1}$  can be interpreted as a *weighted  $\ell_1$ -norm of  $\tilde{\mathbf{W}}$  which enforces*  
170 *weighted sparsity*.

171 We further elaborate the relationship between the proposed LeMA model and our  
172 motivation in an easy-understanding way. In general, we aim at finding a common  
173 subspace by learning a pair of projections ( $\Theta_M$  and  $\Theta_H$ ) corresponding to two kinds  
174 of different modalities (e.g., MS and HS), respectively. In order to effectively improve  
175 the discriminative ability of the learned subspace, we make a connection between the  
176 subspace and label information by jointly estimating the regression coefficient  $\mathbf{P}$  and  
177 common projections  $\Theta$ , as formulated in Eq. (1). What's more, the alignment behavior  
178 of different modalities can be represented by  $\mathbf{W}$ 's connectivity, that is, if the  $i^{\text{th}}$  sample

---

**Algorithm 3:** Solving the subproblem for  $\widetilde{\mathbf{W}}_{HU(MU)}$ 


---

**Input:**  $\mathbf{Z}_{H(M)}$ ,  $\mathbf{Z}_U$ ,  $\widetilde{\mathbf{W}}$ ,  $\beta$ ,  $maxIter$ .

**Output:**  $\widetilde{\mathbf{W}}$ .

```

1 Initialization:  $\mathbf{M} = \widetilde{\mathbf{W}}$ ,  $\mathbf{S} = \mathbf{U} = \mathbf{K} = \mathbf{0}$ ,  $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4 = \mathbf{0}$ ,  $\mu = 10^{-2}$ ,
    $\mu_{max} = 10^6$ ,  $\rho = 2$ ,  $\varepsilon = 10^{-6}$ ,  $t = 1$ .
2 Compute Z:  $\mathbf{Z}_{i,j} = \|\mathbf{Z}_{H(M)}^i - \mathbf{Z}_U^j\|_F^2$ .
3 while not converged or  $t > maxIter$  do
4   Fix other variables to update  $\widetilde{\mathbf{W}}$  by
     
$$\widetilde{\mathbf{W}} = (\mathbf{M} + \mathbf{S} + \mathbf{U} + \mathbf{K} + \Lambda_1 + \Lambda_2 + \Lambda_3 + \Lambda_4) / (4\mu)$$

5   Fix other variables to update  $\mathbf{U}$  by  $\mathbf{U} = \max(\widetilde{\mathbf{W}} - \Lambda_1 / \mu, 0)$ .
6   Fix other variables to update  $\mathbf{M}$  by
     
$$\mathbf{M} = \max(\|\widetilde{\mathbf{W}} - \Lambda_2 / \mu\|_{1,1} - (\beta \mathbf{Z} / 4\mu), 0) \odot \text{sign}(\widetilde{\mathbf{W}} - \Lambda_2 / \mu)$$

7   Fix other variables to update  $\mathbf{S}$  by  $\mathbf{S} = \text{prox}(\widetilde{\mathbf{W}} - \Lambda_3 / \mu)$ .
8   Fix other variables to update  $\mathbf{K}$  by  $\mathbf{K} = \min(\widetilde{\mathbf{W}} - \Lambda_4 / \mu, 1 / N_k)$ .
9   Update Lagrange multipliers by
     
$$\Lambda_1 = \Lambda_1 + \mu(\mathbf{U} - \widetilde{\mathbf{W}}), \quad \Lambda_2 = \Lambda_2 + \mu(\mathbf{M} - \widetilde{\mathbf{W}}),$$

     
$$\Lambda_3 = \Lambda_3 + \mu(\mathbf{S} - \widetilde{\mathbf{W}}), \quad \Lambda_4 = \Lambda_4 + \mu(\mathbf{K} - \widetilde{\mathbf{W}})$$

10  Update penalty parameter by  $\mu = \min(\rho\mu, \mu_{max})$ . Check the convergence conditions: if
      $\|\mathbf{U} - \widetilde{\mathbf{W}}\|_F < \varepsilon$  and  $\|\mathbf{M} - \widetilde{\mathbf{W}}\|_F < \varepsilon$  and  $\|\mathbf{S} - \widetilde{\mathbf{W}}\|_F < \varepsilon$  and  $\|\mathbf{K} - \widetilde{\mathbf{W}}\|_F < \varepsilon$  and
      $\|\widetilde{\mathbf{W}}^{t+1} - \widetilde{\mathbf{W}}^t\|_F < \varepsilon$  then
11    Stop iteration;
12  else
13     $t \leftarrow t + 1$ ;
14  end
15 end

```

---

179  $\mathbf{X}_i$  and the  $j^{th}$  sample  $\mathbf{X}_j$  are connected ( $\mathbf{W}_{i,j} = 1$ ), and then the two samples belong  
180 to the same class; *vice versa*. Besides, we construct an extra adjacency matrix based on  
181 those unlabeled samples in order to globally capture the data distribution. The matrix  
182 is usually obtained by a Gaussian kernel function (semi-supervised CoSpace) and also  
183 can be learned from the data (LeMA as formulated in Eq. (2)).

### 184 2.3. Model Optimization

185 Considering the complexity of the non-convex problem (4), an iterative alternating  
186 optimization strategy is adopted to solve the convex subproblems of each variable  $\mathbf{P}$ ,  
187  $\Theta$ , and  $\mathbf{W}$ . An implementation of LeMA is given in **Algorithm 1**.

188 *Optimization with respect to P:* This is a typical least-squares problem with Tikhonov

189 regularization, which can be formulated as

$$\min_{\mathbf{P}} \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P}\Theta\tilde{\mathbf{X}}\|_{\mathbb{F}}^2 + \frac{\alpha}{2} \|\mathbf{P}\|_{\mathbb{F}}^2, \quad (5)$$

190 which has a closed-form solution

$$\mathbf{P} = (\tilde{\mathbf{Y}}\mathbf{E}^{\text{T}})(\mathbf{E}\mathbf{E}^{\text{T}} + \alpha\mathbf{I})^{-1}, \quad (6)$$

191 where  $\mathbf{E} = \Theta\tilde{\mathbf{X}}$ .

192 *Optimization with respect to  $\Theta$* : the optimization problem for  $\Theta$  can be formulated  
193 as

$$\min_{\Theta} \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P}\Theta\tilde{\mathbf{X}}\|_{\mathbb{F}}^2 + \frac{\beta}{2} \text{tr}(\mathbf{H}\tilde{\mathbf{L}}\mathbf{H}^{\text{T}}) \text{ s.t. } \mathbf{H} = \Theta\tilde{\mathbf{X}}', \quad \Theta\Theta^{\text{T}} = \mathbf{I}. \quad (7)$$

194 In order to solve (7) effectively with ADMM, we consider an equivalent form by intro-  
195 ducing auxiliary variables  $\mathbf{J}$  and  $\mathbf{G}$  to replace  $\Theta\tilde{\mathbf{X}}$  and  $\Theta$ , respectively.

$$\begin{aligned} \min_{\Theta, \mathbf{J}, \mathbf{G}} \quad & \frac{1}{2} \|\tilde{\mathbf{Y}} - \mathbf{P}\mathbf{J}\|_{\mathbb{F}}^2 + \frac{\beta}{2} \text{tr}(\Theta\tilde{\mathbf{X}}'\tilde{\mathbf{L}}(\Theta\tilde{\mathbf{X}}')^{\text{T}}) \\ \text{s.t.} \quad & \mathbf{J} = \Theta\tilde{\mathbf{X}}, \quad \mathbf{G} = \Theta, \quad \mathbf{G}\mathbf{G}^{\text{T}} = \mathbf{I}. \end{aligned} \quad (8)$$

196 **Algorithm 2** lists the more detailed procedures for solving the problem (8).

197 *Optimization with respect to  $\tilde{\mathbf{W}}$* :  $\tilde{\mathbf{W}}$  is a joint adjacency matrix and consists mainly  
198 of nine parts as shown in Fig. 2. Among the nine parts,  $\tilde{\mathbf{W}}_{HH}$ ,  $\tilde{\mathbf{W}}_{HM}$ ,  $\tilde{\mathbf{W}}_{MH}$  and  
199  $\tilde{\mathbf{W}}_{MM}$  can be directly inferred from label information in the form of the LDA-like  
200 graph [33]:

$$\tilde{\mathbf{W}}_{i,j} = \begin{cases} 1/N_k, & \text{if } \mathbf{X}_i \text{ and } \mathbf{X}_j \text{ belong to the } k\text{-th class;} \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

201 Given the symmetry of  $\tilde{\mathbf{W}}$ , (i.e.,  $\tilde{\mathbf{W}}_{HM} = \tilde{\mathbf{W}}_{MH}$ ,  $\tilde{\mathbf{W}}_{MU} = \tilde{\mathbf{W}}_{UM}$ , and  $\tilde{\mathbf{W}}_{MU} =$   
202  $\tilde{\mathbf{W}}_{UM}$ ), we only need to update three of out nine parts, namely  $\tilde{\mathbf{W}}_{HU}$ ,  $\tilde{\mathbf{W}}_{MU}$ , and

---

**Algorithm 4:** Solving the subproblem for  $\widetilde{\mathbf{W}}_{UU}$ 


---

**Input:**  $\mathbf{Z}_U, \widetilde{\mathbf{W}}, \gamma, \text{maxIter}$ .  
**Output:**  $\widetilde{\mathbf{W}}$ .

- 1 **Initialization:**  $\mathbf{M} = \widetilde{\mathbf{W}}, \mathbf{U} = \mathbf{V} = \mathbf{S} = \mathbf{K} = \mathbf{T} = \mathbf{0}$ ,  
 $\Lambda_1 = \Lambda_2 = \Lambda_3 = \Lambda_4 = \Lambda_5 = \Lambda_6 = \Lambda_7 = \mathbf{0}, \mu = 10^{-2}, \mu_{\max} = 10^6, \rho = 2, \varepsilon = 10^{-6}$ ,  
 $t = 1$ .
- 2 **Compute  $\mathbf{Z}$ :**  $\mathbf{Z}_{i,j} = \|\mathbf{Z}_U^i - \mathbf{Z}_U^j\|_{\mathbb{F}}^2$ .
- 3 **while** not converged or  $t > \text{maxIter}$  **do**
- 4     Fix other variables to update  $\widetilde{\mathbf{W}}$  by  

$$\widetilde{\mathbf{W}} = (\mathbf{V} + \mathbf{U}^T + \mathbf{M} + \mathbf{S} + \mathbf{K} + \mathbf{T} + \Lambda_1 + \Lambda_2^T + \Lambda_3 + \Lambda_4 + \Lambda_5 + \Lambda_7)/(6\mu).$$
- 5     Fix other variables to update  $\mathbf{U}$  by  $\mathbf{U} = (\widetilde{\mathbf{W}}^T + \mathbf{V} - (\Lambda_1 + \Lambda_6))/(2\mu)$ .
- 6     Fix other variables to update  $\mathbf{V}$  by  $\mathbf{V} = (\widetilde{\mathbf{W}} + \mathbf{U} - (\Lambda_2 + \Lambda_6))/(2\mu)$ .
- 7     Fix other variables to update  $\mathbf{M}$  by  

$$\mathbf{M} = \max(\|\widetilde{\mathbf{W}} - \Lambda_3/\mu\|_{1,1} - \gamma\mathbf{Z}/(4\mu), 0) \odot \text{sign}(\widetilde{\mathbf{W}} - \Lambda_3/\mu).$$
- 8     Fix other variables to update  $\mathbf{S}$  by  $\mathbf{S} = \text{prox}(\widetilde{\mathbf{W}} - \Lambda_4/\mu)$ .
- 9     Fix other variables to update  $\mathbf{K}$  by  $\mathbf{K} = \max(\widetilde{\mathbf{W}} - \Lambda_5/\mu, 0)$ .
- 10     Fix other variables to update  $\mathbf{T}$  by  $\mathbf{T} = \min(\widetilde{\mathbf{W}} - \Lambda_7/\mu, 1/N_k)$ .
- 11     Update Lagrange multipliers by  

$$\begin{aligned} \Lambda_1 &= \Lambda_1 + \mu(\mathbf{U} - \widetilde{\mathbf{W}}^T), & \Lambda_2 &= \Lambda_2 + \mu(\mathbf{V} - \widetilde{\mathbf{W}}), \\ \Lambda_3 &= \Lambda_3 + \mu(\mathbf{M} - \widetilde{\mathbf{W}}), & \Lambda_4 &= \Lambda_4 + \mu(\mathbf{S} - \widetilde{\mathbf{W}}), \\ \Lambda_5 &= \Lambda_5 + \mu(\mathbf{K} - \widetilde{\mathbf{W}}), & \Lambda_6 &= \Lambda_6 + \mu(\mathbf{U} - \mathbf{V}), \\ \Lambda_7 &= \Lambda_7 + \mu(\mathbf{T} - \widetilde{\mathbf{W}}). \end{aligned}$$
- 12     Update penalty parameter by  $\mu = \min(\rho\mu, \mu_{\max})$ .
- 13     Check the convergence conditions: **if**  $\|\mathbf{U} - \widetilde{\mathbf{W}}^T\|_{\mathbb{F}} < \varepsilon$  **and**  $\|\mathbf{V} - \widetilde{\mathbf{W}}\|_{\mathbb{F}} < \varepsilon$  **and**  
 $\|\mathbf{M} - \widetilde{\mathbf{W}}\|_{\mathbb{F}} < \varepsilon$  **and**  $\|\mathbf{S} - \widetilde{\mathbf{W}}\|_{\mathbb{F}} < \varepsilon$  **and**  $\|\mathbf{K} - \widetilde{\mathbf{W}}\|_{\mathbb{F}} < \varepsilon$  **and**  $\|\mathbf{U} - \mathbf{V}\|_{\mathbb{F}} < \varepsilon$  **and**  
 $\|\mathbf{T} - \widetilde{\mathbf{W}}\|_{\mathbb{F}} < \varepsilon$  **and**  $\|\widetilde{\mathbf{W}}^{t+1} - \widetilde{\mathbf{W}}^t\|_{\mathbb{F}} < \varepsilon$  **then**  
    | Stop iteration;  
- 14     **else**  
- 15     |  $t \leftarrow t + 1$ ;  
- 16     **end**  
- 17     **end**  
- 18 **end**

---

203  $\widetilde{\mathbf{W}}_{UU}$ . The optimization problems of  $\widetilde{\mathbf{W}}_{HU}$  and  $\widetilde{\mathbf{W}}_{MU}$  can be formulated by

$$\min_{\widetilde{\mathbf{W}}_{HU(MU)}} \frac{\beta}{4} \|\widetilde{\mathbf{W}} \odot \mathbf{Z}\|_{1,1} \text{ s.t. } 1/N_k \succeq \widetilde{\mathbf{W}}_{i,j} \succeq 0, \|\widetilde{\mathbf{W}}\|_{1,1} = s, \quad (10)$$

204 which can be solved by ADMM. More details can be found in **Algorithm 3**, where  
 205  $\mathbf{Z}_{H(M)}$  and  $\mathbf{Z}_U$  represent respectively the subspace features of  $\mathbf{X}_{H(M)}$  and  $\mathbf{X}_U$ ,  $\text{prox}$   
 206 stands for the proximal operator for  $\|\widetilde{\mathbf{W}}\|_{1,1} = s$  [34]. We technically add the con-  
 207 straint  $\widetilde{\mathbf{W}}_{i,j} \preceq 1/N_k$  in order to share the same unit level with LDA-like graph.

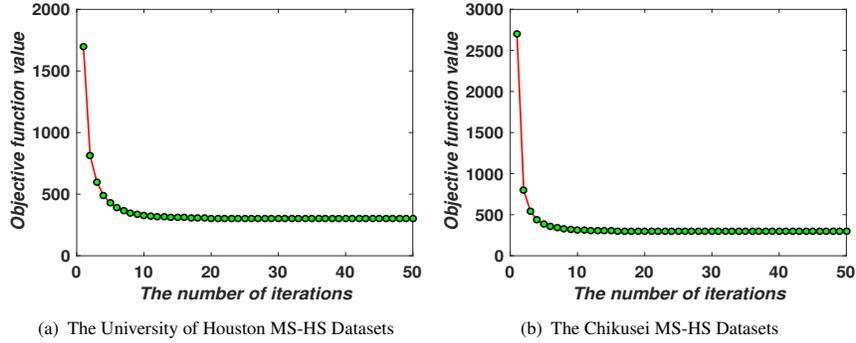


Figure 3: Convergence analysis of LeMA are experimentally performed on the two MS-HS datasets.

208 For  $\widetilde{\mathbf{W}}_{UU}$ , the objective function can be written as

$$\min_{\widetilde{\mathbf{W}}_{UU}} \frac{\beta}{4} \|\widetilde{\mathbf{W}} \odot \mathbf{Z}\|_{1,1} \text{ s.t. } \widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}^T, 1/N_k \succeq \widetilde{\mathbf{W}}_{i,j} \succeq 0, \|\widetilde{\mathbf{W}}\|_{1,1} = s, \quad (11)$$

209 which can be effectively solved using **Algorithm 4**.

210 Finally, we repeat these optimization procedures until a stopping criterion is satis-  
211 fied.

#### 212 2.4. Convergence Analysis

213 The alternative alternating strategy used in **Algorithm 1** is nothing but a block  
214 coordinate descent (BCD), which has been theoretically supported to converge to a  
215 stationary point as long as each subproblem in Eq. (4) is exactly minimized [35]. As  
216 observed, these subproblems with respect to the variables  $\mathbf{P}$ ,  $\Theta$  and  $\widetilde{\mathbf{W}}$  are strongly  
217 convex, and hence each independent task can ideally find an unique minimum when the  
218 Lagrangian parameter is updated within finitely iterative steps [36]. Besides, ADMM  
219 used in each subproblem optimization is actually generalized to *inexact* Augmented  
220 Lagrange Multiplier (ALM) [37], whose convergence has been well studied when the  
221 number of block is less than three [38] (e.g. **Algorithm 2**). Although there is still not a  
222 *generally and strictly* theoretical proof in multi-blocks case, yet the convergence anal-  
223 ysis for some common cases such as our **Algorithm 3** and **Algorithm 4** has been well  
224 conducted in [39][40][41][42]. We also experimentally record the objective function

225 values in each iteration to draw the convergence curves of LeMA on two used HS-MS  
226 datasets (see Fig. 3).

### 227 **3. Experiments**

228 In this section, we quantitatively and qualitatively evaluate the performance of the  
229 proposed method on two simulated HS-MS datasets (University of Houston and Chiku-  
230 sei) and a real multispectral-lidar and hyperspectral dataset provided by 2018 IEEE  
231 GRSS data fusion contest (DFC2018), by the form of classification using two com-  
232 monly used and high-performance classifiers, namely linear support vector machines  
233 (LSVM), and canonical correlation forest (CCF) [43]. Three indices: overall accuracy  
234 (OA), average accuracy (AA), kappa coefficient ( $\kappa$ ), are calculated to quantitatively  
235 assess the classification performance. Moreover, we compare the performance of the  
236 proposed LeMA and several other state-of-art algorithms, i.e. GLP [31], SMA, S-  
237 SMA [29], CoSpace and Semi-supervised CoSpace (S-CoSpace). The original MS  
238 data is used as a baseline. SMA constructs an LDA-like joint graph using label in-  
239 formation. Besides label information, S-SMA method also uses unlabeled samples to  
240 generate the joint graph by computing the similarity based on Euclidean distance. The  
241 same strategy of graph construction is adopted for CoSpace and S-CoSpace.

#### 242 *3.1. The Simulated MS-HS Datasets over the University of Houston*

##### 243 *3.1.1. Data Description*

244 The HS data in the simulated *Houston MS-HS datasets* was acquired by the ITRES-  
245 CASI-1500 sensor with the size of  $349 \times 1905$  at a ground sampling distance (GSD) of  
246 2.5m over the University of Houston campus and its neighboring urban areas. This data  
247 was provided for the 2013 IEEE GRSS data fusion contest, with 144 bands covering  
248 the wavelength range from 364nm to 1046nm. Spectral simulation is performed to  
249 generate the MS image by degrading the HS image in the spectral domain using the  
250 MS spectral response functions (SRFs) of Sentinel-2 as filters (for more details refer to  
251 [6]). The MS data we used is generated with dimensions of  $349 \times 1905 \times 10$ .

Table 1: The number of training and testing samples for the two used MS-HS datasets.

Class No.	Houston MS-HS dataset			Chikusei MS-HS dataset		
	Class Name	Training	Testing	Class Name	Training	Testing
1	Healthy Grass	537	699	Water	301	858
2	Stressed Grass	61	1154	Bare Soil (School)	992	1867
3	Synthetic Grass	340	357	Bare Soil (Farmland)	455	4397
4	Tree	209	1035	Natural Plants	150	4272
5	Soil	74	1168	Weeds in Farmland	928	1108
6	Water	22	303	Forest	486	11904
7	Residential	52	1203	Grass	989	5526
8	Commercial	320	924	Rice Field (Grown)	813	8816
9	Road	76	1149	Rice Field (First Stage)	667	1268
10	Highway	279	948	Row Crops	377	5961
11	Railway	33	1185	Plastic House	165	475
12	Parking Lot1	329	904	Manmade (Non-dark)	170	568
13	Parking Lot2	20	449	Manmade (Dark)	1291	6373
14	Tennis Court	266	162	Manmade (Blue)	111	431
15	Running Track	279	381	Manmade (Red)	35	187
16	/	/	/	Manmade Grass	21	1019
17	/	/	/	Asphalt	384	417
	Total	2897	12021	Total	8335	55447

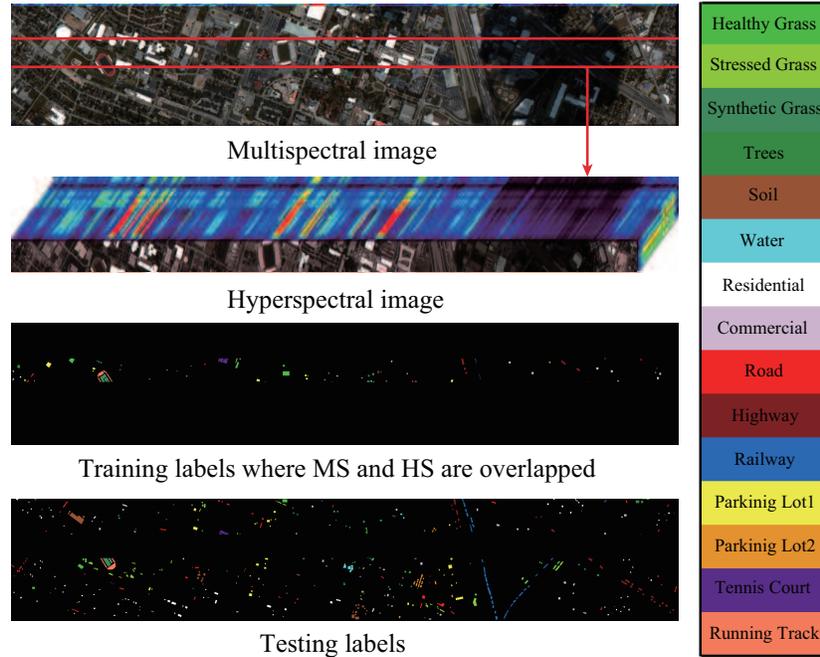


Figure 4: The multispectral image and its corresponding hyperspectral image that partially covers the same area, as well as training and testing labels, for University of Houston dataset.

### 252 3.1.2. Experimental Setup

253 To meet our problem setting, a HS image partially overlapping MS image and a  
254 whole MS image are used in our experiments, and meanwhile the corresponding train-  
255 ing and test samples can be re-assigned, as shown in Fig. 4. In detail, since the total  
256 labels are available, we seek out a region where all kinds of classes are involved. The  
257 labels in the region are selected as the training set and the rest are seen as the test set,  
258 as shown in Fig. 4 and specifically quantified in Table 1.

259 The parameters of the different methods are determined by a 10-fold cross-validation  
260 on the training data. More specifically, we tune the parameters of the different algo-  
261 rithms to maximize their performances, e.g. dimension ( $d$ ), penalty parameters ( $\alpha, \beta$ ),  
262 etc. The dimension ( $d$ ) is a common parameter for all compared algorithms, and it can  
263 be determined covering the range from 10 to 50 at an interval of 10. For the number  
264 of nearest neighbors ( $k$ ) and the standard deviation of Gaussian kernel function ( $\sigma$ )  
265 in artificially computing the adjacency matrix ( $\mathbf{W}$ ) of GLP, SMA, and S-SMA, we se-  
266 lect them in the range of  $\{10, 20, \dots, 50\}$  and  $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$ , respectively,  
267 Similarly to CoSpace, S-CoSpace and LeMA, we set the two regularization parameters  
268 ( $\alpha, \beta$ ) ranging from  $\{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$ .

### 269 3.1.3. Results and Analysis

270 Fig.5 shows the classification maps of compared algorithms using LSVM and CCF  
271 classifiers, while Table 2 lists the specific quantitative assessment results with optimal  
272 parameters obtained by 10-fold cross-validation.

273 Overall, the methods based on manifold alignment outperform baseline and GLP  
274 using the different classifiers. This means that the limited amount of HS data can guide  
275 the corresponding MS data towards better discriminative feature representations. More  
276 specifically when compared with S-SMA, SMA yields a relatively poor performance  
277 since it only considers the correspondences of MS-HS labeled data. This indicates that  
278 reasonably embedding unlabeled samples into the manifold alignment framework can  
279 effectively help us capture the real data distribution, and thereby obtain more accurate  
280 decision boundaries. Unfortunately, these approaches only attempt to align different  
281 data in a common subspace, but they hardly take the connections between the common

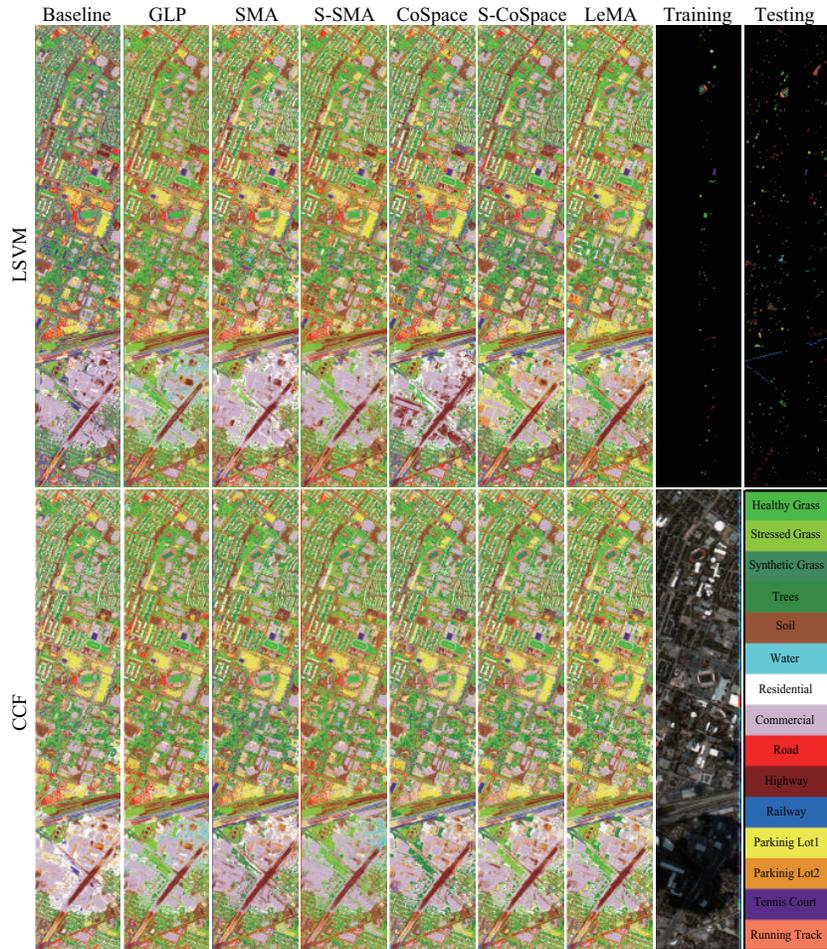


Figure 5: Classification maps of the different algorithms obtained using two kinds of classifiers on the University of Houston dataset.

282 subspace and label information into account<sup>2</sup>, which leads to a lack of discriminative  
 283 ability. With regards to this, our proposed joint learning framework “CoSpace” and  
 284 its semi-supervised version “S-CoSpace” achieve the desired results on the the given  
 285 MS-HS datasets.

286 By fully considering the connectivity of the common subspace, label information,  
 287 and unlabeled information encoded by the learned graph structure, the performance

<sup>2</sup>The connectivity in manifold alignment is not strictly equivalent to the similarity of the two samples.

Table 2: Quantitative performance comparison with the different algorithms on the University of Houston data. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	$d$		$(k, \sigma, d)$		$d$		$(k, \sigma, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$	
	10		(10, 1, 10)		30		(10, 0.1, 30)		(0.01, 0.01, 30)		(0.1, 0.01, 30)		(0.01, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	62.12	68.21	64.71	70.01	68.01	69.59	69.29	70.10	69.38	72.17	70.41	73.75	73.42	<b>76.35</b>
AA	65.97	70.47	68.18	72.18	70.50	71.02	72.00	72.88	71.69	73.56	73.12	75.61	74.76	<b>77.18</b>
$\kappa$	0.5889	0.6543	0.6164	0.6728	0.6520	0.6695	0.6659	0.6754	0.6672	0.6975	0.6784	0.7146	0.7110	<b>0.7428</b>
Class1	76.39	67.95	77.83	77.97	75.25	68.53	74.25	73.53	75.54	69.96	<b>91.85</b>	87.98	89.56	85.84
Class2	80.59	78.08	93.85	<b>98.01</b>	97.57	77.9	97.57	93.67	73.74	77.99	90.12	91.59	93.67	93.85
Class3	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>
Class4	85.51	92.27	89.66	96.62	94.78	98.74	95.85	98.55	98.74	98.26	92.75	97.29	97.49	<b>99.61</b>
Class5	99.06	99.4	99.49	<b>99.66</b>	98.97	99.14	99.32	99.4	99.4	99.4	99.4	<b>99.66</b>	99.49	99.57
Class6	86.14	86.14	96.37	99.01	86.47	70.96	<b>99.67</b>	99.67	85.48	85.15	<b>99.67</b>	96.70	86.47	86.47
Class7	50.62	63.76	48.63	64.01	72.32	77.14	72.15	69.66	73.98	80.05	75.06	80.96	83.21	<b>88.03</b>
Class8	56.49	56.06	56.60	59.85	62.01	62.23	<b>64.61</b>	63.85	63.53	62.01	55.84	60.39	62.77	62.01
Class9	56.22	70.58	69.63	69.02	49.96	61.27	50.57	45.00	59.79	64.93	65.8	<b>71.54</b>	64.49	61.88
Class10	45.36	45.25	45.46	49.89	58.12	52.32	58.33	63.61	<b>64.14</b>	57.70	58.97	51.79	60.97	53.59
Class11	27.43	43.88	22.45	38.65	28.86	36.46	36.46	34.77	36.54	47.26	35.78	38.65	41.27	<b>49.96</b>
Class12	31.64	56.08	31.75	37.83	35.84	62.50	34.18	55.2	46.79	62.72	34.29	58.52	45.02	<b>76.88</b>
Class13	0.00	0.67	0.00	1.11	0.00	0.00	0.00	0.45	0.00	0.45	0.00	0.89	0.00	<b>1.78</b>
Class14	97.53	98.77	94.44	92.59	<b>100.00</b>	<b>100.00</b>	99.38	98.15	<b>100.00</b>	99.38	99.38	<b>100.00</b>	99.38	<b>100.00</b>
Class15	96.59	98.16	96.59	<b>98.43</b>	97.38	98.16	97.64	97.64	97.64	98.16	97.90	98.16	97.64	98.16

288 of LeMA is much more superior to that of any other methods as can be observed in  
 289 Table 2. This demonstrates that LeMA is likely to learn a more discriminative feature  
 290 representation and to find a better decision boundary.

291 As observed from Fig. 4 and Table 2, the training samples are relatively a few and  
 292 meanwhile the distribution between different classes is extremely unbalanced. While  
 293 training the classifier, more attentions are paid on those classes with large-size sam-  
 294 ples, and some small-scale classes possibly play less and even nothing. For this reason,  
 295 we propose to consider those large-scale unlabeled data, achieving a semi-supervised  
 296 learning. Using this strategy, the semi-supervised methods, i.e. GLP, S-SMA, S-  
 297 CoSpace, obviously perform better than baseline and their supervised ones (SMA and  
 298 CoSpace). Moreover, we can see from Table 2 that there is a significant improvement of  
 299 classification performance in some classes (e.g. *Stressed Grass*, *Water*) after account-  
 300 ing for unlabeled samples, particularly between SMA and S-SMA as well as CoSpace  
 301 and S-CoSpace. However, these aforementioned semi-supervised methods carry out  
 302 the label propagation on a given graph manually computed by gaussian kernel function,  
 303 limiting the adaptiveness and discriminability of the algorithms. LeMA can adaptively  
 304 learn a data-driven graph structure where the labels tend to spread more smoothly,  
 305 which can result in a more effective material identification for those challenging classes  
 306 (few training samples), such as *Trees*, *Residential*, *Railway*, *Parking Lot1*. In addi-

307 tion, we can also observe an easily overlooked phenomenon that the LeMA’s ability  
308 in identifying certain classes still remains limited, such as *Parking Lot2*(only 1.78%)  
309 and *Railway* (49.96%). *Parking Lot2* is basically classified to *Commercial* and *Park-*  
310 *ing Lot1*, while *Railway* is largely identified as *Road* and *Commercial*. This might be  
311 explained by the limited number of training samples as well as fairly similar spectral  
312 properties between several classes.

### 313 3.2. The Simulated MS-HS Datasets over Chikusei

#### 314 3.2.1. Data Description

315 Similarly to Houston data, the MS data with dimensions of  $2517 \times 2335 \times 10$  at a  
316 GSD of 2.5 m was simulated by the HS data acquired by the Headwall’s Hyperspec-  
317 VNIR-C sensor over Chikusei area, Ibaraki, Japan. It consists of 128 bands in the  
318 spectral range from 363nm to 1018nm with the 10nm spectral resolution. The dataset  
319 has been made available to the scientific research [44].

#### 320 3.2.2. Experimental Setup

321 Fig. 6 shows the corresponding MS and partial HS images as well as selected train-  
322 ing labels and test labels. Again, the overlapped region between MS and HS, which  
323 should include all the classes listed in Table 1, is chosen based on the given ground  
324 truth [44]. Additionally, the parameters configuration for all algorithms can be adap-  
325 tively completed by a 10-fold cross-validation on the training set, which is more gen-  
326 eralized to different datasets. Regarding how to run the cross-validation for parameters  
327 setting, please refer to section 3.1.2 for more details.

#### 328 3.2.3. Results and Analysis

329 We assess the classification performance of the different algorithms for the Chiku-  
330 sei MS-HS data both quantitatively and visually, as shown in Fig.7 and Table 3.

331 Similarly to the University of Houston MS-HS data, there is a basically consistent  
332 trend for the different algorithms in the Chikusei MS-HS data. On the whole, the  
333 original MS data (baseline) fails to identify some specific materials such as *Plastic*  
334 *House*, *Manmade (Dark)*, *Rice Field (Grown)*, *Bare Soil (Farmland)*, and *Forest*, due to

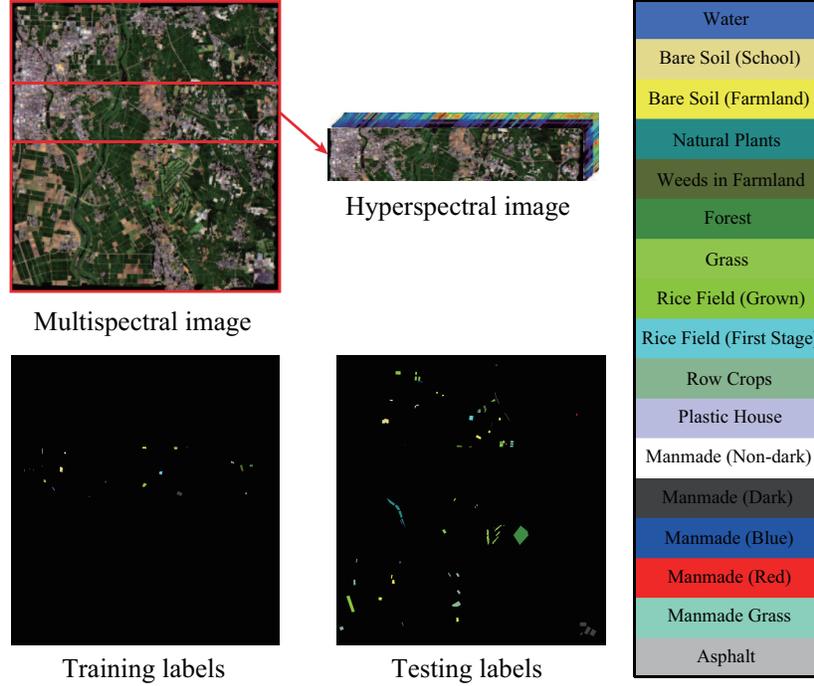


Figure 6: The multispectral image and its corresponding hyperspectral image that partially covers the same area, as well as training and testing labels, for Chikusei Dataset.

Table 3: Quantitative performance comparison with the different algorithms on the Chikusei data. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
	$d$		$(k, \sigma, d)$		$d$		$(k, \sigma, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$	
Parameter	10		(10, 1, 10)		20		(10, 0.1, 20)		(0.1, 0.01, 30)		(0.1, 0.01, 30)		(0.1, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	60.20	71.11	62.30	72.26	67.90	71.53	69.68	73.27	71.12	75.69	72.60	77.11	75.11	<b>81.71</b>
AA	69.42	70.40	69.80	70.71	70.79	66.47	72.27	70.01	73.96	71.46	71.64	71.33	75.29	<b>75.73</b>
$\kappa$	0.5523	0.6761	0.5784	0.6894	0.6391	0.6802	0.6602	0.6818	0.6746	0.7260	0.6911	0.7420	0.7194	<b>0.7933</b>
Class1	78.21	80.54	78.09	80.42	98.72	82.52	<b>99.53</b>	97.90	92.54	79.25	98.83	98.37	98.25	98.83
Class2	94.43	82.70	94.11	93.84	93.20	92.50	93.20	93.09	93.47	<b>94.91</b>	87.04	93.63	93.20	93.79
Class3	23.54	50.06	37.75	76.87	62.57	55.31	68.41	76.55	80.40	77.71	80.65	77.23	89.29	<b>89.90</b>
Class4	92.13	92.56	92.23	95.72	90.57	91.53	92.51	88.76	90.59	96.23	94.64	92.49	95.11	<b>96.96</b>
Class5	<b>97.65</b>	94.68	96.84	88.45	28.43	16.06	24.01	32.85	83.94	66.52	51.81	43.32	60.74	67.78
Class6	62.01	81.48	57.47	69.67	62.52	78.91	68.27	79.67	63.61	79.02	72.34	<b>88.48</b>	76.34	87.27
Class7	99.67	99.93	99.66	<b>100.00</b>	96.87	97.79	95.40	99.37	97.74	99.75	98.41	99.87	97.63	99.80
Class8	57.11	93.40	69.06	98.93	95.59	93.49	96.88	96.53	95.05	92.72	<b>99.48</b>	98.45	99.27	99.18
Class9	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	99.92	99.53	99.13	99.45	99.21	98.66	99.76	99.21	98.34	99.76	<b>100.00</b>
Class10	24.81	19.56	<b>26.64</b>	19.06	21.39	15.48	20.94	13.09	22.35	18.00	22.75	14.83	26.47	26.46
Class11	0.00	2.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	5.47	0.63	<b>5.68</b>
Class12	<b>90.32</b>	88.91	<b>90.32</b>	89.61	90.14	85.92	90.14	89.44	90.32	80.46	89.96	89.44	88.38	90.14
Class13	33.11	33.09	33.11	36.50	32.61	56.25	31.32	30.88	33.11	67.90	33.11	54.93	33.11	<b>68.73</b>
Class14	<b>94.20</b>	85.38	79.12	59.40	72.85	59.40	<b>94.20</b>	86.31	59.40	52.44	14.39	49.19	45.01	53.60
Class15	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	93.58	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	93.58	97.86	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>
Class16	74.88	88.62	74.19	93.52	99.71	99.51	99.80	98.82	97.84	<b>100.00</b>	97.35	97.25	98.04	95.78
Class17	58.03	3.84	58.03	0.24	65.23	7.91	62.11	7.67	64.75	0.00	77.70	11.27	<b>78.66</b>	13.43

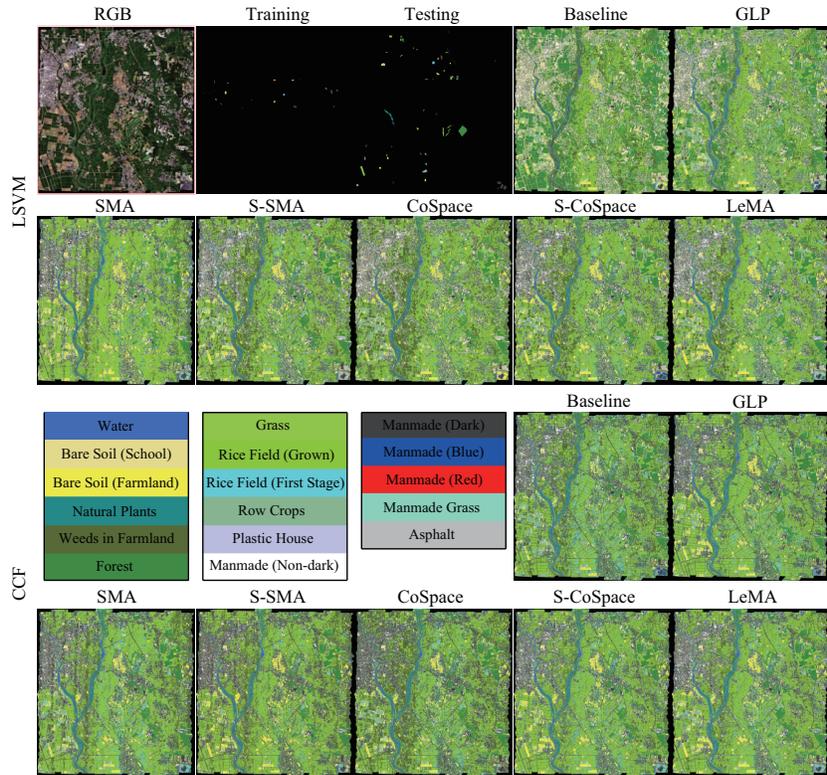


Figure 7: Classification maps of the different algorithms obtained using two kinds of classifiers on the Chikusei dataset.

335 its poor spectral information and a limited number of training samples. GLP utilizes the  
 336 unlabeled samples to augment the training samples in a semi-supervised way, yet it is  
 337 still limited by the low-discriminative spectral signatures. By aligning the MS and HS  
 338 data, these alignment-based approaches (e.g. SMA, S-SMA, CoSpace, S-CoSpace, and  
 339 LeMA) are able to find a common subspace in which the learnt features are expected to  
 340 absorb the different properties from two modalities, resulting in a better performance.  
 341 Compared to the supervised methods (SMA and CoSpace), their corresponding semi-  
 342 supervised versions (S-SMA and S-CoSpace) obtain higher classification accuracies  
 343 on both classifiers, which is detailed in Table 3. As expected, the performance of  
 344 the LeMA is significantly superior to that of others, thanks to the great contributions  
 345 of a common subspace learning from MS-HS data, a data-driven graph learning and

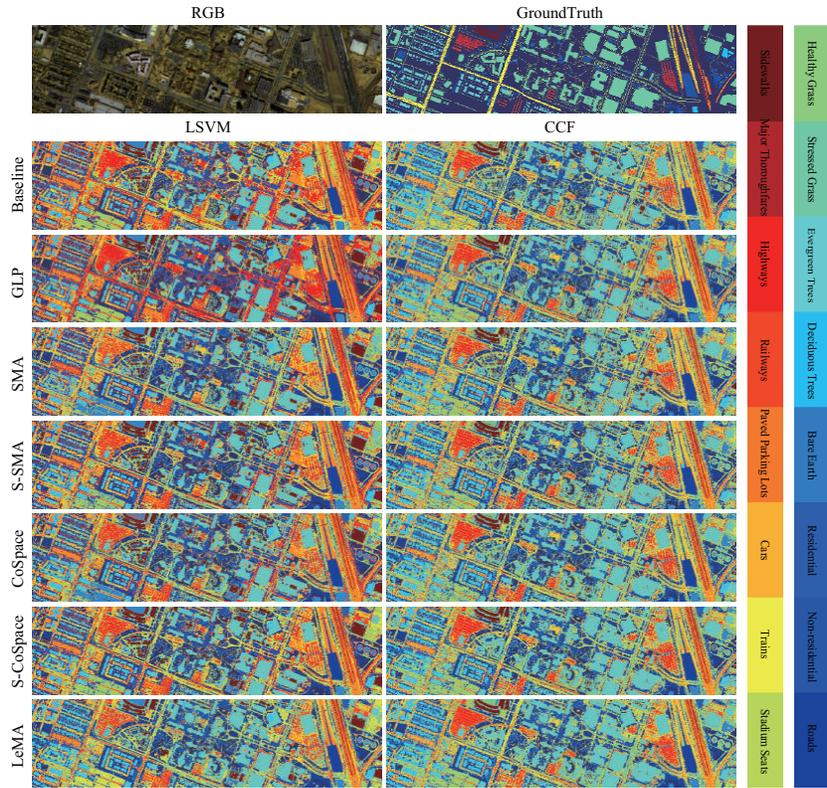


Figure 8: Classification maps of the different algorithms obtained using two kinds of classifiers on the real dataset of DFC2018 (Multispectral-Lidar and Hyperspectral data).

346 the semi-supervised learning strategy. Despite so, the LeMA still fails to recognize  
 347 some challenging classes, such as *Weeds in Farmland*, *Row Crops*, *Plastic House*, and  
 348 *Asphalt*. The reasons could be two-fold. On one hand, the performance of LeMA  
 349 is limited, to some extent, by the unbalanced data sets. On the other hand, LeMA'  
 350 transferring ability would sharply degrade when a great spectral variability between  
 351 training and test samples exists.

### 352 3.3. The Real Multispectral-Lidar and Hyperspectral Datasets in DFC2018

353 Although we follow strict simulation procedures, yet the two MS-HS datasets used  
 354 above (Houston and Chikusei) essentially originate from a similar data source (ho-  
 355 mogeneous), which means there is a strong correlation in their spectral features. This

Table 4: Quantitative performance comparison with the different algorithms on the DFC2018 data. The best one is shown in bold.

Methods	Baseline (%)		GLP (%)		SMA (%)		S-SMA (%)		CoSpace (%)		S-CoSpace (%)		LeMA (%)	
Parameter	$d$		$(k, \sigma, d)$		$d$		$(k, \sigma, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$		$(\alpha, \beta, d)$	
	7		(10, 1, 7)		30		(10, 1, 30)		(0.1, 0.1, 30)		(0.1, 0.01, 30)		(0.1, 0.01, 30)	
Classifier	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF	LSVM	CCF
OA	51.35	72.84	52.28	73.15	52.73	70.37	54.69	72.13	55.56	74.04	58.65	76.59	61.69	<b>79.98</b>
AA	59.46	78.64	60.57	81.64	58.06	77.78	65.34	78.72	66.16	80.46	67.72	83.67	65.54	<b>88.82</b>
$\kappa$	0.4194	0.6534	0.4289	0.6587	0.4366	0.6256	0.4598	0.6441	0.4670	0.6682	0.4987	0.6990	0.5284	<b>0.7414</b>
Class1	91.70	84.62	96.15	93.12	84.01	85.43	94.13	90.89	95.14	89.07	94.74	95.14	92.31	<b>100.00</b>
Class2	33.90	80.17	35.62	80.74	73.00	82.40	69.57	80.17	61.32	80.37	69.73	81.52	78.09	<b>87.90</b>
Class3	94.92	96.16	96.02	96.57	95.06	95.06	96.30	96.30	93.83	97.26	94.79	96.30	96.57	<b>99.45</b>
Class4	83.00	92.50	85.50	97.50	85.50	90.00	84.50	94.00	83.00	91.00	85.50	98.00	79.00	<b>100.00</b>
Class5	43.71	90.42	30.54	87.43	53.29	87.43	52.10	85.03	61.08	92.22	45.51	92.22	30.54	<b>100.00</b>
Class6	80.44	90.60	81.32	91.82	78.79	87.77	82.80	87.98	83.94	90.35	85.24	91.27	89.71	<b>96.50</b>
Class7	59.26	82.01	61.11	81.52	57.62	78.21	58.66	82.45	59.89	82.37	63.95	85.14	69.56	<b>87.47</b>
Class8	14.07	31.98	10.75	36.00	21.71	28.00	20.83	35.16	26.64	38.71	11.77	39.51	31.43	<b>49.96</b>
Class9	48.54	54.14	50.77	58.40	44.87	56.96	52.60	53.49	47.94	63.30	53.69	<b>68.55</b>	40.47	62.26
Class10	10.16	42.07	8.00	31.70	6.77	37.82	5.55	29.21	11.02	36.67	24.21	<b>38.40</b>	12.93	38.04
Class11	23.54	72.03	25.96	79.07	79.07	74.45	45.88	75.45	34.21	76.26	54.12	81.49	62.58	<b>100.00</b>
Class12	93.85	85.85	92.92	94.46	92.00	87.08	85.85	90.15	85.54	86.15	74.15	95.38	66.46	<b>100.00</b>
Class13	60.50	74.96	57.31	87.56	59.33	73.45	60.17	77.98	63.03	79.33	64.71	87.06	70.59	<b>99.83</b>
Class14	39.93	87.15	55.21	90.63	17.71	86.11	47.22	85.76	66.32	89.58	75.69	90.63	55.21	<b>99.65</b>
Class15	95.39	96.77	97.70	<b>100.00</b>	93.55	98.16	99.54	97.70	99.54	98.62	99.54	<b>100.00</b>	95.85	<b>100.00</b>
Class16	78.39	96.77	84.19	99.68	77.74	96.13	89.68	97.74	86.13	96.13	86.13	98.06	77.42	<b>100.00</b>

356 makes the information of the different modalities transferred more effectively, but could  
357 limit the generalization ability in practice. To this end, we apply a real bi-modal dataset  
358 – multispectral-lidar and hyperspectral (heterogeneous) provided by the latest IEEE  
359 GRSS data fusion contest 2018 (DFC2018).

### 360 3.3.1. Data Description

361 Multi-source optical remote sensing data, such as multispectral-lidar data, hyper-  
362 spectral data, and very high-resolution RGB data, is provided in the contest. More  
363 specifically, the multispectral-lidar imagery consists of  $1202 \times 4768$  pixels with 7 bands  
364 (3 intensity bands and 4 DSMs-related bands [45]) collected from 1550nm, 1064nm,  
365 and 532nm at a 0.5m GSD, while the hyperspectral data comprises 48 bands covering  
366 a spectral range from 380nm to 1050nm at 1m GSD, and its size is  $601 \times 2384$ . In  
367 our case, our LeMA model is trained on partial multispectral-lidar and hyperspectral  
368 correspondences and tested only using multispectral-lidar data, in order to meet the  
369 requirement of our cross-modality learning task. The first row of Fig.8 shows the RGB  
370 image of this scene and the labeled ground truth image.

### 371 3.3.2. Experimental Setup

372 Our aim is, once again, to investigate whether the limited amount of hyperspectral  
373 data can improve the performance of another modality, e.g., multispectral data (homo-

374 geneous) or multispectral-lidar data (heterogeneous). Therefore, we randomly assign  
375 10% of total labeled samples as training set and the rest of it as test set in the ex-  
376 periment. Moreover, 16 main classes are selected out of 20 (see Fig.8), by removing  
377 several small classes with too few samples, e.g. *Artificial Turf*, *Water*, *Crosswalks*,  
378 and *Unpaved Parking Lots*. Likewise, we automatically configure the parameters of  
379 the proposed LeMA and the compared algorithms by a 10-fold cross-validation on the  
380 training set, which is detailed in section 3.1.2.

### 381 3.3.3. Results and Analysis

382 We show the averaged results of the different algorithms out of 10 runs to obtain  
383 a relatively stable and meaningful performance comparison, because the training and  
384 test sets are randomly generated from total samples in each round, as listed in Table 4.  
385 Correspondingly, Fig. 8 visually highlights the differences of classification maps for  
386 the different methods.

387 Generally speaking, hyperspectral information embedding can effectively improve  
388 the classification performance of the multispectral-lidar data, which implies that the  
389 models based common subspace learning (e.g., SMA, S-SMA, CoSpace, S-CoSpace,  
390 and LeMA) can transfer the knowledge from one modality to another modality to some  
391 extent. We also observe from Table 4 that the semi-supervised methods which consider  
392 the unlabeled samples (e.g., GLP, S-SMA, S-CoSpace, and LeMA) always perform  
393 better than those purely supervised ones. Not unexpectedly, LeMA integrating rich  
394 spectral information and unlabeled samples achieves a superior performance, which  
395 demonstrates that the learning-based graph structure is more applicable to capturing  
396 the data distribution and further find a potential optimal decision boundary.

397 One thing to be noted, however, is that compared to the performance of the different  
398 algorithms in the simulated MS-HS datasets from similar sources (homogeneous), the  
399 knowledge transferring ability of these algorithms in handling the real multispectral-  
400 lidar and hyperspectral datasets from different sources (heterogeneous) remains lim-  
401 ited, since all listed methods including our LeMA are modeled in a linearized way.  
402 Unfortunately, a single linear transformation fails to fit the gap between heterogeneous  
403 modalities well, despite a limited performance improvement.

#### 404 **4. Conclusions**

405 In real-world problems, a large amount of low-quality data (e.g. MS data) can  
406 often be easily collected. On the contrary, high-quality data (e.g. HS data) are usu-  
407 ally expensive and difficult to obtain. This motivates us to investigate whether a lim-  
408 ited amount of high-quality data can contribute to relevant tasks with a large amount  
409 of low-quality data. For this purpose, we propose a novel semi-supervised learning  
410 framework called LeMA, which effectively connects the common subspace and label  
411 information, and automatically embeds the unlabeled information into the proposed  
412 framework by adaptively learning a Laplacian matrix from the data. Extensive exper-  
413 iments are conducted using the LeMA on two homologous MS-HS simulated datasets  
414 and a heterogenous multispectral-lidar and hyperspectral real dataset in comparison  
415 with the other state-of-arts algorithms, demonstrating the superiority and effectiveness  
416 of the LeMA in the knowledge transferring ability. We have to admit, however, that de-  
417 spite a significant performance improvement in LeMA, yet its representative ability is  
418 still limited by linearly modeling way, especially facing highly-nonlinear heterogenous  
419 data. Towards this issue, we will continue to improve our model to a nonlinear version  
420 and simultaneously consider the spatial information (e.g., morphological profiles) to  
421 further strengthen the feature representation ability.

#### 422 **5. Acknowledgements**

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

428 This work was supported by funding from the European Research Council (ERC)  
429 under the European Union’s Horizon 2020 research and innovation program (grant  
430 agreement No [ERC-2016-StG-714087]) and from Helmholtz Association under the  
431 framework of the Young Investigators Group ”SiPEO” (VH-NG-1018, www.sipeco.bgu.  
432 tum.de). The work of N. Yokoya was supported by Japan Society for the Promotion

433 of Science (JSPS) KAKENHI 15K20955 and Alexander von Humboldt Fellowship for  
434 postdoctoral researchers.

- 435 [1] X. Huang, Q. Lu, L. Zhang, A multi-index learning approach for classification of  
436 high-resolution remotely sensed images over urban areas, *ISPRS J. Photogram-*  
437 *metry Remote Sens.* 90 (2014) 36–48.
- 438 [2] D. Hong, N. Yokoya, X. Zhu, The k-lle algorithm for nonlinear dimensionality  
439 reduction of large-scale hyperspectral data, in: *Hyperspectral Image and Signal*  
440 *Processing: Evolution in Remote Sensing (WHISPERS)*, 2016 8th Workshop on,  
441 *IEEE*, 2016, pp. 1–5.
- 442 [3] J. Kang, M. Körner, Y. Wang, H. Taubenböck, X. Zhu, Building instance classifi-  
443 cation using street view images, *ISPRS J. Photogrammetry Remote Sens.*
- 444 [4] C. Yang, J. H. Everitt, Q. Du, B. Luo, J. Chanussot, Using high-resolution air-  
445 borne and satellite imagery to assess crop growth and yield variability for preci-  
446 sion agriculture, *Proc. IEEE* 101 (3) (2013) 582–592.
- 447 [5] F. D. V. der Meer, H. M. A. V. der Werff, F. J. A. V. Ruitenbeek, Potential of  
448 esa’s sentinel-2 for geological applications, *Remote Sens. Environ.* 148 (2014)  
449 124–133.
- 450 [6] N. Yokoya, C. Grohnfeldt, J. Chanussot, Hyperspectral and multispectral data  
451 fusion: a comparative review, *IEEE Geosci. Remote Sens. Mag.* 5 (2) (2017)  
452 29–56.
- 453 [7] D. Hong, N. Yokoya, X. Zhu, Learning a robust local manifold representation for  
454 hyperspectral dimensionality reduction, *IEEE J. Sel. Topics Appl. Earth Observ.*  
455 *Remote Sens.* 10 (6) (2017) 2960–2975.
- 456 [8] J. Li, H. Zhang, L. Zhang, Column-generation kernel nonlocal joint collaborative  
457 representation for hyperspectral image classification, *ISPRS J. Photogrammetry*  
458 *Remote Sens.* 94 (2014) 25–36.

- 459 [9] Y. Tarabalka, J. Benediktsson, J. Chanussot, Spectral-spatial classification of  
460 hyperspectral imagery based on partitional clustering techniques, *IEEE Trans.*  
461 *Geosci. Remote Sens.* 47 (8) (2009) 2973–2987.
- 462 [10] L. Zhang, L. Zhang, D. Tao, X. Huang, On combining multiple features for hyper-  
463 spectral remote sensing image classification, *IEEE Trans. Geosci. Remote Sens.*  
464 50 (3) (2012) 879–893.
- 465 [11] D. Hong, N. Yokoya, J. Xu, X. Zhu, Joint and progressive learning from high-  
466 dimensional data for multi-label classification, in: *European Conference on Com-*  
467 *puter Vision (ECCV)*, Springer, 2018, pp. 478–493.
- 468 [12] J. Xia, J. Chanussot, P. Du, X. He, Semi-supervised probabilistic principal com-  
469 ponent analysis for hyperspectral remote sensing image classification, *IEEE J.*  
470 *Sel. Topics Appl. Earth Observ. Remote Sens.* 7 (6) (2014) 2224–2236.
- 471 [13] D. Tuia, M. Volpi, M. Trollet, G. Camps-Valls., Semisupervised manifold align-  
472 ment of multimodal remote sensing images, *IEEE Trans. Geosci. Remote Sens.*  
473 52 (12) (2014) 7708–7720.
- 474 [14] L. Bruzzone, M. Marconcini, Domain adaptation problems: A dasvm classifica-  
475 tion technique and a circular validation strategy, *IEEE Trans. Pattern Anal. Mach.*  
476 *Intell.* 32 (5) (2010) 770–787.
- 477 [15] B. Banerjee, F. Bovolo, A. Bhattacharya, L. Bruzzone, S. Chaudhuri, K. M. Bud-  
478 dhiraju, A novel graph-matching-based approach for domain adaptation in classi-  
479 fication of remote sensing image pair, *IEEE Trans. Geosci. Remote Sens.* 53 (7)  
480 (2015) 4045–4062.
- 481 [16] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, D. Tuia, Semisupervised trans-  
482 fer component analysis for domain adaptation in remote sensing image classifi-  
483 cation, *IEEE Trans. Geosci. Remote Sens.* 53 (7) (2015) 3550–3564.
- 484 [17] D. Tuia, C. Persello, L. Bruzzone, Domain adaptation for the classification of re-  
485 mote sensing data: An overview of recent advances, *IEEE Geosci. Remote Sens.*  
486 *Mag.* 4 (2) (2016) 41–57.

- 487 [18] A. Samat, P. Gamba, J. Abuduwaili, S. Liu, Z. Miao, Geodesic flow kernel support  
488 vector machine for hyperspectral image classification by unsupervised subspace  
489 feature transfer, *Remote Sens.* 8 (3) (2016) 234.
- 490 [19] A. Samat, C. Persello, P. Gamba, S. Liu, J. Abuduwaili, E. Li, Supervised and  
491 semi-supervised multi-view canonical correlation analysis ensemble for hetero-  
492 geneous domain adaptation in remote sensing image classification, *Remote Sens.*  
493 9 (4) (2017) 337.
- 494 [20] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, A. Torralba, Undoing the damage  
495 of dataset bias, in: *European Conference on Computer Vision (ECCV)*, Springer,  
496 2012, pp. 158–171.
- 497 [21] C. Woodcock, S. A. Macomber, M. Pax-Lenney, W. B. Cohen, Monitoring large  
498 areas for forest change using landsat: Generalization across space, time and land-  
499 sat sensors, *Remote Sens. Environ.* 78 (1-2) (2001) 194–203.
- 500 [22] J. Jiang, X. Zhai, Instance weighting for domain adaptation in nlp, in: *Proceed-*  
501 *ings of ACL, 2007*, pp. 264–271.
- 502 [23] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, M. Kawanabe, Direct im-  
503 portance estimation with model selection and its application to covariate shift  
504 adaptation, in: *Advances in neural information processing systems (NIPS)*, 2008,  
505 pp. 1433–1440.
- 506 [24] A. Samat, P. Gamba, S. Liu, P. Du, J. Abuduwaili, Jointly informative and man-  
507 ifold structure representative sampling based active learning for remote sensing  
508 image classification, *IEEE Trans. Geosci. Remote Sens.* 54 (11) (2016) 6803–  
509 6817.
- 510 [25] C. C. Persello, L. Bruzzone, Active learning for domain adaptation in the super-  
511 vised classification of remote sensing images, *IEEE Trans. Geosci. Remote Sens.*  
512 50 (11) (2012) 4468–4483.
- 513 [26] C. Wang, P. Krafft, S. Mahadevan, Chapter of *Manifold Learning: Theory and*  
514 *Applications-Manifold alignment*, CSC Press, 2011.

- 515 [27] D. Tuia, D. Marcos, G. Camps-Valls, Multi-temporal and multi-source remote  
516 sensing image classification by nonlinear relative normalization, *ISPRS J. Pho-*  
517 *togrammetry Remote Sens.* 120 (2016) 1–12.
- 518 [28] D. Liao, Y. Qian, J. Zhou, Y. Tang, A manifold alignment approach for hyperspec-  
519 tral image visualization with natural color, *IEEE Trans. Geosci. Remote Sens.*  
520 54 (6) (2016) 3151–3162.
- 521 [29] C. Wang, S. Mahadevan, A general framework for manifold alignment, in: *AAAI*  
522 *Fall Symposium on Manifold Learning and its Applications (AAAI)*, 2009.
- 523 [30] C. Wang, S. Mahadevan, Heterogeneous domain adaptation using manifold align-  
524 ment, in: *Proceedings of the 22th International Joint Conference on Artificial*  
525 *Intelligence (IJCAI)*, 2011, pp. 1541–1546.
- 526 [31] X. Zhu, Z. Ghahramani, J. D. Lafferty, Semi-supervised learning using gaussian  
527 fields and harmonic functions, in: *Proceedings of the 20th International Confer-*  
528 *ence on Machine learning (ICML)*, 2003, pp. 912–919.
- 529 [32] S. Ji, J. Ye, Linear dimensionality reduction for multi-label classification, in: *Pro-*  
530 *ceedings of the 21th International Joint Conference on Artificial Intelligence (IJ-*  
531 *CAI)*, 2009, pp. 1077–1082.
- 532 [33] Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: *Proceed-*  
533 *ings of the 22th International Joint Conference on Artificial Intelligence (IJCAI)*,  
534 2011, pp. 1294–1299.
- 535 [34] F. Heide, W. Heidrich, G. Wetzstein, Fast and flexible convolutional sparse cod-  
536 ing, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,  
537 2015, pp. 5135–5143.
- 538 [35] D. P. Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.
- 539 [36] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and  
540 statistical learning via the alternating direction method of multipliers, *Founda-*  
541 *tions and Trends® in Machine learning* 3 (1) (2011) 1–122.

- 542 [37] L. Chen, X. Li, D. Sun, K. Toh, On the equivalence of inexact proximal  
543 alm and admm for a class of convex composite programming, arXiv preprint  
544 arXiv:1803.10803.
- 545 [38] Z. Lin, M. Chen, Y. Ma, The augmented lagrange multiplier method for exact  
546 recovery of corrupted low-rank matrices, arXiv preprint arXiv:1009.5055.
- 547 [39] D. Hong, N. Yokoya, J. Chanussot, X. Zhu, Learning low-coherence dictionary to  
548 address spectral variability for hyperspectral unmixing, in: Proceedings of IEEE  
549 International Conference on Image Processing (ICIP), 2017, pp. 1–5.
- 550 [40] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace struc-  
551 tures by low-rank representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1)  
552 (2013) 171–184.
- 553 [41] Y. Zhong, X. Wang, L. Zhao, R. Feng, L. Zhang, Y. Xu, Blind spectral unmixing  
554 based on sparse component analysis for hyperspectral remote sensing imagery,  
555 *ISPRS J. Photogrammetry Remote Sens.* 119 (2016) 49–63.
- 556 [42] P. Zhou, C. Zhang, Z. Lin, Bilevel model based discriminative dictionary learning  
557 for recognition, *IEEE Trans. Image Process.* 26 (3) (2017) 1173–1187.
- 558 [43] R. Tom, W. Frank, Canonical correlation forests, arXiv preprint  
559 arXiv:1507.05444.
- 560 [44] N. Yokoya, A. Iwasaki, Airborne hyperspectral data over chikusei, Tech. Rep.  
561 SAL-2016-05-27.
- 562 [45] B. L. Saux, N. Yokoya, R. Hansch, S. Prasad, 2018 ieee grss data fusion contest:  
563 Multimodal land use classification [technical committees], *IEEE Geosci. Remote*  
564 *Sens. Mag.* 6 (1) (2018) 52–54.