

Semi-supervised Superpixel-based Multi-Feature Graph Learning for Hyperspectral Image Data

Madeleine S. Kotzagiannidis, Carola-Bibiane Schönlieb

Abstract—Graphs naturally lend themselves to model the complexities of Hyperspectral Image (HSI) data as well as to serve as semi-supervised classifiers by propagating given labels among nearest neighbours. In this work, we present a novel framework for the classification of HSI data in light of a very limited amount of labelled data, inspired by multi-view graph learning and graph signal processing. Given an a priori superpixel-segmented hyperspectral image, we seek a robust and efficient graph construction and label propagation method to conduct semi-supervised learning (SSL). Since the graph is paramount to the success of the subsequent classification task, particularly in light of the intrinsic complexity of HSI data, we consider the problem of finding the optimal graph to model such data. Our contribution is two-fold: firstly, we propose a multi-stage edge-efficient semi-supervised graph learning framework for HSI data which exploits given label information through pseudo-label features embedded in the graph construction. Secondly, we examine and enhance the contribution of multiple superpixel features embedded in the graph on the basis of pseudo-labels in an extension of the previous framework, which is less reliant on excessive parameter tuning. Ultimately, we demonstrate the superiority of our approaches in comparison with state-of-the-art methods through extensive numerical experiments.

Index Terms—Hyperspectral image classification, graph Laplacian learning, semi-supervised learning

I. INTRODUCTION

THE problem of determining the optimal graph representation of a given dataset has been considered for various tasks in different fields, ranging from signal processing to machine learning, yet remains largely unresolved. Hyperspectral image (HSI) data, with its rich and descriptive spatial and spectral information contained in several hundred bands, encapsulates several layers of dependencies between the high-dimensional pixels and their corresponding labels [1], and a plethora of methods have sought to exploit these under different modelling assumptions.

In hyperspectral image classification, a classifier is sought to assign a class label to each pixel, in light of arising difficulties including high spectral dimensionality, large spatial variability and limited availability of labels. While the majority of frameworks involve a supervised classifier, due to the time and cost associated with obtaining labelled samples, semi-supervised learning (SSL) has experienced a rapid development by tackling the small sample problem through effective

exploitation of both labeled and unlabelled samples [2].

Multiple views or features of data can provide rich insights into the underlying data structure, while exploiting diverse information and thus gaining robustness for subsequent tasks. The identification of suitable features, however, represents a problem in itself. Inspired by the field of multi-view clustering [3] and in an effort to incorporate more sophisticated model priors, we adopt the perspective of learning a graph from multiple data features (or views) in order to effectively capture the complexity of hyperspectral data, while further integrating the label information into the graph so as to exploit prior knowledge of label dependencies. We further resort to superpixel segmentation of the data in order to effectively reduce the complexity of the classification task, while simultaneously capturing a first level of spectral-spatial dependencies through the clustering into local homogeneous regions.

In this paper, we introduce a novel semi-supervised framework for HSI classification which involves the design of a graph classification function which is smooth with respect to both the intrinsic structure of the data, as described via superpixel features, as well as the label space. The proposed graph learning framework is special due to its edge-efficient analytic solution, known to satisfy graph-optimality constraints, and ability to incorporate multiple features, as well as due to its optimization of both the graph and classification function, resulting in pseudo-labels. In addition, in light of the range of parameters required to tune a multi-feature superpixel graph, we propose a variation of the framework which, instead of incorporating pre-set feature weights, learns them by imposing an additional smoothness functional. We summarize the main contributions as follows:

- We extend multi-view graph learning to the domain of superpixels and HSI data, with tunable pseudo-label generation incorporated into an updateable graph. In particular, we employ an initial soft graph on which labels are firstly propagated among nearest neighbours to generate pseudo-label features, which are subsequently utilized to inform an improved graph.
- We propose a pseudo-label-guided framework for HSI feature selection, weighting and subsequent graph construction, enhanced by dynamic pseudo-label-features. To the best of our knowledge, the issue of feature contribution on a multi-feature graph for HSI-data has not yet been tackled, beyond a simple parameter search.
- We extensively validate our proposed approaches on the basis of three benchmark datasets and demonstrate their superiority with respect to comparable state-of-the-art

M. S. Kotzagiannidis and C.-B. Schönlieb are with the Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, UK. e-mail: mk2060@cam.ac.uk, cbs31@cam.ac.uk. This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

approaches.

The proposed framework facilitates edge-efficient graph and label learning, while flexibly incorporating multiple features which capture spectral-, spatial- and label-dependencies within the given data. The remainder of this paper is organized as follows: Section II discusses related work, Section III explores the preliminaries, and details of the proposed methods are stated in Section IV. Section V presents the experimental results of our methods in comparison with state-of-the-art approaches, and Section VI contains concluding remarks.

II. RELATED WORK

Works in SSL can be categorized into generative models, which employ probabilistic generative mixture models, co-and self-training methods, low-density separation methods, which seek a decision boundary through low density regions, and graph-based methods [2]. Popular modelling assumptions and concepts have included spectral clustering, data manifolds and local-global consistency [4], according to which nearby points and points residing on the same cluster or manifold are likely to have the same label [2]. In the context of HSI data, earlier approaches comprise kernel methods such as the purely spectral-based SVM [5] and spectral-spatial multiple kernel learning [6]. Feature extraction methods, such as [7], [8], [9], have sought to characterize a lower dimensional subspace which best captures the spectral-spatial information of the data. Further, classical signal processing concepts such as sparse representation, low-rank, and wavelet analysis have been notably incorporated into SSL methods [10], [11], [1]. More recent methods have gained in complexity by generating multi-stage workflows with different pre- and post-processing levels as well as by combining the strengths of different classifiers in an effort to create more sophisticated models which exploit multiple dependencies of HSI data and counteract the small sample problem through a gradual learning process [1]. Graph-based SSL methods have become increasingly popular due to the superior modelling capabilities of graphs, particularly as a means to counteract the limited amount of labels available, and have included pixel-based [12] as well as superpixel-based graph constructions [13], [14], [15], ranging from end-to-end approaches, that utilize graphs for both data modelling and label propagation, to hybrid approaches. We note that many existing methods, such as EPF [16], which makes use of the bilateral filter, exhibit underlying graph-like qualities, while not directly or only in part employing graphs, and as a result have exhibited superior performance. In addition, approaches such as [17], have sought to incorporate label information early on in the workflow, known as *pseudo-labelling*, which is progressively refined, in an effort to inform data modelling; nevertheless, the graph construction is not necessarily analytic. While identification of the correct model graph is essential for SSL [18], extensive parameter analysis is still inevitable in order to study their influence, as performance is strongly affected by all components of the graph; nevertheless, previous studies have generally not found explorable patterns [18], and stability is preferred over excellence in a narrow parameter range.

A preceding body of work on multi-view clustering has sought to unify the tasks of data similarity learning (i.e. graph construction) and clustering/label propagation in a joint optimization framework [3], [19], [20], which, to the best of our knowledge, has not been fully investigated for the joint challenge of HSI data and SSL. Graph learning and label propagation are separate tasks but recent efforts have opted to combine them, thereby updating and incorporating label information into the graph, under the driving assumption that a single static graph is not sufficient to solve the entire SSL task successfully. Nevertheless, and not least of all for complex data such as HSI, the issue of propagating error noise into each task needs to be addressed. There are multiple ways in which the label information can be incorporated into the graph, which include space fusion approaches, i.e. the graph and label space are fused via the addition of a label correlation matrix [21] or the removal of differently labelled edges [22], and implicit approaches, which instead consider the graph as a function of the distances between labels [3].

Deep learning (DL) methods have penetrated graph-based classification in the form of Graph Convolutional Networks (GCNs), which usually require a given pre-constructed graph, such as the spectral-spatial GCN [23], and have been employed to extract features automatically [24]. Under the assumption that the mapping from the feature to the label space is sufficiently smooth, the theoretical relation between label propagation (LP) and GCNs, as instances which propagate labels and features respectively, has been notably shown to satisfy a smoothness inequality [25]. Nevertheless, DL-, and non-DL methods alike, ultimately rely on the optimality of the graph while often performing worse with limited labels. Neural networks have further been employed to learn graphs from scratch (see e.g. [26]), which inevitably comes at a high computational cost, however, for complex and rich datasets such as HSI the formulation of sophisticated (graph) model priors is paramount to the successful performance of any classifier.

Graphs ultimately capture model constraints in the form of linear dependencies between data points (as per the graph Laplacian [27]) and, as such, present versatile modelling tools which can help simplify more complex modelling assumptions. Most graph-based methods operate under the assumption that the given graph is optimal and/or that the given data naturally resides on a known graph, so any subsequent tasks and operations are subject to noise pertaining to an imperfect model. Given a data set, approaches have opted to either hand-craft the graph, or learn it automatically through the minimization of a chosen optimization function. In the former case, this has included local and/or adaptive graph weighting and connectivity schemes, which take into account variations in i.a. (spectral) data density, and are increasingly refined but also bear a risk of distorting neighborhood information [28]. While handcrafted approaches usually lead to higher accuracy, automated ones can be more robust to variable datasets, requiring less parameter tuning while nevertheless being more costly due to lack of an analytic solution. In an effort to combine the advantages of both, we therefore seek an analytic solution to a well-defined optimization problem

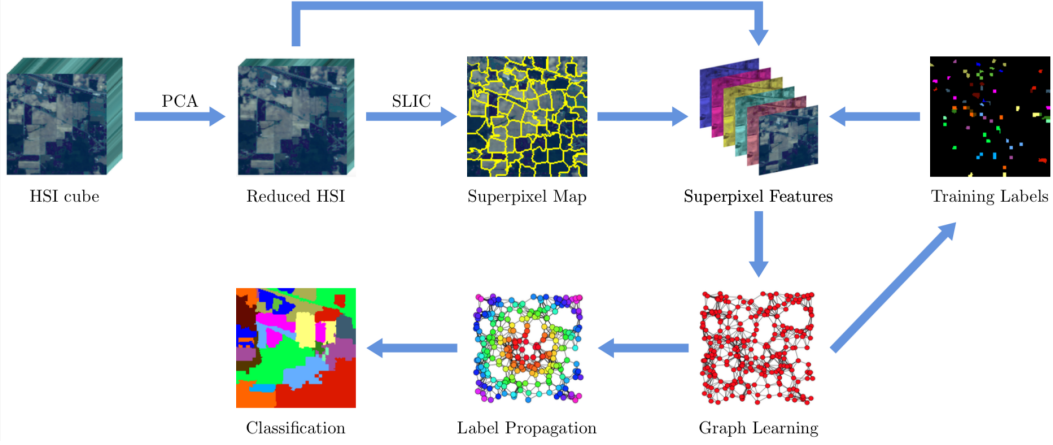


Fig. 1. Main Workflow: The proposed MGL and PMGL methods are mainly composed of spectral dimensionality reduction, superpixel segmentation, superpixel feature extraction and superpixel label regularization, followed by a pseudo-label driven graph learning stage, and culminating in the final classification result via label propagation.

with the possibility of sophisticated parameter learning (i.e. reduction) through regularization.

III. PRELIMINARIES AND PRIOR WORK: GRAPH (LAPLACIAN) LEARNING

An undirected graph $G = (V, E)$ is characterized by a set of vertices V and a set of edges E , and its connectivity is encapsulated in the symmetric adjacency matrix \mathbf{W} , with $W_{i,j} > 0$ if there is an edge between vertices i and j , and $W_{i,j} = 0$ otherwise. The non-normalized graph Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{1})$ is the diagonal degree matrix, with $\mathbf{1}$ denoting the vector of 1's. The construction of a graph which optimally represents a given dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is a multi-part task, which generally establishes relations of the form $W_{i,j} = A_{i,j}w(\mathbf{x}_i, \mathbf{x}_j)$ between data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d$, where $w : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a similarity function and \mathbf{A} an adjacency matrix, for which $A_{i,j} = 1$ iff $j \in \mathcal{N}_i$ for some pre-determined neighborhood \mathcal{N}_i of node i , and $A_{i,j} = 0$ otherwise. From the smoothness of w to the range of influence \mathcal{N} , each component is vital in ensuring effective discriminative data representation.

In graph signal processing, Graph Laplacian Learning (GLL) considers the minimization of the graph Laplacian quadratic form with respect to data (aka graph signal) matrix \mathbf{X} :

$$\text{Tr}(\mathbf{X}^T \mathbf{L} \mathbf{X}) = \frac{1}{2} \sum_{i,j} W_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \frac{1}{2} \|\mathbf{W} \circ \mathbf{Z}\|_{1,1},$$

with trace operator $\text{Tr}(\cdot)$ and $Z_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$, and can be alternatively expressed as a weighted sparsity l_1 -norm of \mathbf{W} , whose minimization enforces connectivity (large $W_{i,j}$) between similar features \mathbf{x}_i and \mathbf{x}_j . Previous works (see e.g. [29], [30]), have considered variations of the general framework:

$$\min_{\mathbf{W} \in \mathcal{W}} \|\mathbf{W} \circ \mathbf{Z}\|_{1,1} + f(\mathbf{W}), \quad (1)$$

consisting of the graph Laplacian quadratic form and a, possibly sparsity-promoting, regularization term $f(\mathbf{W})$, subject to graph constraints, e.g. $\mathcal{W} = \{\mathbf{W} \in \mathbb{R}_{\geq 0}^{N \times N} : \mathbf{W} =$

$\mathbf{W}^T, \text{diag}(\mathbf{W}) = 0\}$. We select $f(\mathbf{W}) = \alpha \|\mathbf{W}\|_F^2, \alpha \in \mathbb{R}$, with Frobenius norm $\|\mathbf{W}\|_F = \sqrt{\sum_{i,j} |W_{i,j}|^2}$, which controls sparsity by preventing the occurrence of strong edges, and due to the edge-locality of the functional, facilitates the decomposition of the problem into a sum over graph edges, and hence an analytic solution.

Specifically, the optimization problem of Eq. (1) becomes separable for $f(\mathbf{W}) = \alpha \|\mathbf{W}\|_F^2$, and can be rewritten in row-wise form [3], for row \mathbf{W}_i of \mathbf{W} , as

$$\min_{\mathbf{W}_i \mathbf{1} = 1, \mathbf{W}_i \geq 0, W_{i,i} = 0} \sum_{j=1}^N Z_{i,j} W_{i,j} + \alpha \sum_{j=1}^N W_{i,j}^2 \quad (2)$$

yielding the closed-form solution

$$W_{i,j} = \left(\eta_i - \frac{Z_{i,j}}{2\alpha} \right)_+,$$

where $(x)_+ = \max(0, x)$ and scalar η_i . Since the second term creates a dense edge pattern, one can further enforce kNN connectivity by determining the maximal α_i per row s.t. the optimal \mathbf{W}_i has exactly k non-zeros (see [3] for details). This leads to

$$W_{i,j} = \begin{cases} \frac{Z_{i,k+1} - Z_{i,j}}{kZ_{i,k+1} - \sum_{h=1}^k Z_{i,h}}, & j \leq k \\ 0, & j > k. \end{cases} \quad (3)$$

where entries $\{Z_{i,1}, \dots, Z_{i,N}\}$ are assumed to be ordered from small to large wlog, and we have $\eta_i = \frac{1}{k} + \frac{\sum_{h=1}^k Z_{i,h}}{2k\alpha_i}$ and $\alpha_i = (k/2)Z_{i,k+1} - (1/2)\sum_{h=1}^k Z_{i,h}$. Here, $W_{i,i} = 0$ is enforced and $Z_{i,i} = 0$ appended at the end. Symmetrization is achieved through $\mathbf{L} = \mathbf{D} - \frac{\mathbf{W} + \mathbf{W}^T}{2}$.

It has been noted that this approach is computationally efficient due to its analytic solution and in-built sparsity which does not require oblique tuning of α (instead only requiring the straight-forward number of edges k) and is further scale-invariant w.r.t. feature vectors \mathbf{x}_i .

For different features (views) of type v , we consider $Z_{i,j} = \sum_v c_v Z_{i,j}^v$ with feature coefficients $c_v \in \mathbb{R}$. In order

to reduce parameters, it has been proposed to constrain the coefficients to be proportional to pre-set feature-dependent functionals and subject to regularization [19], [20].

IV. PROPOSED METHOD: SUPERPIXEL-BASED MULTI-FEATURE GRAPH LEARNING

In the proposed method, several instances of data dependencies are exploited in a multi-level workflow: after conducting an initial spectral dimensionality reduction using PCA [31], we consider a priori the segmentation of the hyperspectral image into superpixels to define local regions of homogeneous spectral content and to reduce spatial dimensionality. Subsequently, we compute analytic superpixel features, which capture different image properties, and extrapolate given pixel labels to superpixel labels via a simple averaging filter. We then compute an initial graph G_0 and form pseudo-label features through a soft label propagation to nearest neighbors on G_0 . This is subsequently updated and refined, before a final graph classifier is applied (see Fig. 1).

A. Superpixel-based Feature Extraction

Given the raw HSI data cube $\mathbf{I} \in \mathbb{R}^{X \times Y \times B}$, we apply dimensionality reduction in the first instance in the spectral domain using PCA to obtain the reduced $\tilde{\mathbf{I}} \in \mathbb{R}^{X \times Y \times b}$, $b \ll B$. Subsequently, the first PC component is used to conduct superpixel segmentation via SLIC [32], resulting in the 2D superpixel labelling map $\tilde{\mathbf{S}} \in \mathbb{R}^{X \times Y}$ with N superpixels:

$$\mathbf{S}_k \text{ s.t. } \mathbf{S}_k = \{\mathbf{S}_{(i,j)} | \mathbf{S}_{(i,j)} = k\}, \tilde{\mathbf{S}} = \bigcup_{k=1}^N \mathbf{S}_k.$$

While it has been established that the goodness and scale of the superpixel segmentation is foundational for the success of subsequent data modelling and classification tasks, it is not the objective of this work to optimize this particular instance of the workflow; as such, we select SLIC [32] as the base superpixel segmentation algorithm and determine the number of superpixels approximately according to [33], which takes into account both the size and resolution of the HSI image (albeit not the scene complexity).

Let $\mathbf{Y} \in \mathbb{R}^{XY \times c}$ denote the initial class indicator matrix with $Y_{i,j} = 1$ if pixel i belongs to class j for c classes. Following superpixel segmentation, we regularize this to $\mathbf{Y}^S \in \mathbb{R}^{N \times c}$ by averaging over the existing pixel labels per superpixel. Specifically, $Y_{i,j}^S$ records the number of pixels per superpixel \mathbf{S}_i which belong to class j divided by the total number of pixels in \mathbf{S}_i .

As statistical descriptors for the superpixels, we consider the features as proposed in [6], comprising the mean feature vector \mathbf{s}_k^M

$$\mathbf{s}_k^M = \frac{\sum_{i,j} \tilde{\mathbf{I}}_{(i,j)}}{N_k}, \mathbf{S}_{(i,j)} = k, k = 1, \dots, N$$

which takes a simple average of the N_k pixels per superpixel k , and the spatial-mean feature vector \mathbf{s}_k^S

$$\mathbf{s}_k^S = \sum_{i=1}^J w_{k,a_i} \mathbf{s}_{a_i}^M, w_{k,a_i} = \frac{\exp(-\|\mathbf{s}_{a_i}^M - \mathbf{s}_k^M\|_2^2/h)}{\sum_{i=1}^J \exp(-\|\mathbf{s}_{a_i}^M - \mathbf{s}_k^M\|_2^2/h)}$$

which constitutes a weighted sum of the mean feature vectors of adjacent superpixels, given by index set $\mathcal{A}_k = \{a_1, \dots, a_J\}$ for the k -th superpixel and with pre-set scalar $h \in \mathbb{R}$. Further, we consider the centroidal location of each superpixel as:

$$\mathbf{s}_k^C = \frac{\sum_{j=1}^{N_k} l_{k,j}}{N_k}$$

where $l_{k,j}$ denotes the 2D image coordinate of the k -th superpixel. We note that an optimized graph can only be as good as the extracted features it is built upon, however, the task of feature optimization, in line with superpixel segmentation, represents a problem in itself and is not the focus of this work.

B. Dynamic Graph Learning and Label Propagation

In the following, we wish to learn a superpixel HSI graph and conduct classification through label propagation. Consider the joint optimization problem over the graph \mathbf{W} and the labelling function \mathbf{F}

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{F}} \sum_{i,j} (Z_{i,j} + \gamma Z_{i,j}^F) W_{i,j} + \alpha W_{i,j}^2, \\ \text{s.t. } \sum_j W_{i,j} = 1, W_{i,j} \geq 0, \mathbf{F}_l = \mathbf{Y}_l^S. \end{aligned} \quad (4)$$

with $Z_{i,j}^F = \|\mathbf{f}_i - \mathbf{f}_j\|_2^2$, where \mathbf{f}_i denotes the i -th row of \mathbf{F} , and $\mathbf{Y}_l^S \in \mathbb{R}^{l \times c}$ is the labelled submatrix of \mathbf{Y}^S . Via alternating optimization, the optimal graph \mathbf{W} can be computed according to Eq. (3) with $Z_{i,j} \rightarrow Z_{i,j} + \gamma Z_{i,j}^F$ and by absorbing α into k , the number of edges per row, while for \mathbf{F} , this yields the solution

$$\mathbf{F}_u = -\mathbf{L}_{u,u}^{-1} \mathbf{L}_{u,l} \mathbf{Y}_l^S \quad (5)$$

for the unlabelled superpixel nodes, which is also known as harmonic label propagation [20], [34]. Here, $\mathbf{L}_{u,u}$ denotes the graph Laplacian submatrix with rows and columns corresponding to unlabelled nodes. The final superpixel labels are assigned via the decision function

$$y_i = \arg \max_j F_{i,j}, \quad \forall j = 1, \dots, c. \quad (6)$$

Accordingly, we require the optimal graph to exhibit smoothness with respect to both the pre-designed superpixel features, as extracted from the HSI data, as well as the labelling function \mathbf{F} . The functional $\text{Tr}(\mathbf{F}^T \mathbf{L} \mathbf{F}) = \frac{1}{2} \sum_{i,j} Z_{i,j}^F W_{i,j}$, given \mathbf{L} , has been employed as a standalone graph classifier (e.g. [4], [34]), while in the present framework, it is further utilized to enrich the graph construction process, rendering it dynamic.

Let \mathbf{Z}^v with $Z_{i,j}^v = \|\mathbf{s}_i^v - \mathbf{s}_j^v\|_2^2$ denote the Euclidean distance matrix between superpixel feature vectors of type (or view) $v \in \{M, S, C\}$, as previously defined in Sect. IV-A, and c_v the feature weight. In a variation of the above, we propose to employ the superpixel multi-feature dynamic graph G of the basic form as in Eq. (3) with weighted pairwise distances $Z_{i,j} = \sum_v c_v Z_{i,j}^v$ and $Z_{i,j}^{\tilde{F}} = \|\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_j\|_2^2$, the latter of which we define as *pseudo-label features*

$$\tilde{\mathbf{f}}_j[j] = \sum_k P_{i,k}^0 Y_{k,j}^S, \quad j = 1, \dots, c$$

where $\tilde{\mathbf{f}}_i$ denotes the i -th row of $\tilde{\mathbf{F}} \in \mathbb{R}^{N \times c}$, \mathbf{W}_0 the initial graph based on $Z_{i,j}$ and $\mathbf{P}^0 = \mathbf{D}_0^{-1}\mathbf{W}_0$ its random-walk normalized version. In particular, $\tilde{\mathbf{F}}$ constitutes one instance of a random walk, as opposed to the fully converged solution in Eq. (5). The motivation behind this construction is to provide a soft pre-labelling approach by propagating given labels among their nearest neighbors, as determined through an initial superpixel graph \mathbf{W}_0 based solely on HSI superpixel features, thereby merging information from the label and superpixel dependencies. The resulting distance matrix $\mathbf{Z}^{\tilde{\mathbf{F}}}$ is then utilized as an additional component to rebuild the graph, with large values penalizing nodes not in the same class, further rendering the graph construction dynamic and implicit. Final label propagation on this graph is conducted via the converged harmonic solution in Eq. (5) and class assignment via Eq. (6). We summarize the approach in Algorithm 1.

Algorithm 1 Multi-Feature Graph SSL for HSI Data

- 1: **INPUT:** raw HSI cube \mathbf{I} , label matrix \mathbf{Y} , parameters: $k, \{c_v\}_v, \gamma$.
 - 2: **OUTPUT:** Classification map \mathbf{F} .
 - 3: Apply PCA on \mathbf{I} to obtain $\tilde{\mathbf{I}}$.
 - 4: Conduct superpixel segmentation to obtain $\tilde{\mathbf{S}}$ and label regularization to obtain \mathbf{Y}^S .
 - 5: Extract superpixel features $\{\mathbf{s}^M, \mathbf{s}^S, \mathbf{s}^C\}$.
 - 6: Compute initial superpixel-feature graph \mathbf{W}_0 with $Z_{i,j} = \sum_v c_v Z_{i,j}^v$, pre-set c_v , $v \in \{S, M, C\}$, via Eq. (3) & symmetrize.
 - 7: Compute pseudo-label features $\tilde{\mathbf{F}} = \mathbf{P}^0 \mathbf{Y}^S$ and $\mathbf{Z}^{\tilde{\mathbf{F}}}$.
 - 8: Update graph with $\tilde{Z}_{i,j} = Z_{i,j} + \gamma Z_{i,j}^{\tilde{\mathbf{F}}}$ and symmetrize.
 - 9: Compute unknown labels with graph classifier $\mathbf{F}_u = -\mathbf{L}_{u,u}^{-1} \mathbf{L}_{u,l} \mathbf{Y}_l^S$.
 - 10: Assign final classes via Eq. (6).
-

Remark: The RBF kernel with $W_{i,j} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{\sigma^2}\right)$ constitutes a prominent approach to model graph weights [18] and is incidentally the result of the optimization problem in Eq. (1) with $f(\mathbf{W}) = \sigma^2 \sum_{i,j} W_{i,j} \log W_{i,j}$ in normalized form. Nevertheless, performance is strongly affected by the tuning of σ and graph connectivity, the latter of which is not embedded into the graph solution and both of which are non-trivial.

While the proposed approach simplifies the issue of tuning edge connectivity and neighbourhood range, thus increasing robustness, it still comprises a range of parameters, which are categorized into: k (the number of nearest neighbour edges), $\{c_v\}_v$ (superpixel feature weights), and γ (pseudo-label feature weight). As the goodness of the graph is dependent on the goodness of its features, we proceed to simultaneously learn feature weights and update the graph with the goal to help guide as well as minimize uninformed parameter-tuning.

C. Parameter-optimal Multi-feature Graph Learning

While a reduction of parameters generally occurs at the sacrifice of performance and cannot replace a thorough parameter search, we investigate the possibility of a pseudo-label guided parameter reduction and further propose a variation of the

preceding framework, inspired by [19], in an effort to facilitate training and generalizability to diverse datasets.

Consider individual superpixel feature graphs, denoted with \mathbf{A}^v , whose entries are computed from $Z_{i,j}^v$ using Eq. (3) with k edges per row, and initialize the global graph $\mathbf{W} = \sum_{v=1}^V c_v \mathbf{A}^v$ with $c_v = \frac{1}{V}$. We assume that the deviation $r^v = \|\mathbf{W} - \mathbf{A}^v\|_F^2$ from \mathbf{W} is inversely related to the feature importance c_v . After an initial pseudo-label computation $\tilde{\mathbf{F}}$ using Eq. (5), we update the global graph \mathbf{W} by solving

$$\min_{\mathbf{W}_i \geq 0, \mathbf{W}_i \mathbf{1} = 1} \left\| \mathbf{W}_i + \frac{(\gamma_1/2) \mathbf{Z}_i^{\mathbf{W}^F} - \sum_v c_v \mathbf{A}_i^v}{\sum_v c_v} \right\|_2^2 \quad (7)$$

In contrast to the previous method, we determine the pseudo-labels $\tilde{\mathbf{F}}$ here via the converged harmonic solution in Eq. (5) and apply a binary mask W which holds the nonzero locations of the current global graph estimate \mathbf{W} , giving the Hadamard product $\mathbf{Z}^{\mathbf{W}^F} = W \circ \mathbf{Z}^{\tilde{\mathbf{F}}}$. As such, when solving Eq. (7) the positions of the non-zero graph weights remain fixed, while their values are perturbed, i.e. re-weighted or eliminated according to pseudo-label information. This prevents the formation of noisy edges when $\mathbf{Z}^{\tilde{\mathbf{F}}}$ is dense, and constitutes an alternative to soft pseudo-labelling (for which we previously employed one random walk). Subsequently, we regularize the weights c_v via an l_2 -norm term:

$$\min_{\mathbf{c}} \sum_v c_v \|\mathbf{W} - \mathbf{A}^v\|_F^2 + \gamma_2 \|\mathbf{c}\|_2^2, \text{ s.t. } c_v \geq 0, \mathbf{c}^T \mathbf{1} = 1 \quad (8)$$

which can be simplified to

$$\min_{c_v \geq 0, \mathbf{c}^T \mathbf{1} = 1} \left\| \mathbf{c} + \frac{\mathbf{r}}{2\gamma_2} \right\|_2^2. \quad (9)$$

Notably, Eq. (2) can be written in the same form as Eqs. (7) and (9), which constitute the Euclidean projection onto the probabilistic simplex [35]; however, as we cannot apply the same kNN simplification, we solve the latter two iteratively, whereby we employ Newton's method to enforce the unity sum constraint. As such, both the graph edge learning and feature weight learning stage are essentially the same.

By recomputing $\mathbf{Z}^{\mathbf{W}^F}$ and then \mathbf{W} with corresponding parameter γ_3 , we obtain a pseudo-label enhanced graph which is used by the graph classifier of Eq. (5) to obtain the final solution.

Overall, γ_2 controls the disparity between feature weights, while replacing V parameters with one, while γ_1 and γ_3 regulate the pseudo-label contribution at different stages. Here we employ pseudo-labels in a two-fold way to inform feature contribution as well as to form a separate feature embedded in the graph. While this can be tuned with a single parameter γ_1 , in practice, we observe that performance benefits from weighting the steps separately by introducing γ_3 , as we will demonstrate in Sect. V. We summarize the graph learning stage of the approach in Algorithm 2; here, each computed graph is a posteriori symmetrized via $\frac{\mathbf{W} + \mathbf{W}^T}{2}$. The approach is similar to solving the joint optimization problem of Eqs. (4) (with respect to \mathbf{F}) and (8) alternately, as each step constitutes an optimal closed-form solution; however, we refrain from further iterations between steps 4 and 6 to limit possible noise resulting from pseudo-labelling.

We note that large deviations in (feature) scales, and thus in \mathbf{r} , result in binary weights (i.e. single-feature selection), which can be remedied, in part, by tuning γ_2 , as well as by refining the feature selection. For this approach, we introduce two composite superpixel feature measures, which merge the centroidal feature, which is less informative for a standalone feature graph, with either of the spectral-content features. Specifically, we consider the multiplicative $Z_{i,j}^v \circ Z_{i,j}^C$ and additive $Z_{i,j}^v + \lambda Z_{i,j}^C$, $v \in \{M, S\}$, with v chosen as per dataset and $\lambda \sim \sigma_v / \sigma_C$, where $\sigma_v = \sum_{i,j} Z_{i,j}^v / N^2$ denotes the scale per feature.

Algorithm 2 Parameter-optimal Multi-Feature Graph SSL

- 1: Initialize superpixel-feature graphs \mathbf{A}^v with \mathbf{Z}^v via Eq. (3) and $\mathbf{W}^0 = \sum_v c_v \mathbf{A}^v$ with $c_v = \frac{1}{V}$, then symmetrize.
 - 2: Compute pseudo-label features $\tilde{\mathbf{F}}$ via Eq. (5) and $\mathbf{Z}^{WF} = \mathbf{W} \circ \mathbf{Z}^{\tilde{\mathbf{F}}}$.
 - 3: Update graph with $\tilde{Z}_{i,j} = (\frac{\gamma_1}{2} Z_{i,j}^{WF} - \sum_v c_v A_{i,j}^v) / \sum_v c_v$ with pre-set γ_1 by solving Eq. (7), symmetrize.
 - 4: Compute feature weights c_v with pre-set γ_2 via Eq. (9).
 - 5: Update pseudo-labels $\tilde{\mathbf{F}}$ via Eq. (5).
 - 6: Update graph with pre-set pseudo-label weight γ_3 in Eq. (7), then symmetrize.
 - 7: Compute final labels via graph classifier $\mathbf{F}_u = -\mathbf{L}_{u,u}^{-1} \mathbf{L}_{u,l} \mathbf{Y}_l^S$ and Eq. (6).
-

Remark: One could further consider the constraint $r^v = \sum_{i,j} Z_{i,j}^v W_{i,j}$ in Eq. (8) as a non-separable way to estimate the feature weights, however, we observe that discrepancies in scaling render the parameter γ_2 more difficult to tune. Instead, we opt to separate the graph into the sum of individual feature graphs. While this bears the bias of reduced global interaction between the features (i.e. in Eq. (3) the sum $Z_{i,j} = \sum_v c_v Z_{i,j}^v$ drives the assignment of the nearest k edges), we remedy this by incorporating pseudo-labels into the framework as a means to perturb the solution and reinforce inter-and intra-class relations as well as by introducing composite features.

V. EXPERIMENTAL RESULTS

A. Dataset Description

We validate our approach on three benchmark HSI datasets: **Indian Pines:** This data set was gathered by an Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) over an agricultural site in Indiana and consists of 145×145 pixels with a spatial resolution of 20 m per pixel. The AVIRIS sensor has a wavelength range from 0.4 to $2.5 \mu\text{m}$, which is divided into 224 bands, of which 200 are retained for experiments. There are 16 classes, the distribution of which is imbalanced, with Alfalfa, Oats and Grass/Pasture-mowed containing relatively few labeled samples.

Salinas: The second data set was similarly acquired by the AVIRIS sensor over Salinas Valley, California, comprising 512×217 pixels with a notably higher spatial resolution of 3.7 m per pixel. Further, 204 bands are retained. The scene contains 16 classes, covering i.a. soils and fields.

University of Pavia: The third data set was collected by the Reflective Optics System Imaging Spectrometer (ROSIS),

containing 610×340 pixels with a spatial resolution of 1.3 m per pixel, the highest of the three datasets. The spectral range from 0.43 to $0.86 \mu\text{m}$ is divided into 115 spectral bands, of which 103 are retained. The urban site contains 9 classes and covers the Engineering School at the University of Pavia.

B. Experimental Design

In the following, we conduct the experimental evaluation of our method on the benchmarking datasets and demonstrate its superiority compared to state-of-the-art algorithms. For evaluation, experiments are repeated 10 times and performance is assessed on the basis of the average and standard deviation of three quality indices: the overall accuracy (OA), as the percentage of correctly classified pixels, the average accuracy (AA), as the mean of the percentage of correctly classified pixels per class, and the Kappa Coefficient (κ), as the percentage of correctly classified pixels corrected by the number of agreements expected purely by chance. Further, we compare performance with the Local Covariance Matrix Representation (LCMR) [9], the Edge-Preserving Filtering (EPF) [16], the Image Fusion and Recursive Filtering (IFRF) [7], and the SVM [5] methods, which, as established state-of-the-art methods, were specifically chosen as comparisons due to their inherent spectral-spatial modelling techniques (with exception of the purely spectral SVM). The SVM algorithm is implemented in the LIBSVM library [36], adopting the Gaussian kernel with fivefold cross validation for the classifier. We further adopt model-parameters as specified in these works. The proposed methods are abbreviated as Multi-Feature Graph Learning (MGL) and Parameter-optimal Multi-Feature Graph Learning (PMGL) respectively.

C. Parameter Specification

For both proposed methods, we conduct PCA on the standardized data to explain 99.8% of the data's variance and employ SLIC [32] for superpixel segmentation with a compactness of 10, where we fix the number K of superpixels per dataset to $K = 1287$ (Indian Pines), $K = 2237$ (Salinas) and $K = 3080$ (University of Pavia). For all graph constructions, we set $k = 10$ as the number of edges per node, and for the spatial mean feature construction, we select $h = 15$.

For the first method (MGL), we employ pseudo-label feature weight $\gamma = 10$ and superpixel feature weights $c_S = 1$, $c_M = 0.5$, $c_C = 10^{-2}$ (Indian Pines), $c_S = 0.5$, $c_M = 5$, $c_C = 10^{-5}$ (University of Pavia), and $c_S = 1$, $c_M = 0.1$, $c_C = 10^{-4}$ (Salinas).

For the second method (PMGL), we employ $\gamma_1 = 0$, $\gamma_2 = 30$, $\gamma_3 = 1$ (Indian Pines), $\gamma_1 = 20$, $\gamma_2 = 40$, $\gamma_3 = 0$ (University of Pavia) and $\gamma_1 = 1$, $\gamma_2 = 30$, $\gamma_3 = 1$ (Salinas). Further, for the latter method we construct three individual feature graphs respectively based on the following feature distances: $\{\mathbf{Z}^M, \mathbf{Z}^S, \mathbf{Z}^C \circ \mathbf{Z}^S\}$ (Indian Pines), $\{\mathbf{Z}^M, \mathbf{Z}^S, \mathbf{Z}^M + \lambda \mathbf{Z}^C\}$ (University of Pavia), $\{\mathbf{Z}^S, \mathbf{Z}^M + \lambda \mathbf{Z}^C, \mathbf{Z}^S + \lambda \mathbf{Z}^C\}$ (Salinas), which were deemed to summarize most effectively the main properties of the different HSI scenes.

D. Experimental Results & Discussion

Performance is evaluated using a very limited amount of labels for training, with rates of 3–20 samples per class, which are randomly selected, in several stages of experiments. In the first instance, we conduct numerical evaluations and comparisons of classification accuracy of the proposed methods with state-of-the-art approaches, as detailed above, followed by an evaluation of the corresponding visual classification maps.

E1: We begin by evaluating the OA and Kappa coefficient in comparison using a reduced label rate of 3-20 randomly selected labels per class, whose results are graphically displayed in Fig. 2 for the three benchmark data sets. We observe that both proposed methods consistently outperform the other methods over the entire range of label rates for all three benchmark data sets. In particular, the performance gain is highest for lower label rates, signifying that the proposed pseudo-label guided graph-based methods perform strongly even when extremely few labels are available as a result of their superior model. LCMR and IFRF form the closest competitors for the Indian Pines and Salinas data sets with a maximum gap of approximately 10% and 8% respectively to the closest competitor at 3 labelled samples, with LCMR being the dominant competitor for the more complex Pavia University data set with a gap of 10% to PMGL. Further, among the two proposed methods, in the lower label limit of the Pavia dataset, PMGL exhibits up to 4% gain in OA performance, while for the other data sets, this gain is vanishingly small with the two methods performing comparably. This indicates that a more refined parameter selection can be beneficial for structurally complex data sets.

E2: To demonstrate the influence of the pseudo-labels and feature coefficients in PMGL on performance, we consider the overall accuracy at a fixed label rate of 7 samples per class with varying feature parameters over 10 trials. In particular, in order to illustrate the interaction between pseudo-label contribution and feature contribution, we fix pseudo-label parameter γ_3 , while varying γ_1 , along with the feature weight-regularization parameter γ_2 , and consider the mean OA in a 3D plot. We report results for the University of Pavia data set, as the most structurally complex of the three, in Fig. 3 with the OA plotted against γ_1 and γ_2 and $\gamma_3 = 0$ in (a), and $\gamma_3 = 1$ in (b), with (c) showing the resulting feature coefficient distribution for (a). We observe that in (a), OA is highest when both γ_1 and γ_2 are increased, generating a perturbation toward more evenly distributed coefficients, which, as shown in (c), corresponds to the gradual matching of composite coefficient $c_{M+\lambda C}$ and spectral mean coefficient c_M , while the contribution of the spatial mean coefficient c_S is negligible. When additionally, the pseudo-label feature is incorporated into the graph via γ_3 , the coefficient distribution for the best OA changes, instead overall moving toward one dominant superpixel feature. While we observe interactions between pseudo-label and superpixel features which drive performance, the feature regularization parameter is ultimately dependent on the data set and the selected features at hand.

E3: For each data set we use 7 labeled samples per class and run the methods again to calculate the OA, Kappa coefficient,

AA, and a full class by class accuracy breakdown over 10 trials. For all three data sets, the proposed methods MGL and PMGL consistently outperform the competitors in OA, Kappa coefficient and AA, and for the majority of per class accuracies. For the Indian Pines data set, PMGL performs only slightly better than MGL in the first three measures, and around 10% improvement in OA over its closest competitor, LCMR. In the case of the University of Pavia data set, the gain of PMGL over MGL is even larger with 3-4%, followed by LCMR with 10% improvement in OA. Lastly, for the Salinas data set, MGL and PMGL perform comparatively well, with a gain of 3% over their closest competitor, IFRF. Overall, we observe that while PMGL utilizes more intricate relations and selective feature contributions, MGL is still close in performance, with a gain of the former becoming more evident for increasingly complex data sets, such as Pavia University.

E4: At last, we show the full classification maps produced for training with 7 samples per class for all methods in comparison. In Fig. 4, classification maps for the Indian Pines data set illustrate increased smoothness and local homogeneity for the proposed graph-based MGL and PMGL, exemplifying their superiority. Their closest competitor, LCMR exhibits more noisy regions. Fig. 5 shows the University of Pavia classification maps, which as the most structurally complex scene of the three with some scattered classes, similarly exhibits smooth yet spatially refined classification results for the proposed MGL and PMGL, despite the inherent crudeness of superpixel segmentation and label regularization. At last, for the Salinas data set in Fig. 6, which due to its locally homogeneous regions and overall spatial regularity represents a simpler scene, the proposed methods still to manage to improve over existing methods, achieving near perfect classification.

It becomes evident that the use of superpixels, ensuring local homogeneity, as well as that of graphs, for refined local and global modelling, which incorporates pixel- and superpixel-level as well as spectral-spatial-label dependencies, facilitates a significant performance gain. Notably, EPF and IFRF employ a spectral-spatial filtering approach which can be likened to graph filtering, with the distinction that the latter is more flexible and versatile to model; nevertheless, the use of the SVM classifier in all competitors, as opposed to a graph-filter in the latter case, contributes to a comparatively decreased accuracy owing to the noise of the spectral-based SVM.

VI. CONCLUSION

In this work, we have developed a pseudo-label-guided and superpixel-based graph learning framework for semi-supervised classification of HSI data. Specifically, we have presented two methods: while the former constructs a single dynamic graph by fusing different superpixel features along with a pseudo-label feature, the latter constructs a global graph as the sum of individual feature graphs whose contribution is informed through pseudo-label regularization. We have demonstrated on the basis of benchmark data sets the superiority in performance through quantitative and qualitative results in comparison to state-of-the-art methods, particularly

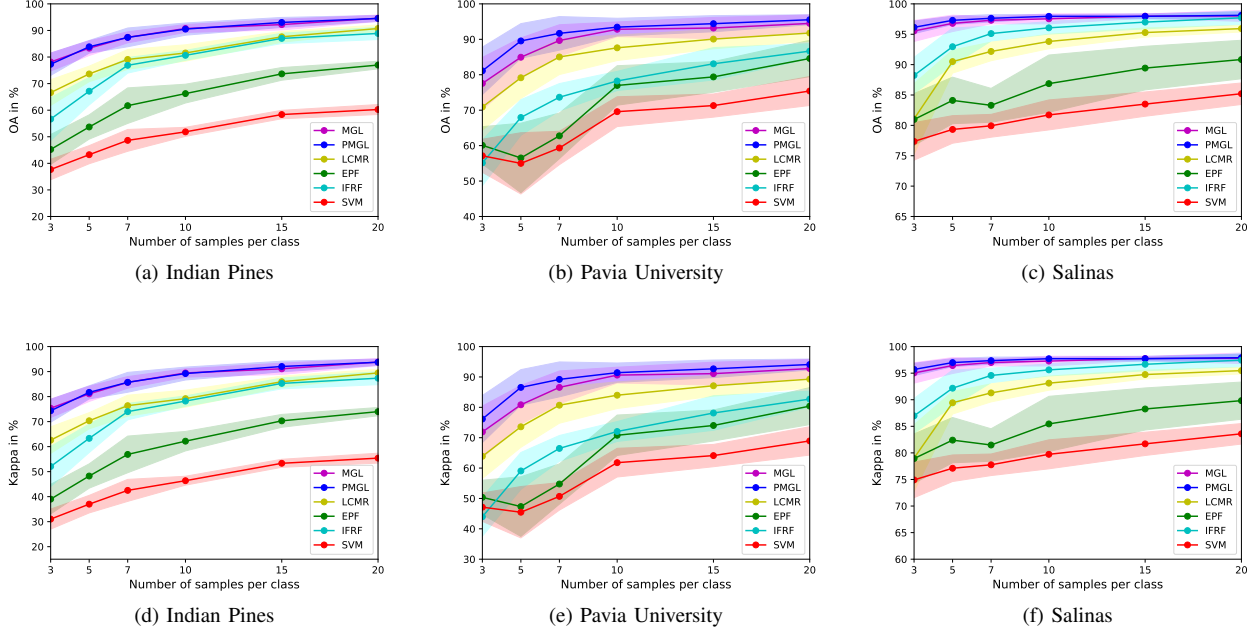


Fig. 2. Comparison of the classification accuracy (OA) and Kappa coefficient of different methods with varying number of training samples over 10 trials. The solid lines represent the mean while the shaded area covers the standard deviation from the mean.

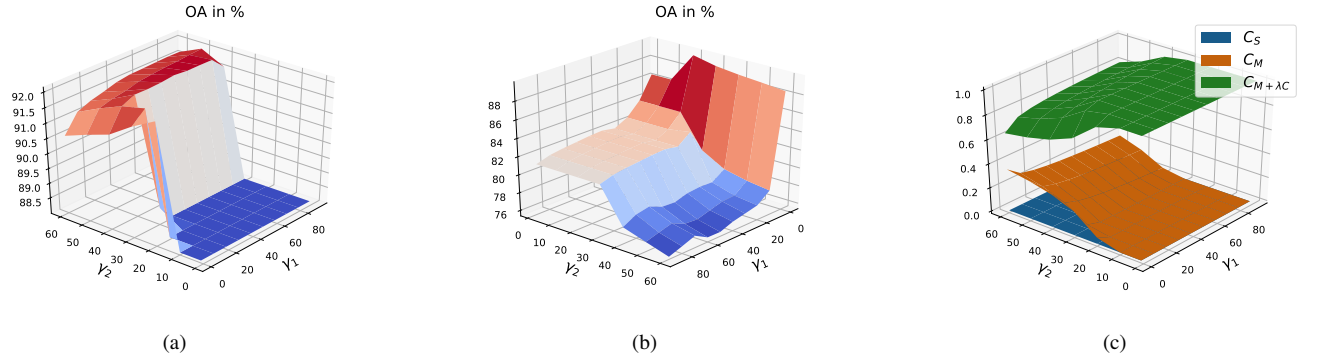


Fig. 3. University of Pavia: results for 7 training samples per class, 3D plot of pseudo-label weight γ_1 vs feature regularization weight γ_2 against OA of PMGL with (a) $\gamma_3 = 0$, (b) $\gamma_3 = 1$, and (c) feature coefficient distribution for $\gamma_3 = 0$.

in the small labelled sample limit.

The incorporation of multiple features as well as label information through pseudo-labels into the graph facilitate refined modelling of the complex dependencies present in HSI data, and ultimately leverage these for an improved classification performance. Furthermore, the multi-stage workflow, which employs superpixels and a flexible, inherently sparse graph design with the option to reduce parameters through regularization, is versatile by allowing for multiple components and exploiting multiple levels of spectral-spatial and label-dependencies. Nevertheless, it remains a challenge to completely eliminate parameters whilst maintaining competitive performance as well as to select the ideal features upon which the goodness of subsequent steps depends; ultimately, the process of feature selection/extraction and feature weighting presented is not exhaustive. In our future work, we wish to explore automated DL approaches for both feature selection

and graph construction which can be guided by sophisticated model-based priors.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe from Purdue University and the NASA Jet Propulsion Laboratory for providing the hyperspectral data sets. They would also like to thank Prof D. Coomes from the Department of Plant Sciences, University of Cambridge, for his advice and support. This research was carried out as part of the INTEGRAL project funded by GCRF (EPSRC) grant EP/T003553/1. In addition, CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Award RG98755, the Leverhulme Trust project Unveiling the invisible, the European Union Horizon 2020 research and innovation programme under the Marie

TABLE I
OA, KAPPA, AA AND PER CLASS BREAKDOWN IN % WITH 7 TRAINING LABELS PER CLASS

Indian Pines						
CLASS	MGL	PMGL	LCMR [9]	EPF [16]	IFRF [7]	SVM [5]
C1	97.95 ± 1.03	97.95 ± 1.03	99.23 ± 1.24	95.13 ± 7.69	99.23 ± 1.24	72.05 ± 13.54
C2	66.17 ± 10.17	70.63 ± 11.86	73.36 ± 8.24	39.32 ± 9.52	65.58 ± 9.98	33.12 ± 4.41
C3	76.11 ± 9.73	86.12 ± 8.26	62.08 ± 8.98	55.65 ± 12.92	76.1 ± 9.16	44.63 ± 6.07
C4	96.17 ± 4.01	92.3 ± 7.54	93.48 ± 6.45	78.13 ± 20.28	87.3 ± 7.99	48.96 ± 12.01
C5	88.84 ± 9.09	88.47 ± 8.83	88.55 ± 8.04	83.8 ± 10.94	80.8 ± 6.47	70.82 ± 10.22
C6	94.54 ± 10.67	93.43 ± 8.9	86.43 ± 8.34	90.06 ± 8.05	86.02 ± 13.87	68.49 ± 10.61
C7	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	98.1 ± 2.46	100.0 ± 0.0	83.33 ± 15.27
C8	100.0 ± 0.0	100.0 ± 0.0	97.62 ± 4.47	68.3 ± 16.46	99.62 ± 0.71	56.28 ± 9.54
C9	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	89.23 ± 12.67
C10	88.05 ± 7.41	89.97 ± 7.04	76.29 ± 11.93	61.06 ± 17.8	79.19 ± 8.29	45.67 ± 10.66
C11	87.53 ± 6.94	82.07 ± 10.0	61.55 ± 9.07	48.35 ± 13.16	64.08 ± 12.69	38.08 ± 7.69
C12	86.64 ± 12.78	93.16 ± 10.15	79.68 ± 7.3	57.59 ± 25.11	68.45 ± 12.47	40.39 ± 11.65
C13	99.49 ± 0.0	99.49 ± 0.0	99.19 ± 0.9	99.49 ± 0.0	98.59 ± 1.52	90.86 ± 4.36
C14	97.02 ± 8.58	94.42 ± 8.77	95.86 ± 5.15	76.92 ± 23.07	82.22 ± 7.23	60.51 ± 16.01
C15	94.51 ± 9.39	94.64 ± 9.29	92.32 ± 9.22	41.42 ± 13.57	86.15 ± 7.41	31.77 ± 7.52
C16	93.72 ± 1.74	93.72 ± 1.74	99.19 ± 2.57	95.23 ± 6.79	97.09 ± 4.04	85.58 ± 10.11
OA	86.75 ± 2.31	86.93 ± 2.79	77.88 ± 4.07	60.7 ± 6.92	75.77 ± 4.65	47.84 ± 3.29
Kappa	84.96 ± 2.6	85.19 ± 3.12	75.14 ± 4.51	56.16 ± 7.59	72.84 ± 5.06	42.07 ± 3.6
AA	91.67 ± 1.31	92.27 ± 1.11	87.8 ± 2.27	74.28 ± 4.98	85.65 ± 2.77	59.99 ± 3.48
University of Pavia						
CLASS	MGL	PMGL	LCMR [9]	EPF [16]	IFRF [7]	SVM [5]
C1	76.23 ± 8.91	78.42 ± 6.35	79.92 ± 9.14	76.0 ± 11.24	53.47 ± 9.97	62.63 ± 8.75
C2	92.53 ± 7.9	96.12 ± 5.36	81.38 ± 11.88	59.17 ± 10.17	78.64 ± 6.42	55.55 ± 8.96
C3	87.5 ± 10.43	90.19 ± 8.26	83.98 ± 10.72	60.88 ± 13.2	55.76 ± 10.56	55.05 ± 9.57
C4	84.03 ± 7.83	84.76 ± 6.85	93.72 ± 4.43	86.49 ± 5.0	65.45 ± 16.27	88.46 ± 5.49
C5	93.27 ± 6.35	93.68 ± 5.89	97.2 ± 7.54	96.76 ± 10.23	99.13 ± 0.62	96.63 ± 8.49
C6	94.94 ± 6.51	98.23 ± 2.63	84.82 ± 8.41	63.46 ± 13.17	83.4 ± 7.23	57.34 ± 9.54
C7	96.02 ± 1.84	99.46 ± 0.26	85.15 ± 12.58	94.32 ± 9.97	77.41 ± 14.24	86.22 ± 8.86
C8	82.41 ± 19.95	90.71 ± 5.06	71.03 ± 9.2	74.57 ± 15.13	71.8 ± 10.57	64.32 ± 10.18
C9	92.96 ± 3.37	95.9 ± 2.03	94.46 ± 2.06	99.78 ± 0.56	50.54 ± 10.7	99.66 ± 0.58
OA	88.7 ± 3.56	92.08 ± 2.48	82.58 ± 5.74	68.81 ± 4.31	72.63 ± 3.09	63.15 ± 3.64
Kappa	85.3 ± 4.44	89.61 ± 3.11	77.77 ± 6.74	61.27 ± 4.75	64.86 ± 3.7	54.59 ± 3.77
AA	88.87 ± 2.35	91.94 ± 1.6	85.74 ± 2.64	79.05 ± 4.05	70.62 ± 2.78	73.98 ± 2.75
Salinas						
CLASS	MGL	PMGL	LCMR [9]	EPF [16]	IFRF [7]	SVM [5]
C1	100.0 ± 0.0	100.0 ± 0.0	99.89 ± 0.21	99.77 ± 0.5	98.23 ± 5.61	97.85 ± 1.26
C2	100.0 ± 0.0	100.0 ± 0.0	89.79 ± 5.67	99.74 ± 0.4	96.36 ± 4.47	98.14 ± 1.42
C3	100.0 ± 0.0	100.0 ± 0.0	98.55 ± 2.5	86.26 ± 19.94	99.97 ± 0.08	81.24 ± 18.02
C4	94.18 ± 5.92	94.17 ± 8.91	100.0 ± 0.0	99.88 ± 0.13	98.61 ± 2.63	99.15 ± 0.59
C5	93.28 ± 4.31	94.72 ± 0.0	96.53 ± 0.69	97.62 ± 1.81	92.43 ± 5.1	96.07 ± 2.37
C6	99.59 ± 0.07	99.59 ± 0.06	99.02 ± 0.99	99.51 ± 0.79	99.53 ± 0.65	98.02 ± 2.1
C7	99.89 ± 0.06	100.0 ± 0.0	97.9 ± 2.15	99.79 ± 0.07	98.82 ± 3.33	98.57 ± 0.93
C8	98.45 ± 0.71	97.46 ± 1.84	85.58 ± 5.54	64.96 ± 18.35	85.93 ± 7.01	57.35 ± 11.09
C9	100.0 ± 0.0	100.0 ± 0.0	92.23 ± 9.51	99.49 ± 0.61	99.94 ± 0.15	98.23 ± 0.91
C10	89.47 ± 7.41	90.37 ± 5.68	97.44 ± 1.37	89.24 ± 9.66	97.16 ± 3.84	80.48 ± 9.14
C11	95.57 ± 5.16	97.67 ± 0.85	99.91 ± 0.08	97.15 ± 2.03	96.03 ± 3.32	87.51 ± 3.7
C12	97.59 ± 0.31	97.59 ± 0.31	99.23 ± 2.38	100.0 ± 0.0	98.3 ± 1.45	96.18 ± 3.08
C13	97.67 ± 0.62	97.48 ± 0.36	98.26 ± 0.85	98.92 ± 0.38	95.73 ± 5.84	98.22 ± 0.6
C14	92.91 ± 4.8	94.65 ± 2.95	93.13 ± 5.84	95.3 ± 1.66	96.32 ± 3.28	88.86 ± 2.69
C15	96.32 ± 1.28	98.23 ± 0.97	82.71 ± 7.54	71.74 ± 21.97	93.98 ± 3.76	59.05 ± 12.78
C16	100.0 ± 0.0	100.0 ± 0.0	91.23 ± 7.49	94.71 ± 4.17	98.12 ± 3.02	92.33 ± 4.68
OA	97.67 ± 0.48	97.93 ± 0.45	91.99 ± 1.34	87.14 ± 3.22	94.89 ± 1.7	81.98 ± 1.54
Kappa	97.41 ± 0.54	97.7 ± 0.5	91.1 ± 1.49	85.73 ± 3.57	94.32 ± 1.89	80.01 ± 1.69
AA	97.18 ± 0.69	97.62 ± 0.65	95.09 ± 1.01	93.38 ± 2.24	96.59 ± 0.85	89.2 ± 1.5

Skodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

REFERENCES

- [1] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 3, pp. 1579–1597, 2018.
- [2] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005. [Online]. Available: http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
- [3] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 977–986. [Online]. Available: <https://doi.org/10.1145/2623330.2623726>
- [4] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information*

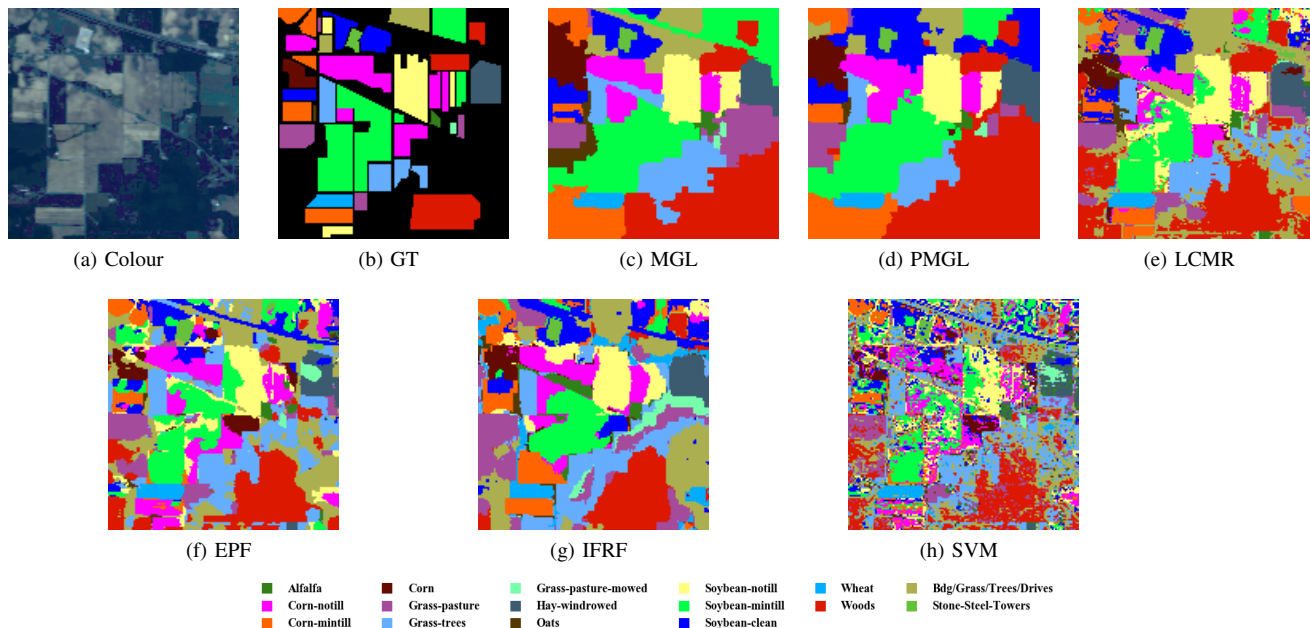


Fig. 4. Indian Pines: (a) Colour composite, (b) Ground truth, (c)-(h) classification maps produced using 7 labelled samples per class.

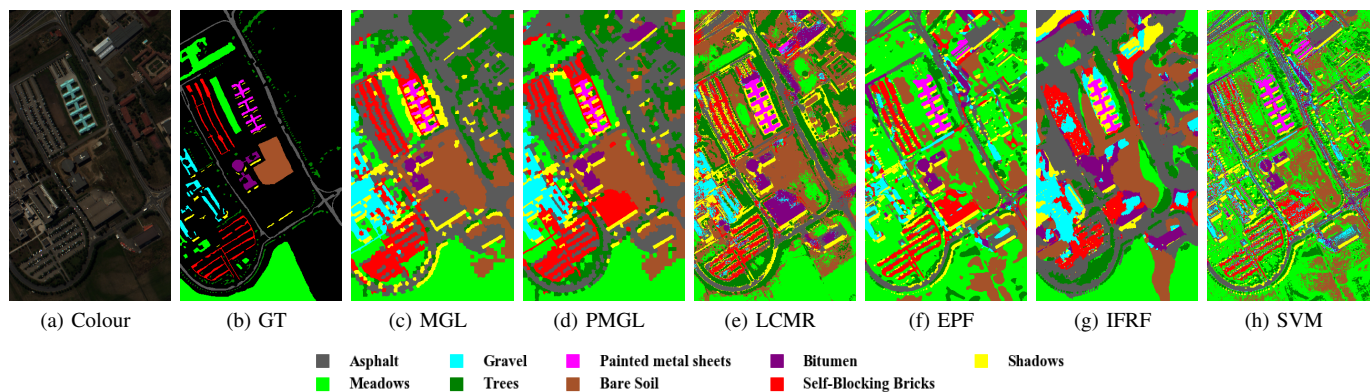


Fig. 5. Pavia University: (a) Colour composite, (b) Ground truth, (c)-(h) classification maps produced using 7 labelled samples per class.

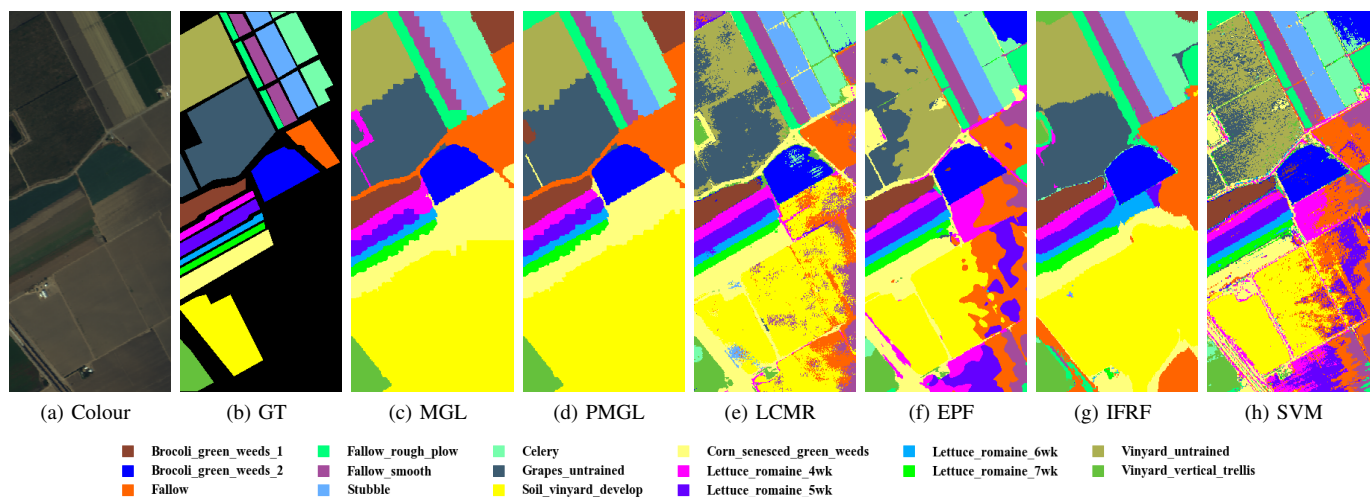


Fig. 6. Salinas: (a) Colour composite, (b) Ground truth, (c)-(h) classification maps produced using 7 labelled samples per class.

- Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16. MIT Press, 2004.
- [5] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
 - [6] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 12, pp. 6663–6674, 2015.
 - [7] X. Kang, S. Li, and J. A. Benediktsson, "Feature extraction of hyperspectral images with image fusion and recursive filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 6, pp. 3742–3752, 2014.
 - [8] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3681–3693, 2015.
 - [9] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial–spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3534–3546, 2018.
 - [10] Z. He, Y. Wang, and J. Hu, "Joint sparse and low-rank multitask learning with laplacian-like regularization for hyperspectral classification," *Remote Sensing*, vol. 10, no. 2, 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/2/322>
 - [11] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3973–3985, 2011.
 - [12] G. Camps-Valls, T. V. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044–3054, 2007.
 - [13] P. Sellars, A. I. Aviles-Rivero, and C. Schönlieb, "Superpixel contracted graph-based learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4180–4193, 2020.
 - [14] B. Cui, X. Xie, S. Hao, J. Cui, and Y. Lu, "Semi-supervised classification of hyperspectral images based on extended label propagation and rolling guidance filtering," *Remote Sensing*, vol. 10, no. 4, 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/4/515>
 - [15] B. Cui, X. Xie, X. Ma, G. Ren, and Y. Ma, "Superpixel-based extended random walker for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3233–3243, 2018.
 - [16] X. Kang, S. Li, and J. A. Benediktsson, "Spectral–spatial hyperspectral image classification with edge-preserving filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2666–2677, 2014.
 - [17] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 35–49, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271619302199>
 - [18] C. A. R. de Sousa, S. O. Rezende, and G. E. A. P. A. Batista, "Influence of graph construction on semi-supervised learning," in *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–175.
 - [19] F. Nie, J. Li, and X. Li, "Self-weighted multiview clustering with multiple graphs," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2564–2570. [Online]. Available: <https://doi.org/10.24963/ijcai.2017/357>
 - [20] F. Nie, G. Cai, J. Li, and X. Li, "Auto-weighted multi-view learning for image clustering and semi-supervised classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1501–1511, 2018.
 - [21] B. Wang and J. Tsotsos, "Dynamic label propagation for semi-supervised multi-class multi-label classification," *Pattern Recognition*, vol. 52, pp. 75–84, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320315003738>
 - [22] W. Liu and S. Chang, "Robust multi-class transductive learning with graphs," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 381–388.
 - [23] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 241–245, 2019.
 - [24] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8246–8257, 2020.
 - [25] H. Wang and J. Leskovec, "Unifying graph convolutional neural networks and label propagation," 2020.
 - [26] B. Jiang, Z. Zhang, D. Lin, J. Tang, and B. Luo, "Semi-supervised learning with graph learning-convolutional networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 305–11 312.
 - [27] M. S. Kotzagiannidis and M. E. Davies, "Analysis vs synthesis with structure – an investigation of union of subspace models on graphs," <https://arxiv.org/pdf/1811.04493.pdf>, 2018.
 - [28] J. R. Stevens, R. G. Resmini, and D. W. Messinger, "Spectral-density-based graph construction techniques for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5966–5983, 2017.
 - [29] V. Kalofolias, "How to learn a graph from smooth signals," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Gretton and C. C. Robert, Eds., vol. 51. Cadiz, Spain: PMLR, 09–11 May 2016, pp. 920–929. [Online]. Available: <http://proceedings.mlr.press/v51/kalofolias16.html>
 - [30] X. Dong, D. Thanou, M. Rabbat, and P. Frossard, "Learning graphs from data: A signal representation perspective," *IEEE Signal Processing Magazine*, vol. 36, no. 3, pp. 44–63, 2019.
 - [31] I. Jolliffe, *Principal Component Analysis*. American Cancer Society, 2005. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa501>
 - [32] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012. [Online]. Available: <https://doi.org/10.1109/TPAMI.2012.120>
 - [33] S. Jia, X. Deng, J. Zhu, M. Xu, J. Zhou, and X. Jia, "Collaborative representation-based multiscale superpixel fusion for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7770–7784, 2019.
 - [34] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, pp. 912–919.
 - [35] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 272–279. [Online]. Available: <https://doi.org/10.1145/1390156.1390191>
 - [36] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, May 2011. [Online]. Available: <https://doi.org/10.1145/1961189.1961199>