# Salient Band Selection for Hyperspectral Image Classification via Manifold Ranking

Qi Wang, *Senior Member, IEEE*, Jianzhe Lin, *Student Member, IEEE*, and Yuan Yuan, *Senior Member, IEEE*

*Abstract*—Saliency detection has been a hot topic in recent years, and many efforts have been devoted in this area. Unfortunately, the results of saliency detection can hardly be utilized in general applications. The primary reason, we think, is unspecific definition of salient objects, which makes that the previously published methods cannot extend to practical applications. To solve this problem, we claim that saliency should be defined in a context and the salient band selection in hyperspectral image (HSI) is introduced as an example. Unfortunately, the traditional salient band selection methods suffer from the problem of inappropriate measurement of band difference. To tackle this problem, we propose to eliminate the drawbacks of traditional salient band selection methods by manifold ranking. It puts the band vectors in the more accurate manifold space and treats the saliency problem from a novel ranking perspective, which is considered to be the main contributions of this paper. To justify the effectiveness of the proposed method, experiments are conducted on three HSIs, and our method is compared with the six existing competitors. Results show that the proposed method is very effective and can achieve the best performance among the competitors.

*Index Terms*—Band selection, deep learning, hyperspectral image (HSI) classification, manifold ranking (MR), saliency, stacked autoencoders (SAEs).

## I. Introduction

SALIENCY detection [1]–[3] has been a hot topic in the field of vision community. The widespread attention of this technique is mainly due to its importance and effectiveness in image/video processing. Since the aim of saliency detection is to extract the noticeable objects, subsequent processing based on the detection results can focus the limited computational resources on the interesting targets and ignore the unnecessary backgrounds. Therefore, saliency detection is generally considered as an efficient preprocessing step or high-level feature extraction procedure in many applications [4]–[6], such as multimedia retrieval, object detection, video compression, and image resizing.

Though tremendous effort toward saliency detection has been made, and many works achieve manifest performance on public data sets, the general-purpose saliency detection is still far from satisfying. One primary reason, we think, is the unclear definition of saliency, which makes the task-specific application less compatible. Considering this deficiency, we restrict the scope of saliency interpretation to a specific area, hyperspectral image (HSI) classification [7], [8]. In this case, the examined HSI consists of hundreds of spectral bands for the same scene, enabling accurate discrimination of different land cover materials. Generally speaking, the finer spectral resolution will lead to better discriminative ability. But that does not mean more data are always needed. For one thing, the huge volume of data implies large computational complexity. For another, the low number of labeled data in HSI is prone to result in the Hughes phenomenon [9]. Therefore, a balance between the spectral number and the classification performance is highly desired. In order to make a good compromise, the salient bands from the entire hyperspectral volume should be selected, extracting the most informative clues and abandoning the less important ones [10]. Clearly, the saliency formulation in this context has a meaningful explanation and the salient band definition is the key to success.

Actually, the salient band selection is not the only means for reducing the hyperspectral data redundancy. Feature extraction is also an alternative. But differently, feature extraction techniques transform the original high-dimensional data to low-dimensional data through certain kinds of projections, such as principle component analysis [11], independent component analysis [12], and discrete wavelet transform [13]. This treatment cannot preserve the physical meaning of the original bands, which might be very crucial for some geology analysis. Considering this fact, the salient band selection is usually preferred for the HSI processing.

Although many efforts toward the salient band selection have been made during the past decades, the selection criterions are mainly based on the similarity metrics. One disadvantage of this strategy is that most of these criterions are computed under the assumption that the original data lie on the Euclidean space. But this is not the case from time-to-time because the data may lie along a low-dimensional manifold embedded in a high-dimensional space, where the

low-dimensional space reflects the underlying parameters and a high-dimensional space is the feature space [14]. Therefore, the traditional treatment is inappropriate. In order to make a more fair measurement, we tackle this problem in the intrinsic manifold structure collectively revealed by a great amount of data. The proposed method is named as salient band selection via manifold ranking (MR). Three contributions are claimed in this paper.

1) Propose a novel method of MR-based band selection. Instead of rating the similarities in the Euclidean space, the manifold structure is taken into consideration to properly assess the hyperspectral data structure. The associated measurement is input to a ranking operation and a subsequent band selection is based on the obtained ranking score. This is a novel alternative that reformulates the hyperspectral band selection as a ranking problem.

2) Estimate the interband distance in a batch manner. Most existing techniques for band selection always compute the distance between two individual bands. The calculated results then serve as guidance for band selection. However, this strategy is not suitable for the sequential selection because the selected band at this time might resemble the one selected at previous time. In our implementation, we treat the already selected batch of bands as the query, and the examined band is compared with the whole batch. This can ensure the further selected band is distinct with the previously selected ones.

3) Provide a thorough comparison using different band selection methods and classifiers. In order to validate the effectiveness of the proposed method, we compare it with several recently presented methods. Besides, we also test these methods on typical classifiers that are frequently used for HSI classification. These experimental comparisons are meaningful references for other researchers. Furthermore, the possibility of popular deep learning technique as a classifier is also discussed in this paper.

The rest of this paper is organized as follows. Section II briefly reviews the existing works on the topic of salient band selection. Section III gives a detailed description of the proposed method, which is salient band selection via MR. Section IV presents the experimental results to justify the effectiveness of this paper. Section V extends the work to the deep learning framework and discusses the associated problems. Finally, the conclusion is drawn in Section VI.

## II. RELATED WORK

In this section, we will review three topics that are closely related to this paper. First, salient band selection is primarily investigated, which is directly the related work of the proposed method. Then, saliency detection in computer vision and HSI classification are briefly introduced. These two points are not the main focus of this paper but have certain relations.

### A. Salient Band Selection

The salient band selection is enabled by two key elements. The first one is an appropriate criterion, which measures the difference between bands. For this purpose, the existing methods investigate various indices [15]–[17], such as mutual information (MI), spectral angle mapper distance, information divergence, correlation coefficient, and volume-based distance. Apart from these traditional measures, other new criteria are also presented recently, including sparse reconstruction error and approximate MI.

The second one is the selection strategy. According to the band selection procedure, existing works can be divided into two categories: 1) sequential selection and 2) batch selection. For the sequential selection, the desired bands are chosen one by one, instead of all at a time. For example, Chang et al. [18] construct a loading factor by eigenvalue and eigenvector analysis to rank bands according to the effectiveness of their classification abilities. After this band prioritization step, the spectral association is then decorrelated to eliminate the similar bands among the selected bands. Ball and Bruce [19], [20] first use a forward selection to choose the band set enabling maximum Receiver Operating Characteristic curve (ROC) curve area $A_z$. Then, a backward rejection is enforced to remove the bands that cannot help the $A_z$ criterion. In the end, the selected bands are input to the level-set segmentation process. Du and Yang [21] employ a strategy similar to that in [18]–[20]. But the criterion for selecting band is different, in which the potential band is determined as the one with the largest linear reconstruction error from existing bands. Yang et al. [22] select the bands with an incremental manner. They initially choose one band and project others to its orthogonal subspace. Then, the band with maximum projection, represented by a stochastic feature, is selected. This procedure is repeated until the desired number of bands is obtained.

As for the batch selection, more distinctive techniques are designed. Sun et al. [23] pay more attention to the band quality instead of band information. They introduce a new index to measure the quality of a data cube by combining the noise adjusted principle components with maximum determinant of covariance matrix. Based on this, the minimum noise band selection method is proposed, aiming at selecting bands with high quality. The selection process begins with full bands set, followed by removing bands successively. Venkataraman et al. [24] use the manual grouping and the automated grouping to divide the original bands into subsets. Subsequently, features are extracted by the supervised and unsupervised methods. Martínez-Usó et al. [16] utilize a similar strategy. But the clustering method is hierarchical, which can ensure producing the minimum variance partition. Then, the representative band for each group is selected as the one that is least correlated with the others. Xia et al. [25] divert from the general idea of selecting the bands that are distinctive from others. Instead, they construct a graph network using the image pixels and choose bands that can form the most approximate network construction compared with the original data. Chang and Wang [26] propose a constrained energy minimization (CEM) for band selection. It linearly constrains the examined band image while minimizing the interfering effects caused by the remaining bands. Yuan et al. [27] present an evolutionary immune clone strategy to handle
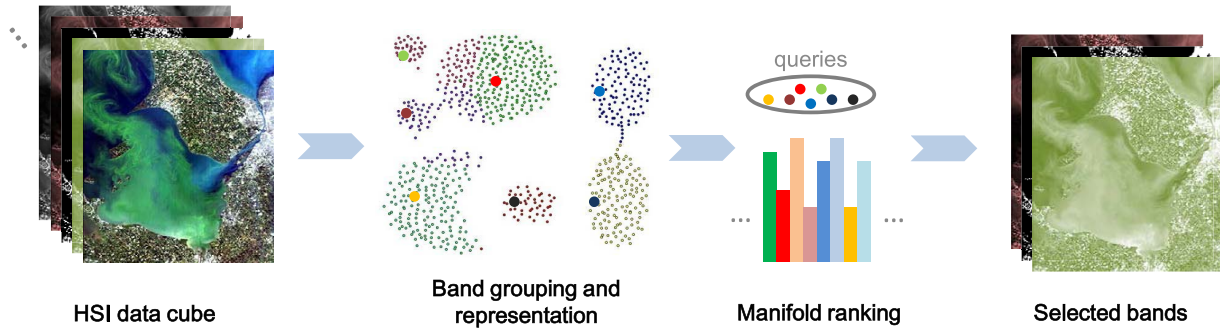
Fig. 1.   Flowchart of the proposed salient band selection method via MR.

the computational burden of possible band combinations. Besides, a novel multitask sparsity pursuit-based criterion is adopted to evaluate the performance of each candidate band set. Similar work can be found in [17], who use the trivariate MI to approximate the MI measure. In addition, a clone search strategy is taken to find a good solution with low time and space cost. The work of Chang *et al.* [28] is also of this type, while parallel simulated annealing is adopted.

Besides the two aspects, other researchers dig alternatively from the related topics of a salient band selection. Demir and Erturk [29] concentrate on reducing the computational complexity before formal band selection. They use one-bit transform of each band to remove the noisy and less discriminative bands, which is decided via the number of transitions in the one-bit map. Yang *et al.* [30] focus on the parallel implementations via emerging general-purpose graphics processing units.

### B. Saliency Detection in Computer Vision

Saliency detection is a hot topic in recent years in the computer vision community. Since the pioneer work of Koch and Ullman [53], numerous methods have been presented. In general, these methods can be categorized into two classes [31]: 1) contrast-based computational model and 2) learning-based adaptive model. For the first type, the saliency model is defined beforehand in terms of the color/texture contrast. No matter what the input image is, the saliency value is calculated with the same predefined formula. For the second type, the machine learning techniques are usually employed to adaptively train a saliency model in a particular data set. With respect to different training set, the learned model parameters might differ with each other. However, despite the extensive research in these two paradigms, few works have concentrated on the specific definition of saliency detection.

### C. Hyperspectral Image Classification

HSI classification has been researched a lot in remote sensing field. Traditional methods primarily focus on the spectral feature with adjusted classifier, such as support vector machines (SVMs) [32] and *k*-nearest neighborhood (kNN) [33], [34]. However, only utilizing the spectral clue achieves a limited performance. The joint spectral and spatial classification techniques [35] have recently attracted

much more attention because they can considerably overcome the salt and pepper noise that often exists in HSI. For example, morphological filters [36], segmentation [37], Markov random fields [38], and empirical mode decomposition with spectral gradient enhancement [39] are recently presented and they demonstrate a superior performance. Though these methods differ in their ways of utilizing the HSI data, they all try to explore the abundant clues from the original HSI volume. Unfortunately, they neglect a fact that the large number of bands might be redundant from time-to-time. Trying to select the most critical bands may still get a good result, at the same time reducing the storage burden.

### III. Proposed Method

In this paper, we present a salient band selection method based on MR for HSI classification. The flowchart of the method is shown in Fig. 1. First, the original band set is grouped into subsets, within which each band image has similar characteristics to the others. Then, the representative of each group is chosen by clone selection algorithm, with the principle that the representatives should be far to each other. After that, the representative bands are treated as queries and the other bands will be ranked according to the queries. Finally, the most dissimilar band will be added to the query set and the whole procedure repeats again until the desired number is achieved. In the following, we will focus on the queries generation and the MR steps.

### A. Query Generation

Our aim is to select the most distinctive bands to support classification. For this purpose, the interband similarity metrics are not suitable because they only reflect a local perspective. For instance, the examined two bands may have large distance, and selecting them seems to be a good choice. But putting them in the whole band set, they might highly resemble the other selected ones, which is undesirable. The main reason inducing this fact is that we improperly check the relationship of band pairs and lack the global view to systematically balance all the selected bands. To overcome this limitation, an MR technique is adopted [40]. Instead of treating the bands individually, we adopt a batch manner. We first get several query band representatives, and then the remaining bands are ranked according to their similarities to all the queries. Every time the ranking is conducted globally, considering the full data relationships.

The general idea of MR is to list the data in order with respect to the similarities to the predefined queries. In this paper, the queries are initially specified as the original band set representatives. This can be achieved by a simple clustering method, such as $k$-means, and then choosing the representatives in each cluster. However, we think the selected representatives should not only have maximum interdistances (distinctiveness among clusters) but also minimum intradistances (representativeness with a cluster). The formal formulation is as follows.

In this paper, each band is denoted as a point in the high-dimensional space. Suppose there are $k$ mean centers $\{\mu_1, \mu_2, \ldots, \mu_k\}$ after the clustering of the original bands and the desired representatives are $\{r_1, r_2, \ldots, r_k\}$. The sum of interdistances are then defined as $D_{\text{inter}} = (1/2) \sum_{i,j=1:k} d(r_i, r_j)$, where $d(,)$ measures the Euclidean distance between two points, and sum of intradistances are similarly calculated as $D_{\text{intra}} = \sum_{i=1:k} \sum_{j=1:n_i} d(\mu_i, e_{i,j})$, where $e_{i,j}$ is the $j$th element in cluster $i$ and $n_i$ is the number of elements within it. The final criterion is the combination of the two terms

$$\arg \max_{\{r_1, r_2, \ldots, r_k\}} \frac{D_{\text{inter}}}{D_{\text{intra}}}. \tag{1}$$

By searching for the solution of the objective function, we can get the desired representatives. However, maximizing this criterion is not an easy task because there are numerous candidate combinations. A traversal search is impossible and impractical. Fortunately, inspired by the success of natural computation [41], the clone selection algorithm [42], [43] is utilized to solve this problem.

This algorithm is motivated by the immunology and is a typical paradigm of artificial immune systems. It uses the basic immune principles to help solve complex engineering tasks. To be specific, when an animal is exposed to an antigen, it will produce particular antibodies with different affinities. Those antibodies with high affinity values will respond more adequately to the antigen. After the interaction, the antibodies will be cloned with a number proportional to its affinity values and at the same time, the obtained clones are also accelerated by mutation, the probability of which is inversely proportional to its affinity. With this strategy, the animal body can effectively eliminate the antigen infection.

Motivated by the efficiency of the clone selection principle, this work employs it to get the solution of (1). Similarly, the original problem of establishing the band representatives is treated as the antigen, and the chosen set of bands is taken as the antibody. By maximizing the affinity function defined in (1), the best antibody can be selected. Initially, we randomly select $l$ sets of representatives, with each set containing one band from each cluster. This means there are $l$ antibodies in the beginning. Then, these antibodies will subject to three steps of processing, clone, mutation, and selection.

1) *Clone:* The antibodies will be cloned according to their affinities to the antigen. Higher affinity indicates more copies will be obtained.
2) *Mutation:* The antibodies after the above step will mutate to generate new antibodies, which means any band contained in an antibody may change to the other band in the same cluster. The probability of mutation is inversely proportional to the affinity of the antibody.
3) *Selection:* After the clone and mutation procedure, there will be a larger number of antibodies compared with the initial set, which demonstrates a more various diversity and possibility. We then select the $l$ most promising ones to start the next round of processing.

### B. Manifold Ranking

With the obtained $k$ queries, the other bands are ranked according to them. Suppose the band with the lowest ranking score is found, which means it is the most dissimilar one to the queries. Then, it will be added to the original query set, the number of which now becomes $k+1$. With the updated query set, the remaining bands are further ranked with respect to it. This operation repeats until the desired $K$ bands are identified (in general, $K$ is predefined by the user). This strategy can ensure the newly selected band is much different from what have been selected. Clearly, this strategy is fundamentally different from traditional calculation of pairs of bands.

As for the ranking procedure, the manifold structure of data is considered because it can explore the intrinsic data nature [40], [44]. In this step, the goal is to learn a ranking function, which defines the relevance between the queries (representative bands) and the unlabeled data (remaining bands). In the following, a detailed introduction is followed.

Suppose a set of given points (hyperspectral bands) $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \in R^m$, where $m$ is the dimensionality of the data (the number of pixels for a single band image) and $n$ is the number of the data (the band number). Some of these points are labeled beforehand as queries and the others are unlabeled. The aim is to rank the unlabeled ones according to their relevance to the queries. Before detailed explanation of the method, notations are first introduced [45]. Let $f : \mathcal{X} \to \mathbb{R}$ denote a ranking function that assigns every point $x_i$ a ranking score $f_i$, leading to a vector of $\mathbf{f} = [f_1, f_2, \ldots, f_n]$. We also define an indicator vector $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ with $y_i = 1$ means $x_i$ is a query and $y_i = 0$ otherwise. For an appropriate measurement of the ranking function, we define a graph network $G = (V, E)$ on the data points $\mathcal{X}$, where $V$ is the vertex set, and $E$ is the edge set. We also define an affinity matrix $\mathbf{W} = [w_{ij}]_{n \times n}$ with

$$w_{ij} = e^{-d^2(x_i, x_j)/2\sigma^2} \tag{2}$$

if $x_i$ and $x_j$ are connected; otherwise $w_{ij} = 0$. For the HSI context, if two points (bands) are neighboring relationship, they are assumed to be connected. The distance between two connected points $d(x_i, x_j)$ is computed as the Euclidean distance between them. Consequently, the degree matrix is denoted as $\mathbf{D} = \text{diag}\{d_{11}, \ldots, d_{nn}\}$, where $d_{ii} = \Sigma_j w_{ij}$.

With these definitions, the optimal ranking is derived by solving the following optimization problem [45], [46]:

$$\mathbf{f}^* = \arg\min_f \frac{1}{2} \left( \sum_{i,j=1}^{n} w_{ij} \left\| \frac{f_i}{d_{ii}} - \frac{f_j}{d_{jj}} \right\| + \mu \sum_{i=1}^{n} \| f_i - y_i \|^2 \right) \tag{3}$$

---

**Algorithm 1** Salient Band Selection via Manifold Ranking

---

**Input**: $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, $K$.

**Initialize**: $k$, $\sigma$, $\mu$, $l$.

Step1: Get the $k$ clustering groups.

Step2: Use clone selection strategy to establish the group representatives as the initial queries.

Step3:

**while** not enough bands are selected **do**

 1: Rank the other bands according to the queries.

 2: Select the most dissimilar one as the new band.

 3: Update the query set with the newly established band.

**end while**

**Output**: $K$ bands.

---

where the first term is the smoothness constraint indicating the neighboring points should not differ greatly, and the second term the fitting constraint implying the ranking score should not divert too much from the initial query assignment. $\mu$ balances the contributions of the two terms.

To solve the above problem, we set the derivative of (3) to be zero. Let $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ be the normalized Laplacian matrix. The desired ranking function can be obtained as

$$\mathbf{f}^* = (\mathbf{I} - \alpha\mathbf{S})^{-1}\mathbf{y} \qquad (4)$$

where $\mathbf{I}$ is the identity matrix and $\alpha = 1/(1 + \mu)$.

With this solution, the ranking for each point is finally acquired, which indicates all the other hyperspectral bands are ranked with reference to the query bands. Therefore, the one with the lowest score is set as the most dissimilar one compared with the queries and it is added into the query set to start another round of ranking. The whole procedure is outlined in Algorithm 1.

## IV. Experiments

In this section, intensive experiments are conducted to prove the effectiveness of the proposed method. First, the data set used in the experiments is introduced. Then, the comparative methods are selected, with a brief analysis. After that, the experimental setup is detailed and the results are shown and analyzed.

### A. Data Sets

To verify the effectiveness of the proposed framework, we complete our experiments on three traditional data sets, Indian Pines, Salinas Scene, and The Pavia University.

1) Indian Pines image was gathered by AVIRIS sensor over the Indian Pines test site in North-Western Indiana in 1992. It consists of $145 \times 145$ pixels and 224 spectral reflectance bands in the wavelength range of 0.4–2.5 $\mu$m. The spatial resolution is 20 m/pixel. The 16 classes of vegetation and forests are included in the image and the ground truth labels are publicly available. In general, the water absorbtion bands are removed, leading to a total of 200 bands.

2) The Salinas Scene was also captured by AVIRIS sensor in 1998, but at a different location in Salinas Valley, California. The image size is $512 \times 217$, with a spatial resolution of 3.7 m/pixel and spectral coverage within 0.4–2.5 $\mu$m. There are also 224 spectral bands and 16 classes of interest, including vegetables, bare soils, and vineyard fields.

3) The Pavia University was acquired by the ROSIS sensor during a flight campaign over Pavia, Northern Italy in 2002. The sensor generates 115 spectral bands ranging from 0.43 to 0.86 $\mu$m. Removing the 12 noisiest bands, the 103 bands are retained, and the image size is $610 \times 340$. The geometric resolution is 1.3 m/pixel and nine classes of land cover objects are included.

### B. Comparative Methods

In order to justify the effectiveness of the proposed MR method, several competitors are employed to conduct comparison. They are of two prototypes: 1) CEM selection [26] and 2) clustering-based band selection (CBBS) [16].

1) For the CEM selection, each band image is represented as a column vector. Then, an energy function reflecting the band image correlation is defined. For solving this function, a linear constraint is enforced on one band and all the bands are used to calculate the correlation matrix. With the obtained solution, two criteria for band selection are defined: 1) band correlation minimization and 2) band correlation constraint. The induced algorithms are denoted by CEM-Band Correlation Minimization (BCM) and CEM-Band Correlation Constraint (BCC). However, the enormous size of band vectors can cause tremendous computing time. In order to mitigate this problem, the band image is treated as matrix instead of vector conversion. This idea can be traced back to the linearly constrained minimum variance (LCMV) [47] and the induced algorithms are similarly denoted as LCMV-BCM and LCMV-BCC.

2) For the CBBS, a hierarchical clustering structure, Ward's linkage method [48] to be specific, is used to group bands. Then, the representative for each group is chosen as the one having the highest correlation with the other bands in the group. It can minimize the intracluster variance and maximize the intercluster variance. The criterion for measuring the band similarity is MI and Kullback–Leibler (KL) divergence. Therefore, the resulted comparative methods are denoted as CBBS-MI and CBBS-KL.

### C. Experimental Setup

To evaluate the effectiveness of the proposed method, comparisons should be conducted. Two factors need to be considered here, the influence of selected band number and the effect of different classifiers. For one specific classifier, we set the selected band number every five intervals from small to large. This can test the ability of different band selection methods under fixed band numbers. In order to see the robustness of band selection methods, we also vary the classifiers to repeat the above processing. Four widely used classifiers are adopted in our experiments, Naive Bayes, kNN,
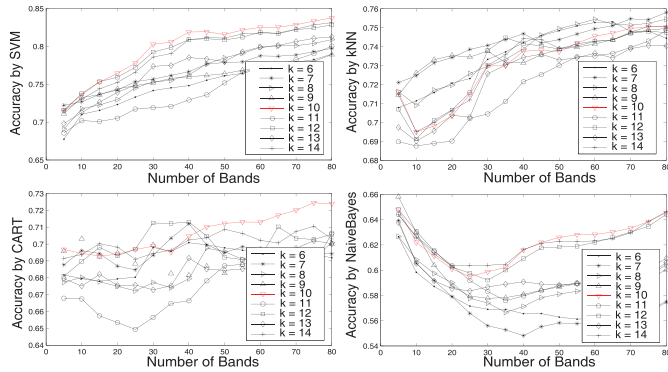
Fig. 2. Classification performance under different choices of $k$ for SVM, kNN, CART, and Naive Bayes classifiers. Each curve is averaged on the three HSIs (Indian Pines, Salinas Scene, and Pavia University).
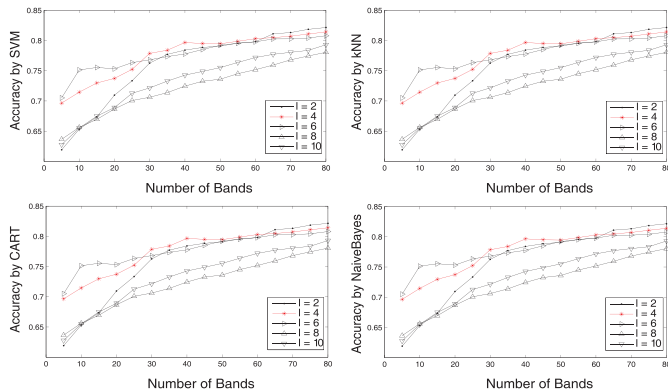


Fig. 3. Classification performance under different choices of $l$ for SVM, kNN, CART, and Naive Bayes classifiers. Each curve is averaged on the three HSIs (Indian Pines, Salinas Scene, and Pavia University).

classification and regression trees (CARTs), and SVMs, which are the benchmark classifiers [49] used in HSI classification.

There are four parameters in the experiments to be determined. They are the edge weight related $\sigma$, the balance between the smoothness term and the fitting term $\alpha$, the initial clustering number $k$, and the number of antibodies $l$. The first two are empirically set as $\sigma^2 = 0.1$, $\alpha = 0.99$ according to [45]. As for $k$ and $l$, a lot of experiments have been done to choose the best parameters. We set the two variables as different values and then check their performance under various classifiers and images. From Figs. 2 and 3, we can see that $k = 10$ and $l = 4$ is the best choice. With these two values, the averaged performance on the three images is satisfying for each classifier.

Note the initial query number $k$ is generally smaller than the desired band number. For one reason, this means fewer initial clusters and more representative and discriminative queries. For another, if we have enough query bands, there is no need to conduct the MR procedure.

### D. Results

In this section, we will discuss the band selection results on different HSIs, with respect to the different classifiers. Two kinds of results are shown in the experiments. The first kind is the band number–accuracy curves and the second kind is the averaged accuracy bars. Please note the accuracy is

defined as the proportion of correctly classified pixels to all the corresponding class pixels in the image.

For the Indian Pines image, the performance of our MR method differs from classifier to classifier. This is shown in Figs. 4 and 5. With the SVM and kNN classifiers, MR generally achieves the best classification accuracy. This superiority is more manifest when the selected band number is small, which is also the case for the CART and Naive Bayes classifiers. But when the selected bands increase, the CART and Naive Bayes are no better than the other competitors, especially when the band number surpasses 50.

For the Salinas Scene, the superiority of our MR method is more obvious as shown in Figs. 6 and 7, particularly for the SVM, kNN, and CART. At every selected band, MR acquires much higher accuracy than all the other competitors. When it comes to the Naive Bayes classifier, the result is similar to that of the Indian Pines image. If the selected band number is small, MR method works very well. But when the band number goes beyond 50, LCMV-BCM ranks first. However, the difference between MR and LCMV-BCM is only 3% and MR still outperforms all the other ones.

For the Pavia University, the results are more different, as shown in Figs. 8 and 9. The proposed MR method can classify the pixels more precisely on the kNN and CART classifiers. But for the SVM classifier, the difference from the other competitors except CEM-BCM and CEM-BCC is not so apparent. Unfortunately, the results of the Naive Bayes classifier fluctuate heavily and cannot see an absolute superiority.

After introducing the experimental results, we will give some in-depth analysis. The first question is about the performance on different images. In general, the proposed MR method is more effective than the other competitors. This is mainly due to the MR method and batch comparison strategy. But this superiority is not equally demonstrated on the three images. The Salinas Scene is a typical example, while the Pavia University and the Indian Pines image are less obvious. This can be understandable because the performance of one specific method is actually related to the training and testing data. Since the experimental subject varies with each other, the performances have more or less disparities.

The second question is about the performance on different classifiers. The proposed MR method is more stable and robust on the SVM, kNN, and CART. For the Naive Bayes classifier, MR has a vibrate result, especially on the Pavia University. We also find with the increase of band number, the MR method with Naive Bayes will have a decreasing performance. This is because the Naive Bayes models each feature dimension as an independent and normal distribution. If we sample the feature space randomly, the obtained samples can reflect the statistics rightly. But our operation is to select the samples that are dissimilar with each other, using the MR technique. Therefore, the more the selected bands, the more the true distribution will be distorted.

The third question is about the band number. There is a phenomenon that the performance is better when the band number is larger. But the superiority of the proposed MR method is more evident when the band number is small. This is because for the purpose of hyperspectral band selection,
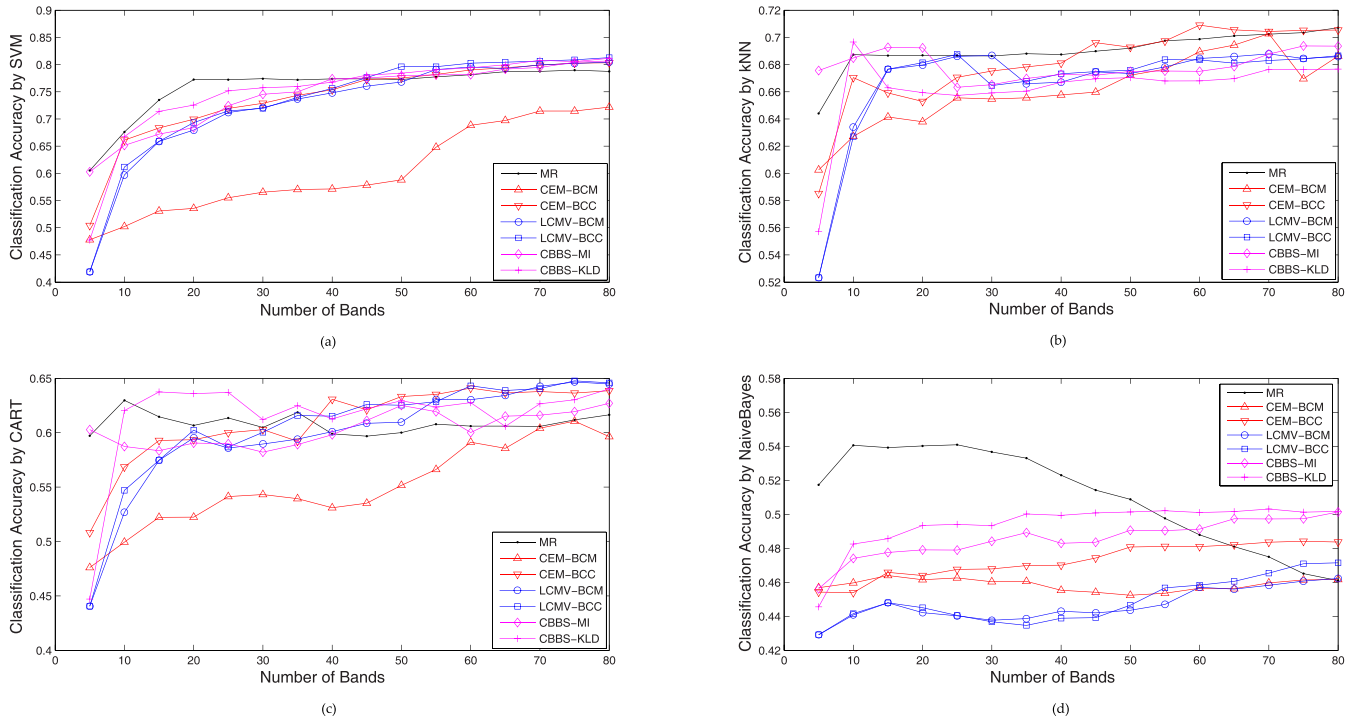
Fig. 4. Band selection results on the Indian Pines image. (a)–(d) Results by SVM, kNN, CART, and Naive Bayes classifiers.
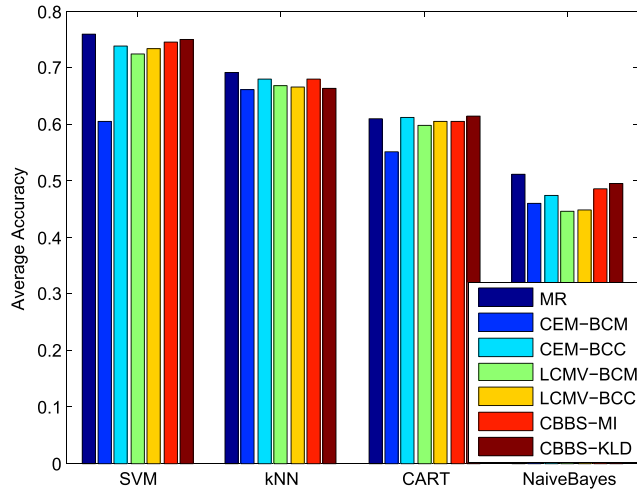


Fig. 5. Average band selection results on the Indian Pines image.

we aim to choose the most representative and the distinctive bands from the original large volume of bands. This operation can enhance the computational efficiency and relieve the storage burden at the same time. From this point of view, if the selected band number is not large but the performance of classification is satisfying, we can say the band selection method is effective and of great value. Fewer bands while good classification performance is encouraged for a band selection method. Therefore, we think the performance with a smaller band number can reflect the real ability and meaning of the band selection method. The fact that our MR method is more effective when the selected number is small just proves the success of MR compared with the other competitors. In fact,

the small number of bands is of great use in practice. With the increase of data volume, the requirement for storage and processing ability is demanding. If we can reduce the original data volume while retaining its acceptable performance, the designed method might be helpful.

The fourth question is about the comparison with the full band classification. We conduct experiments using the whole image cube and the results are as follows: Salinas Scene [SVM(0.9078) kNN(0.8594) CART(0.8248) NaiveBayes(0.7741)], Indian Pines [SVM(0.7736) kNN(0.6543) CART(0.5400) NaiveBayes(0.5509)], and Pavia University [SVM(0.8717) kNN(0.7379) CART(0.7012) NaiveBayes(0.6622)]. Comparing the band selection method (Figs. 5, 7, and 9) with the full band method, we can find that the proposed band selection method does not decrease the performance very much. For each image and classifier, abandoning most redundant bands only leads to a small amount of accuracy drop ($<5\%$) for the classification task. This means our band selection method is very effective and useful. Though we only select a limited number of bands, we can achieve an acceptable performance.

To sum up, the proposed method has distinct performance on the three images and with the various classifiers. However, we can say our method is more superior than the other competitors of the salient band selection methods. This superiority is more obvious when the selected band number is small.

## V. EXTENSION TO DEEP LEARNING

In the previous sections, we have analyzed the effect of band selection to the popular HSI classification task. The employed classifiers are all traditional ones that are mostly used in the
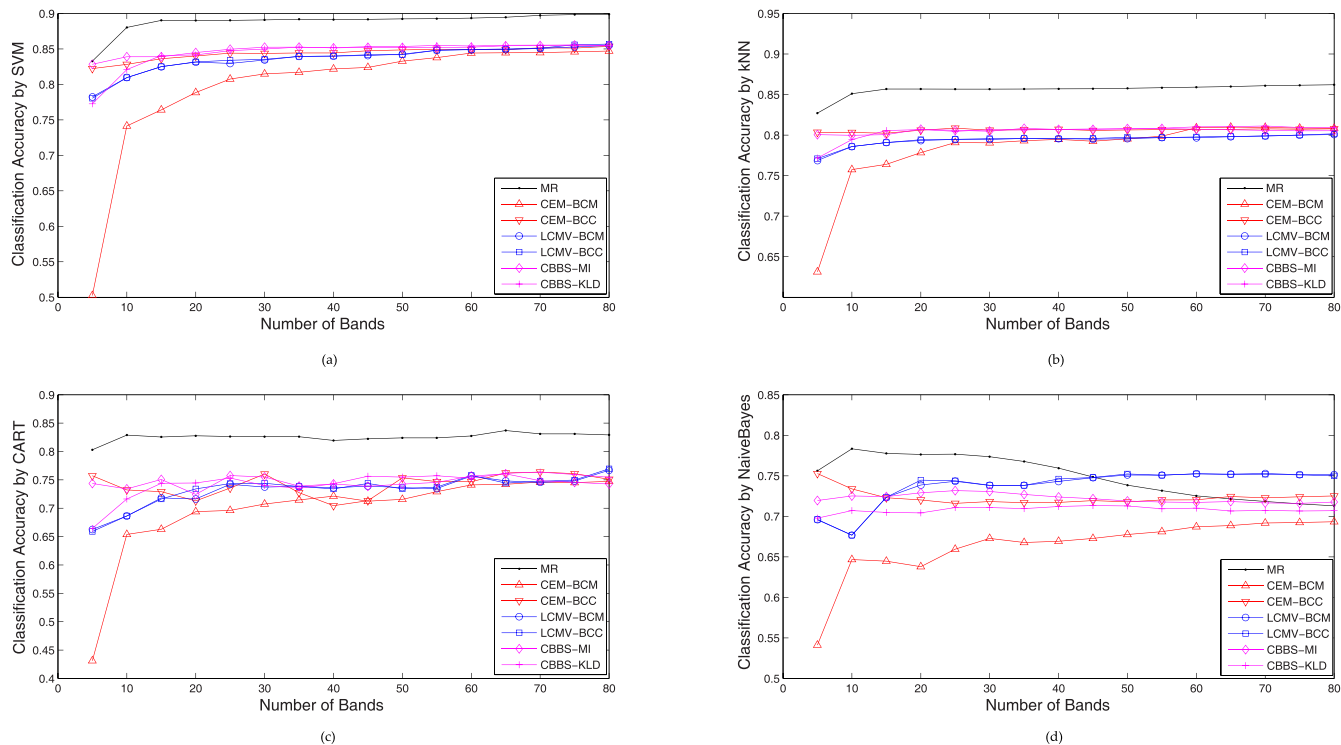
Fig. 6.   Band selection results on the Salinas Scene image. (a)–(d) Results by SVM, kNN, CART, and Naive Bayes classifiers.
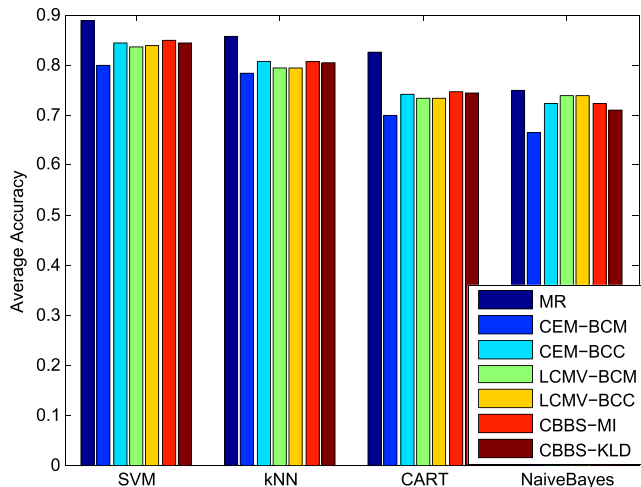


Fig. 7.   Average band selection results on the Salinas Scene image.

classification process. In this section, we will talk about the state of the art classifier—deep neural networks [50], [51] for HSI classification combined with our band selection process. For this purpose, there are several types of deep architectures, such as deep belief networks, deep Boltzmann machines, convolutional neural networks, and stacked autoencoders (SAEs). In this paper, we mainly focus on the SAEs [52].

An AE has the ability to self-express the original input. It includes two parts: 1) encoder and 2) decoder. The first part reexpresses the input vector in a higher level through the hidden layers and the second part uses the same parameter to reconstruct the original input. After the training procedure, the reconstruction layers are removed and the learned high-level features lie in the hidden layers. This can be utilized for direct classification or as the input of a higher layer to produce a deeper architecture. The SAEs are actually a concatenation of single AE, with the output of each AE as the input of another one. Each layer is a higher level expression of the previous layer, and the finally learned features have certain kinds of semantic meanings and are more appropriate for the future classification.

Unfortunately, deep networks including SAE have a disadvantage of the overfitting problem, due to the small number of training samples and the large number of model parameters. Deeper networks generally have more powerful ability, but are harder to train. Another serious problem is that the architecture of deep networks is not easy to determine. Most of the success heavily relies on the designer's experience. All these factors make the deep learning related tasks formidable.

In this paper, the flowchart of the processing steps are shown in Fig. 10. The input HSI first goes through the band selection operation. Then, the resulted data vectors are input to the SAEs. After that, the logistic regression that is learned together with the deep parameters, is used to classify the data. The SAEs have a structure of $K$-50-50-20 nodes, where $K$ is the number of selected bands. The experimental results are shown in Fig. 11.

From Fig. 11, we can see clearly that the proposed deep networks perform stably for each HSI. But the performance is not the best compared with the results shown in Figs. 4, 6, and 8. The reasons are analyzed as follows. First, the performance does not improve with the increase of the band number. Generally speaking, more selected bands
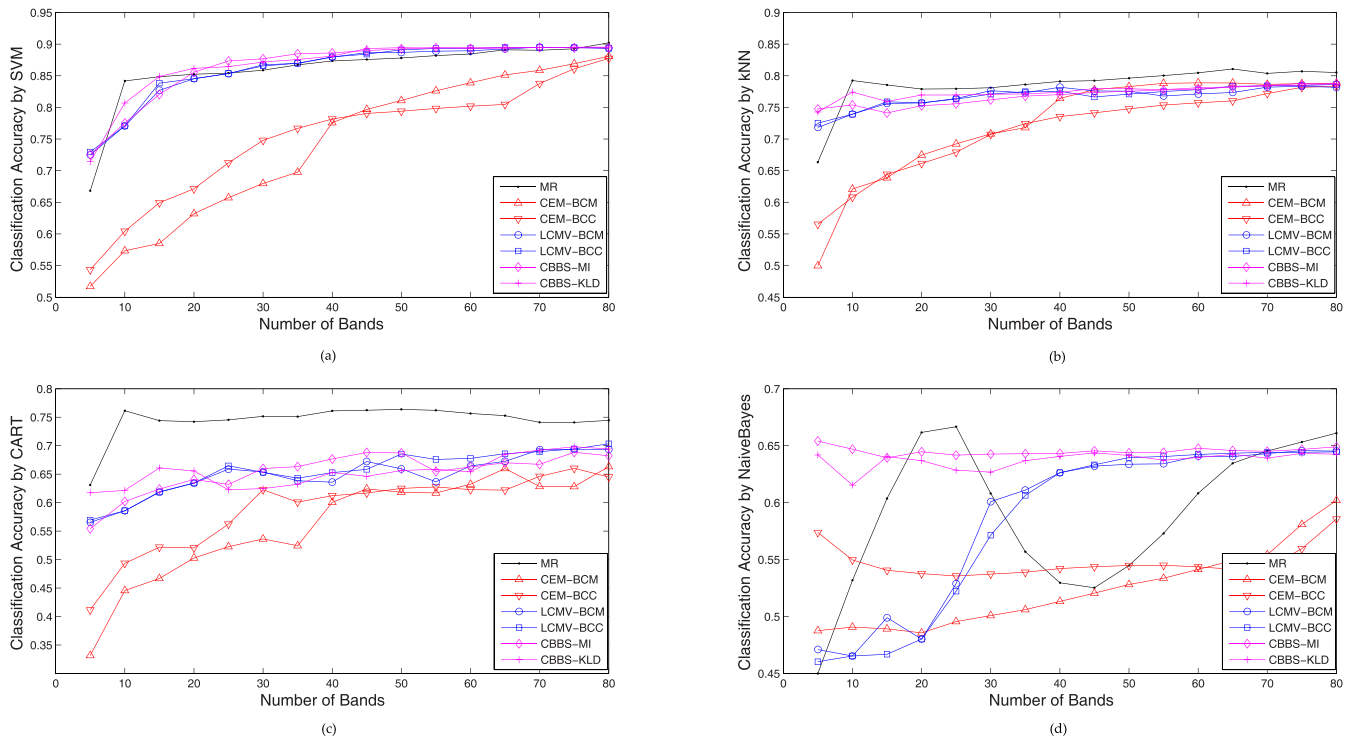
Fig. 8.   Band selection results on the Pavia University image. (a)–(d) Results by SVM, kNN, CART, and Naive Bayes classifiers.
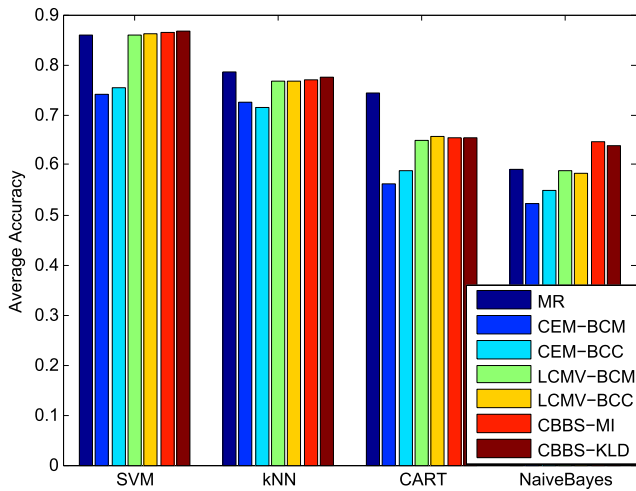


Fig. 9.   Average band selection results on the Pavia University image.
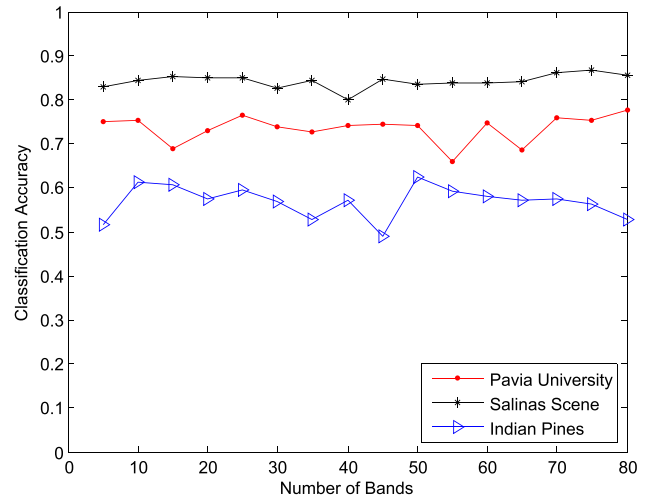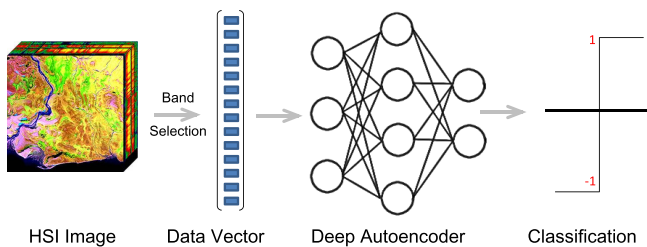


Fig. 11.   Band selection results by the SAE.



Fig. 10.   Flowchart using SAE for HSI classification.

will lead to better classification accuracy. This is true for the traditional classifiers, but is not fully right for the deep networks. The reason is that the dimensionality of the

visible layer in SAE equals to the dimensionality of the input vector. The increased band number will lead to higher input dimensionality, implying more parameters and connections in the networks. The effect of increased bands compromises with that of the added parameters. Therefore, the performance does not improve much, as shown in Fig. 11. Second, the model structure is not necessarily the best. In our experiments, we find different layers and nodes can affect the final results greatly. There is no recognized principle to determine these factors. Therefore, it is actually very hard to get a properly configured deep networks. Sometimes, we have to try many times to establish an acceptable structure. Third, we think the SAEs have certain kind of overfitting. There are all

together $50 \times 50 \times 20 \times K$ parameters in the model. Unfortunately, only 10% pixels of the input HSI are used for training. At the same time, it is difficult to do data augmentation for HSI. All these factors make the performance not very satisfying.

To summarize, the deep networks are extraordinarily popular in the vision community. But there are several unresolved problems associated with them. To make the deep structure more rationale, more deep research is needed.

## VI. Conclusion

Saliency definition is a critical factor that influences the saliency detection performance. In this paper, we formulate saliency in the context of band selection in the HSI classification and propose an MR-based selection method. In order to prove its rightness, intensive experiments and comparisons are conducted. Results show that the proposed method is effective and outperforms the competitors.

The contributions are threefolds. First, the band selection is alternatively treated as a novel MR problem. Second, the interband distance is measured in a batch manner. Third, we have provided a meaningful reference for other researchers with the abundant experimental results and comparisons, together with the discussion on the popular deep learning extension. In the future, how to effectively and systematically design a deep learning model for HSI classification is the next step.

## References

[1] H. Zheng, G. Lefebvre, and C. Laurent, "Fast-learning adaptive-subspace self-organizing map: An application to saliency-based invariant image feature construction," *IEEE Trans. Neural Netw.*, vol. 19, no. 5, pp. 746–757, May 2008.

[2] D. Chen and H. Chu, "Scale-invariant amplitude spectrum modulation for visual saliency detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1206–1214, Aug. 2012.

[3] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009–2021, Dec. 2013.

[4] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[5] Q. Wang and Y. Yuan, "High quality image resizing," *Neurocomputing*, vol. 131, pp. 348–356, Jan. 2014.

[6] Q. Zhu, L. Shao, X. Li, and L. Wang, "Targeting accurate object extraction from an image: A comprehensive study of natural image matting," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 2, pp. 185–207, Feb. 2015.

[7] E. Merényi, W. H. Farrand, J. V. Taranik, and T. B. Minor, "Classification of hyperspectral imagery with neural networks: Comparison to conventional tools," *EURASIP J. Appl. Signal Process.*, vol. 2014, no. 1, pp. 1–19, Dec. 2014.

[8] P. Zhong and R. Wang, "Jointly learning the hybrid CRF and MLR model for simultaneous denoising and classification of hyperspectral imagery," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1319–1334, Jul. 2014.

[9] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[10] N. Keshava, "Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 7, pp. 1552–1565, Jul. 2004.

[11] A. Agarwal, T. El-Ghazawi, H. El-Askary, and J. Le-Moigne, "Efficient hierarchical-PCA dimension reduction for hyperspectral imagery," in *Proc. 7th IEEE Int. Symp. Signal Process. Inf. Technol.*, Cairo, Egypt, Dec. 2007, pp. 353–356.

[12] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1586–1600, Jun. 2006.

[13] S. Jia, Y. Qian, J. Li, W. Liu, and Z. Ji, "Feature extraction and selection hybrid algorithm for hyperspectral imagery classification," in *Proc. 29th IEEE Int. Geosci. Remote Sens. Symp.*, Honolulu, HI, USA, Jul. 2010, pp. 72–75.

[14] L. Cayton, "Algorithms for manifold learning," School Comput. Sci., Univ. California, San Diego, CA, USA, Tech. Rep. CS2008-0923, Jun. 2005.

[15] M. J. Mendenhall and E. Merenyi, "Relevance-based feature extraction for hyperspectral images," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 658–672, Apr. 2008.

[16] A. Martínez-Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007.

[17] J. Feng, L. C. Jiao, X. Zhang, and T. Sun, "Hyperspectral band selection based on trivariate mutual information and clonal selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 4092–4105, Jul. 2014.

[18] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 6, pp. 2631–2641, Nov. 1999.

[19] J. Ball and L. M. Bruce, "Accuracy analysis of hyperspectral imagery classification using level sets," in *Proc. ASPRS Annu. Conf.*, Reno, NV, USA, May 2006, pp. 1–12.

[20] J. E. Ball and L. M. Bruce, "Level set hyperspectral image classification using best band analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3022–3027, Oct. 2007.

[21] Q. Du and H. Yang, "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 564–568, Oct. 2008.

[22] H. Yang, Q. Du, H. Su, and Y. Sheng, "An efficient method for supervised hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 138–142, Jan. 2011.

[23] K. Sun, X. Geng, L. Ji, and Y. Lu, "A new band selection method for hyperspectral image based on data quality," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2697–2703, Jun. 2014.

[24] S. Venkataraman, L. M. Bruce, A. Cheriyadat, and A. Mathur, "Hyperspectral dimensionality reduction via localized discriminant bases," in *Proc. 24th IEEE Int. Geosci. Remote Sens. Symp.*, Seoul, Korea, Jul. 2005, pp. 1245–1248.

[25] W. Xia, Z. Dong, H. Pu, B. Wang, and L. Zhang, "Network topology analysis: A new method for band selection," in *Proc. 31st IEEE Int. Geosci. Remote Sens. Symp.*, Munich, Germany, Jul. 2012, pp. 3062–3065.

[26] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006.

[27] Y. Yuan, G. Zhu, and Q. Wang, "Hyperspectral band selection by multitask sparsity pursuit," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 2, pp. 631–644, Feb. 2015.

[28] Y.-L. Chang, J.-P. Fang, W.-Y. Liang, L. Chang, H. Ren, and K.-S. Chen, "A parallel simulated annealing approach to band selection for hyperspectral imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Boston, MA, USA, Jul. 2008, pp. II-994–II-997.

[29] B. Demir and S. Erturk, "Reducing the computational load of hyperspectral band selection using the one-bit transform of hyperspectral bands," in *Proc. 27th IEEE Int. Geosci. Remote Sens. Symp.*, Boston, MA, USA, Jul. 2008, pp. II-919–II-922.

[30] H. Yang, Q. Du, and G. Chen, "Unsupervised hyperspectral band selection using graphics processing units," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 3, pp. 660–668, Sep. 2011.

[31] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. (2014). "Salient object detection: A survey." [Online]. Available: http://arxiv.org/abs/1411.5878

[32] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[33] M. Hasanlou and F. Samadzadegan, "Comparative study of intrinsic dimensionality estimation and dimension reduction techniques on hyperspectral images using K-NN classifier," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1046–1050, Nov. 2012.

[34] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[35] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.

[36] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[37] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[38] B. Demir and S. Erturk, "Empirical mode decomposition of hyperspectral images for support vector machine classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4071–4084, Nov. 2010.

[39] A. Erturk, M. K. Gullu, and S. Erturk, "Hyperspectral image classification using empirical mode decomposition with spectral gradient enhancement," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2787–2798, May 2013.

[40] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *Proc. 16th Adv. Neural Inf. Process. Syst.*, Whistler, BC, Canada, Dec. 2003, pp. 169–176.

[41] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[42] D. Simon, *Evolutionary Optimization Algorithms*. NJ, USA: Wiley, 2013.

[43] L. N. de Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Trans. Evol. Comput.*, vol. 6, no. 3, pp. 239–251, Jun. 2002.

[44] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. 15th Int. Conf. Multimedia*, Augsburg, Germany, Sep. 2007, pp. 862–871.

[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3166–3173.

[46] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. 16th Adv. Neural Inf. Process. Syst.*, Whistler, BC, Canada, Dec. 2003, pp. 321–328.

[47] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[48] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.

[49] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2001.

[50] A. Stuhlsatz, J. Lippel, and T. Zielke, "Feature extraction with deep neural networks by a generalized discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.

[51] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2303–2308, Dec. 2014.

[52] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Jan. 2010.

[53] C. Koch and S. Ulman, "Shifts in selective visual attention towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

**Qi Wang** (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently an Associate Professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His current research interests include computer vision and pattern recognition.



**Jianzhe Lin** (S'15) received the B.E. degree in optoelectronic information engineering and the B.A. degree in English from the Huazhong University of Science and Technology, Wuhan, China, in 2013. He is currently pursuing the master's degree with the State Key Laboratory of Transient Optics and Photonics, Center for OPTical IMagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

His current research interests include computer vision and machine learning.

**Yuan Yuan** (M'05–SM'09) is currently a Full Professor with the Chinese Academy of Sciences, Beijing, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS and *Pattern Recognition* and conference papers, such as the Computer Vision and Pattern Recognition Conference, the British Machine Vision Conference, the International Conference on Image Processing, and the International Conference on Acoustics, Speech and Signal Processing. Her current research interests include visual information processing and image/video content analysis.