Feature Selection for Cross-Scene Hyperspectral Image Classification Using Cross-Domain I-ReliefF

Chengjie Zhang, Minchao Ye, *Member, IEEE*, Ling Lei, and Yuntao Qian, *Senior Member, IEEE*

***Abstract*—In the classification of hyperspectral images (HSIs), too many spectral bands (features) cause feature redundancy, resulting in a reduction in classification accuracy. In order to solve this problem, it is a good method to use feature selection to search for a feature subset which is useful for classification. Iterative ReliefF (I-ReliefF) is a traditional single-scene-based algorithm, and it has good convergence, efficiency, and can handle feature selection problems well in most scenes. Most single-scene-based feature selection methods perform poorly in some scenes (domains) which lack labeled samples. As the number of HSIs increases, the cross-scene feature selection algorithms which utilize two scenes to deal with the high dimension and low sample size problem are more and more desired. The spectral shift is a common problem in cross-scene feature selection. It leads to difference in spectral feature distribution between source and target scenes even though these scenes are highly similar. To solve the above problems, we extend I-ReliefF to a cross-scene algorithm: cross-domain I-ReliefF (CDIRF). CDIRF includes a cross-scene rule to update feature weights, which considers the separability of different land-cover classes and the consistency of the spectral features between two scenes. So CDIRF can effectively utilize the information of source scene to improve the performance of feature selection in target scene. The experiments are conducted on three cross-scene datasets for verification, and the experimental results demonstrate the superiority and feasibility of the proposed algorithm.**

***Index Terms*—Cross-domain I-ReliefF, cross-scene feature selection, hyperspectral images.**

I. Introduction

MOST of the hyperspectral scenes include dozens or hundreds of spectral bands which may cause the Hughes phenomenon [1]. The high dimension and low sample size problem is a great challenge for hyperspectral images (HSIs) classification [2]–[4]. Some features are useful for pixel classification, while the other features have negative effects on the classification accuracy. Therefore, how to select a feature subset required by the HSI classification through a certain algorithm must be considered. An intuitive interpretation is that the feature selection problem is a combinatorial optimization problem which is used to compute the scores of different feature subsets. Feature selection obtains a most useful feature subset by eliminating irrelevant features. Irrelevant features will lead to decreased accuracies and overfitting problems [5], [6]. According to the relationship with the learning model, existing feature selection algorithms can be generally categorized as filter methods, wrapper methods and embedded methods [7], [8]. There are also some algorithms which combine the filter and wrapper methods [9], [10]. Recently, Ghosh *et al.* [11] proposed a wrapper-filter combination of ant colony optimization, which introduces the subset evaluation using a filter method instead of using a wrapper method to reduce computational complexity. As for wrapper methods, the feature selection and classification are separated, and conducted iteratively. Typical examples are dynamic classifier [12], hybrid whale optimization algorithm with simulated annealing [13], and modified ant lion optimizer [14]. In the embedded methods, the feature selection and classification are unified into one model. Typical examples are sparse rescaled linear square regression [15] and weighted Gini index feature selection [16]. However, the computational cost of wrapper methods and embedded methods are high in application. In contrast to the other methods, using filter methods for feature selection has two advantages: efficiency and robustness. And many filter-based feature selection methods have been proposed, such as clustering-based, sparsity-based, and ranking-based methods. The clustering-based methods construct the feature subsets for the HSIs by grouping the similar features and separating dissimilar features within the clustering framework. The features around the cluster centroids are considered as the most representative features and selected to constitute the final feature subset. Typical examples are spectral clustering [17], optimal clustering [18], dual clustering [19], and kernel-based probabilistic clustering [20]. However, the performance of clustering-based algorithm is usually sensitive to the number of clusters. On the other hand, the best feature subset in the cluster may not be the global best one. In recent years, sparsity-based methods have been applied to solve the feature selection problem, which make full use of sparse coefficients of all features to select features [17]. Some typical algorithms are as follows: Laplacian-regularized low-rank subspace clustering [21] and dissimilarity-weighted sparse self-representation [22]. However, the sparse coefficients are sensitive to the convergence of defined optimization program.

Manuscript received December 7, 2020; revised February 16, 2021 and April 29, 2021; accepted May 24, 2021. Date of publication June 3, 2021; date of current version June 21, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61701468 and 62071421, and in part by the National Key Research and Development Program of China under Grants 2018AAA0100500 and 2018YFB0505000. *(Corresponding author: Minchao Ye.)*

Chengjie Zhang, Minchao Ye, and Ling Lei are with the Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, Hangzhou 310018, China (e-mail: s1903081208@cjlu.edu.cn; yeminchao@cjlu.edu.cn; lling5@cjlu.edu.cn).

Yuntao Qian is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China (e-mail: ytqian@zju.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3086151

The ranking-based methods evaluate the contribution of each feature to HSI classification, and select the top-ranked features. Examples of ranking-based methods are minimum redundancy maximum relevance [23], mutual information [24], [25], and similarity-based ranking method [26]. In this article, we focus on the ranking-based methods, because every selected feature contains more useful information for classification.

It is conceivable that using abundant labeled samples can improve the performance of feature selection. However, there are some HSIs which lack labeled samples, because labeling samples is a costly, time-consuming, and labor-intensive task. On these HSIs with a small number of labeled samples, unsupervised learning and semi-supervised learning have played huge roles [12], [18], [27]. As the number of HSIs increases, it can be found that many HSIs are related. For example, cities always have the same land-cover classes, like land, trees, rivers, etc. Therefore, it is useful to utilize the strongly related scene with abundant labeled samples (named source scene) to improve the classification accuracy of the scene which lacks labeled samples (named target scene), that is, transfer learning. But when the labeled samples from the source scene are directly merged with a limited number of labeled samples from the target scene, the classification results are often poor. This is caused by spectral shift which is a common problem between different scenes [28]. The spectral features of the land-cover classes are affected in many ways, e.g., the difference on illumination, atmosphere, humidity, sensor, and even the angle of image acquisition. So how to reduce the impact of the spectral shift on HSI classification and promote the consistency of selected features between two related HSI scenes are huge challenges. Transfer learning is usually applied to classification and feature dimensionality reduction. Domain adaptation is a popular research direction in transfer learning, state-of-the-art classification algorithms include discriminative transfer joint matching [29], joint correlation alignment-based graph neural network [30], multiple domain adaptation fusion method, and the multiple base classifier fusion method [31]. And the typical examples of cross-domain feature selection are corss-domain feature selection using clustering (CDFSC) [32] and cross-domain information gain [33].

The measure of distance between samples is a commonly used evaluation criterion in filter methods. Based on this evaluation criterion, the following conditions need to be considered. The feature distance between samples in the same class should be as small as possible, while the distance between samples in different classes should be larger. Following this principle, Kira and Rendell proposed the representative feature selection algorithm named Relief, which is simple and effective. Relief is a filter method for feature selection to deal with binary classification problems. Relief is also regarded as one of the successful algorithms [34]. And the principle of this feature selection algorithm is concluded as follows: in each iteration, a sample $\mathbf{x}_n$ is randomly selected by Relief, and the feature weights are calculated depending on the distance between samples. The feature weight will increase when the pairwise distance between $\mathbf{x}_n$ and the nearest hit is smaller than pairwise distance between $\mathbf{x}_n$ and the nearest miss; otherwise, the feature weight decreases. Since Relief is not applicable to deal with the feature selection of multiclass classification problems, Kononenko extended Relief algorithm to ReliefF [35]. Though ReliefF is more effective, it is still not perfect enough. ReliefF cannot completely eliminate the influence of outliers (mislabeled samples or samples highly corrupted by noise) on feature selection. So on the basis of ReliefF, Sun [36] and Chen and Chen [37] proposed a new improvement: iterative ReliefF (I-ReliefF). By introducing I-ReliefF, the feature selection is based on a weighted feature space and considers the influence of outliers.

I-ReliefF can solve the feature selection problems well on the single HSI scenes. But the performance of each iteration is often related to the number of labeled samples, so it does not perform well in some scenes which lack labeled samples. To solve this problem, a cross-scene feature selection algorithm is proposed in this article based on I-ReliefF, which is named cross-domain I-ReliefF (CDIRF). The contributions of this article include the following.

- When the number of labeled samples in a scene is insufficient, we need to use the information contained in other scenes to improve the feature selection performance of target scene. In this case, the single-scene-based feature selection algorithm is no longer applicable, and the method we proposed can solve this problem.
- The original I-ReliefF is extended to a cross-scene feature selection algorithm named CDIRF and a new cross-domain rule of updating feature weights is proposed.
- There is spectral shift between different scenes. In order to reduce the influence of spectral shift, CDIRF considers the separability of selected features between different land-cover classes and the consistency of the selected features between two scenes at the same time.
- The influences of two different distance measures on feature selection are evaluated.
- The proposed CDIRF is robust to the outlier samples which is proved by the experiments.
- The influence of the number of labeled samples on the cross-scene algorithm is analyzed, and CDIRF is compared with the state-of-the-art cross-scene feature selection algorithms. The experimental results prove the effectiveness of CDIRF.

The rest of this article is organized as follows. In Section II, we introduce the concepts and related rules of the Relief, ReliefF, and I-ReliefF algorithms, and explain the advantages and disadvantages of these algorithms. In Section III, we propose a new feature selection algorithm called CDIRF, which is extended form the original single-scene-based I-ReliefF. In Section IV, experimental results on three datasets are presented to demonstrate the feasibility of the newly proposed method. Finally, Section V concludes this article.

II. Relief and I-ReliefF

The multiclass feature selection algorithm ReliefF assigns relevance to features based on their ability to disambiguate similar samples, where similarity is defined by proximity in feature space. In each iteration, ReliefF randomly selects a labeled sample $\mathbf{x}_n$ from dataset. For each class, ReliefF finds

k-nearest neighbor samples by calculating the pairwise distances between samples and $\mathbf{x}_n$. These nearest neighbor samples are divided into two types. One is from the same class with $\mathbf{x}_n$ (Nearest Hit, H) and the other is from different classes (Nearest Miss, M). Then, the contributions of features are determined by the following principles: 1) The average distance between $\mathbf{x}_n$ and H of this feature is smaller than the average distance between $\mathbf{x}_n$ and M of this feature. It means this feature is beneficial to classification and the weight of this feature should be increased; 2) Otherwise, the feature has a negative effect on classification and the weight of this feature is reduced. Following the above principles, ReliefF selects N samples to iteratively modify the feature weights.

ReliefF does not consider following aspects. First, although ReliefF selects k nearest neighbor samples of each $\mathbf{x}_n$ to reduce impact of outliers on feature selection, it still performs poorly in some datasets with a lot of outliers. Second, the main weakness of ReliefF is that H and M are selected on the original feature space. It means that ReliefF treats each feature with the same weight. If the feature weights are calculated in the weighted feature space and the pairwise distance calculations are updated each time with the feature weights $\mathbf{w}$ in the previous iteration, the low scoring feature from the previous iteration will have less impact on feature distance in the current iteration [34]. In order to solve these problems, a new way to select the nearest neighbor samples was proposed in [36]. I-ReliefF regards all samples, except $\mathbf{x}_n$, as the potential nearest neighbor samples, rather than only k-nearest neighbor samples. In each iteration, I-ReliefF calculates the probability of each sample being the nearest neighbor sample of $\mathbf{x}_n$. This method also avoids the consideration of whether the nearest neighbor sample is an outlier, because the impact of an outlier on the whole dataset is extremely limited.

Since I-ReliefF is applicable to multi-class datasets, it is necessary to define some sets to distinguish the samples which have the same and different class labels with $\mathbf{x}_n$: $\mathbf{H}_n = \{i : 1 \leq i \leq N, y_i = y_n, i \neq n\}$, $\mathbf{M}_n(C) = \{i : 1 \leq i \leq N, C \neq y_n\}$, where y_n denotes the label of $\mathbf{x}_n$. First, the objective function is defined which needs to be optimized. Suppose that the nearest hits and misses of each $\mathbf{x}_n$ are known. The indices of these nearest neighbors are saved in the set $\mathbf{S}_n = \{(S_h, S_m(C))\}$, where $S_h \in \mathbf{H}_n$ and $S_m(C) \in \mathbf{M}_n(C)$. To represent whether $\mathbf{x}_n$ is an outlier, a set of binary parameters $\mathbf{o} = [o_1, o_2, \ldots, o_n]$ is defined. Such that $o_n = 0$, if $\mathbf{x}_n$ is an outlier; or $o_n = 1$ otherwise. It should be mentioned that I-ReliefF uses two distance measures.

1) Absolute distance, which is calculated as

$$d_{\mathbf{w}}^{Abs}\left(\mathbf{x}_n, \mathbf{x}_i\right) = \sum_f w_f \left|\left(\mathbf{x}_n\right)_f - \left(\mathbf{x}_i\right)_f\right|. \tag{1}$$

2) Squared Euclidean distance [38], which is calculated as

$$d_{\mathbf{w}}^{Euc}\left(\mathbf{x}_n, \mathbf{x}_i\right) = \sum_f w_f \left|\left(\mathbf{x}_n\right)_f - \left(\mathbf{x}_i\right)_f\right|^2. \tag{2}$$

In this section, these two distance measures are represented by $d_{\mathbf{w}}(\mathbf{x}_n, \mathbf{x}_{S_h})$. And $(\cdot)_f$ means the fth element of vector. Under the above assumptions, the average difference between within-class distance and between-class distance can be calculated by

$$U\left(\mathbf{w}\right) = \frac{1}{N}\sum_{n=1}^{N} o_n \left(\sum_{C \neq R} \eta_C d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_{S_m(C)}\right) - d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_{S_h}\right)\right). \tag{3}$$

$P(C)$ denotes the probability of class C in the whole dataset, and $\eta_C = \frac{P(C)}{1-P(R)}$ is the weight of class C. Among them, R represents the class of $\mathbf{x}_n$. Since set $\mathbf{S}_n = \{(S_h, S_m(C))\}$ and the vector $\mathbf{o}$ are unknown, I-ReliefF can proceed by deriving the probability distributions of the unobserved data if I-ReliefF assumes the elements of $\mathbf{S}_n = \{(S_h, S_m(C))\}$ and $\mathbf{o}$ are random variables. In order to optimize (3), the following works need to be completed. By calculating the pairwise distances between the $\mathbf{x}_n$ and the potential nearest hits or misses, the probability of the ith sample in $\mathbf{M}_n(C)$ being the nearest miss of $\mathbf{x}_n$ can be defined as

$$P_m^C\left(i|\mathbf{x}_n, \mathbf{w}\right) = \frac{g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_i\right)\right)}{\sum_{j \in \mathbf{M}_n(C)} g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_j\right)\right)}, \forall i \in \mathbf{M}_n(C). \tag{4}$$

In (4), $g(\cdot)$ is a kernel function. One commonly used example is $g(d) = \exp(-d/\sigma)$, where the kernel width σ is a user-defined parameter. Other kernel functions can also be used. Similarly, the probability of the ith sample in $\mathbf{H}_n$ being the nearest hit of $\mathbf{x}_n$ can be defined as

$$P_h\left(i|\mathbf{x}_n, \mathbf{w}\right) = \frac{g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_i\right)\right)}{\sum_{j \in \mathbf{H}_n} g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_j\right)\right)}, \forall i \in \mathbf{H}_n. \tag{5}$$

For brevity of notation, I-ReliefF calculates the pairwise distances between $\mathbf{x}_n$ and other samples, and obtains the average distance in each class. The number of samples in each class is denoted as $N(C)$. Among them, the average distance between $\mathbf{x}_n$ and the samples in different classes with $\mathbf{x}_n$ is denoted as

$$dist\left(C \neq R\right) = \sum_{C \neq R}\left(\frac{\sum_{i \in \mathbf{M}_n(C)} g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_i\right)\right)}{N(C)}\right). \tag{6}$$

Similarly, the average distance between $\mathbf{x}_n$ and its within-class samples is denoted as

$$dist(R) = \frac{\sum_{i \in \mathbf{H}_n} g\left(d_{\mathbf{w}}\left(\mathbf{x}_n, \mathbf{x}_i\right)\right)}{N(R)}. \tag{7}$$

Then, the probability P_o that $\mathbf{x}_n$ is not an outlier can be defined as

$$P_o\left(o_n = 1 \mid \mathbf{w}\right) = \frac{dist(R)}{dist\left(C \neq R\right) + dist(R)}. \tag{8}$$

The purpose of I-ReliefF is to find a feature set following the large margin principle. So based on above definitions, I-ReliefF uses the idea of expectation maximization (EM) algorithm to calculate feature weights. In order to introduce this algorithm more clearly, we have following notations: let $\alpha_{i,n}(C) = P_m^C(i|\mathbf{x}_n, \mathbf{w}^{(t)})$, $\beta_{i,n} = P_h(i|\mathbf{x}_n, \mathbf{w}^{(t)})$, $\gamma_n = P_o(o_n = 1|\mathbf{w}^{(t)})$. Then, we denotes $\gamma_n = E(o_n = 1)$,

$\sum_{i \in M_n(C)} \alpha_{i,n}(C)\mathbf{x}_i = E(\mathbf{x}_{s_m^c})$, $\sum_{i \in H_n} \beta_{i,n}\mathbf{x}_i = E(\mathbf{x}_{s_h})$. Then, the algorithm of I-ReliefF can be completed through two steps.

Step 1: Before selecting the $\mathbf{x}_n$, the feature weights need to be initialized. Assuming that the weight of each feature $w_f = 1/\sqrt{F}$, where F denotes the feature dimension of samples. In each iteration, I-ReliefF selects N samples, and calculates the average difference between within-class distance and between-class distance of each $\mathbf{x}_n$. Here we approximate the expectation by the average U value along limited number of samples. The average difference can be defined as function Q in the $(t+1)$th iteration. Then, function Q can be expressed as

$$\begin{aligned} Q\left(\mathbf{w} \mid \mathbf{w}^{(t)}\right) &= E_{\{\mathbf{S},\mathbf{o}\}}\left[U\left(\mathbf{w}\right)\right] \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\gamma_n\left[\sum_{C\neq R}\left(\eta_C \sum_{i\in\mathbf{M}_n(C)}\alpha_{i,n}(C)d_{\mathbf{w}}\left(\mathbf{x}_n,\mathbf{x}_i\right)\right)\right. \\ &\quad \left. - \sum_{i\in\mathbf{H}_n}\beta_{i,n}d_{\mathbf{w}}\left(\mathbf{x}_n,\mathbf{x}_i\right)\right]. \end{aligned} \tag{9}$$

If $i \in \mathbf{M}_n(C)$, $\mathbf{m}_{i,n} = |\mathbf{x}_n - \mathbf{x}_i|$, and if $i \in \mathbf{H}_n$, $\mathbf{h}_{i,n} = |\mathbf{x}_n - \mathbf{x}_i|$. Then, (9) is simplified to (10) before applying it to update the weights.

$$\begin{aligned} &Q\left(\mathbf{w} \mid \mathbf{w}^{(t)}\right) \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\gamma_n\left[\sum_{C\neq R}\left(\eta_C \sum_{i\in\mathbf{M}_n(C)}\alpha_{i,n}(C)d_{\mathbf{w}}\left(\mathbf{x}_n,\mathbf{x}_i\right)\right)\right. \\ &\quad \left. - \sum_{i\in\mathbf{H}_n}\beta_{i,n}d_{\mathbf{w}}\left(\mathbf{x}_n,\mathbf{x}_i\right)\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\gamma_n\left[\sum_{C\neq R}\left(\eta_C \sum_{i\in\mathbf{M}_n(C)}\alpha_{i,n}(C)\sum_f w_f\left(\mathbf{m}_{i,n}\right)_f\right)\right. \\ &\quad \left. - \sum_{i\in\mathbf{H}_n}\beta_{i,n}\sum_f w_f\left(\mathbf{h}_{i,n}\right)_f\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\gamma_n\left[\sum_{C\neq R}\left(\sum_f w_f\eta_C \underbrace{\sum_{i\in\mathbf{M}_n(C)}\alpha_{i,n}(C)\left(\mathbf{m}_{i,n}\right)_f}_{\left(\bar{\mathbf{m}}_n(C)\right)_f}\right)\right. \\ &\quad \left. - \sum_f w_f \underbrace{\sum_{i\in\mathbf{H}_n}\beta_{i,n}\left(\mathbf{h}_{i,n}\right)_f}_{\left(\bar{\mathbf{h}}_n\right)_f}\right] \\ &= \mathbf{w}^T\frac{1}{N}\sum_{n=1}^{N}\gamma_n\left(\sum_{C\neq R}\eta_C\bar{\mathbf{m}}_n(C) - \bar{\mathbf{h}}_n\right) \\ &= \mathbf{w}^T\frac{1}{N}\sum_{n=1}^{N}\gamma_n\left(\bar{\mathbf{m}}_n - \bar{\mathbf{h}}_n\right). \end{aligned} \tag{10}$$

```
Algorithm 1: I-ReliefF.
Input:
  Dataset $\mathbf{X}$.
  Number of iterations $T$.
  Kernel width $\sigma$.
  Parameter of stop criterion $\theta$.
Output:
  Feature weight vector $\mathbf{w}$.
Set $w_f^{(0)} = 1/\sqrt{F}, f \in [1, 2, \ldots, F]$.
for $t = 1, 2, \ldots, T$ do
    Calculate pairwise distances with respect to $\mathbf{w}^{(t-1)}$
      using (1) or (2).
    Calculate the probability that each sample becomes
      the nearest neighbor sample $P_m^C$, $P_h$, and $P_o$
      using (4), (5), (8).
    Update weights using (13).
    if $\|\mathbf{w}^{(t+1)} - \mathbf{w}^t\| \leq \theta$ then
        break
    end if
end for
```

Some detailed denotations are as follows:

$$\bar{\mathbf{h}}_n = \sum_f \sum_{i\in\mathbf{H}_n}\beta_{i,n}\left(\mathbf{h}_{i,n}\right)_f \tag{11}$$

$$\begin{aligned} \bar{\mathbf{m}}_n &= \sum_{C\neq R}\eta_C\sum_f\sum_{i\in\mathbf{M}_n(C)}\alpha_{i,n}(C)\left(\mathbf{m}_{i,n}\right)_f \\ &= \sum_{C\neq R}\eta_C\bar{\mathbf{m}}_n(C). \end{aligned} \tag{12}$$

Step 2: Finally, on the basis of (10), feature weights can be obtained by (13).

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \arg\max_{\mathbf{w}\in\mathbf{W}} Q\left(\mathbf{w} \mid \mathbf{w}^{(t)}\right) \\ &= \arg\max \mathbf{w}^T\frac{1}{N}\sum_{n=1}^{N}\gamma_n\left(\bar{\mathbf{m}}_n - \bar{\mathbf{h}}_n\right). \end{aligned} \tag{13}$$

Denoting $\mathbf{v} = \frac{1}{N}\sum_{n=1}^{N}\gamma_n(\bar{\mathbf{m}}_n - \bar{\mathbf{h}}_n)$, we get

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \arg\max_{\mathbf{w}\in\mathbf{W}} Q\left(\mathbf{w} \mid \mathbf{w}^{(t)}\right) \\ &= \arg\max \mathbf{w}^T\mathbf{v} \\ &= (\mathbf{v})^+ / \left\|(\mathbf{v})^+\right\|_2. \end{aligned} \tag{14}$$

It should be mentioned that $\mathbf{w}$ cannot increase without bound, and a limit is given: $\mathbf{W} = \{\mathbf{w} : \|\mathbf{w}\|_2^2 = 1, \mathbf{w} \geq 0\}$. The above two steps alternatively iterate until the convergence: $\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \theta$. The detailed derivation in (14) about $(\mathbf{v})^+/\|(\mathbf{v})^+\|_2$ can be referred to [36]. $(\mathbf{v})^+$ means process v_f as $(v_f)^+ = \max(v_f, 0)$. The pseudocode of I-ReliefF is shown in Algorithm 1.

III. The Proposed Cross-Domain I-ReliefF

Traditional single-scene-based algorithms are often affected by the insufficient number of samples. In order to solve this problem, we need to extend I-ReliefF to a cross-scene feature selection algorithm CDIRF. In order to increase the reliability of the algorithm, the following constraints need to be imposed on the cross-scene algorithm: 1) The separability of different classes; and 2) the consistency of selected features between two scenes should be guaranteed [33].

Suppose there are a large number of labeled samples in the source scene, and only a small number of labeled samples in the target scene. When extending I-ReliefF to CDIRF, in order to mitigate the effect of spectral shift and improve the reliability of feature weights, we need to give several definitions to distinguish samples between source and target scenes.

- $H^{\mathcal{TT}}$: the samples belonging to the same class with $\mathbf{x}_n^{\mathcal{T}}$ from the target scene;
- $M^{\mathcal{TT}}(C)$: the samples belonging to the different classes with $\mathbf{x}_n^{\mathcal{T}}$ from the target scene;
- $H^{\mathcal{ST}}$: the samples belonging to the same class with $\mathbf{x}_n^{\mathcal{T}}$ from the source scene;
- $M^{\mathcal{ST}}(C)$: the samples belonging to the different classes with $\mathbf{x}_n^{\mathcal{T}}$ from the source scene;
- $H^{\mathcal{TS}}$: the samples belonging to the same class with $\mathbf{x}_n^{\mathcal{S}}$ from the target scene;
- $M^{\mathcal{TS}}(C)$: the samples belonging to the different classes with $\mathbf{x}_n^{\mathcal{S}}$ from the target scene;
- $H^{\mathcal{SS}}$: the samples belonging to the same class with $\mathbf{x}_n^{\mathcal{S}}$ from the source scene;
- $M^{\mathcal{SS}}(C)$: the samples belonging to the different classes with $\mathbf{x}_n^{\mathcal{S}}$ from the source scene.

In the above definitions, the superscript $\mathcal{S}$ (or $\mathcal{T}$) of the $\mathbf{x}_n$ represents that $\mathbf{x}_n$ is taken from the source scene (or the target scene). In the superscript of the samples M (or H), the first item represents which scene the M (or H) is taken. The second item has the same meaning as the superscript of $\mathbf{x}_n$. The first four definitions mean that once CDIRF selects a sample $\mathbf{x}_n$ from the target scene, the potential H and M need to be found in both source and target scenes. To improve the consistency between the two scenes, we not only select $\mathbf{x}_n$ from target scene, but also select $\mathbf{x}_n$ from source scene. So the average difference between within-class distance and between-class distance can be calculated by

$$U(\mathbf{w}) = U^{\mathcal{T}}(\mathbf{w}) + U^{\mathcal{S}}(\mathbf{w}). \tag{15}$$

In (15), $U^{\mathcal{T}}(\mathbf{w})$ represents that the selected samples $\mathbf{x}_n^{\mathcal{T}}$ are from target scene. $U^{\mathcal{T}}(\mathbf{w})$ is defined as

$$\begin{aligned} U^{\mathcal{T}}(\mathbf{w}) = \frac{1}{N^{\mathcal{T}}} \sum_{n=1}^{N^{\mathcal{T}}} o_n^{\mathcal{T}} \Bigg(& \sum_{C \neq R} \eta_C^{\mathcal{T}} d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{T}}, \mathbf{x}_{S_m^{\mathcal{T}}(C)}^{\mathcal{T}} \right) \\ & - d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{T}}, \mathbf{x}_{S_h^{\mathcal{T}}}^{\mathcal{T}} \right) + \sum_{C \neq R} \eta_C^{\mathcal{S}} d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{T}}, \mathbf{x}_{S_m^{\mathcal{S}}(C)}^{\mathcal{S}} \right) \\ & - d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{T}}, \mathbf{x}_{S_h^{\mathcal{S}}}^{\mathcal{S}} \right) \Bigg). \end{aligned} \tag{16}$$

In (15), $U^{\mathcal{S}}(\mathbf{w})$ represents that the selected samples $\mathbf{x}_n^{\mathcal{S}}$ are from source scene. $U^{\mathcal{S}}(\mathbf{w})$ is defined as

$$\begin{aligned} U^{\mathcal{S}}(\mathbf{w}) = \frac{1}{N^{\mathcal{S}}} \sum_{n=1}^{N^{\mathcal{S}}} o_n^{\mathcal{S}} \Bigg(& \sum_{C \neq R} \eta_C^{\mathcal{S}} d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{S}}, \mathbf{x}_{S_m^{\mathcal{S}}(C)}^{\mathcal{S}} \right) \\ & - d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{S}}, \mathbf{x}_{S_h^{\mathcal{S}}}^{\mathcal{S}} \right) + \sum_{C \neq R} \eta_C^{\mathcal{T}} d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{S}}, \mathbf{x}_{S_m^{\mathcal{T}}(C)}^{\mathcal{T}} \right) \\ & - d_{\mathbf{w}} \left(\mathbf{x}_n^{\mathcal{S}}, \mathbf{x}_{S_h^{\mathcal{T}}}^{\mathcal{T}} \right) \Bigg). \end{aligned} \tag{17}$$

One of the differences between CDIRF and Target domain iterative ReliefF (TDIRF) is that the possibility of $\mathbf{x}_n^{\mathcal{T}}$ ($\mathbf{x}_n^{\mathcal{S}}$) being the outlier needs to be considered in both source and target scene. Combined with the source scene, the possibility that $\mathbf{x}_n^{\mathcal{T}}$ is not an outlier can be calculated by

$$P_o^{\mathcal{T}}(o_n = 1 \mid \mathbf{w}) = \frac{dist^{\mathcal{S}}(R) + dist^{\mathcal{T}}(R)}{totaldist\left(\mathbf{x}_n^{\mathcal{T}}\right)}. \tag{18}$$

In (18), $totaldist(\cdot)$ is defined by

$$\begin{aligned} totaldist\left(\mathbf{x}_n^{\mathcal{T}}\right) = {} & dist^{\mathcal{S}}(C \neq R) + dist^{\mathcal{T}}(C \neq R) \\ & + dist^{\mathcal{S}}(R) + dist^{\mathcal{T}}(R). \end{aligned} \tag{19}$$

The possibility that $\mathbf{x}_n^{\mathcal{S}}$ is not an outlier can be calculated by

$$P_o^{\mathcal{S}}(o_n = 1 \mid \mathbf{w}) = \frac{dist^{\mathcal{S}}(R) + dist^{\mathcal{T}}(R)}{totaldist\left(\mathbf{x}_n^{\mathcal{S}}\right)}. \tag{20}$$

For brevity of notation, let $\gamma_n^{\mathcal{T}} = P_o^{\mathcal{T}}(o_n = 1 \mid \mathbf{w})$, and $\gamma_n^{\mathcal{S}} = P_o^{\mathcal{S}}(o_n = 1 \mid \mathbf{w})$. In order to incorporate the samples from source and target scenes, we extend (10) to (21).

$$\begin{aligned} & Q(\mathbf{w} \mid \mathbf{w}^{(t)}) = E_{\{\mathbf{S}^{\mathcal{T}}, \mathbf{S}^{\mathcal{S}}, \mathbf{o}^{\mathcal{T}}, \mathbf{o}^{\mathcal{S}}\}} \left[U(\mathbf{w}) \right] \\ & \approx \frac{1}{N^{\mathcal{T}}} \sum_{n=1}^{N^{\mathcal{T}}} \gamma_n^{\mathcal{T}} \left(NM^{\mathcal{TT}} - NH^{\mathcal{TT}} + NM^{\mathcal{ST}} - NH^{\mathcal{ST}} \right) \\ & \quad + \frac{1}{N^{\mathcal{S}}} \sum_{n=1}^{N^{\mathcal{S}}} \gamma_n^{\mathcal{S}} \left(NM^{\mathcal{TS}} - NH^{\mathcal{TS}} + NM^{\mathcal{SS}} - NH^{\mathcal{SS}} \right) \\ & = \mathbf{w}^T \Bigg[\frac{1}{N^{\mathcal{T}}} \sum_{n=1}^{N^{\mathcal{T}}} \gamma_n^{\mathcal{T}} \left(\overline{\mathbf{m}}_n^{\mathcal{TT}} - \overline{\mathbf{h}}_n^{\mathcal{TT}} + \overline{\mathbf{m}}_n^{\mathcal{ST}} - \overline{\mathbf{h}}_n^{\mathcal{ST}} \right) \\ & \quad + \frac{1}{N^{\mathcal{S}}} \sum_{n=1}^{N^{\mathcal{S}}} \gamma_n^{\mathcal{S}} \left(\overline{\mathbf{m}}_n^{\mathcal{TS}} - \overline{\mathbf{h}}_n^{\mathcal{TS}} + \overline{\mathbf{m}}_n^{\mathcal{SS}} - \overline{\mathbf{h}}_n^{\mathcal{SS}} \right) \Bigg]. \end{aligned} \tag{21}$$

$N^{\mathcal{S}}$ (or $N^{\mathcal{T}}$) means the number of selected samples $\mathbf{x}_n$ in source scene (or target scene). For the new cross-domain rule in (21), an intuitive explanation is that CDIRF tends to increase the weights of features with large average differences in both target and source scenes. Otherwise, the weights of features are reduced. At the same time, the weights of features whose average difference is inconsistent in the two scenes are also reduced. It is also beneficial to increase the consistency of the sample distribution of the two scenes. The following formulas are the first four items of (21), and $\mathbf{x}_n^{\mathcal{T}}$ is selected in the target scene.

$NM^{\mathcal{ST}}$ and $NH^{\mathcal{ST}}$ consider the cross-scene consistency, which make the sample distributions more similar between two scenes. $NM^{\mathcal{ST}}$ means that the samples in different classes also need to be far away from the $\mathbf{x}_n^{\mathcal{T}}$ even if they are in different scenes. And $NH^{\mathcal{ST}}$ encourages within-class similarity.

$$\begin{cases} NM^{\mathcal{TT}} = \sum_{C\neq R}\left(\eta_C^{\mathcal{T}}\sum_{i\in\mathbf{M}_n^{\mathcal{T}}(C)}\alpha_{i,n}^{\mathcal{T}}(C)d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{T}},\mathbf{x}_i^{\mathcal{T}}\right)\right) \\ NH^{\mathcal{TT}} = \sum_{i\in\mathbf{H}_n^{\mathcal{T}}}\beta_{i,n}^{\mathcal{T}}d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{T}},\mathbf{x}_i^{\mathcal{T}}\right), \\ NM^{\mathcal{ST}} = \sum_{C\neq R}\left(\eta_C^{\mathcal{S}}\sum_{i\in\mathbf{M}_n^{\mathcal{S}}(C)}\alpha_{i,n}^{\mathcal{S}}(C)d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{T}},\mathbf{x}_i^{\mathcal{S}}\right)\right) \\ NH^{\mathcal{ST}} = \sum_{i\in\mathbf{H}_n^{\mathcal{S}}}\beta_{i,n}^{\mathcal{S}}d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{T}},\mathbf{x}_i^{\mathcal{S}}\right) \end{cases}. \tag{22}$$

The probability of each sample becoming the nearest neighbor sample of $\mathbf{x}_n$ needs to be calculated in the scene which the sample belongs to. We use $\alpha_{i,n}^{\mathcal{T}}(C)$, $\beta_{i,n}^{\mathcal{T}}$, $\alpha_{i,n}^{\mathcal{S}}(C)$, and $\beta_{i,n}^{\mathcal{S}}$ to represent these probabilities, respectively, e.g., $\beta_{i,n}^{\mathcal{T}}$ indicates that the probability of whether the $\mathbf{x}_i^{\mathcal{T}}$ is the nearest neighbor sample of $\mathbf{x}_n^{\mathcal{T}}$ should be calculated in the target scene. There are two reasons for choosing $\mathbf{x}_n^{\mathcal{S}}$ in the source scene. Abundant labeled samples will improve the performance of the feature selection algorithms. In addition, the source and target scenes are strongly correlated. And the following formulas are the last four items of (21), and $\mathbf{x}_n^{\mathcal{S}}$ is selected in the source scene. In (21), $NM^{\mathcal{SS}}$ and $NH^{\mathcal{SS}}$ denote that no matter the samples from the source scene or target scene, the pairwise distances between M and the $\mathbf{x}_n$ should be far, and the pairwise distances between H and the $\mathbf{x}_n$ should be close. $NM^{\mathcal{TS}}$ and $NH^{\mathcal{TS}}$ also improve the cross-scene consistency, which makes the sample distributions more similar across source and target scenes.

$$\begin{cases} NM^{\mathcal{TS}} = \sum_{C\neq R}\left(\eta_C^{\mathcal{T}}\sum_{i\in\mathbf{M}_n^{\mathcal{T}}(C)}\alpha_{i,n}^{\mathcal{T}}(C)d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{S}},\mathbf{x}_i^{\mathcal{T}}\right)\right) \\ NH^{\mathcal{TS}} = \sum_{i\in\mathbf{H}_n^{\mathcal{T}}}\beta_{i,n}^{\mathcal{T}}d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{S}},\mathbf{x}_i^{\mathcal{T}}\right), \\ NM^{\mathcal{SS}} = \sum_{C\neq R}\left(\eta_C^{\mathcal{S}}\sum_{i\in\mathbf{M}_n^{\mathcal{S}}(C)}\alpha_{i,n}^{\mathcal{S}}(C)d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{S}},\mathbf{x}_i^{\mathcal{S}}\right)\right) \\ NH^{\mathcal{SS}} = \sum_{i\in\mathbf{H}_n^{\mathcal{S}}}\beta_{i,n}^{\mathcal{S}}d_{\mathbf{w}}\left(\mathbf{x}_n^{\mathcal{S}},\mathbf{x}_i^{\mathcal{S}}\right) \end{cases}. \tag{23}$$

The proportion of the number of samples in each land-cover class to the total number of samples in the scene is defined as the class weight. It should be noted that the values of the class weight on different scenes are different, that is, $\eta_C^{\mathcal{T}} \neq \eta_C^{\mathcal{S}}$. In (21),

$$\begin{cases} \overline{\mathbf{m}}_n^{\mathcal{TT}} = \sum_{C\neq R}\eta_C^{\mathcal{T}}\sum_f\sum_{i\in\mathbf{M}_n^{\mathcal{T}}(C)}\alpha_{i,n}^{\mathcal{T}}(C)\left(\mathbf{m}_{i,n}^{\mathcal{TT}}\right)_f \\ \overline{\mathbf{h}}_n^{\mathcal{TT}} = \sum_f\sum_{i\in\mathbf{H}_n^{\mathcal{T}}}\beta_{i,n}^{\mathcal{T}}\left(\mathbf{h}_{i,n}^{\mathcal{TT}}\right)_f \\ \overline{\mathbf{m}}_n^{\mathcal{ST}} = \sum_{C\neq R}\eta_C^{\mathcal{S}}\sum_f\sum_{i\in\mathbf{M}_n^{\mathcal{S}}(C)}\alpha_{i,n}^{\mathcal{S}}(C)\left(\mathbf{m}_{i,n}^{\mathcal{ST}}\right)_f \\ \overline{\mathbf{h}}_n^{\mathcal{ST}} = \sum_f\sum_{i\in\mathbf{H}_n^{\mathcal{S}}}\beta_{i,n}^{\mathcal{S}}\left(\mathbf{h}_{i,n}^{\mathcal{ST}}\right)_f \\ \overline{\mathbf{m}}_n^{\mathcal{TS}} = \sum_{C\neq R}\eta_C^{\mathcal{T}}\sum_f\sum_{i\in\mathbf{M}_n^{\mathcal{T}}(C)}\alpha_{i,n}^{\mathcal{T}}(C)\left(\mathbf{m}_{i,n}^{\mathcal{TS}}\right)_f \\ \overline{\mathbf{h}}_n^{\mathcal{TS}} = \sum_f\sum_{i\in\mathbf{H}_n^{\mathcal{T}}}\beta_{i,n}^{\mathcal{T}}\left(\mathbf{h}_{i,n}^{\mathcal{TS}}\right)_f \\ \overline{\mathbf{m}}_n^{\mathcal{SS}} = \sum_{C\neq R}\eta_C^{\mathcal{S}}\sum_f\sum_{i\in\mathbf{M}_n^{\mathcal{S}}(C)}\alpha_{i,n}^{\mathcal{S}}(C)\left(\mathbf{m}_{i,n}^{\mathcal{SS}}\right)_f \\ \overline{\mathbf{h}}_n^{\mathcal{SS}} = \sum_f\sum_{i\in\mathbf{H}_n^{\mathcal{S}}}\beta_{i,n}^{\mathcal{S}}\left(\mathbf{h}_{i,n}^{\mathcal{SS}}\right)_f \end{cases}. \tag{24}$$

Algorithm 2: Cross-Domain I-ReliefF (CDIRF).

```
Input:
  Dataset $\mathbf{X}^{\mathcal{T}}$, $\mathbf{X}^{\mathcal{S}}$.
  Number of iterations $T$.
  Kernel width $\sigma$.
  Parameter of stop criterion $\theta$.
Output:
  Feature weight vector $\mathbf{w}$.
Set $w_f^{(0)} = 1/\sqrt{F}, f \in [1, 2, \ldots, F]$.
for $t = 1, 2, \ldots, T$ do
    Calculate pairwise distances with respect to $\mathbf{w}^{(t-1)}$
    using (1) or (2).
    Calculate the probability that each sample in the two
    scenes becomes the nearest neighbor sample $P_m^C$,
    $P_h$, $P_o^{\mathcal{T}}$, $P_o^{\mathcal{S}}$ as (4), (5), (18) and (20).
    Update weights using (25).
    if $\|\mathbf{w}^{(t+1)} - \mathbf{w}^t\| \leq \theta$ then
        break
    end if
end for
```

Following the large margin principle, the feature weights can be calculated by

$$\begin{aligned} \mathbf{w}^{(t+1)} &= \arg\max_{\mathbf{w}\in W} Q\left(\mathbf{w} \mid \mathbf{w}^{(t)}\right) \\ &= \arg\max_{\mathbf{w}\in W} \mathbf{w}^T\mathbf{v} \\ &= (\mathbf{v})^+ / \left\|(\mathbf{v})^+\right\|_2. \end{aligned} \tag{25}$$

The $\mathbf{v}$ in (25) is expressed as

$$\begin{aligned} \mathbf{v} = &\frac{1}{N^{\mathcal{S}}}\sum_{n=1}^{N^{\mathcal{S}}}\gamma_n^{\mathcal{S}}\left(\overline{\mathbf{m}}_n^{\mathcal{TS}} - \overline{\mathbf{h}}_n^{\mathcal{TS}} + \overline{\mathbf{m}}_n^{\mathcal{SS}} - \overline{\mathbf{h}}_n^{\mathcal{SS}}\right) \\ &+ \frac{1}{N^{\mathcal{T}}}\sum_{n=1}^{N^{\mathcal{T}}}\gamma_n^{\mathcal{T}}\left(\overline{\mathbf{m}}_n^{\mathcal{TT}} - \overline{\mathbf{h}}_n^{\mathcal{TT}} + \overline{\mathbf{m}}_n^{\mathcal{ST}} - \overline{\mathbf{h}}_n^{\mathcal{ST}}\right). \end{aligned} \tag{26}$$

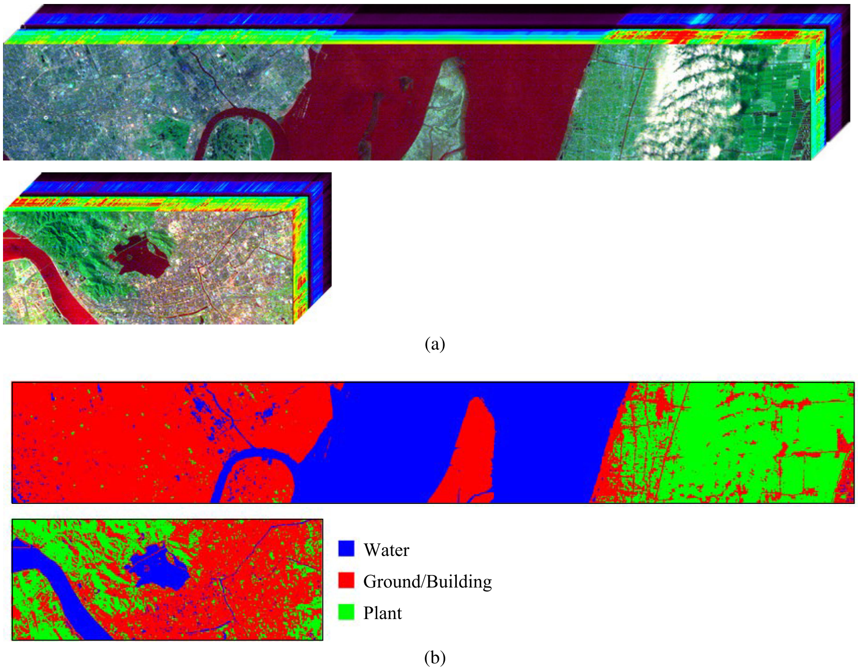
The pseudocode of CDIRF is shown in Algorithm 2.

IV. Experiments

A. Datasets

In this section, we choose three representative cross-scene HSI datasets for verification. The first dataset is EShanghai-EHangzhou dataset. Shanghai and Hangzhou are two big cities in the east of China. Both scenes were captured by the EO-1 Hyperion hyperspectral sensor. As the source scene, the EO-1 Shanghai (EShanghai) scene is sized $1600 \times 230 \times 198$. The EO-1 Hangzhou (EHangzhou) scene is the target scene sized $590 \times 230 \times 198$. The data cubes and groundtruth maps are illustrated in Fig. 1. Labeled samples of three land-cover classes are collected in EShanghai-EHangzhou dataset, which are listed in Table I.

The second dataset is DPaviaU-DPaviaC, which was taken by hyperspectral airborne sensor DAIS in Italy. It contains two


(a)

(b)

Fig. 1. The data cubes and groundtruth maps of EShanghai-EHangzhou dataset. The upper one is the source scene (EShanghai), while the lower one is the target scene (EHangzhou). (a) Data cubes. (b) Groundtruth maps.

TABLE I
NUMBER OF LABELED SAMPLES IN EACH LAND-COVER CLASS WITHIN ESHANGHAI-EHANGZHOU DATASET AND THE NUMBER OF USED SAMPLES

Class		Labeled samples	
	Name	Source scene	Target scene
1	Water	123123 (200 samples for training / 0 sample for testing)	18043 (5 samples for training / 18038 samples for testing)
2	Ground/Building	161689 (200 samples for training / 0 sample for testing)	77450 (5 samples for training / 77445 samples for testing)
3	Plant	83188 (200 samples for training / 0 sample for testing)	40207 (5 samples for training / 40202 samples for testing)

TABLE II
NUMBER OF LABELED SAMPLES IN EACH LAND-COVER CLASS WITHIN DPAVIAU-DPAVIAC DATASET AND THE NUMBER OF USED SAMPLES

Class		Labeled samples	
	Name	Source scene	Target scene
1	Trees	266 (200 samples for training / 0 sample for testing)	2424 (5 samples for training / 2419 samples for testing)
2	Asphalt	266 (200 samples for training / 0 sample for testing)	1704 (5 samples for training / 1699 samples for testing)
3	Parking lot	265 (200 samples for training / 0 sample for testing)	287 (5 samples for training / 282 samples for testing)
4	Bitumen	206 (200 samples for training / 0 sample for testing)	685 (5 samples for training / 680 samples for testing)
5	Meadow	273 (200 samples for training / 0 sample for testing)	1251 (5 samples for training / 1246 samples for testing)
6	Soil	213 (200 samples for training / 0 sample for testing)	1475 (5 samples for training / 1470 samples for testing)

scenes. The source scene is DAIS Pavia University (DPaviaU) with the size of $243 \times 243 \times 72$, and the target scene is DAIS Pavia Center (DPaviaC) with the size of $400 \times 400 \times 72$. The groundtruth maps and data cubes are illustrated in Fig. 2. DPaviaU and DPaviaC scenes contain six land-cover classes. And the number of labeled samples in each class is listed in Table II.

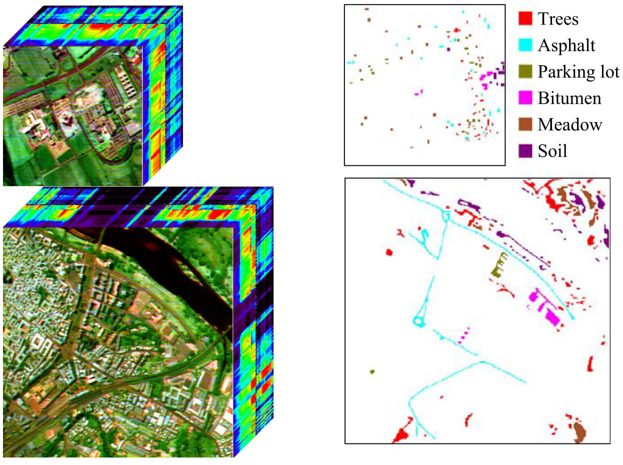
The third dataset is RPaviaC-RPaviaU, and the data were collected by the ROSIS (reflective optics system imaging spectrometer) sensor. RPaviaC-RPaviaU dataset still contains two scenes: the ROSIS Pavia University (RPaviaU) and ROSIS Pavia Center (RPaviaC). RPaviaC is the source scene sized $1096 \times 715 \times 102$. As for the target scene, RPaviaU originally has 103 bands. In order to have the same spectral bands as RPaviaC, we dropped the last band, so RPaviaU is sized $610 \times 340 \times 102$. In addition, RPaviaU and RPaviaC originally contain nine land-cover classes, and we select seven related land-cover classes. The groundtruth maps and data cubes are illustrated in Fig. 3. And the number of labeled samples in each class is listed in Table III.

B. Compared Algorithms and Parameter Settings

On these datasets, we apply six feature selection algorithms, and compare their classification accuracies when the dimension

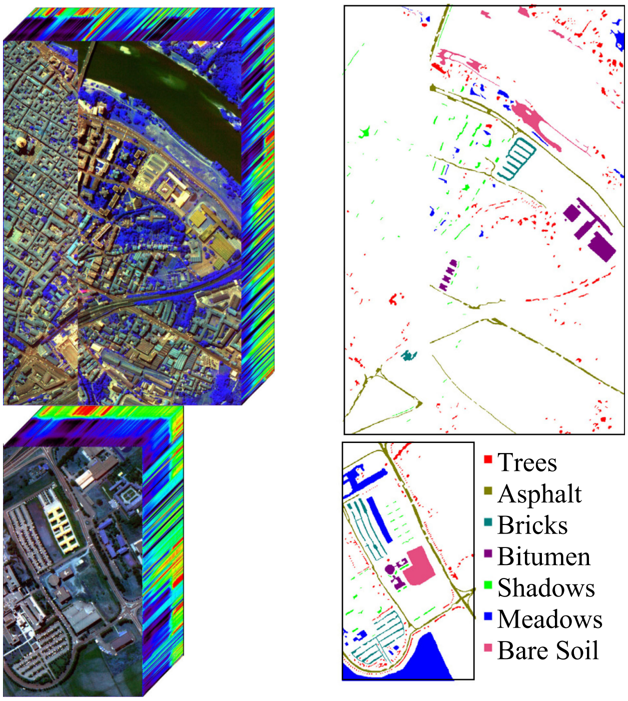
TABLE III
NUMBER OF LABELED SAMPLES IN EACH LAND-COVER CLASS WITHIN RPAVIAC-RPAVIAU DATASET AND THE NUMBER OF USED SAMPLES

Class		Labeled samples	
	Name	Source scene	Target scene
1	Trees	7598 (200 samples for training / 0 sample for testing)	3064 (5 samples for training / 3059 samples for testing)
2	Asphalt	9248 (200 samples for training / 0 sample for testing)	6631 (5 samples for training / 6626 samples for testing)
3	Bricks	2685 (200 samples for training / 0 sample for testing)	3682 (5 samples for training / 3677 samples for testing)
4	Bitumen	7287 (200 samples for training / 0 sample for testing)	1330 (5 samples for training / 1325 samples for testing)
5	Shadows	2863 (200 samples for training / 0 sample for testing)	947 (5 samples for training / 942 samples for testing)
6	Meadows	3090 (200 samples for training / 0 sample for testing)	18649 (5 samples for training / 18644 samples for testing)
7	Bare Soil	6584 (200 samples for training / 0 sample for testing)	5029 (5 samples for training / 5024 samples for testing)



(a) (b)

Fig. 2. Data cubes and groundtruth maps of DPaviaU-DPaviaC dataset. The upper one is the source scene (DPaviaU), while the lower one is the target scene (DPaviaC). (a) Data cubes. (b) Groundtruth maps.



(a) (b)

Fig. 3. Data cubes and groundtruth maps of RPaviaC-RPaviaU dataset. The upper one is the source scene (RPaviaC), while the lower one is the target scene (RPaviaU). (a) Data cubes. (b) Groundtruth maps.

of selected features is the same. The overview of the six algorithms is as follows.

- Cross-domain information gain (CDIG) [33]: It is a cross-scene extension of Information Gain (IG), and a feature with larger IG contributes more to the classification.
- Cross-domain feature selection using clustering (CDFSC) [32]: CDFSC combines density-based clustering method and IG to solve cross-scene feature selection problems.
- Target domain ReliefF (TDRF) [35]: The original ReliefF is applied to the samples from target scene.
- Cross-domain ReliefF (CDRF) [35]: It is a cross-scene extension of TDRF, which can be executed on two scenes.
- Target domain iterative ReliefF (TDIRF): The original I-ReliefF algorithm applied to only target scene. The detailed introduction can be seen in Section II. And TDIRF updates feature weights by (13). In these experiments, we use two distance measures: 1) TDIRF_1 using absolute distance, as seen in (1); 2) TDIRF_2 using squared Euclidean distance, as seen in (2).
- Cross-domain iterative ReliefF (CDIRF): CDIRF is the proposed algorithm in this work, which considers the separability of different land-cover classes and the consistency of the selection features between two scenes. In these experiments, we use two distance measures: 1) CDIRF_1 using absolute distance, as seen in (1); 2) CDIRF_2 using squared Euclidean distance, as seen in (2). And feature weights are updated by (25).

We do the same preprocessing on the three datasets to ensure the reliability of experiments. In order to reduce the impact of adverse factors, such as light and weather on the source and target scenes, we normalize all spectral vectors $\mathbf{x}$ to unit ℓ_2 norm ($\mathbf{x} \leftarrow \mathbf{x}/\|\mathbf{x}\|_2$) as data preprocessing. After this, we randomly select 200 labeled samples for each class in the source scene, and select five labeled samples for each class in the target scene, which are listed in Tables I–III. These samples are used in feature selection. The remaining samples in the target scene are utilized as a test subset. It is worth noting that the samples in the source scene are only used for feature selection, and the samples in the target scene are used to train a model and test the accuracies. The EShanghai-EHangzhou dataset is easy for classification. So we test the classification accuracy of each algorithm separately when the dimension of selected features is taken as $N_F \in \{2, 4, 6, \ldots, 20\}$. Compared with EShanghai-EHangzhou dataset, DPaviaU-DPaviaC dataset and RPaviaC-RPaviaU dataset are more difficult to classify.

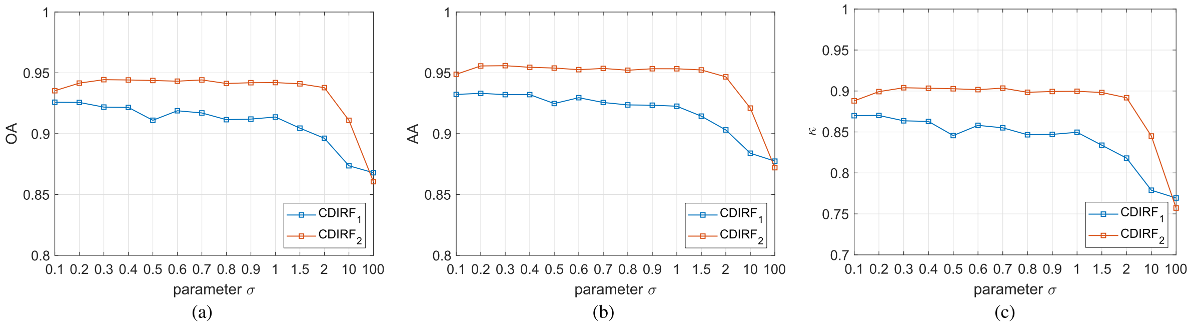


Fig. 4. Accuracies on EShanghai-EHangzhou dataset obtained by CDIRF$_1$ and CDIRF$_2$. (a) OA. (b) AA. (c) κ.

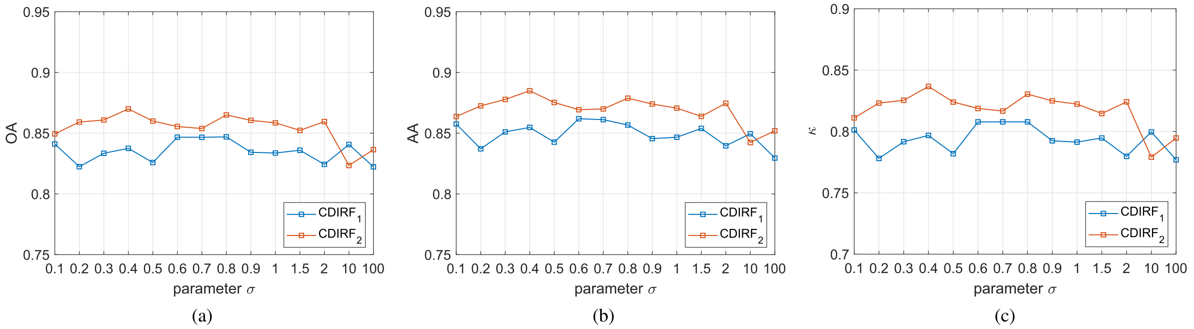


Fig. 5. Accuracies on DPaviaU-DPaviaC dataset obtained by CDIRF$_1$ and CDIRF$_2$. (a) OA. (b) AA. (c) κ.

So more features are required and we take the dimension of selected features as $N_F \in \{5, 10, 15, \ldots, 60\}$.

In addition to CDIRF$_1$ and CDIRF$_2$, there are six algorithms included in the experiments. CDRF and TDRF have two parameters: the number of the nearest neighbor samples k and the number of iterations T. We set $k \in \{1, 2, 3, 4\}$ and $T \in \{50, 100, \ldots, 300\}$, respectively. After traversing all combinations of k and T values, we choose the highest accuracy as the final result. CDIRF and TDIRF set the same parameters $\sigma \in \{0.1, 0.2, \ldots, 0.9, 1, 1.5, 2, 10, 100\}$, $T = 100$, $\theta = 10^{-5}$. To validate the performance of feature selection, we choose to use support vector machine (SVM) with radial basis function kernel as the classifier [39]. SVM is an effective way to verify the results of feature selection algorithms, and has been applied in various algorithm verifications [9], [17]. The parameters are set as $\gamma \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ and $C \in \{10^{-2}, 10^{-1}, \ldots, 10^{3}\}$ [40]. Since the randomly obtained sample subsets have a great impact on the accuracies of the experiments, we repeat the experiment for 10 times. Then, we calculate the average accuracy as the final result. In addition, we used three different accuracy evaluation criteria: overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ).

C. Parametric Analysis

The analysis of parameter σ is also involved in the algorithm TDIRF proposed by Sun [36]. In TDIRF, σ is not a critical parameter and has little effect on the result of feature selection. In this article, we also analyzed the influence of σ on CDIRF. On the EShanghai-EHangzhou dataset, when we take the feature dimension $N_F = 4$, the classification accuracies of CDIRF under different σ are shown in Fig. 4. In order to reduce the impact of randomly selected sample subsets on experimental accuracy, we repeat experiment for 10 times. It can be seen that when $\sigma \in \{0.1, 0.2, \ldots, 2\}$, the accuracies of CDIRF do not have much difference. We also experimented on DPaviaU-DPaviaC and RPaviaC-RPaviaU datasets. When the feature dimension $N_F = 20$, the classification accuracies of CDIRF using different σ are shown in Figs. 5 and 6. It can be seen that the classification accuracies of CDIRF on the these two datasets are floating, but there is no much difference in general. It means that the performance of CDIRF is not sensitive to the choice of σ. In summary, we suggest $\sigma \in \{0.1, 0.2, \ldots, 2\}$.

D. Experimental Results

The first experiment is conducted on EShanghai-EHangzhou dataset. Classification accuracies with respect to the dimension of selected features on EShanghai-EHangzhou dataset are shown in Fig. 7. In this experiment, each algorithm takes the most accurate result among different parameters. First of all, we compared algorithms based on I-ReliefF. As we can see, the accuracy of CDIRF$_2$ is always higher than CDIRF$_1$ especially when the dimension of selected features is small. CDIRF$_2$ obtains the highest accuracy at $N_F = 4$, which is also the highest accuracy among all compared algorithms. The accuracy of CDIRF$_1$ reaches its peak at $N_F = 8$, which means that CDIRF$_2$ can use

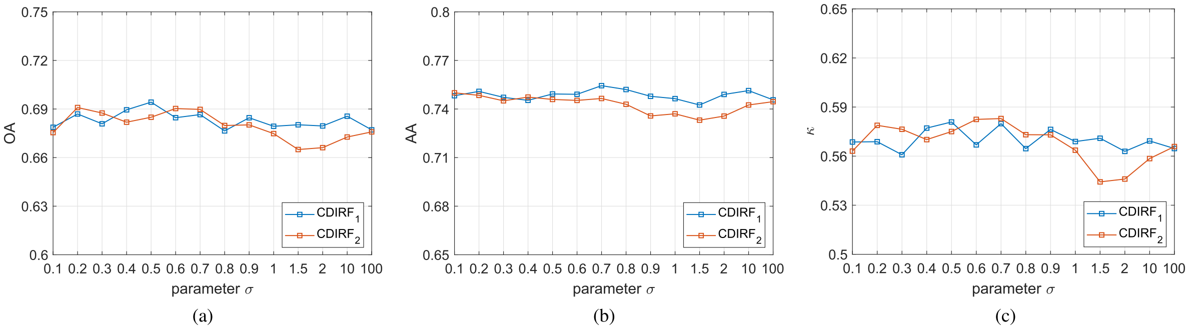


Fig. 6. Accuracies on RPaviaC-RPaviaU dataset obtained by CDIRF_1 and CDIRF_2. (a) OA. (b) AA. (c) κ.

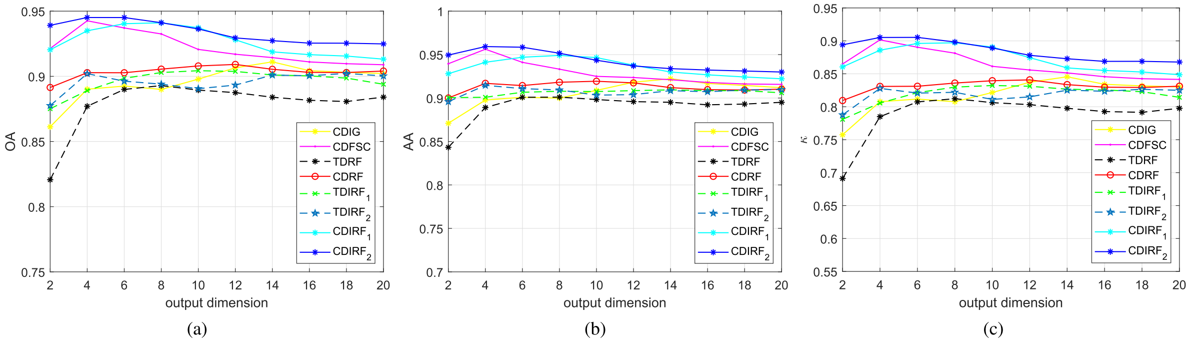


Fig. 7. Accuracies on EShanghai-EHangzhou dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF_1, TDIRF_2, CDIRF_1, and CDIRF_2. (a) OA. (b) AA. (c) κ.

fewer features to get better classification accuracy. It means that squared Euclidean distance is more suitable for CDIRF. We also compared CDIRF_2 with TDIRF_2 which also uses squared Euclidean distance. The accuracy of TDIRF_2 is floating, but still lower than CDIRF_2. In addition, we compared CDIRF_1 with TDIRF_1. And the accuracy of TDIRF_1 is always lower than CDIRF_1. It can be seen that CDIRF is superior to TDIRF on the EShanghai-EHangzhou dataset. Compared with absolute distance, squared Euclidean distance can select a better feature subset when the dimension of selected features is low.

We also compared cross-scene algorithms in Fig. 7. When $N_F = 12$, the accuracy of CDRF reaches the peak, which is still lower than proposed CDIRF_1 and CDIRF_2. Especially when the dimension of selected features is low, the accuracy gaps between these algorithms are larger. As for CDIG, the accuracy of CDIG reaches the peak when $N_F = 14$, but still lower than CDIRF_1 and CDIRF_2. The accuracy of CDFSC is higher than CDIRF_1 when the dimension of selected features is low. But as the feature dimension increases, the accuracy of CDIRF_1 is higher than CDFSC. And the accuracy of CDIRF_2 is always higher than CDFSC. The accuracy of TDRF has a great rise when the dimension of selected features is low, but still far worse than CDIRF_1 and CDIRF_2. It can be seen that CDIRF has a certain superiority. For the eight feature selection algorithms, the feature selection results with feature size $N_F = 20$ from one running experiment are illustrated in Fig. 8. It can be seen that the features selected by CDIRF_2 are more dispersed. The other algorithms relatively prefer to select continuous bands. As for a good feature selection algorithm, the selected features should be irrelevant, so we believe that CDIRF_2 is better than remaining algorithms. In addition, Pearson correlation coefficient is usually used to analyze the correlation between feature subsets of HSIs [41]. So we also perform correlation coefficient analysis on the feature subsets selected by eight algorithms in Fig. 9. It can be seen that most of the features selected in the EShanghai-EHangzhou dataset are related. CDIRF_1 and CDIRF_2 selected both positively related features and negatively related features on the EShanghai-EHangzhou dataset. In addition, Fig. 10 contains the classification maps obtained by each algorithm on a set of experimental data when $N_F = 6$. And we can find CDIRF_2 has advantages on the land cover class Water.

TABLE IV
ACCURACIES ON ESHANGHAI-EHANGZHOU DATASET

Feature selection	OA	AA	κ
CDIG	0.8959	0.9062	0.8185
CDFSC	0.9215	0.9289	0.8636
TDRF	0.8788	0.8904	0.7884
CDRF	0.9035	0.9127	0.8311
TDIRF_1	0.8967	0.9064	0.8190
TDIRF_2	0.8958	0.9073	0.8184
CDIRF_1	0.9266	0.9353	0.8720
CDIRF_2	**0.9339**	**0.9426**	**0.8848**

In Table IV, we list the average accuracies of the eight algorithms across different feature dimensions. It can be seen that the feature subset selected by CDIRF_2 performs well on

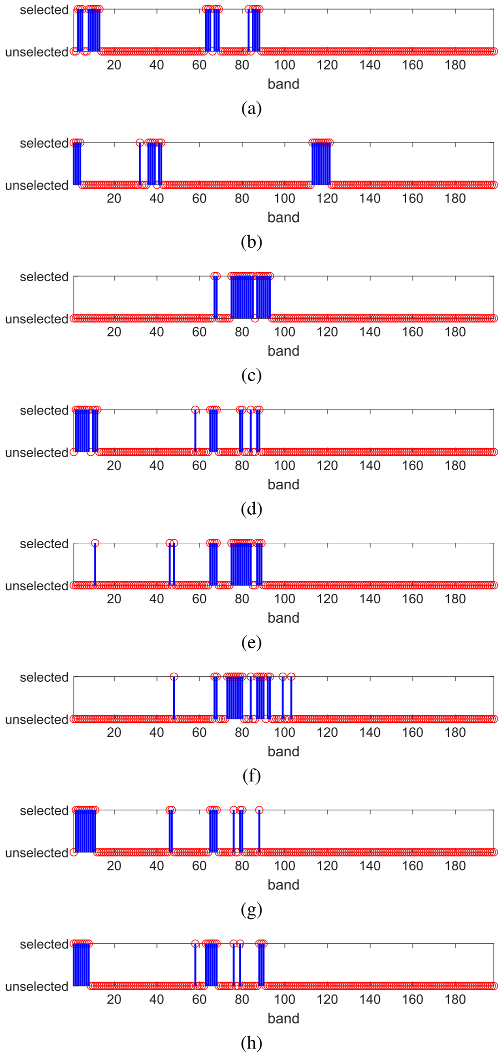


Fig. 8. Selected features on EShanghai-EHangzhou dataset. (a) CDIG. (b) CDFSC. (c)TDRF. (d) CDRF. (e) TDIRF_1. (f) TDIRF_2. (g) CDIRF_1. (h) CDIRF_2.

classification in EShanghai-EHangzhou dataset. The mean accuracies (AA, OA, and κ) of CDIRF_2 are higher than the other seven algorithms. CDIRF_2 obtains the highest accuracy 0.9339 at OA, which has the increases of 0.0381 and 0.0551 compared with the two single-scene-based algorithms TDIRF_2 and TDRF, respectively. As for cross-scene feature selection algorithms, the OA of CDIRF_2 has the increases of 0.0304, 0.0124, and 0.0380 compared with CDRF, CDFSC, and CDIG. The experiments are based on the same dimension of selected features and training samples, so it has a significant improvement even if the promotion of accuracy is small.

TABLE V
ACCURACIES ON DPAVIAU-DPAVIAC DATASET

Feature selection	OA	AA	κ
CDIG	0.8285	0.8363	0.7850
CDFSC	0.8546	0.8661	0.8175
TDRF	0.8335	0.8452	0.7913
CDRF	0.8376	0.8509	0.7961
TDIRF_1	0.8495	0.8630	0.8111
TDIRF_2	0.8409	0.8522	0.8004
CDIRF_1	0.8531	0.8639	0.8155
CDIRF_2	**0.8617**	**0.8735**	**0.8262**

For the second cross-scene HSI dataset DPaviaU-DPaviaC, the classification accuracies with respect to the dimension of selected features are shown in Fig. 11. First of all, we compared algorithms based on I-ReliefF. By comparing CDIRF_1 with CDIRF_2, it can be seen that their accuracies in different feature dimensions are similar, but CDIRF_2 always performs better. It can be seen that using squared Euclidean distance makes classification results better. We also compared CDIRF_2 with TDIRF_2. When the dimension of selected features is small, the accuracies of CDIRF_2 and TDIRF_2 are close. But with the increase of feature dimension, the classification accuracy of the CDIRF_2 increases rapidly, and enlarges the accuracy gap between CDIRF_2 and TDIRF_2. It reflects the effectiveness of the cross-domain feature selection. In addition, TDIRF_1 performs well than CDIRF_1 when the dimension of selected features is small, e.g., $N_F \in \{5, 10\}$. But with the increase of feature dimension, the accuracy of CDIRF_1 rises quickly and keeps a certain advantage over TDIRF_1.

In addition, the classification accuracy of CDRF is higher than the accuracy of TDRF, but they are still lower than CDIRF_1 and CDIRF_2. As for CDIG, CDIG performs poorly when the feature dimension is $N_F \in \{5, 10, \ldots, 40\}$, and is always lower than CDIRF_1 and CDIRF_2. The accuracies of CDIRF_1 and CDIRF_2 are always higher than CDFSC when the feature dimension is $N_F \in \{5, \ldots, 40\}$. In general, although the accuracy curves are partially overlapping when the dimension of selected features is small, it can be seen that CDIRF_2 performs best in different feature dimensions among these algorithms. For these feature selection algorithms, the selected feature subsets sized $N_F = 20$ from one running experiment are illustrated in Fig. 12. We also perform correlation coefficient analysis on these feature subsets in Fig. 13. It can be seen that the correlation of feature subsets selected by CDIRF on the DPaviaU-DPaviaC dataset is lower than other algorithms. In addition, Fig. 14 contains the classification maps obtained by eight algorithms on a set of experimental data when $N_F = 20$. And we can find CDIRF_2 has advantages on the land cover class Bitumen.

The mean accuracies (OA, AA, and κ) of these algorithms across different feature dimensions are summarized in Table V. It can be seen that among the eight algorithms, the proposed CDIRF_2 obtains the highest accuracies in the three accuracy evaluation criteria. Compared with single-scene-based algorithms TDIRF_2 and TDRF, OA of CDIRF_2 is increased by 0.0208 and 0.0282, respectively. In the comparison of cross-scene algorithms, the OA of CDIRF_2 has the increases of 0.0241, 0.0071, and 0.0332 compared with CDRF, CDFSC, and CDIG.

For the third cross-scene HSI dataset RPaviaC-RPaviaU, the classification accuracies with respect to the dimension of selected features are shown in Fig. 15. At first, we compared algorithms based on I-ReliefF. By comparing CDIRF_1 with CDIRF_2, it can be seen that when the feature dimension is low, e.g., $N_F \in \{5, 10\}$, the accuracy of CDIRF_1 is lower than

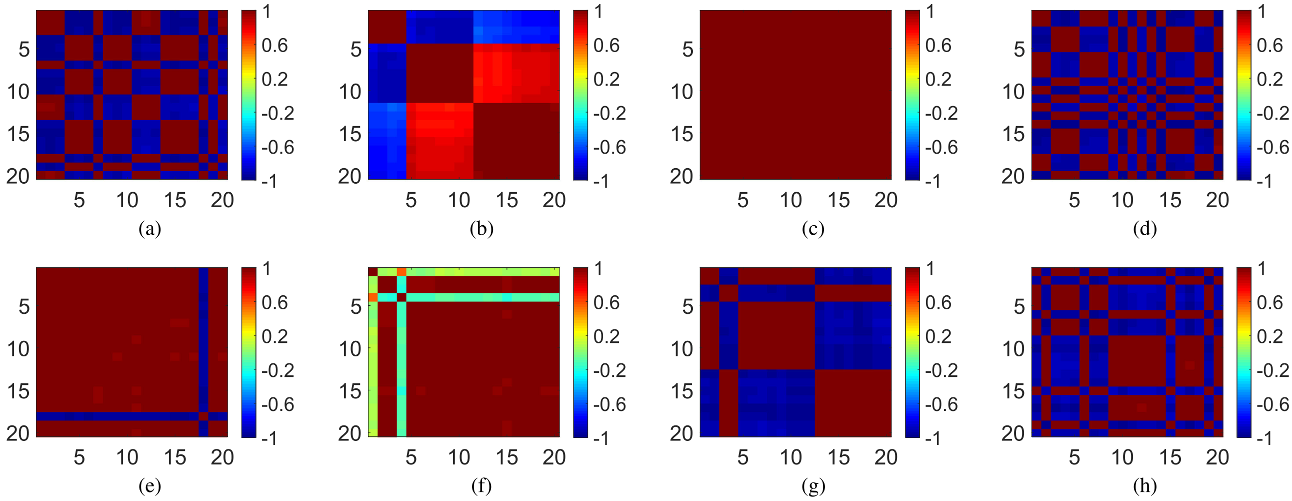


Fig. 9. Correlation coefficient matrices of feature subsets selected by eight algorithms on EShanghai-EHangzhou dataset. (a) CDIG. (b) CDFSC. (c) TDRF. (d) CDRF. (e) TDIRF$_1$. (f) TDIRF$_2$. (g) CDIRF$_1$. (h) CDIRF$_2$.

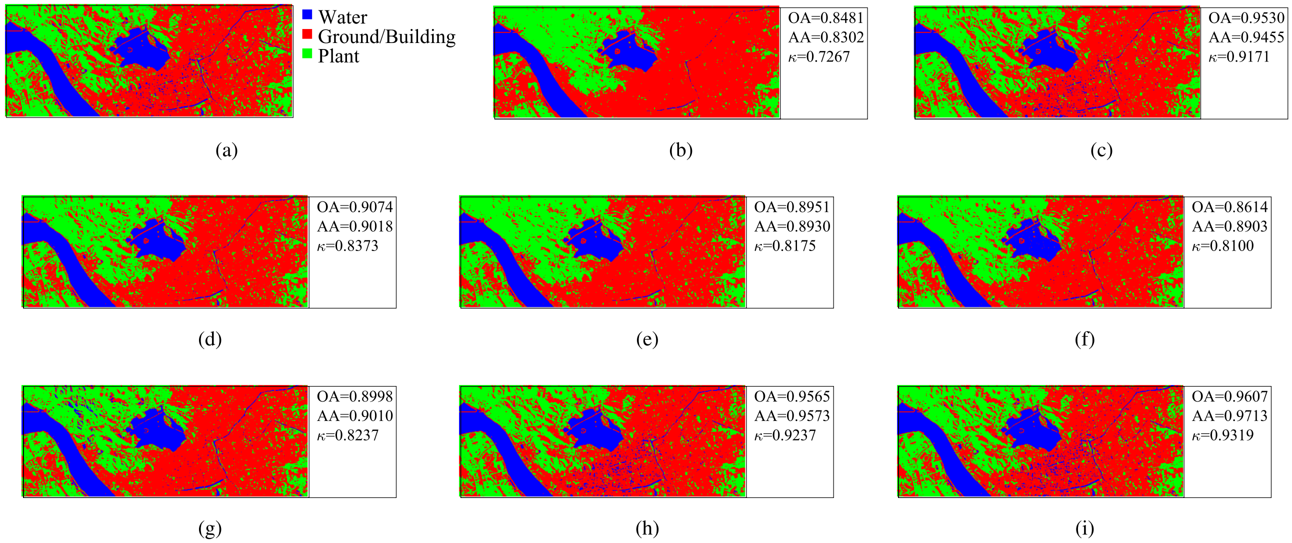


Fig. 10. Classification maps of EShanghai-EHangzhou dataset. (a) Groundtruth map. (b) The classification map over the selected features of CDIG. (c) The classification map over the selected features of CDFSC. (d) The classification map over the selected features of TDRF. (e) The classification map over the selected features of CDRF. (f) The classification map over the selected features of TDIRF$_1$. (g) The classification map over the selected features of TDIRF$_2$. (h) The classification map over the selected features of CDIRF$_1$. (i) The classification map over the selected features of CDIRF$_2$.

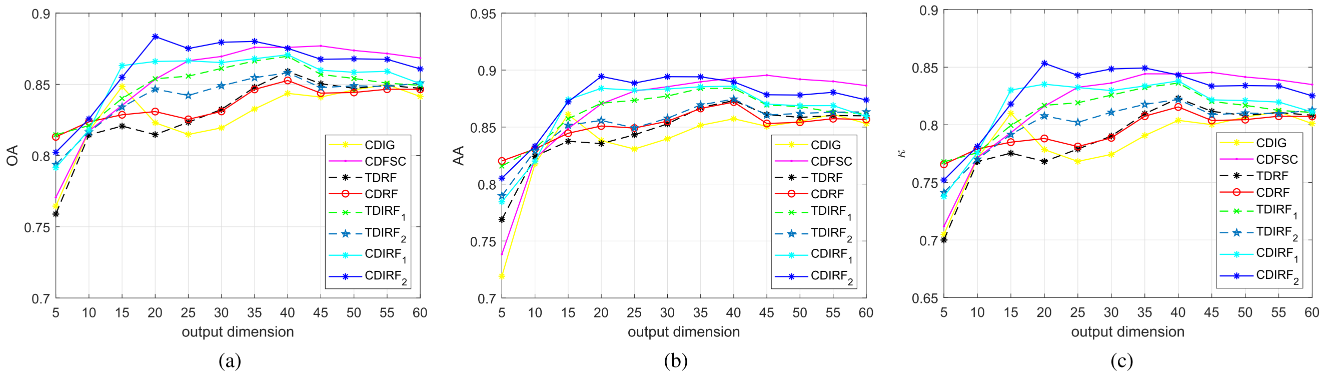
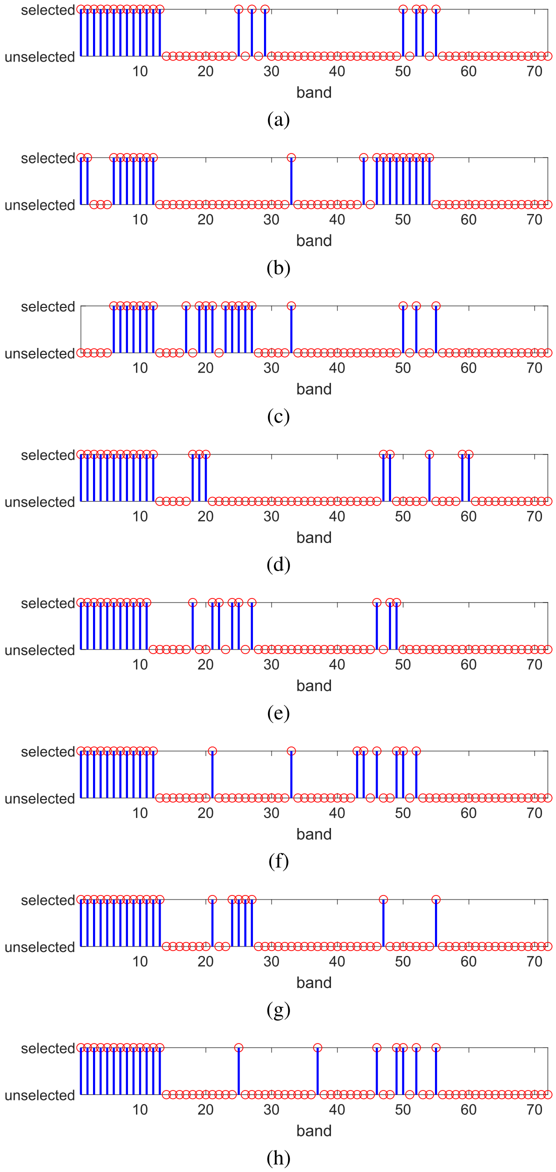


Fig. 11. Accuracies on DPaviaU-DPaviaC dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF$_1$, TDIRF$_2$, CDIRF$_1$, and CDIRF$_2$. (a) OA. (b) AA. (c) κ.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 12. Selected features on DPaviaU-DPaviaC dataset. (a) CDIG. (b) CDFSC. (c)TDRF. (d) CDRF. (e) TDIRF_1. (f) TDIRF_2. (g) CDIRF_1. (h) CDIRF_2.

TABLE VI
ACCURACIES ON RPAVIAC-RPAVIAU DATASET

Feature selection	OA	AA	κ
CDIG	0.6726	0.7341	0.5551
CDFSC	0.6312	0.7160	0.5071
TDRF	0.6792	0.7421	0.5644
CDRF	0.6830	0.7446	0.5666
TDIRF_1	0.6868	0.7532	0.5760
TDIRF_2	0.6942	0.7513	0.5795
CDIRF_1	0.6952	0.7515	0.5868
CDIRF_2	**0.7026**	**0.7597**	**0.5954**

CDIRF_2. As the feature dimension increases, CDIRF_1 gradually reduces the gap with CDIRF_2, but it is still lower than CDIRF_2. This means that the squared Euclidean distance has a better impact on the performance of the CDIRF algorithm. We compared CDIRF_2 with TDIRF_2. The accuracy of TDIRF_2 is always lower than CDIRF_2. We also compared CDIRF_1 with TDIRF_1. When the feature dimension is low, the accuracy of TDIRF_1 is slightly higher than CDIRF_1. But with the increase of the feature dimension, the performance of CDIRF_1 is better than TDIRF_1. Overall, CDIRF performs better than TDIRF on the RPaviaC-RPaviaU dataset.

In addition, We compared CDIRF_2 with CDRF. The accuracy of CDRF is always lower than CDIRF_2. The accuracy of CDIRF_1 is similar to CDRF when the feature dimension is low, but as the feature dimension increases, the accuracy of CDIRF_1 is much higher than CDRF. As for CDIG, the accuracy of CDIG is floating, and is always lower than CDIRF_1 and CDIRF_2. The accuracy of CDFSC is also lower than CDIRF_1 and CDIRF_2. We also compared CDIRF_2 with TDRF; we can see that TDRF is significantly lower than CDIRF_2. For these feature selection algorithms, the feature subsets sized $N_F = 20$ from one running experiment are illustrated in Fig. 16. And the correlation coefficient matrices on these feature subsets are shown in Fig. 17. It can be seen that CDIRF selects more weakly correlated features on the RPaviaC-RPaviaU dataset, while other algorithms prefer to select a feature subset of strongly correlated features. In addition, Fig. 18 contains the classification maps obtained by eight algorithms on a set of experimental data when $N_F = 20$. And we can find CDIRF_2 has advantages on the land-cover classes Meadows and Bitumen.

In Table VI, we list the average accuracies of the eight algorithms across different feature dimensions. It can be seen that the accuracy of CDIRF_2 is the highest accuracy in OA, AA, and κ. Comparing with TDIRF_1 and TDIRF_2, the OA of CDIRF_2 increased by 0.0158 and 0.0084, respectively. In addition, the OA of CDIRF_2 is 0.0196, 0.0300, and 0.0714 higher than CDRF, CDIG, and CDFSC, respectively. Comparing CDIRF_2 with TDRF, the OA of CDIRF_2 has increased by 0.0234.

In summary, CDIRF_2 not only has the highest classification accuracies on the three datasets, but also more accurate in selecting features. This is very meaningful in the HSI classification.

E. Robustness Analysis

To study the robustness of these algorithms, we still use the training set and testing set of the three datasets used in Section IV-D. For each dataset, we add outliers to the training set taken from the target scene. Specifically, we modify the label of one sample to an incorrect class in each class, so there is an outlier in each class. The experimental results on algorithms are listed in Table VII. It can be seen that when there are outliers in the training set, the accuracies of CDIRF_1 and CDIRF_2 are still higher than other compared algorithms. It means that the proposed algorithms CDIRF_1 and CDIRF_2 perform better than other compared algorithms in dealing with outliers. As for CDIRF_1 and CDIRF_2, the accuracies of CDIRF_1 are higher than CDIRF_2 on the EShanghai-EHangzhou and DPaviaU-DPaviaC datasets. Combined with the results in Section IV-D, the performance of CDIRF_2 is better than CDIRF_1 on datasets with a few outliers, while CDIRF_1 has better robustness.

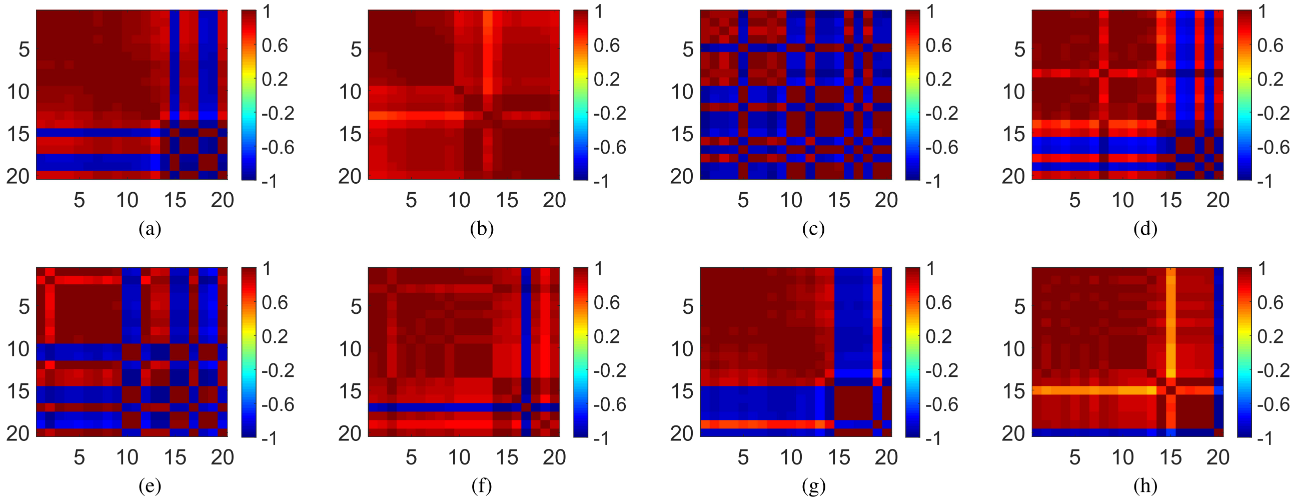


Fig. 13. Correlation coefficient matrices of feature subsets selected by eight algorithms on DPaviaU-DPaviaC dataset. (a) CDIG. (b) CDFSC. (c) TDRF. (d) CDRF. (e) TDIRF_1. (f) TDIRF_2. (g) CDIRF_1. (h) CDIRF_2.

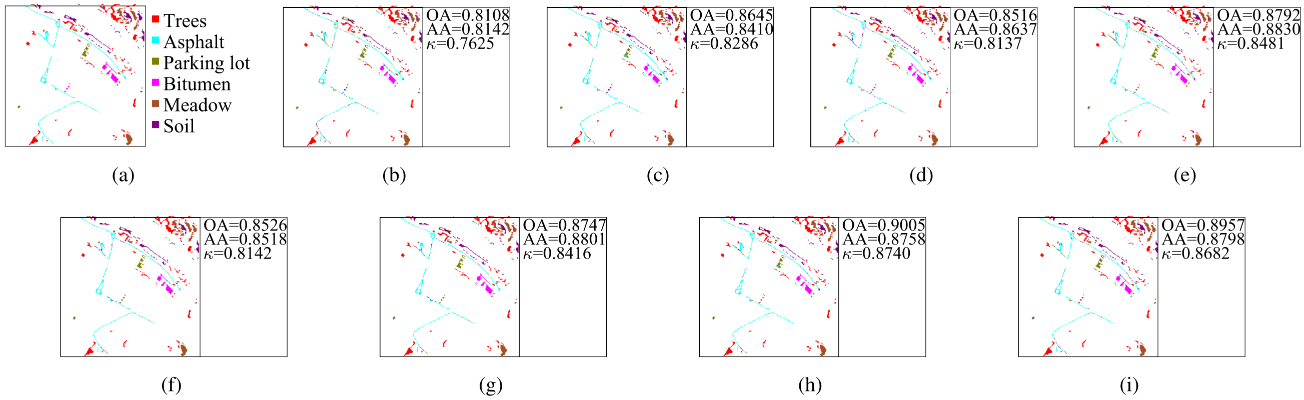


Fig. 14. Classification maps of DPaviaU-DPaviaC dataset. (a) Groundtruth map. (b) The classification map over the selected features of CDIG. (c) The classification map over the selected features of CDFSC. (d) The classification map over the selected features of TDRF. (e) The classification map over the selected features of CDRF. (f) The classification map over the selected features of TDIRF_1. (g) The classification map over the selected features of TDIRF_2. (h) The classification map over the selected features of CDIRF_1. (i) The classification map over the selected features of CDIRF_2.

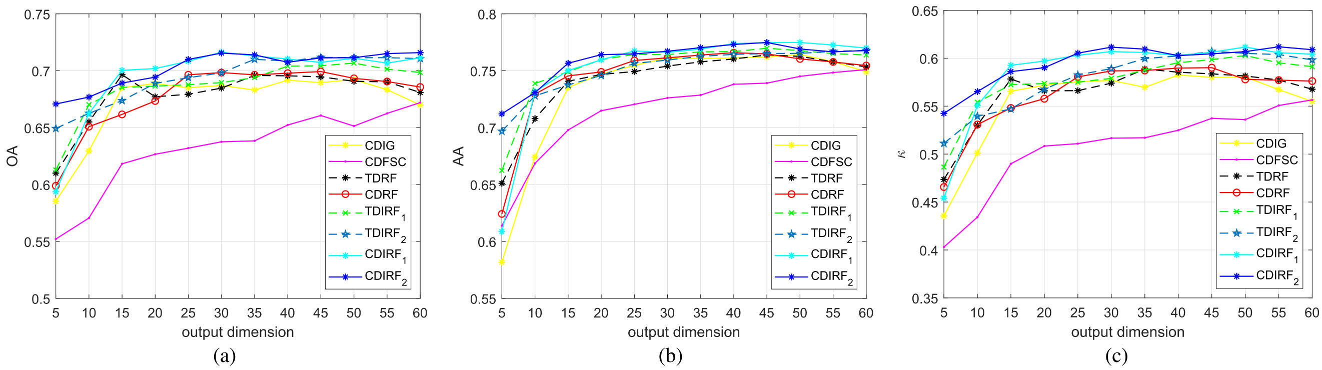
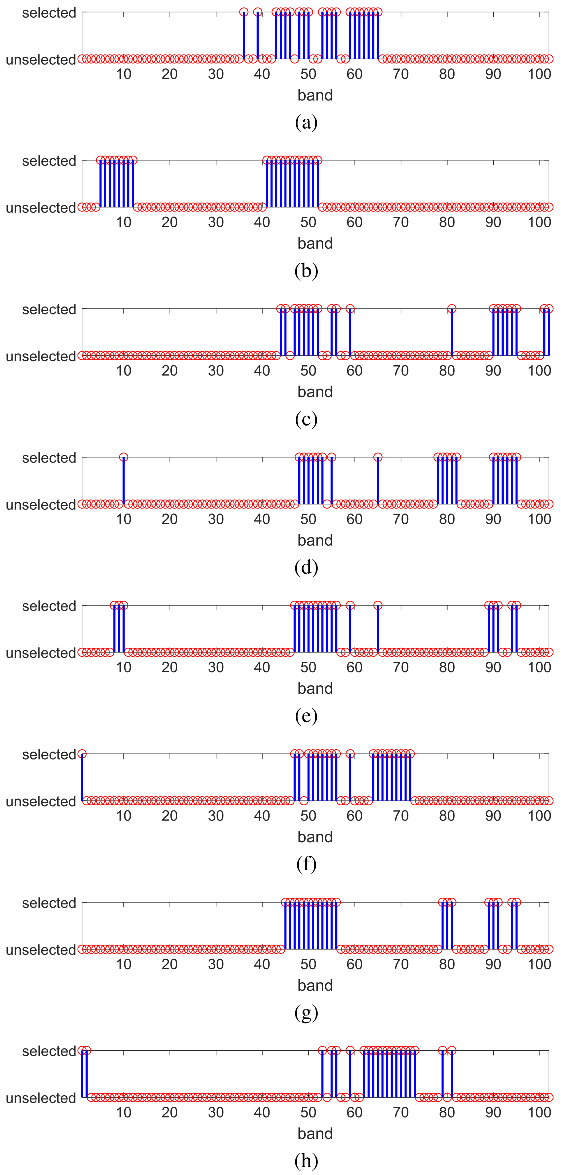


Fig. 15. Accuracies on RPaviaC-RPaviaU dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF_1, TDIRF_2, CDIRF_1, and CDIRF_2. (a) OA. (b) AA. (c) κ.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 16. Selected features on RPaviaC-RPaviaU dataset. (a) CDIG. (b) CDFSC. (c)TDRF. (d) CDRF. (e) TDIRF_1. (f) TDIRF_2. (g) CDIRF_1. (h) CDIRF_2.

TABLE VII
ROBUSTNESS ANALYSIS ON OUTLIER SAMPLES. SIGNIFICANCE OF BOLD ENTITIES ARE HIGHEST ACCURACY AMONG COMPARED METHODS.

Datasets	Algorithm	OA	AA	κ
EShanghai-EHangzhou	CDIG	0.8809	0.8872	0.7929
	CDFSC	0.7735	0.8001	0.6184
	TDRF	0.8336	0.8548	0.7152
	CDRF	0.8766	0.8865	0.7855
	TDIRF_1	0.8605	0.8785	0.7594
	TDIRF_2	0.8477	0.8716	0.7408
	CDIRF_1	**0.9208**	**0.9288**	**0.8619**
	CDIRF_2	0.8988	0.9151	0.8258
DPaviaU-DPaviaC	CDIG	0.7725	0.7797	0.7161
	CDFSC	0.7781	0.7860	0.7224
	TDRF	0.7771	0.7898	0.7212
	CDRF	0.77.58	0.7930	0.7197
	TDIRF_1	0.7871	0.7991	0.7335
	TDIRF_2	0.7821	0.7923	0.7270
	CDIRF_1	**0.8032**	0.8137	**0.7536**
	CDIRF_2	0.8030	**0.8178**	0.7531
RPaviaC-RPaviaU	CDIG	0.5852	0.6744	0.4535
	CDFSC	0.5760	0.6869	0.4425
	TDRF	0.5991	0.6839	0.4589
	CDRF	0.5976	0.6854	0.4638
	TDIRF_1	0.5960	0.6849	0.4565
	TDIRF_2	0.5906	0.6832	0.4501
	CDIRF_1	0.6172	0.6842	0.4863
	CDIRF_2	**0.6177**	**0.6906**	**0.4866**

TABLE VIII
COMPUTATIONAL COMPLEXITY

Name	Computational complexity
CDIG	$O(F)$
CDFSC	$O\left(\left(N^{\mathcal{T}}+N^{\mathcal{S}}\right)F'^2\right)$
TDRF	$O(TNF)$
CDRF	$O\left(T\left(N^{\mathcal{T}}+N^{\mathcal{S}}\right)F\right)$
TDIRF_1	$O\left(TN^2F\right)$
TDIRF_2	$O\left(TN^2F\right)$
CDIRF_1	$O\left(T\left(N^{\mathcal{T}}+N^{\mathcal{S}}\right)^2F\right)$
CDIRF_2	$O\left(T\left(N^{\mathcal{T}}+N^{\mathcal{S}}\right)^2F\right)$

F. Effect of Sample Size on Algorithm Performance

In order to test the performance of compared algorithms in the case of high dimension and small sample size, we reduced the number of samples in the experiment and set the following number of samples: $\{50\text{-}3, 100\text{-}3, 200\text{-}3, 200\text{-}5\}$, where $N^{\mathcal{S}}$-$N^{\mathcal{T}}$ means that we selected $N^{\mathcal{S}}$ samples from source scene and selected $N^{\mathcal{T}}$ samples from target scene. The experimental results can be seen in Figs. 19–21. First, we compared 200-3 with 200-5. As the number of labeled samples in the target scene decreases, the accuracies of the compared algorithms are greatly reduced. But the accuracies of the proposed CDIRF_1 and CDIRF_2 are always higher than other compared algorithms. In addition, we also analyzed the three cases of 50-3, 100-3, and 200-3. It can be seen that the accuracies of compared algorithms have not changed significantly. And the accuracies of CDIRF_1 and CDIRF_2 are still higher than other compared algorithms. In conclusion, the performance of CDIRF_1 and CDIRF_2 is better than other compared algorithms when dealing with high dimension and small sample size problem.

G. Computational Complexity

The complexity of each algorithm is shown in the Table VIII, where T is the total number of iterations, F is the feature dimensionality, F' is size of selected feature subset specified by algorithm, and N, $N^{\mathcal{S}}$, $N^{\mathcal{T}}$ are the numbers of samples.

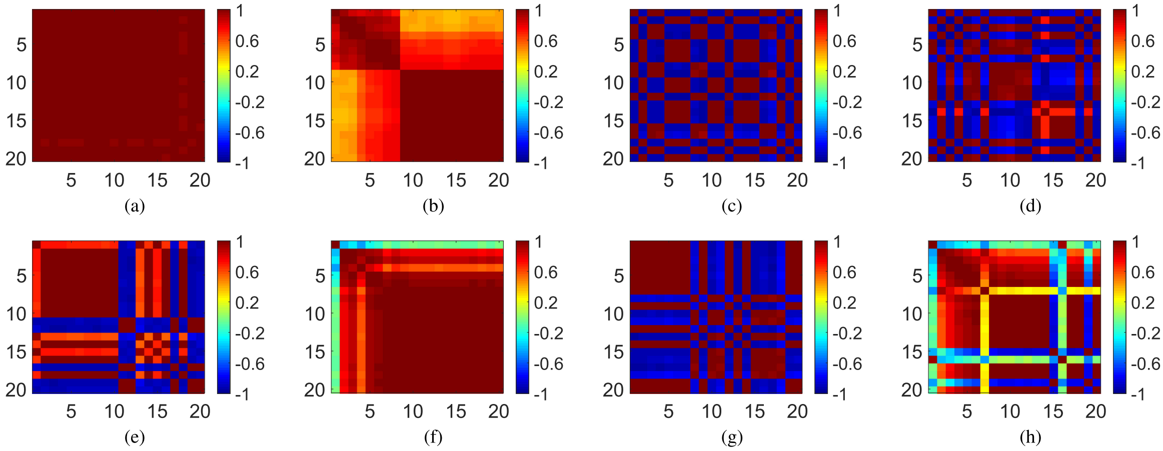


Fig. 17. Correlation coefficient matrices of feature subsets selected by eight algorithms on RPaviaC-RPaviaU dataset. (a) CDIG. (b) CDFSC. (c) TDRF. (d) CDRF. (e) TDIRF_1. (f) TDIRF_2. (g) CDIRF_1. (h) CDIRF_2.

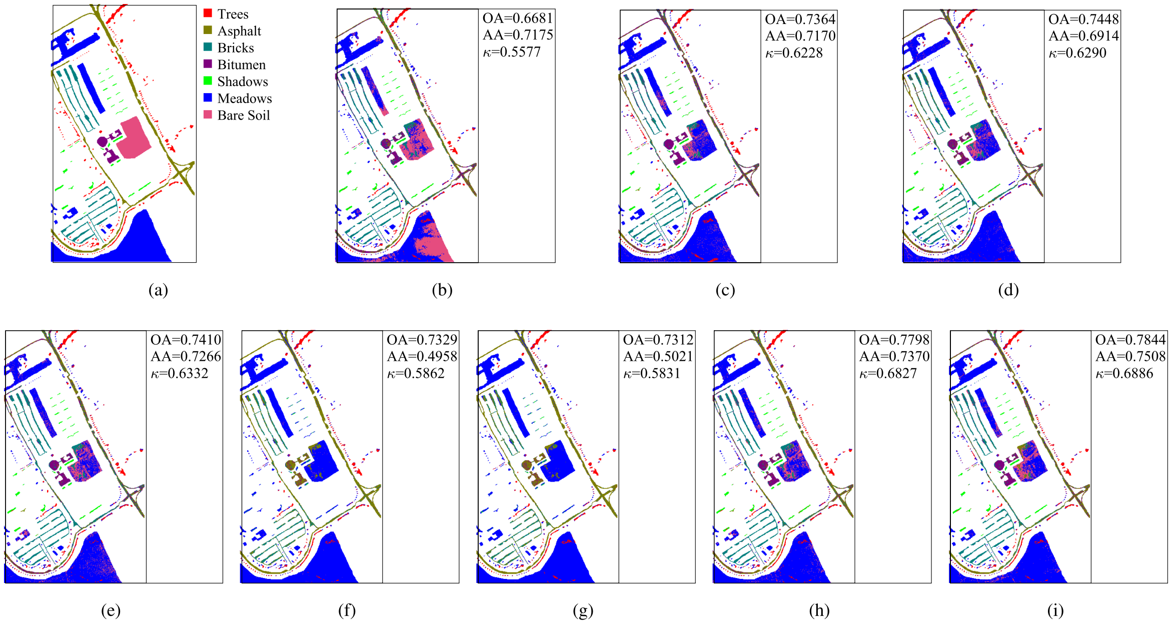


Fig. 18. Classification maps of RPaviaC-RPaviaU dataset. (a) Groundtruth map. (b) The classification map over the selected features of CDIG. (c) The classification map over the selected features of CDFSC. (d) The classification map over the selected features of TDRF. (e) The classification map over the selected features of CDRF. (f) The classification map over the selected features of TDIRF_1. (g) The classification map over the selected features of TDIRF_2. (h) The classification map over the selected features of CDIRF_1. (i) The classification map over the selected features of CDIRF_2.

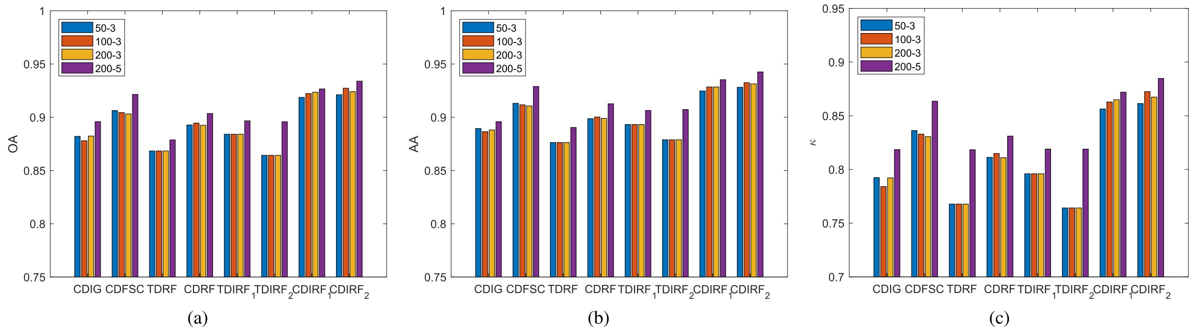
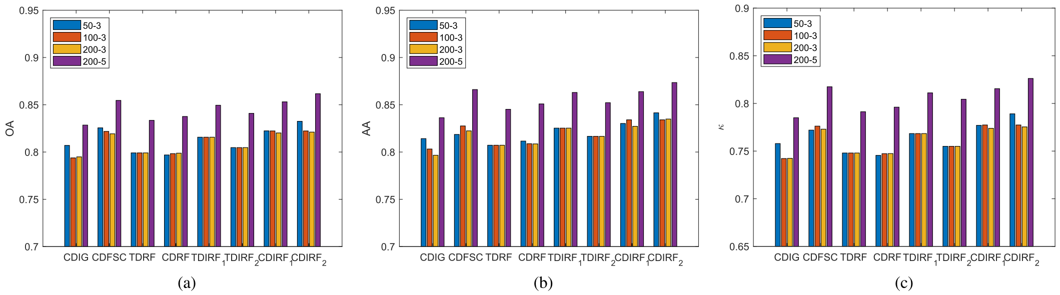
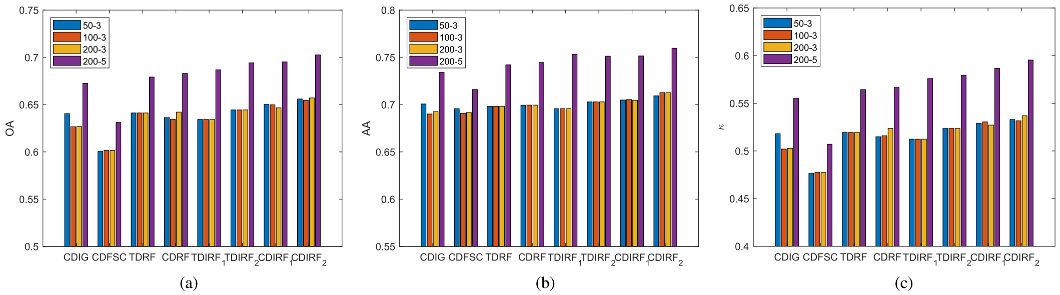


Fig. 19. Accuracies on EShanghai-EHangzhou dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF_1, TDIRF_2, CDIRF_1, and CDIRF_2. (a) OA. (b) AA. (c) κ.



(a) (b) (c)

Fig. 20. Accuracies on DPaviaU-DPaviaC dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF$_1$, TDIRF$_2$, CDIRF$_1$, and CDIRF$_2$. (a) OA. (b) AA. (c) κ.



(a) (b) (c)

Fig. 21. Accuracies on RPaviaC-RPaviaU dataset obtained by CDIG, CDFSC, TDRF, CDRF, TDIRF$_1$, TDIRF$_2$, CDIRF$_1$, and CDIRF$_2$. (a) OA. (b) AA. (c) κ.

V. Conclusion

In this article, we developed a cross-scene feature selection algorithm called CDIRF. Under the premise of supervised learning, CDIRF uses the information of source scene to help the feature selection of target scene. Hence, the feature selection performs well in the target scene, which lacks labeled samples. As a cross-scene extension of the traditional I-ReliefF, CDIRF updates the weights on the basis of the weighted feature space, and also considers the influence of spectral shift between source and target scenes. In order to improve the separability of selected features between different land-cover classes and the consistency of the selected features between different scenes, CDIRF considers eight types of samples. CDIRF also combines the information from two scenes to determine the possibility of a sample being an outlier. Under this novel feature weight updating rule, experimental results on three datasets demonstrate that the newly proposed method CDIRF can effectively select feature subsets to improve the classification accuracy. And compared with squared Euclidean distance, absolute distance is more robust for CDIRF. We also analyzed the impact of the number of labeled samples on CDIRF, and it can be seen that CDIRF has great performance in dealing with high dimension and small sample size problem.

However, despite the high classification accuracies achieved by the feature subset selected by CDIRF, CDIRF does not guarantee low redundancy of the selected feature subset. In future work, we may try to incorporate the low redundancy criterion into CDIRF.

References

[1] W. Sun, L. Tian, Y. Xu, D. Zhang, and Q. Du, "Fast and robust self-representation method for hyperspectral band selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 5087–5098, Nov. 2017.

[2] Y. Qian, F. Yao, and S. Jia, "Band selection for hyperspectral imagery using affinity propagation," *IET Comput. Vis.*, vol. 3, pp. 213–222(9), 2009.

[3] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[4] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[5] S. A. Medjahed and M. Ouali, "Band selection based on optimization approach for hyperspectral image classification," *Egypt. J. Remote Sens. Space Sci.*, vol. 21, no. 3, pp. 413–418, 2018.

[6] A. Datta, S. Ghosh, and A. Ghosh, "Combination of clustering and ranking techniques for unsupervised band selection of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2814–2823, Jun. 2015.

[7] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[8] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.

[9] C. Lai, W. Yeh, and C. Chang, "Gene selection using information gain and improved simplified swarm optimization," *Neurocomputing*, vol. 218, pp. 331–338, 2016.

[10] H. Marwa, B. Slim, H. Chih-Cheng, and B. S. Lamjed, "A multi-objective hybrid filter-wrapper evolutionary approach for feature selection,*" Memetic Comput.*, vol. 11, no. 2, pp. 193–208, 2018.

[11] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Comput. Appl.*, vol. 32, no. 12, p. 7839–7857, 2020.

[12] X. Cao, C. Wei, Y. Ge, J. Feng, and J. A. Zhao, "Semi-supervised hyperspectral band selection based on dynamic classifier selection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1289–1298, Apr. 2019.

[13] M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[14] M. Wang, C. Wu, L. Wang, D. Xiang, and X. Huang, "A feature selection approach for hyperspectral image based on modified ant lion optimizer," *Knowl.-Based Syst.*, vol. 168, pp. 39–48, 2019.

[15] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, Jan. 2020.

[16] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 703–715, May 2019.

[17] W. Sun, J. Peng, and G. Yang, "Correntropy-based sparse spectral clustering for hyperspectral band selection," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 3, pp. 484–488, Mar. 2020.

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

[19] Y. Yuan, J. Lin, and W. Qi, "Dual-clustering-based hyperspectral band selection by contextual analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1431–1445, Mar. 2016.

[20] M. Bevilacqua and Y. Berthoumieu, "Multiple-feature kernel-based probabilistic clustering for unsupervised band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6675–6689, Sep. 2019.

[21] H. Zhai, H. Zhang, L. Zhang, and P. Li, "Laplacian-regularized low-rank subspace clustering for hyperspectral image band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1723–1740, Mar. 2019.

[22] W. Sun, L. Zhang, L. Zhang, and Y. M. Lai, "A dissimilarity-weighted sparse self-representation method for band selection in hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4374–4388, Sep. 2016.

[23] S. Mallik, T. Bhadra, and U. Maulik, "Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data," *IEEE Trans. Nanobiosci.*, vol. 16, no. 1, pp. 3–10, Jan. 2017.

[24] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Mutual-information-based semi-supervised hyperspectral band selection with high discrimination, high information, and low redundancy," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2956–2969, May 2015.

[25] C. I. Chang, Y. M. Kuo, S. Chen, C. C. Liang, K. Y. Ma, and P. F. Hu, "Self-mutual information-based band selection for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, early access, Oct. 2, 2020, doi: 10.1109/TGRS.2020.3024602.

[26] B. Xu, X. Li, W. Hou, Y. Wang, and Y. Wei, "A similarity-based ranking method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sens.*, early access, Jan. 14, 2021, doi: 10.1109/TGRS.2020.3048138.

[27] 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.

[28] M. Ye, Y. Qian, J. Zhou, and Y. Y. Tang, "Dictionary learning-based feature-level domain adaptation for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1544–1562, Mar. 2017.

[29] J. Peng, W. Sun, L. Ma, and Q. Du, "Discriminative transfer joint matching for domain adaptation in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 972–976, Jun. 2019.

[30] W. Wang, L. Ma, M. Chen, and Q. Du, "Joint correlation alignment-based graph neural network for domain adaptation of multitemporal hyperspectral remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3170–3184, Mar. 2021.

[31] H. Wei, L. Ma, Y. Liu, and Q. Du, "Combining multiple classifiers for domain adaptation of remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1832–1847, Jan. 2021.

[32] C. Ni, W. Liu, Q. Gu, X. Chen, and D. Chen, "A cluster based feature selection method for cross-project software defect prediction," *J. Comput. Sci. Technol.*, vol. 32, no. 6, pp. 1090–1107, 2017.

[33] M. Ye, Y. Xu, H. Lu, K. Yan, and Y. Qian, "Cross-scene feature selection for hyperspectral images based on cross-domain information gain," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4764–4767.

[34] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Inform.*, vol. 85, pp. 189–203, 2018.

[35] M. Ye, Y. Xu, C. Ji, H. Chen, H. Lu, and Y. Qian, "Feature selection for cross-scene hyperspectral image classification using cross-domain ReliefF," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 17, no. 5, 2019, Art no. 1950039.

[36] Y. Sun, "Iterative RELIEF for feature weighting: Algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, Jun. 2007.

[37] L. Chen and D. Chen, "Alignment based feature selection for multi-label learning," *Neural Process. Lett.*, vol. 50, no. 3, pp. 2323–2344, 2019.

[38] S. P. Patel and S. Upadhyay, "Euclidean distance based feature ranking and subset selection for bearing fault diagnosis," *Expert Syst. with Appl.*, vol. 154, 2020, Art no. 113400.

[39] B. C. Kuo, H. H. Ho, C. H. Li, and C. C. Hung, "A kernel-based feature selection method for SVM with RBF kernel for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 317–326, Jan. 2014.

[40] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 231–240, 2006.

[41] W. Zhang, X. Li, and L. Zhao, "A fast hyperspectral feature selection method based on band correlation analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1750–1754, Nov. 2018.

Chengjie Zhang received the B.E. degree in computer science and technology from China Jiliang University, Hangzhou, China, in 2019. He is currently working toward the master's degree in computer science and technology with China Jiliang University.

His research interests include hyperspectral image processing and machine learning.

Minchao Ye (Member, IEEE) received the B.E. degree in computer science and technology from Sichuan University, China, in 2010 and the Ph.D. degree in computer science and technology from Zhejiang University, China, in 2016.

Since 2016, he has been with the College of Information Engineering, China Jiliang University, Hangzhou, China, where he is currently an Associate Professor in Computer Science and Technology. His research interests include hyperspectral image processing and applications, machine learning, and pattern recognition.

Ling Lei received the B.E. degree in computer science and technology and the M.E. degree in control theory and control engineering from the Taiyuan University of Technology, Taiyuan, China, in 1997 and 2002, respectively.

Since 2002, she has been with the College of Information Engineering, China Jiliang University, Hangzhou, China.

Her current research interests include hyperspectral image processing, machine learning, and pattern recognition.

Yuntao Qian (Senior Member, IEEE) received the B.E. and M.E. degrees in automatic control from Xi'an Jiaotong University, Xi'an, China, in 1989 and 1992, respectively, and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 1996.

During 1996–1998, he was a Postdoctoral Fellow with the Northwestern Polytechnical University, Xi'an, China. Since 1998, he has been with the College of Computer Science, Zhejiang University, Hangzhou, China, where he became a Professor in 2002. During 1999–2001, 2006, 2010, 2013, 2015–2016, and 2018, he was a Visiting Professor at Concordia University, Hong Kong Baptist University, Carnegie Mellon University, the Canberra Research Laboratory of NICTA, Macau University, and Griffith University. His research interests include machine learning, signal and image processing, pattern recognition, and hyperspectral imaging.

Prof. Qian is currently an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.