remote sensing

MDPI

Article

A Novel Tri-Training Technique for the Semi-Supervised Classification of Hyperspectral Images Based on Regularized Local Discriminant Embedding Feature Extraction

Depin Ou [1,†], Kun Tan [1,2,†,*], Qian Du [3], Jishuai Zhu [1,4], Xue Wang [1] and Yu Chen [1,*]

[1] Key Laboratory for Land Environment and Disaster Monitoring of NASG, China University of Mining and Technology, Xuzhou 221116, China; tb17160017b2@cumt.edu.cn (D.O.); zhujishuai@charmingglobe.com (J.Z.); tb16160015b2@cumt.edu.cn (X.W.)

[2] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

[3] Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA; du@ece.msstate.edu

[4] Chang Guang Satellite Technology Co. Ltd., Changchun 130033, China

* Correspondence: tankun@cumt.edu.cn (K.T.); chenyu@cumt.edu.cn (Y.C.); Tel.: +86-051683591309 (K.T.)

† These authors contributed equally to this work.

Received: 1 February 2019; Accepted: 15 March 2019; Published: 18 March 2019

check for updates

Abstract: This paper introduces a novel semi-supervised tri-training classification algorithm based on regularized local discriminant embedding (RLDE) for hyperspectral imagery. In this algorithm, the RLDE method is used for optimal feature information extraction, to solve the problems of singular values and over-fitting, which are the main problems in the local discriminant embedding (LDE) and local Fisher discriminant analysis (LFDA) methods. An active learning method is then used to select the most useful and informative samples from the candidate set. In the experiments undertaken in this study, the three base classifiers were multinomial logistic regression (MLR), *k*-nearest neighbor (KNN), and random forest (RF). To confirm the effectiveness of the proposed RLDE method, experiments were conducted on two real hyperspectral datasets (Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) and Reflective Optics System Imaging Spectrometer (ROSIS)), and the proposed RLDE tri-training algorithm was compared with its counterparts of tri-training alone, LDE, and LFDA. The experiments confirmed that the proposed approach can effectively improve the classification accuracy for hyperspectral imagery.

Keywords: feature extraction; regularized local discriminant embedding (RLDE); semi-supervised tri-training; hyperspectral imagery

1. Introduction

Hyperspectral sensors have hundreds of spectrally contiguous bands, which can provide abundant spectral information [1]. Due to the high spectral resolution, hyperspectral images (HSIs) have been widely used in applications such as agricultural mapping [2], water quality analysis [3], and mineral identification [4]. The key component in these applications is the classification. Some of the conventional supervised classifiers can offer satisfactory classification performances, but the performance is dependent on both the quantity and quality of the training samples. However, labeled training samples can be costly, difficult, and time-consuming to obtain, and it is difficult for the traditional supervised classifiers to obtain good performances when the number of labeled training samples is limited [5]. Despite the fact that deep learning based methods have now been developed for HSI classification, including convolutional neural networks (CNNs) [6–8], 3D convolutional neural networks (3D-CNNs) [9,10], and long short-term memory (LSTM) networks [11,12], these problems still exist. Therefore, how to use unlabeled samples to improve the classification performance has become a hot research topic. The use of unlabeled samples to improve the classification performance is known as semi-supervised learning [13]. Common semi-supervised learning algorithms include multi-view learning algorithms [14], self-learning algorithms [15], tri-training algorithms [16], graph-based approaches [17], and the transductive support vector machine (TSVM) algorithm [18]. High-dimensional data processing needs more storage and computation time [19,20]. In addition, the spectral bands in an HSI are highly correlated, and the classification performance deteriorates as the dimensionality increases (the Hughes phenomenon) with limited training samples [21,22]. Therefore, in order to reduce the time consumption and improve the classification performance, it is necessary to extract the useful spectral information before performing classification.

The basic technique of spectral information extraction is dimension reduction, the goal of which is to embed the high-dimensional data in a low-dimensional space containing the crucial information [23,24]. Research into dimension reduction has experienced rapid development in recent years. Linear dimension reduction methods obtain the spectral information in the low-dimensional space by building a linear model. Typical methods include principal component analysis (PCA) [25], linear discriminant analysis (LDA) [26], direct linear discriminant analysis (DLDA) [27], and the maximum margin criterion (MMC) [28]. These methods are simple to operate, efficient, and have a strong generalization ability for linear datasets. However, these methods cannot obtain satisfactory performances in nonlinear datasets. Therefore, nonlinear dimension reduction methods have been proposed for use with nonlinear datasets [29]. Common nonlinear dimension reduction methods include kernel based approaches [30,31] and manifold learning algorithms [32]. In [33], kernel PCA was first proposed to solve the sparsity and dimensionality problems of nonlinear datasets. In [34], a new nonlinear dimension reduction method combining a kernel function with Fisher discriminant analysis was used in the classification of HSIs. In [35,36], Song et al. proposed models to learn a set of robust hash functions to map the high-dimensional data points into binary hash codes by effectively utilizing the local structural information. However, how to select a suitable kernel function lacks a theoretical basis.

The manifold learning algorithms depict the intrinsic structure of high-dimensional data by constructing a representation of the data lying in a low-dimensional manifold [31]. Tenenbaum [37] tried to preserve the geodesic distances based on multi-dimensional scaling, and proposed the isometric feature mapping (Isomap) method. In [38], locally linear embedding (LLE) was used to embed data points in a low-dimensional space by finding the optimal linear reconstruction in a small neighborhood. He et al. [39] subsequently proposed the neighborhood preserving embedding algorithm based on LLE, and regarded the error minimization as the objective function. In [40], the local discriminant embedding (LDE) algorithm was used to extend global LDA to a local version, so as to perform the local discriminant embedding in a graph embedding framework. However, the aforementioned manifold learning algorithms have singularity and cannot preserve the data diversity in the case of limited training samples.

Therefore, in this paper, we propose a new feature extraction method—regularized local discriminant embedding (RLDE)—to preserve the local feature information and overcome the singularity when training samples are limited. In order to make full use of the unlabeled samples, we select the semi-supervised tri-training algorithm. We also use an active learning method to select the unlabeled samples and use ensemble learning to improve the classification result.

2. Spatial Mean Filtering and Feature Extraction

$X = [x_1, x_2, \cdots, x_m] \in R^{n \times m}$ denotes the training dataset with n-dimensional feature vectors; $Y = [y_1, y_2, \cdots, y_m] \in R$ represents the corresponding labels; m is the number of training samples; and all the datasets are denoted as $\{x'_i\}_{i=1}^{l} \in R^n$, where l is the number of datasets.

2.1. Spatial Mean Filtering

To reduce noise and smooth the homogeneous regions, we first use spatial mean filtering to preprocess the HSIs. The spatial mean filtering of a labeled pixel X_i is denoted as:

$$X'_i = \frac{X_i + \sum_{k=1}^{w^2-1} v_k X_{ik}}{1 + \sum_{k=1}^{w^2-1} v_k}, \tag{1}$$

where w is the width of the neighborhood window; s $= w^2 - 1$ is the number of neighbors of X_i; $v_k = \exp\left\{-\gamma_0 ||X_i - X_{ik}||^2\right\}$ stands for the spectral distance of the neighboring pixels to the central pixel; and γ_0 represents the degree of filtering.

2.2. Local Discriminant Embedding (LDE)

LDE is a nonlinear supervised dimension reduction method. The local information of homogeneous and heterogeneous samples is preserved by defining inter-class graphs and within-class graphs [41,42]. The basic idea is to simultaneously attain between-class separation and within-class local structure preservation. The objective function of LDE is denoted as:

$$\begin{cases} J(V) = argmax \sum_{i,j} \|V^T x_i - V^T x_j\|^2 \omega\prime_{i,j} \\ s.t. \sum_{i,j} \|V^T x_i - V^T x_j\|^2 \omega_{i,j} = 1 \end{cases}, \tag{2}$$

where V is the optimal projection matrix; and ω', ω are the weight matrix of the heterogeneous neighboring sample points and the weight matrix of the nearest-neighbor sample points, which are defined as:

$$\omega\prime_{i,j} = \begin{cases} exp\left[-\|x_i - x_j\|^2/t\right] \; if \; x_i \in N(x_j) \; or \; x_j \in N(x_i) \\ and \; y_{x_i} \neq y_{x_j} \\ 0 \quad otherwise \end{cases}, \tag{3}$$

$$\omega_{i,j} = \begin{cases} exp\left[-\|x_i - x_j\|^2/t\right] \; if \; x_i \in N(x_j) \; or \; x_j \in N(x_i) \\ and \; y_{x_i} = y_{x_j} \\ 0 \quad otherwise \end{cases}, \tag{4}$$

where t is a constant parameter, and the value of t is the square of the mean value of the Euclidean distances between the sample points. $N(x)$ is the k neighborhood samples of training sample $\boldsymbol{x}$.

Equation (2) can be converted into:

$$J = \sum_{i,j} tr\left\{V^T (x_i - x_j)(x_i - x_j)^T V\right\}\omega\prime_{i,j}. \tag{5}$$

After conversion, we can obtain:

$$J = 2tr\left\{V^T X (D' - W\prime) X^T V\right\}. \tag{6}$$

Thus, the objective function can be written as follows:

$$\begin{cases} J(V) = 2tr\{V^T X(D' - W\prime)X^T V\} \\ s.t.\ 2tr\{V^T X(D - W)X^T V\} = 1 \end{cases}, \tag{7}$$

where D' and D are diagonal matrices, and the diagonal elements are $\mathrm{D'}_{i,i} = \sum \omega'_{i,j}$ and $\mathrm{D}_{i,i} = \sum \omega_{i,j}$. W and W' are affinity weight matrices, which are sparse and symmetric, as computed by Equations (3) and (4), respectively.

The optimal LDE projection is obtained by finding the eigenvectors corresponding to nonzero small eigenvalues of the following generalized Eigen-decomposition problem:

$$\mathrm{X}(\mathrm{D'} - \mathrm{W'})\mathrm{X}^{\mathrm{T}}\mathrm{V} = \lambda \mathrm{X}(\mathrm{D} - \mathrm{W})\mathrm{X}^{\mathrm{T}}\mathrm{V}. \tag{8}$$

2.3. Regularized Local Discriminant Embedding (RLDE)

The manifold structure of all the data can be obtained after simulating the manifold structure of the training data through the LDE and local Fisher discriminant analysis (LFDA) algorithms [43,44]. These algorithms can not only detect the internal structure, but can also preserve the discriminative structure of the data [45]. However, the LDE and LFDA algorithms have the following shortcomings: (1) when the number of training samples is smaller than the spectral dimension, the singular value problem occurs in the process of solving the projection vector and (2) in attempting to preserve the local difference information, the over-fitting problem occurs [46]. Therefore, we propose the RLDE method to solve the above problems. The objective function of this method is derived from Equation (2):

$$J(V) = \begin{cases} argmax\left\{\alpha \frac{\sum_{i,j} \|V^T X_i - V^T X_j\|^2 \omega\prime_{i,j}}{\sum_{i,j} \|V^T X_i - V^T X_j\|^2 \omega_{i,j}} + (1-\alpha) R_{reg} f(x)\right\} \\ s.t.\ VV^T = 1 \end{cases}, \tag{9}$$

where

$$R_{reg} f(x) = \frac{\sum_{i,j} \|V^T X_i - V^T X_j\|^2}{\sum_{i,j} \|V^T X_i - V^T X_j\|^2 \omega_{i,j}} \tag{10}$$

is the added regular constraint, and α is a regularization parameter with a value of [0,1]. Equation (10) is equivalent to:

$$\begin{cases} J(V) = argmax\left\{ \begin{array}{c} 2tr\{\alpha V^T X(D' - W')X^T V + (1-\alpha)V^T XX^T V\}/ \\ 2tr\{\alpha V^T X(D - W)X^T V + (1-\alpha)diag(V^T X(D - W)X^T V)XX^T\} \end{array} \right\} \\ s.t.\ VV^T = 1 \end{cases}. \tag{11}$$

The optimized objective of LDE is to maximize $\sum_{i,j} \|V^T X_i - V^T X_j\|^2 \omega\prime_{i,j}$ and minimize $\sum_{i,j} \|V^T X_i - V^T X_j\|^2 \omega_{i,j}$, where XX^T is utilized to preserve the maximal data variance. The diagonal regularization in the denominator improves the stability of the solution, without impacting the local intra-class neighborhood preserving ability. RLDE is suitable for the small-sample-size HSI classification problem. The item $V^T X(D - W)X^T V$ is used to maintain the intra-class relationships. The item XX^T is used to keep the maximal data variance.

The optimal RLDE projection is obtained by finding the eigenvectors corresponding to nonzero small eigenvalues of the following generalized Eigen-decomposition problem:

$$\left(\alpha X(D' - W')X^T + (1-\alpha)XX^T\right)V = \lambda\left(\alpha\left(X(D - W)X^T\right) + (1-\alpha)\left(diag\left(X(D - W)X^T\right)\right)\right)V. \tag{12}$$

2.4. Cooperative Training Strategy Combining Local Features

In [47], the optimal classifier combination selected by the diversity measures was multinomial logistic regression (MLR), *k*-nearest neighbor (KNN), and extreme learning machine (ELM). In this study, the correlation coefficient, disagreement metric, and double-fault measure were implemented to select the optimal classifier combination. It was found that the combination of MLR, KNN, and random forest (RF) achieved the best performance. Hence, the base classifiers were selected as MLR, KNN, and RF in this research. The procedure of the proposed method can be summarized as follows.

(1) A mean filtering process is employed to reduce the noise in the HSI.

(2) The local feature information of training samples L_i is extracted by the RLDE method, and is labeled L'_i.

(3) The classifier h_i is trained with L'_i, to obtain the predicted classification result S_i.

(4) For the classifier h_i, another two classifiers are selected which agree on the labeling of these samples to build the candidate set U'_i.

(5) The active learning method is used to select the most useful and informative samples L'_i from the candidate sets $L_i = L_i \cup L'_i$ and $U_i = U_i \cup U'_i$.

(6) The process is terminated if the stopping condition is met; otherwise, go to Step (2).

The final classification result is obtained by the majority voting method.

Pseudo-code Describing the RLDE Tri-Training Algorithm

```
Algorithm: RLDE tri-training
Input: $L$: Original labeled sample set
    $U$: Unlabeled sample set
        BT: Breaking ties algorithm
        MV: Majority voting algorithm
Process:
        $L$←SMF($L$); $U$←SMF($U$)
        $L_1$←$L$; $L_2$←$L$; $L_3$←$L$
        Repeat until none of $h_i$(i∈{1,2,3}) changes
            $L_1'$←RLDE($L_1$); $L_2'$←RLDE($L_2$); $L_3'$←RLDE($L_3$)
            $h_1$ ←MLR($L_1'$); $h_2$ ←KNN($L_2'$); $h_1$ ←RF($L_3'$)
            $S_1$ ←$h_1(U_1)$; $S_2$ ←$h_2(U_2)$; $S_3$ ←$h_3(U_3)$
            For i ∈ {1,2,3} do
              $S'_i$←$S_j \cap S_k$(i ≠ j ≠ k)
               $L''_i$←BT($S'_i$)
              $L_i \leftarrow L_i \cup L''_i$; $U_i \leftarrow \left(U_i - L''_i\right)$
            End of for
        End of repeat
OUTPUT: $S$ ←MV($S_1 + S_2 + S_3$)
```

3. Experimental Results and Analysis

In the spatial mean filtering (SMF) operation, the parameters for the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) dataset were set as follows: the degree of filtering $\gamma_0 = 0.9$ and the filtering window w = 9. The parameters for the Reflective Optics System Imaging Spectrometer (ROSIS) dataset were set as $\gamma_0 = 0.9$ and w = 7. These parameters can prevent over-filtering and increase the similarity and consistency of the neighboring pixels. In the feature extraction, the parameter in RLDE was selected as α = 0.5 for the AVIRIS dataset and 0.7 for the ROSIS dataset. We selected L = 5, 10, and 15 samples per class as the initial labeled training sets. We set k = 3 for KNN, and the parameter settings of MLR and RF were set as the default values. The number of most useful and informative samples in each iteration was set as 100. All the experiments were carried out 10 times, and the average results are reported. The initial training samples also have an impact on the accuracy (see Section 4). The experiments were therefore performed with the optimal feature number for each dataset.

3.1. Data Used in the Experiments

In the experiments, two real HSIs were used to evaluate the proposed approach. The HSI used in the first experiment was collected by the AVIRIS sensor over the Indian Pines test site in Northwestern Indiana in 1992. This dataset has a spatial size of 145 × 145 pixels and is made up of 224 spectral bands in the wavelength range of 0.4–2.5 um at 10 nm intervals, with a spatial resolution of 20 m. In total, 202 bands were used in the experiment after the noisy and water absorption bands were removed. For illustrative purposes, the image scene in pseudocolor is shown in Figure 1a. The ground-truth map available for the scene with 16 mutually exclusive ground-truth classes is shown in Figure 1b.

The HSI used in the second experiment was collected by the ROSIS sensor over the urban area of the University of Pavia, Italy. This dataset has a spatial size of 610 × 340 pixels and is made up of 115 spectral bands in the wavelength range of 0.43–0.68 um, with a spatial resolution of 1.3 m. In total, 103 bands were used in the experiment after the noisy and water absorption bands were removed. For illustrative purposes, the image scene in pseudocolor is shown in Figure 2a. The ground-truth map available for the scene with nine mutually exclusive ground-truth classes is shown in Figure 2b.

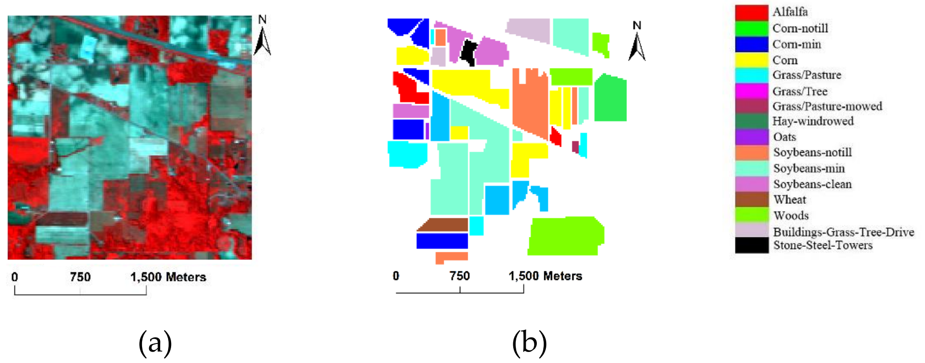


Figure 1. (**a**) Pseudocolor composite of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Indian Pines dataset. (**b**) The test area with 16 mutually exclusive ground-truth classes.

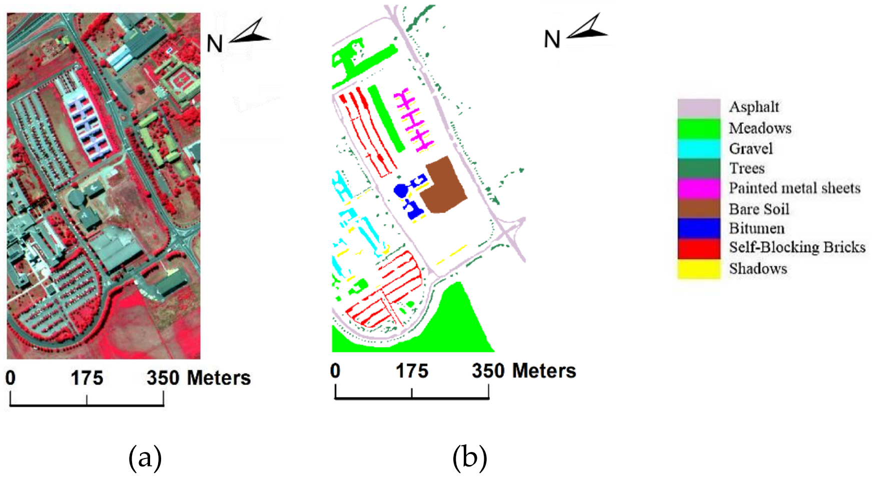


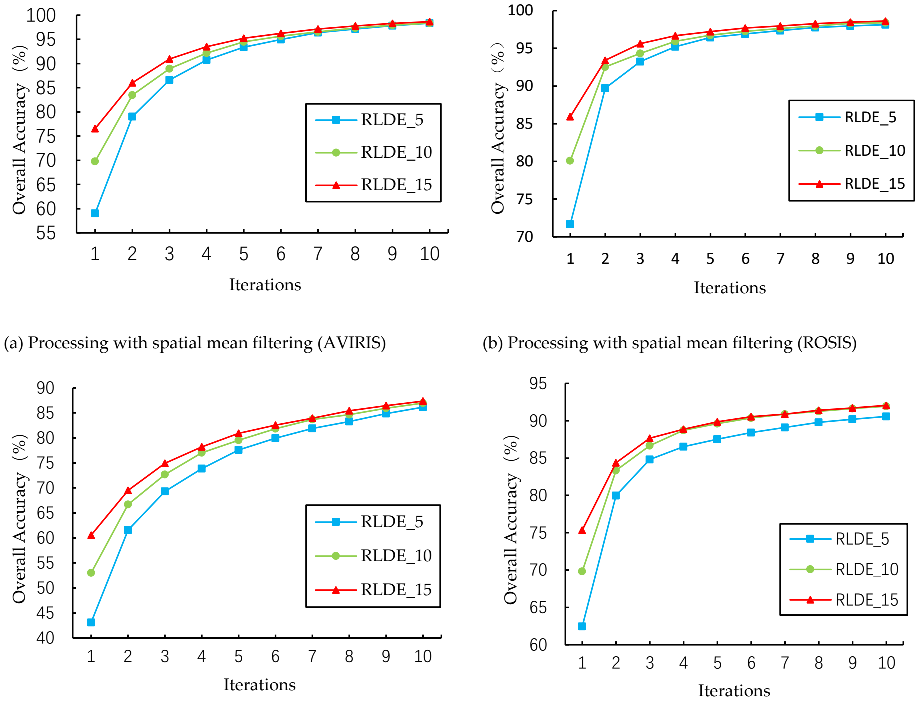
Figure 2. (**a**) Pseudocolor composite of the Reflective Optics System Imaging Spectrometer (ROSIS) Pavia scene. (**b**) The test area with nine mutually exclusive ground-truth classes.

3.2. The Effect of the Spatial Mean Filtering

Table 1 and Figure 3 show the classification results of the tri-training algorithm based on the RLDE method, using spatial mean filtering (SMF) and non-spatial mean filtering (non-SMF). As the unlabeled samples are continuously added, the classification accuracy increases. However, when the iterations reach seven, the classification accuracy starts to level off. In the AVIRIS experiment, with 5, 10, and 15 initial training samples per class, the overall accuracy (OA) of SMF increases by 12.19%, 11.39%, and 11.3% compared with non-SMF. In the ROSIS experiment, the OA of SMF increases by 7.56%, 6.45%, and 6.57% compared with non-SMF. Therefore, we used SMF to process the datasets in the subsequent experiments.

Table 1. Results of cooperative training classification based on the regularized local discriminant embedding (RLDE) local feature extraction method (%).

			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
AVIRIS	Non- SMF	5	43.11	61.59	69.31	73.88	77.58	79.93	81.91	83.29	84.86	86.15
		10	53.01	66.71	72.70	77.04	79.56	81.86	83.69	84.64	85.95	86.96
		15	60.57	69.52	74.92	78.21	80.91	82.56	83.94	85.44	86.45	87.35
	SMF	5	59.01	79.01	86.60	90.75	93.36	94.98	96.37	97.13	97.83	98.34
		10	69.77	83.51	88.93	92.14	94.48	95.67	96.55	97.35	97.92	98.35
		15	76.54	86.00	90.96	93.47	95.23	96.21	97.14	97.79	98.30	98.65
ROSIS	Non- SMF	5	62.45	79.98	84.83	86.53	87.51	88.43	89.10	89.78	90.19	90.58
		10	69.83	83.35	86.68	88.72	89.61	90.36	90.87	91.27	91.63	91.94
		15	75.36	84.35	87.65	88.88	89.86	90.54	90.88	91.38	91.70	92.05
	SMF	5	71.70	89.71	93.24	95.21	96.43	96.92	97.36	97.75	97.96	98.14
		10	80.11	92.52	94.33	95.91	96.73	97.27	97.63	97.96	98.29	98.39
		15	85.94	93.41	95.63	96.69	97.23	97.68	97.97	98.26	98.49	98.62



(a) Processing with spatial mean filtering (AVIRIS)

(b) Processing with spatial mean filtering (ROSIS)

(c) Processing without spatial mean filtering (AVIRIS)

(d) Processing without spatial mean filtering (ROSIS)

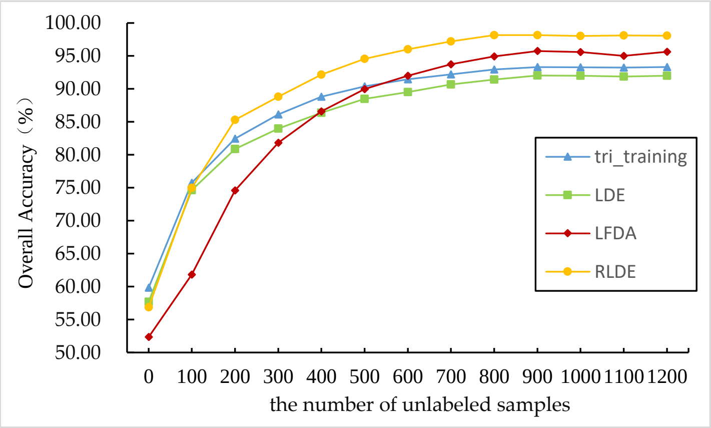
Figure 3. The results of cooperative training classification based on RLDE local feature extraction.

3.3. Comparison between the Different Feature Extraction Methods: AVIRIS Data

Figure 4 shows the classification results of the tri-training algorithm alone and the classification results of the tri-training algorithm based on the RLDE, LDE, and LFDA methods with the AVIRIS data. Specifically, the tri-training algorithm based on the LFDA method was proposed by Zhang and Jia in 2011 [48]. From Table 2 and Figures 4 and 5, we can see that the classification accuracy is not significantly related to the number of initial samples when the number of unlabeled samples reaches 900 or more. For example, the classification accuracy using the LDE feature extraction method is 92.03%, 93.09%, and 94.01% when the number of initial samples is 5, 10, and 15, respectively. This indicates that the proposed algorithm is both reliable and robust. The proposed tri-training classification algorithm based on RLDE feature extraction performs the best among all the methods with different initial training samples. The OA is improved by 4.85%, 6.13%, and 2.42% compared with tri-training alone, LDE, and LFDA when the initial samples are 5. The OA is 4.84 %, 5.75%, and 2.78% higher than that of tri-training alone, LDE, and LFDA when the initial samples are 10. When the initial samples are 15, the classification accuracy is 4.53 %, 4.97%, and 2.48% higher than that of tri-training alone, LDE, and LFDA. Meanwhile, the classification accuracy based on the RLDE feature extraction method reaches 98.98%, which indicates that the proposed tri-training classification algorithm is superior to the other methods.

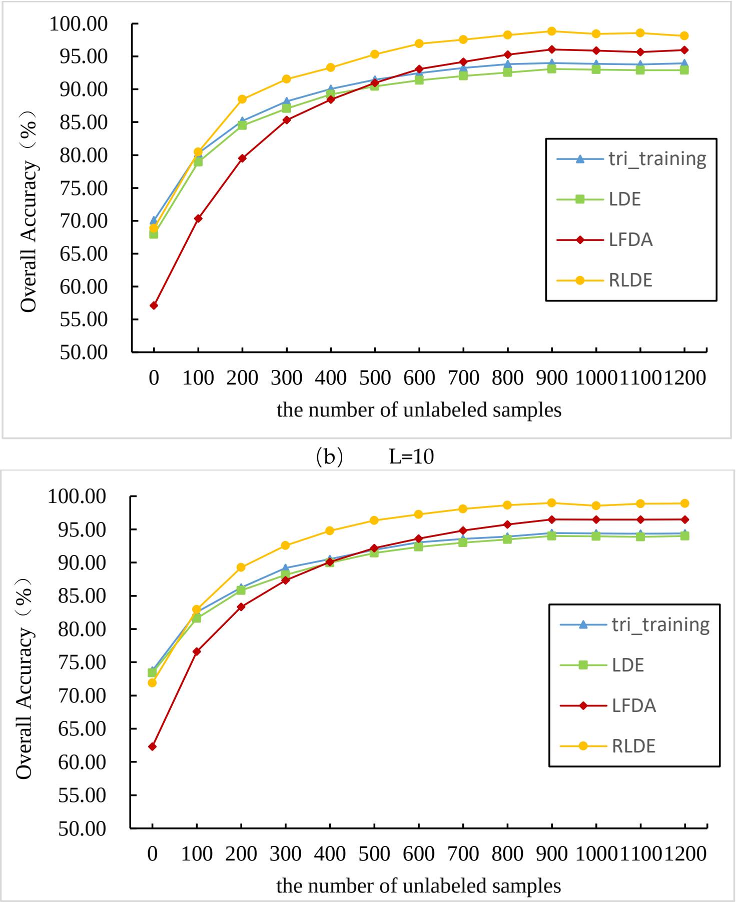
Table 2. Tri-training classification results based on the different feature extraction methods (%).

			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L = 5	Tri-training	OA	59.83	75.76	82.46	86.13	88.80	90.38	91.44	92.21	92.93	**93.31**
		Kappa	55.41	72.38	79.96	84.15	87.21	89.01	90.23	91.11	91.93	92.36
	LDE	OA	57.69	74.65	80.88	83.99	86.36	88.47	89.51	90.67	91.42	**92.03**
		Kappa	56.65	71.93	78.50	81.93	84.57	86.97	88.12	89.44	90.30	90.99
	LFDA	OA	52.34	61.83	74.61	81.84	86.56	89.95	92.01	93.74	94.92	**95.74**
		Kappa	61.09	70.80	79.20	84.25	88.23	90.97	92.56	94.08	95.02	95.67
	RLDE	OA	56.86	74.96	85.29	88.82	92.14	94.56	95.99	97.19	98.16	**98.16**
		Kappa	52.78	71.87	83.37	87.34	91.07	93.81	95.44	96.80	97.90	97.90
L = 10	Tri-training	OA	70.07	80.29	85.21	88.17	90.07	91.47	92.49	93.25	93.81	**94.00**
		Kappa	66.56	77.51	83.10	86.51	88.68	90.27	91.43	92.31	92.94	93.16
	LDE	OA	67.93	78.95	84.48	87.06	89.22	90.46	91.37	92.04	92.54	**93.09**
		Kappa	67.32	77.15	82.80	85.48	87.89	89.28	90.30	91.01	91.58	92.18
	LFDA	OA	57.09	70.36	79.51	85.31	88.42	91.00	93.07	94.16	95.28	**96.06**
		Kappa	69.07	75.21	82.21	87.06	89.50	91.70	93.53	94.32	95.25	96.00
	RLDE	OA	68.85	80.45	88.48	91.53	93.32	95.32	96.96	97.54	98.26	**98.84**
		Kappa	65.59	78.11	86.95	90.41	92.42	94.67	96.53	97.20	98.02	98.68
L = 15	Tri-training	OA	73.75	82.56	86.25	89.17	90.55	91.93	93.04	93.57	93.92	**94.45**
		Kappa	70.60	80.12	84.31	87.64	89.22	90.80	92.07	92.67	93.07	93.68
	LDE	OA	73.43	81.61	85.83	88.15	89.97	91.43	92.36	93.03	93.48	**94.01**
		Kappa	72.68	79.82	84.21	86.70	88.66	90.29	91.32	92.07	92.59	93.21
	LFDA	OA	62.32	76.62	83.36	87.31	90.11	92.17	93.62	94.85	95.74	**96.50**
		Kappa	68.91	80.68	85.36	88.31	90.62	92.39	93.62	94.80	95.60	96.36
	RLDE	OA	71.89	82.96	89.29	92.57	94.77	96.34	97.28	98.08	98.63	**98.98**
		Kappa	68.92	80.82	87.88	91.57	94.05	95.83	96.90	97.82	98.44	98.84



(a) L=5

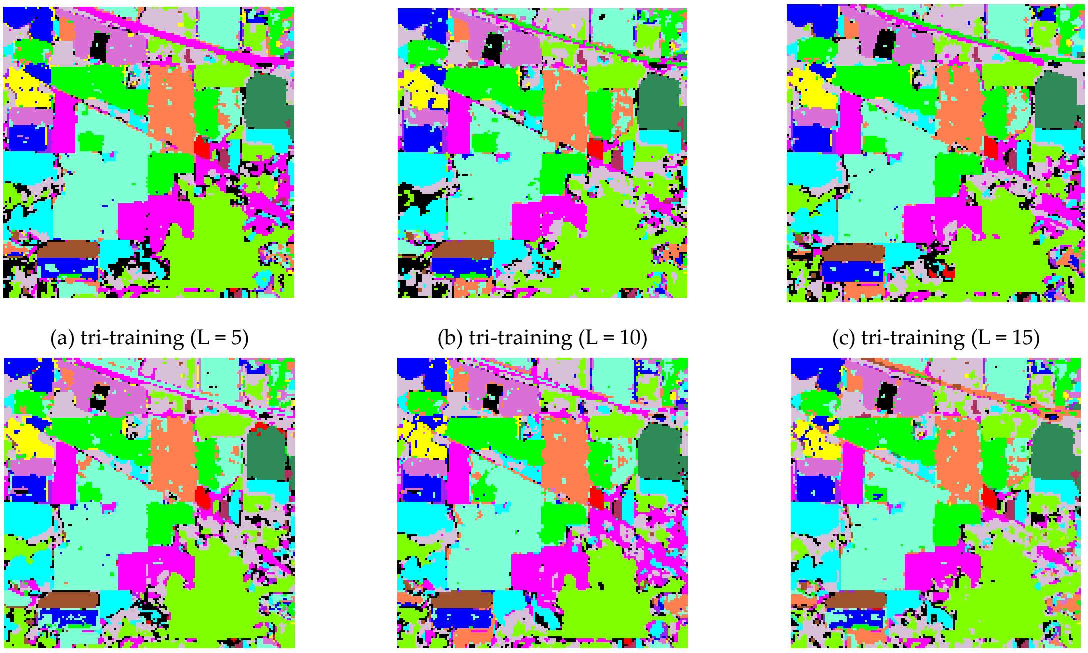
Figure 4. *Cont.*



(b) L=10

(c) L=15

Figure 4. AVIRIS data classification accuracy, as obtained by the different feature extraction methods under different initial training samples.



(a) tri-training (L = 5)

(b) tri-training (L = 10)

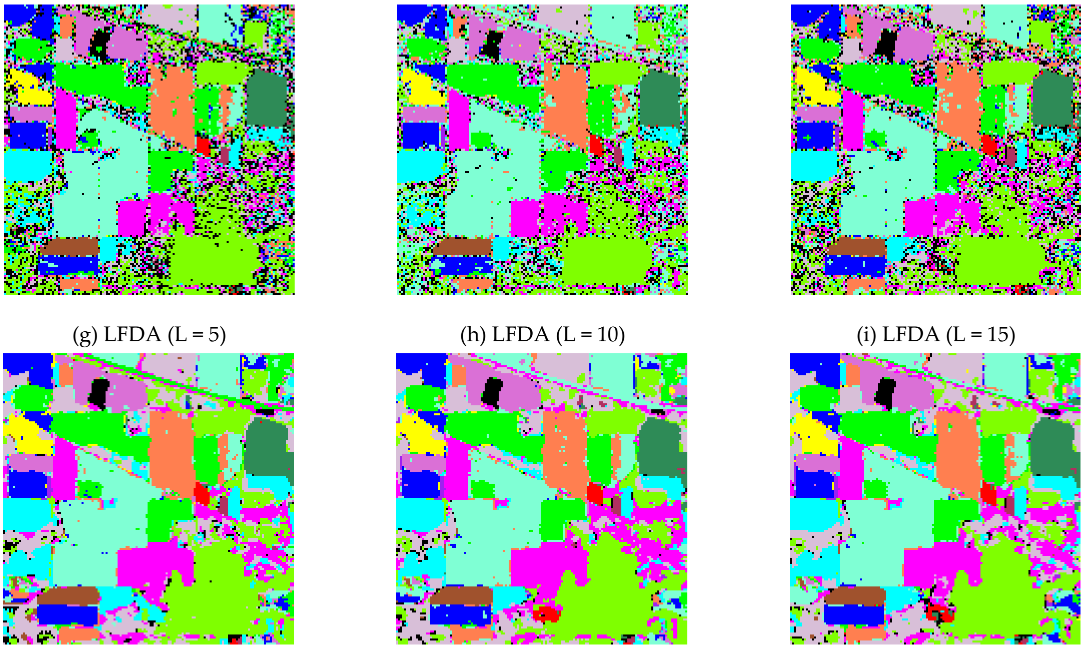
(c) tri-training (L = 15)

(d) LDE (L = 5)

(e) LDE (L = 10)

(f) LDE (L = 15)

Figure 5. *Cont*.



(g) LFDA (L = 5) (h) LFDA (L = 10) (i) LFDA (L = 15)

(j) RLDE (L = 5) (k) RLDE (L = 10) (l) RLDE (L = 15)

Figure 5. Co-training classification results based on the different feature extraction methods.

3.4. Comparison between the Different Feature Extraction Methods: ROSIS Data

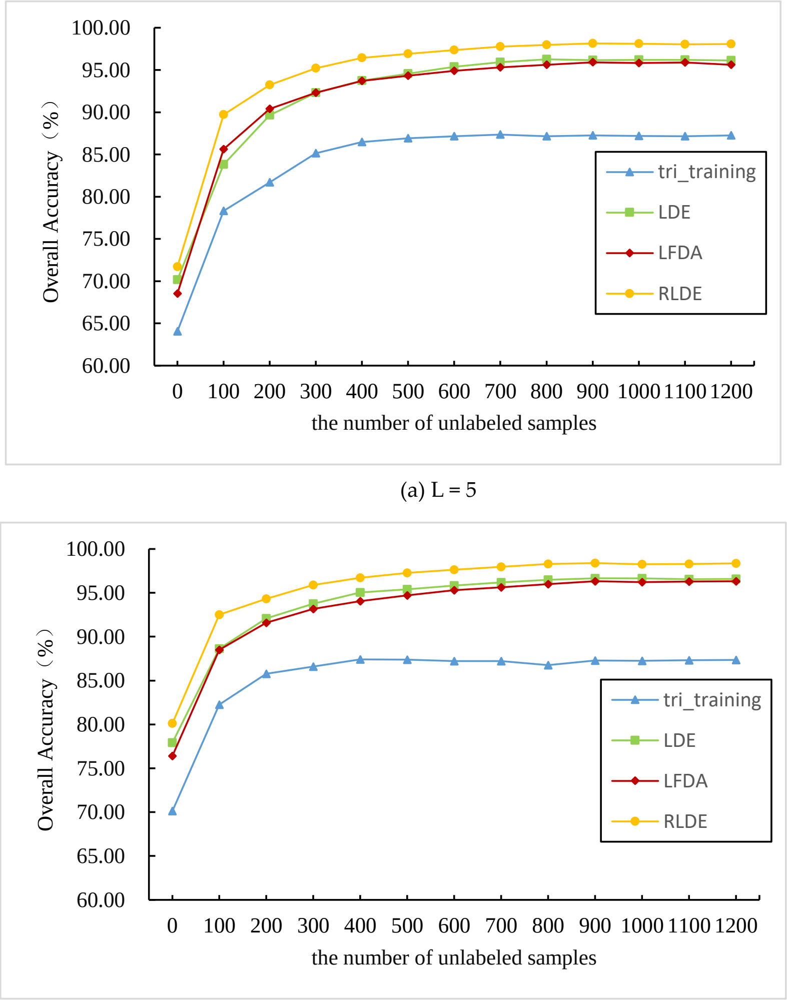
Figure 6 shows the classification results of the tri-training algorithm alone and the classification results of the tri-training algorithm based on the RLDE, LDE, and LFDA methods with the ROSIS data. From Table 3 and Figures 6 and 7, we can see that, as the unlabeled samples are continuously added, the classification accuracy increases. However, when the unlabeled samples reach 700, the OA becomes stable. The classification accuracy is not significantly related to the number of initial samples when the number of unlabeled samples reaches 900 or more. For example, the classification accuracy using the LDE feature extraction method is 96.16%, 96.66%, and 96.66% when the number of initial samples is 5, 10, and 15, respectively. This indicates that the proposed algorithm is both reliable and robust. The proposed tri-training classification algorithm based on RLDE feature extraction performs the best among all the methods under the different initial training samples. The OA is improved by 10.79%, 1.73%, and 2.06% compared with tri-training alone, LDE, and LFDA when the initial samples are five. The OA is 10.97%, 1.73%, and 2.06% higher than that of tri-training alone, LDE, and LFDA when the initial samples are 10. When the initial samples are 15, the OA is 11.36%, 1.96%, and 2.08% higher than that of tri-training alone, LDE, and LFDA, respectively. Meanwhile, the classification accuracy based on the RLDE feature extraction method reaches 98.62%.

Table 3. Tri-training classification results based on the different feature extraction methods (%).

			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L = 5	tri-training	OA	64.05	78.30	81.71	85.13	86.47	86.91	87.16	**87.35**	87.14	87.26
		Kappa	55.62	71.67	76.05	80.25	82.07	82.75	83.09	83.35	83.11	83.26
	LDE	OA	70.15	83.80	89.63	92.29	93.72	94.56	95.37	95.92	**96.26**	96.16
		Kappa	62.78	78.42	86.14	89.67	91.61	92.75	93.82	94.56	95.02	94.89
	LFDA	OA	68.54	85.61	90.40	92.30	93.70	94.33	94.88	95.31	95.60	**95.90**
		Kappa	65.16	81.62	87.14	89.53	91.40	92.22	92.99	93.57	93.97	94.36
	RLDE	OA	71.70	89.71	93.24	95.21	96.43	96.92	97.36	97.75	97.96	**98.14**
		Kappa	67.16	87.22	91.24	93.68	95.22	95.84	96.41	96.93	97.21	97.44

Table 3. *Cont.*

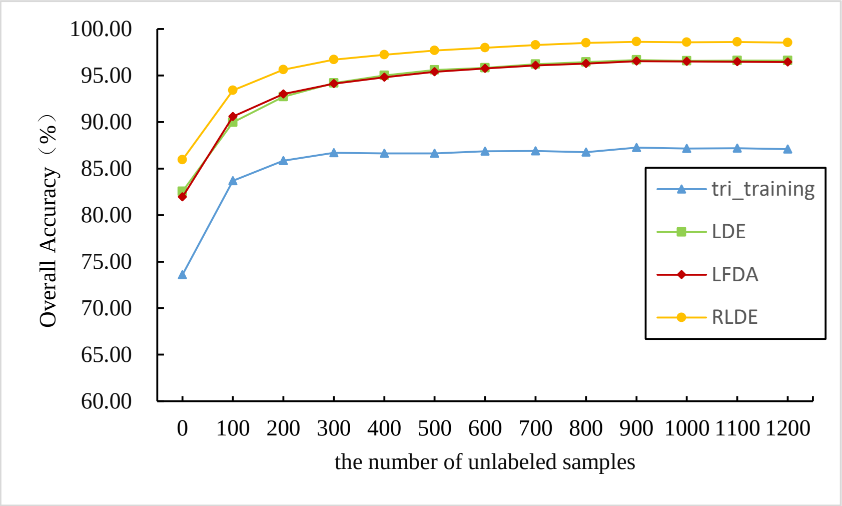
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
L = 10	tri-training	OA	70.12	82.27	85.78	86.59	**87.42**	87.39	87.23	87.22	86.75	87.30
		Kappa	63.03	76.53	80.98	82.21	83.34	83.39	83.25	83.24	82.70	83.37
	LDE	OA	77.92	88.64	92.10	93.76	95.03	95.40	95.84	96.19	96.50	**96.66**
		Kappa	72.27	84.81	89.38	91.64	93.37	93.86	94.45	94.92	95.34	95.55
	LFDA	OA	76.41	88.52	91.59	93.18	94.07	94.73	95.32	95.65	95.98	**96.33**
		Kappa	73.85	86.09	89.41	91.24	92.24	93.02	93.76	94.17	94.57	95.04
	RLDE	OA	80.11	92.52	94.33	95.91	96.73	97.27	97.63	97.96	98.29	**98.39**
		Kappa	76.45	90.38	92.53	94.51	95.58	96.30	96.78	97.21	97.66	97.80
L = 15	tri-training	OA	73.58	83.70	85.85	86.70	86.64	86.62	86.84	86.89	86.75	**87.26**
		Kappa	66.94	78.41	81.24	82.46	82.44	82.48	82.81	82.88	82.74	83.37
	LDE	OA	82.54	89.98	92.71	94.20	95.02	95.58	95.81	96.21	96.45	**96.66**
		Kappa	77.72	86.66	90.24	92.24	93.35	94.10	94.41	94.95	95.28	95.55
	LFDA	OA	81.94	90.59	92.99	94.12	94.82	95.38	95.75	96.09	96.30	**96.54**
		Kappa	79.61	87.84	90.56	92.01	92.97	93.70	94.20	94.64	94.92	95.26
	RLDE	OA	85.94	93.41	95.63	96.69	97.23	97.68	97.97	98.26	98.49	**98.62**
		Kappa	83.61	91.45	94.21	95.56	96.25	96.85	97.22	97.62	97.94	98.10



(a) L = 5

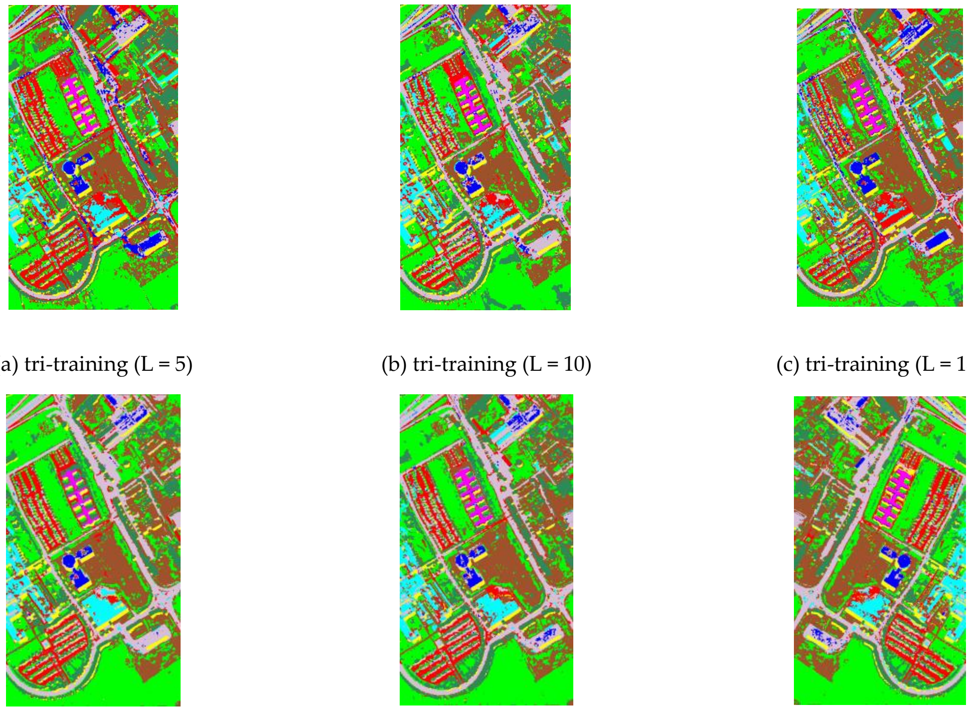
(b) L = 10

Figure 6. *Cont.*



(c) L = 15

Figure 6. AVIRIS data classification accuracy, as obtained by the different feature extraction methods under different initial training samples.



(a) tri-training (L = 5)

(b) tri-training (L = 10)

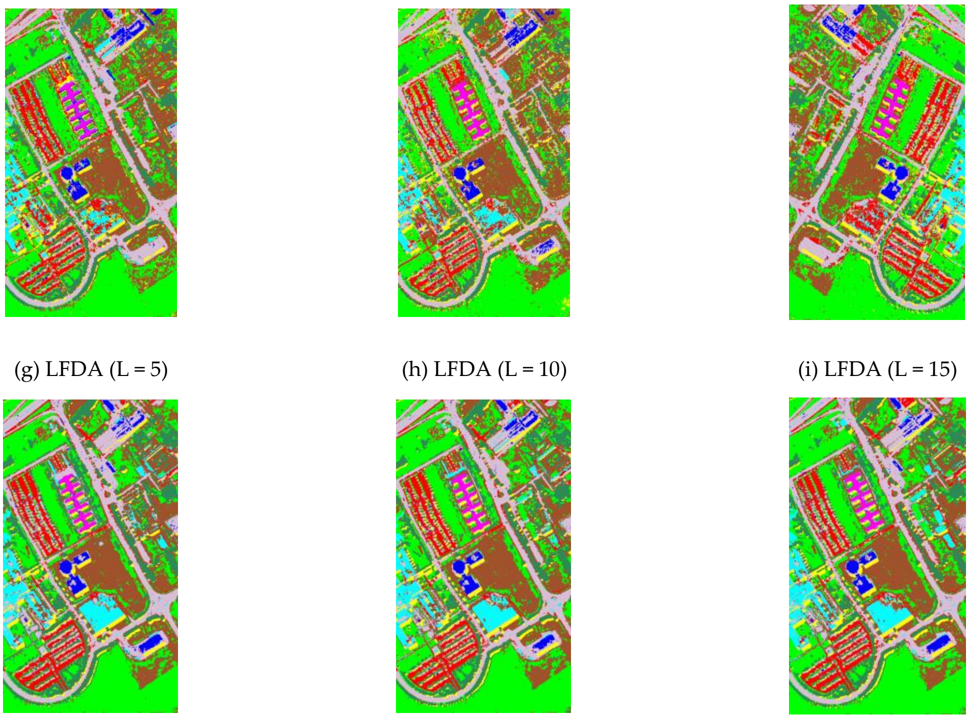
(c) tri-training (L = 15)

(d) LDE (L = 5)

(e) LDE (L = 10)

(f) LDE (L = 15)

Figure 7. *Cont*.



(g) LFDA (L = 5) (h) LFDA (L = 10) (i) LFDA (L = 15)

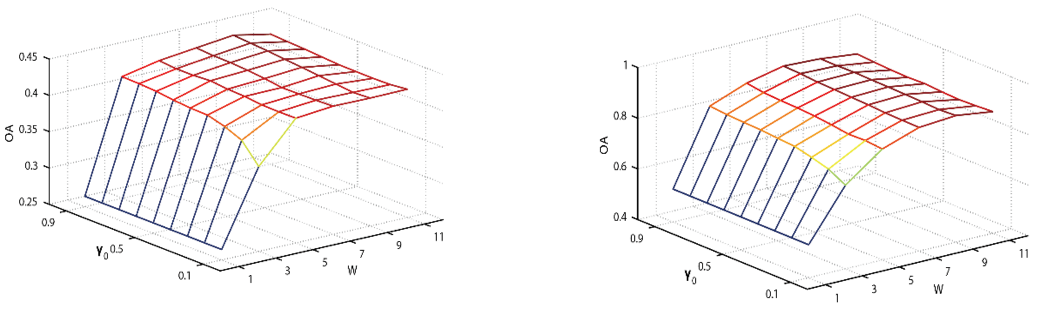
(j) RLDE (L = 5) (k) RLDE (L = 10) (l) RLDE (L = 15)

Figure 7. Co-training classification results based on the different feature extraction methods.

4. Discussion

In this section, the hyperparameters, w, γ_0, and α are experimentally analyzed. In the SMF, both w and γ_0 affect the final precision. Hence, parameter w was chosen from the range of {1, 3, 5, 7, 9, 11}, and parameter γ_0 was chosen from the range of {0.1, 0.2, 0.3, . . . , 0.9}. In this parameter analysis, α was always set to 0.1. In the RLDE feature extraction method, α is the essential parameter, and was chosen from the range of {0, 0.1, 0.2, . . . , 1}. Parameter w was set to 3, and γ_0 was set to 0.2. Fifteen samples in each class were selected as the training set, and no addition operation was conducted with the training samples.

Figure 8 shows the OA versus w and γ_0 for the AVIRIS and ROSIS datasets, where it is shown that γ_0 has less impact on the classification accuracy than w. The optimal value of w is 9 for the AVIRIS dataset and 7 for the ROSIS dataset. The classification accuracy tends to be stable with parameter w within a range from 5 to 9. Figure 9 shows the OA versus α for the AVIRIS and ROSIS datasets. The optimal value of parameter α is 0.5 for the AVIRIS dataset and 0.7 for the ROSIS dataset.



(a) OA versus w and γ_0 for the AVIRIS dataset (b) OA versus w and γ_0 for the ROSIS dataset

Figure 8. Overall accuracy (OA) versus w and γ_0 for the AVIRIS and ROSIS datasets.

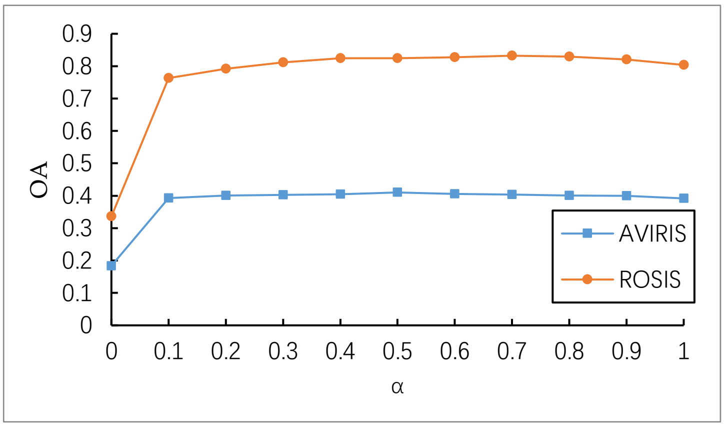


Figure 9. OA versus α for the AVIRIS and ROSIS datasets.

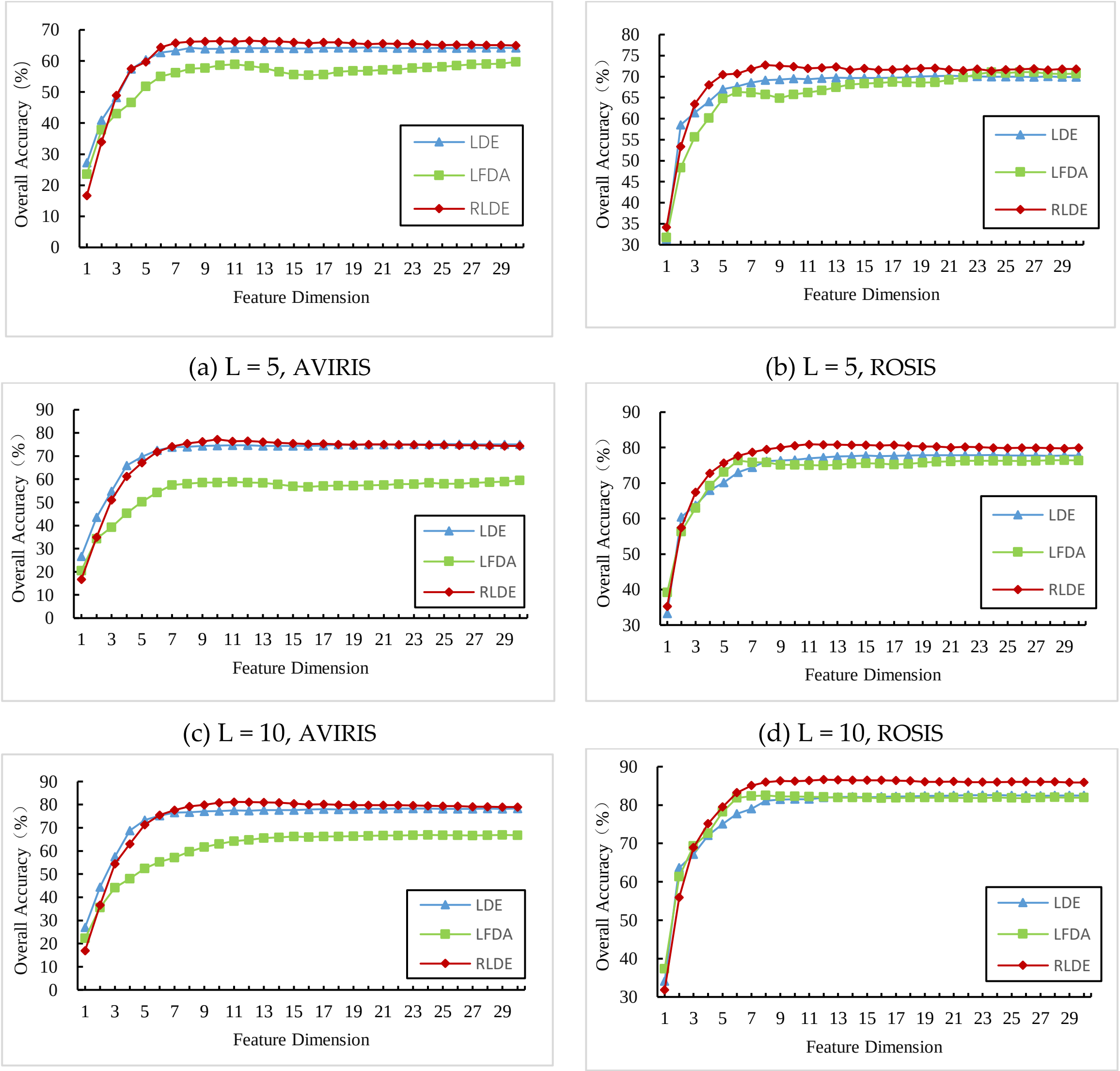
The initial training sample conditions has an impact on the accuracy. In this section, optimal feature selection is discussed. In this analysis, the range of the spectral information dimension was set from 1 to 30. With 5, 10, and 15 initial training samples per class, and different feature extraction methods, we selected the optimal feature information for all the dimensions, as shown in Table 4 and Figure 10.

For the AVIRIS data, when the number of initial training samples per class is 5, the maximum OA and the dimension of LDE are 64.35% and 20, respectively. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 12 and 30, respectively. When the number of initial training samples per class is 10, the maximum OA is obtained (75.16%) and the dimension of LDE is 26. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 10 and 30, respectively. When the number of initial training samples per class is 15, the maximum OA is obtained (78.35%) and the dimension of PCA is 30. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 10 and 24, respectively. Among the four different feature extraction methods, RLDE can obtain the highest classification accuracy and requires the smallest feature information dimension. With 5, 10, and 15 initial training samples per class, the feature information dimensions of all the methods were set as shown in Table 2 in the experiments.

Table 4. The optimal feature number and classification accuracy of the different feature extraction methods under different initial training sample conditions.

	Training Samples Feature Extraction Method	L = 5	L = 10	L = 15
AVIRIS	LDE	64.35%(20)	75.16%(26)	78.35%(30)
	LFDA	59.72%(30)	59.48%(30)	66.90%(24)
	RLDE	66.54%(12)	77.23%(10)	81.20%(11)
ROSIS	LDE	70.20%(21)	77.93%(24)	82.61%(24)
	RLDE	72.76%(8)	80.95%(11)	86.62%(12)
	LFDA	71.09%(24)	76.43%(28)	82.50%(8)

For the ROSIS data, when the number of initial training samples per class is 5, the maximum OA and the dimension of LDE are 70.20% and 21, respectively. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 8 and 24, respectively. When the number of initial training samples per class is 10, the maximum OA and the dimension of LDE are 77.93% and 24, respectively. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 11 and 38, respectively. When the number of initial training samples per class is 15, the maximum OA and the dimension of LDE are 82.61% and 24, respectively. RLDE and LFDA can obtain the maximum OA when the feature information dimension is 12 and 8, respectively. Among the four different feature extraction methods, RLDE can obtain the highest classification accuracy and requires the smallest feature information dimension. With 5, 10, and 15 initial training samples per class, the feature information dimensions of all the methods were set based on Table 1 in the experiments.



(a) L = 5, AVIRIS (b) L = 5, ROSIS

(c) L = 10, AVIRIS (d) L = 10, ROSIS

(e) L = 15, AVIRIS (f) L = 15, ROSIS

Figure 10. AVIRIS data and ROSIS data classification accuracy for different feature dimension, as obtained by the different feature extraction methods under different initial training samples.

Finally, we compared the proposed method with the other state-of-the-art deep learning methods of 1D-CNN, the CNN classifier proposed by Hu et al. [7], the five-layer CNN classifier proposed by Mei et al. [49], and the M3D-DCNN classifier proposed by He et al. [50]. All the methods, were compared under the same experimental settings (number of training samples, patch size, etc.) The OAs achieved by the different methods with the different HSI datasets are listed in Table 5. As can be seen, the proposed method shows a performance that is better than or comparable to the performance of the other four methods.

Table 5. The classification OA of the different deep learning methods with the different hyperspectral image (HSI) datasets.

Dataset	1D-CNN	Hu et al. [7]	Mei et al. [49]	M3D-DCNN [50]	Proposed Method
Indian Pines	82.39%	90.07%	95.70%	97.61%	**98.98%**
Pavia Univ.	93.29%	92.74%	98.00%	98.49%	**98.62%**

5. Conclusions

Hyperspectral sensors acquire hundreds of spectrally contiguous bands and provide abundant (but redundant) spectral information. In order to reduce the time consumption and improve the classification performance, it is necessary to extract the discriminant information before performing classification. In this paper, a novel semi-supervised tri-training algorithm for HSI classification has been proposed in conjunction with RLDE. The RLDE algorithm finds the optimal feature information, preserves the local information, and overcomes the singularity in the case of limited training samples. In the proposed algorithm, active learning is used to select the unlabeled samples, and ensemble learning is used to improve the classification result. In a comparison with other state-of-the-art deep learning methods, the proposed method achieved the highest classification accuracy with the least feature information.

Author Contributions: D.O. and K.T. conceived and designed the experiments; D.O. performed the experiments; D.O., K.T., J.Z., Y.C., and X.W. analyzed the data; D.O., K.T., Q.D., and J.Z. wrote the paper.

Funding: This research was supported in part by the National Natural Science Foundation of China (nos. 41871337, 41471356) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Acknowledgments: The authors would like to thank Paolo Gamba at Pavia University for providing the ROSIS dataset and David Landgrede at Purdue University for providing the AVIRIS dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ben-Dor, E.; Schläpfer, D.; Plaza, A.J.; Malthus, T. Hyperspectral remote sensing. In *Airborne Measurements for Environmental Research: Methods and Instruments*; Wiley-VCH Verlag & Co. KGaA: Weinheim, Germany, 2013; pp. 1249–1259.
2. Groves, P.; Tian, L.F.; Bajwa, S.G.; Bajcsy, P. Hyperspectral image data mining for band selection in agricultural applications. *Trans. ASAE* **2004**, *47*, 895–907.
3. Plaza, J.; Pérez, R.; Plaza, A.; Martínez, P.; Valencia, D. Mapping oil spills on sea water using spectral mixture analysis of hyperspectral image data. In *Chemical and Biological Standoff Detection III*; International Society for Optics and Photonics: Bellingham, WA, USA, 2005; Volume 5995, pp. 79–86.
4. Iranzad, A. *Hyperspectral Mineral Identification Using SVM and SOM*; Brock University: St. Catharines, ON, Canada, 2013.
5. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Conference on Computational Learning Theory, Madisson, WI, USA, 24–26 July 1998; pp. 92–100.
6. Peng, L.; Hui, Z.; Eom, K.B. Active deep learning for classification of hyperspectral images. *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.* **2017**, *10*, 712–724.
7. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep convolutional neural networks for hyperspectral image classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]
8. Song, J.; Zhang, H.; Li, X.; Gao, L.; Wang, M.; Hong, R. Self-supervised video hashing with hierarchical binary auto-encoder. *IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc.* **2018**, *27*, 3210. [CrossRef]
9. Li, Y.; Zhang, H.; Shen, Q. Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network. *Remote Sens.* **2017**, *9*, 67. [CrossRef]
10. Wang, X.; Gao, L.; Wang, P.; Sun, X.; Liu, X. Two-stream 3d convnet fusion for action recognition in videos with arbitrary size and length. *IEEE Trans. Multimed.* **2018**, *20*, 634–644. [CrossRef]
11. Wang, X.; Gao, L.; Song, J.; Shen, H. Beyond frame-level cnn: Saliency-aware 3d cnn with lstm for video action recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 510–514. [CrossRef]

12. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
13. Goldberg, A.B.; Zhu, X.; Singh, A.; Xu, Z.; Nowak, R. Multi-manifold semi-supervised learning. *Ynh Lr on Arfal Nllgn & Mahn Larnng* **2009**, *5*, 169–176.
14. Tan, K.; Zhou, S.; Du, Q. Semisupervised discriminant analysis for hyperspectral imagery with block-sparse graph. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1–5.
15. Tuia, D.; Ratle, F.; Pacifici, F.; Kanevski, M.F. Active learning methods for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2218–2232. [CrossRef]
16. Huang, R.; He, W. Using tri-training to exploit spectral and spatial information for hyperspectral data classification. In Proceedings of the International Conference on Computer Vision in Remote Sensing, Xiamen, China, 16–18 December 2012; pp. 30–33.
17. Zhou, Z.H.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1529–1541. [CrossRef]
18. Tan, K.; Li, E.; Du, Q.; Du, P. An efficient semi-supervised classification approach for hyperspectral imagery. *Isprs J. Photogramm. Remote Sens.* **2014**, *97*, 36–45. [CrossRef]
19. Nixon, M. *Feature Extraction & Image Processing for Computer Vision*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2008; pp. 595–599.
20. Rui, Y.; Huang, T.S.; Chang, S.F. Image retrieval: Current techniques, promising directions, and open issues. *J. Vis. Commun. Image Represent.* **1999**, *10*, 39–62. [CrossRef]
21. Hughes, G.F.; Hughes, G. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* **1968**, *14*, 55–63. [CrossRef]
22. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
23. Pevný, T.; Filler, T.; Bas, P. *Using High-Dimensional Image Models to Perform Highly Undetectable Steganography*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 161–177.
24. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the Twentieth International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; pp. 856–863.
25. Draper, B.A.; Baek, K.; Bartlett, M.S.; Beveridge, J.R. Recognizing faces with pca and ica. *Comput. Vis. Image Underst.* **2003**, *91*, 115–137. [CrossRef]
26. Liu, Z.P. *Linear Discriminant Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2006; pp. 2464–2485.
27. Kukharev, G.; Forczmaski, P.L. Face recognition by means of two-dimensional direct linear discriminant analysis. In Proceedings of the 8th International Conference on Pattern Recognition and Information Processing, Minsk, Belarus, 18–20 May 2005; Volume 280.
28. Li, H.; Jiang, T.; Zhang, K. Efficient and robust feature extraction by maximum margin criterion. *IEEE Trans. Neural Netw.* **2006**, *17*, 157–165. [CrossRef]
29. Bilgin, G.; Erturk, S.; Yildirim, T. Nonlinear dimension reduction methods and segmentation of hyperspectral images. In Proceedings of the IEEE Signal Processing, Communication and Applications Conference, Aydin, Turkey, 20–22 April 2008; pp. 1–4.
30. Camps-Valls, G.; Gomez-Chova, L.; Munoz-Mari, J.; Rojo-Alvarez, J.L. Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1822–1835. [CrossRef]
31. Song, J.; Gao, L.; Nie, F.; Shen, H.; Yan, Y.; Sebe, N. Optimized graph learning with partial tags and multiple features for image and video annotation. *IEEE Trans. Image Process.* **2016**, *25*, 4999–5011. [CrossRef]
32. Zhang, Z.; Wang, J.; Zha, H. Adaptive manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 253–265. [CrossRef]
33. Wu, W.; Massart, D.L.; Jong, S.D. The kernel pca algorithms for wide data. Part i: Theory and algorithms. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 165–172. [CrossRef]
34. Mika, S.; Rätsch, G.; Weston, J.; Schölkopf, B.; Müller, K.R. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop*; The Institute of Electrical and Electronics Engineers, Inc.: New York, NY, USA, 1999; pp. 41–48.
35. Song, J.; Yang, Y.; Li, X.; Huang, Z.; Yang, Y. Robust hashing with local models for approximate similarity search. *IEEE Trans. Cybern.* **2014**, *44*, 1225. [CrossRef]

36. Song, J.; Yang, Y.; Huang, Z.; Shen, H.T.; Luo, J. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Trans. Multimed.* **2013**, *15*, 1997–2008. [CrossRef]
37. Tenenbaum, J.B.; De, S.V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319. [CrossRef]
38. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323. [CrossRef]
39. He, X.; Cai, D.; Yan, S.; Zhang, H.J. Neighborhood preserving embedding. In Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005; pp. 1208–1213.
40. Chen, H.T.; Chang, H.W.; Liu, T.L. Local discriminant embedding and its variants. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 846–853.
41. Zhou, Y.; Peng, J.; Chen, C.L.P. Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 1082–1095. [CrossRef]
42. Liao, W.; Pizurica, A.; Philips, W.; Pi, Y. Feature extraction for hyperspectral images based on semi-supervised local discriminant analysis. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 401–404.
43. Sugiyama, M.; Idé, T.; Nakajima, S.; Sese, J. *Semi-Supervised Local Fisher Discriminant Analysis for Dimensionality Reduction*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 35–61.
44. Hua, G.; Brown, M.; Winder, S. Discriminant embedding for local image descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Rio De Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
45. Wan, M.; Yang, G.; Lai, Z.; Jin, Z. Feature extraction based on fuzzy local discriminant embedding with applications to face recognition. *IET Comput. Vis.* **2011**, *5*, 301–308. [CrossRef]
46. Pang, Y.; Yu, N. Regularized local discrimimant embedding. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 14–19 May 2006; p. III.
47. Tan, K.; Zhu, J.; Du, Q.; Wu, L.; Du, P. A novel tri-training technique for semi-supervised classification of hyperspectral images based on diversity measurement. *Remote Sens.* **2016**, *8*, 749. [CrossRef]
48. Zhang, G.; Jia, X. Feature selection using kernel based local fisher discriminant analysis for hyperspectral image classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Vancouver, BC, Canada, 24–29 July 2011; pp. 1728–1731.
49. Mei, S.; Ji, J.; Bi, Q.; Hou, J.; Qian, D.; Wei, L. Integrating spectral and spatial information into deep convolutional neural networks for hyperspectral classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Beijing, China, 10–15 July 2016.
50. He, M.; Bo, L.; Chen, H. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017.