remote sensing

MDPI

Article

Supervised and Semi-Supervised Multi-View Canonical Correlation Analysis Ensemble for Heterogeneous Domain Adaptation in Remote Sensing Image Classification

Alim Samat [1,2], Claudio Persello [3], Paolo Gamba [4], Sicong Liu [5], Jilili Abuduwaili [1,2,*] and Erzhu Li [6]

[1] State Key Laboratory of Desert and Oasis Ecology, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi 830011, China; alim.smt@gmail.com

[2] CAS Research Center for Ecology and Environment of Central Asia, Chinese Academy of Sciences, Urumqi 830011, China

[3] Department of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7500 AE Enschede, The Netherlands; c.persello@utwente.nl

[4] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, 27100 Pavia, Italy; paolo.gamba@unipv.it

[5] College of Surveying and Geoinformatics, Tongji University, Shanghai 200092, China; sicongliu.rs@gmail.com

[6] Department of Geographical Information Science, Nanjing University, Nanjing 210000, China; lierzhu2008@126.com

* Correspondence: jilil@ms.xjb.ac.cn

Academic Editors: Qi Wang, Nicolas H. Younan, Carlos López-Martínez, Xiaofeng Li and Prasad S. Thenkabail
Received: 7 January 2017; Accepted: 29 March 2017; Published: 1 April 2017

Abstract: In this paper, we present the supervised multi-view canonical correlation analysis ensemble (SMVCCAE) and its semi-supervised version (SSMVCCAE), which are novel techniques designed to address heterogeneous domain adaptation problems, i.e., situations in which the data to be processed and recognized are collected from different heterogeneous domains. Specifically, the multi-view canonical correlation analysis scheme is utilized to extract multiple correlation subspaces that are useful for joint representations for data association across domains. This scheme makes homogeneous domain adaption algorithms suitable for heterogeneous domain adaptation problems. Additionally, inspired by fusion methods such as Ensemble Learning (EL), this work proposes a weighted voting scheme based on canonical correlation coefficients to combine classification results in multiple correlation subspaces. Finally, the semi-supervised MVCCAE extends the original procedure by incorporating multiple speed-up spectral regression kernel discriminant analysis (SRKDA). To validate the performances of the proposed supervised procedure, a single-view canonical analysis (SVCCA) with the same base classifier (Random Forests) is used. Similarly, to evaluate the performance of the semi-supervised approach, a comparison is made with other techniques such as Logistic label propagation (LLP) and the Laplacian support vector machine (LapSVM). All of the approaches are tested on two real hyperspectral images, which are considered the target domain, with a classifier trained from synthetic low-dimensional multispectral images, which are considered the original source domain. The experimental results confirm that multi-view canonical correlation can overcome the limitations of SVCCA. Both of the proposed procedures outperform the ones used in the comparison with respect to not only the classification accuracy but also the computational efficiency. Moreover, this research shows that canonical correlation weighted voting (CCWV) is a valid option with respect to other ensemble schemes and that because of their ability to balance diversity and accuracy, canonical views extracted using partially joint random view generation are more effective than those obtained by exploiting disjoint random view generation.

Keywords: heterogeneous domain adaptation; transfer learning; multi-view canonical correlation analysis ensemble; semi-supervised learning; canonical correlation weighted voting; ensemble learning; image classification

1. Introduction

Supervised learning algorithms predominate over all other land cover mapping/monitoring techniques that use remote sensing (RS) data. However, the performance of supervised learning algorithms varies as a function of labeled training data properties, such as the sample size and the statistically unbiased and discriminative capabilities of the features extracted from the data [1]. As monitoring requires multi-temporal images, radiometric differences, atmospheric and illumination conditions, seasonal variations, and variable acquisition geometries can affect supervised techniques, potentially causing a distribution shift in the training data [2,3]. Regardless of the cause, any distribution change or domain shift that occurs after learning a classifier can degrade performance.

In the pattern recognition (PR) and RS image classification communities, this challenge is commonly referred to as covariate shift [4] or sample selection bias [5]. Many solutions have been proposed to resolve this problem, including image-to-image normalization [6], absolute and relative image normalization [7,8], histogram matching [9], and a multivariate extension of the univariate matching [10]. Recently, domain adaptation (DA) techniques, which attempt to mitigate performance the degradation caused by a distribution shift, has attracted increasing attention and is widely considered to provide an efficient solution [11–16].

According to the technical literature in PR and machine learning (ML), DA is a special case of transductive transfer learning (TTL). Its goal is to learn a function that predicts the label of a novel test sample in the target domain [12,15]. Depending on the availability of the source and the target domain data, the DA problem can result into supervised domain adaptation (SDA), semi-supervised domain adaptation (SSDA), unsupervised domain adaptation (UDA), multisource domain adaptation (MSDA) and heterogeneous domain adaption (HDA) [14–19].

Moreover, according to the "knowledge" transferred across domains or tasks, classical approaches to DA can be grouped into parameter adapting, instance transferring, feature representation, and relational knowledge transfer techniques.

Parameter adapting approaches aim to transfer and adapt a classification model and/or its parameters to the target domain; the model and/or parameters are learned from the source domain (SD) [20]. The seminal work presented by Khosla et al. [5] and Woodcock et al. [7], which features parameter adjustment for a maximum-likelihood classifier in a multiple cascade classifier system by retraining, can be categorized into this group.

In instance transferring, the samples from the SD are reweighted [21] or resampled [22] for their use in the TD. In the RS community, active learning (AL) has also been applied to address DA problems. For example, AL for DA in the supervised classification RS images is proposed by Persello and Bruzzone [23] via iteratively labeling and adding to the training set the minimum number of the most informative samples from the target domain, while removing the source-domain samples that do not fit with the distributions of the classes in the TD.

For the third group, feature representation-based adaptation searches for a set of shared and invariant features using feature extraction (FE), feature selection (FS) or manifold alignment to reduce the marginal, conditional and joint distributions between the domains [16,24–26]. Matasci et al. [14] investigated the semi-supervised transfer component analysis (SSTCA) [27] for both hyperspectral and multispectral high resolution image classification, whereas Samat et al. [16] analyzed a geodesic Gaussian flow kernel based support vector machine (GFKSVM) in the context of hyperspectral image classification, which adopts several unsupervised linear and nonlinear subspace feature transfer techniques.

Finally, relational knowledge transfer techniques address the problem of how to leverage the knowledge acquired in SD to improve accuracy and learning speed in a related TD [28].

Among these four groups, it is easy to recognize the importance of RS image classification of adaptation strategies based on feature representation. However, most previous studies have assumed that data from different domains are represented by the same types of features with the same dimensions. Thus, these techniques cannot handle the problem of data from source and target domains represented by heterogeneous features with different dimensions [18,29]. One example of this scenario is land cover updating using current RS data; each time, there are different features with finer spatial resolution and more spectral bands (e.g., Landsat 8 OLI with nine spectral bands at 15–30 m spatial resolution, and Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) with 224 spectral bands at 20 m spatial resolution), when the training data are only available at coarser spatial and spectral resolutions (e.g., MSS with four spectral bands and 60 m spatial resolution).

One of the simplest feature-based DA approaches is the feature augmentation proposed in [17], whose extended versions, called heterogeneous feature augmentation (HFA) and semi-supervised HFA (SHFA), were recently proposed in [18]. Versions that consider the intermediate domains as being manifold-based were proposed in [30,31]. However, none of these approaches have been considered in RS image classification.

Finding a joint feature representation between the source and target domains requires FS [12,19] or FE [16] to select the most effective feature set. To accomplish this aim, canonical correlation analysis (CCA), which aims to maximize the correlation between two variable sets (in this case, the different domains) could be a very effective technique. Indeed, CCA and kernel CCA (KCCA) have already been applied with promising results in object recognition and text categorization [29], action recognition and image-to-text classification [32]. However, existing joint optimization frameworks such as [32] are limited to scenarios in which the labeled data from both domains are available. This is not the case in many practical situations. To solve this problem, CTSVM was proposed in [29], incorporating the DA ability into the classifier design for a cross-domain recognition scenario of labeled data that is available only in the SD. However, the CTSVM might fail to balance the possible mismatches between the heterogeneous domains.

One solution might be to multi-view learning (MVL), a procedure that implies the splitting of high-dimensional data into multiple "views" [33,34]. If multiple views are available, then multiple classification results must be reconciled, and this step is efficiently performed using Ensemble Learning (EL) [35,36]. Accordingly, this work introduces an EL technique based on supervised multi-view CCA, which is called supervised multi-view canonical correlation analysis ensemble (SMVCCAE), and we prove its effectiveness for DA (and specifically heterogeneous DA) problems.

Additionally, in real applications, it is typical to experience situations in which there are very limited or even no labeled samples available. In this case, a semi-supervised learning (SSL) technique (e.g., [37]), which uses of unlabeled data to improve performance using a small amount of labeled data from the same domain, might be an appropriate solution. As a matter of fact, many SSDAs have been proposed. However, most existing studies, such as asymmetric kernel transforms (AKT) [38], domain-dependent regularization (DDR) [32], TCA, SSTCA [14,27], and co-regularization based SSDA [39], were designed for homogeneous DA. Very recently, Li et al. [18] proposed a semi-supervised heterogeneous DA by convex optimization of standard multiple kernel learning (MKL) with augmented features. Unfortunately, this optimization is quite challenging in real-world applications. This work instead proposes a semi-supervised version of the above-mentioned multi-view canonical correlation analysis ensemble (called SSMVCCAE), incorporating multiple speed-up spectral regression kernel discriminant analysis (SRKDA) [40] into the original supervised algorithm.

2. Related Work

2.1. Notation for HDA

According to the technical literature, feature-based approaches to HDA can be grouped into the following three clusters, depending on the features used to connect the target and the SD:

(1) If data from the source and target domains share the same features [41–43], then latent semantic analysis (LSA) [44], probabilistic latent semantic analysis (pLSA) [45], and risk minimization techniques [46] may be used.
(2) If additional features are needed, "feature augmentation" approaches have been proposed, including the method in [37], HFA and SHFA [18], manifold alignment [31], sampling geodesic flow (SGF) [47], and geodesic flow kernel (GFK) [16,30]. All of these approaches introduce a common subspace for the source and target data so that heterogeneous features from both domains.
(3) If features are adapted across domains through learning transformations, feature transformation-based approaches are considered. This group of approaches includes the HSMap [48], the sparse heterogeneous feature representation (SHFR) [49], and the correlation transfer SVM (CTSVM) [29]. The algorithms that we propose fit into this group.

Although all of the approaches reviewed above have achieved promising results, they also have some limitations of all the approaches reviewed above. For example, the co-occurrence features assumption used in [41–43] may not hold in applications such as object recognition, which uses only visual features [32]. For the feature augmentation based approaches discussed in [18,30,31], the domain-specific copy process always requires large storage space, and the kernel version requires even more space and computational complexity because of the parameter tuning. Finally, for the feature transformation based approaches proposed in [29,32,48], they do not optimize the objective function of a discriminative classifier directly, and the computational complexity is highly dependent on the total number of samples or features used for adaptation [12,19].

In this work, we assume that there is only one SD (S_D) and one TD (T_D). We also define $\mathbf{X}_S = \left[\mathrm{x}_1^S, ..., x_{n_S}^S\right]^\dagger \in \Re^{d_S \times n_S}$ and $\mathbf{X}_T = \left[\mathrm{x}_1^T, ..., x_{n_T}^T\right]^\dagger \in \Re^{d_T \times n_T}$ as the feature spaces in the two domains, with the corresponding marginal distributions $p(\mathbf{X}_S)$ and $p(\mathbf{X}_T)$ for S_D and T_D, respectively. The parameters d_S and d_T represent the size of $\mathrm{x}_i^S, i = 1, ..., n_S$ and $\mathrm{x}_j^T, j = 1, ..., n_T$, n_S and n_T are the sample sizes for $\mathbf{X}_S$ and $\mathbf{X}_T$, and we have $S_D = \{\mathbf{X}_S, P(\mathbf{X}_S)\}$, $T_D = \{\mathbf{X}_T, P(\mathbf{X}_T)\}$. The labeled training samples from the SD are denoted by $\left\{ \left(\mathrm{x}_j^S, y_j^S\right)\Big|_{j=1}^{n_S} \right\}, y_j^S \in \Omega = \{\varpi_l\}_{l=1}^c$, and they refer to c classes. Furthermore, let us consider as "task" Y the task to assign to each element of a set a label selected in a label space by means of a predictive function f, so that $v = \{y, f\}$.

In general, if the feature sets belong to different domains, then either $\mathbf{X}_S \neq \mathbf{X}_T$ or $p(\mathbf{X}_S) \neq p(\mathbf{X}_T)$, or both. Similarly, the condition $v_S \neq v_T$ implies that either $\mathrm{Y}_S \neq \mathrm{Y}_T$ ($\mathrm{Y}_S = [y_1^S, ..., y_{n_S}^S]$, $\mathrm{Y}_T = [y_1^T, ..., y_{n_T}^T]$) or $p(\mathrm{Y}_S|\mathbf{X}_S) \neq p(\mathrm{Y}_T|\mathbf{X}_T)$, or both. In this scenario, a "domain adaptation algorithm" is an algorithm that aims to improve the learning of the predictive function f_T in the TD T_D using the knowledge available in the SD S_D and in the learning task v_S, when either $S_D \neq T_D$ or $v_S \neq v_T$. Moreover, in heterogeneous problems, the additional condition $d_S \neq d_T$ holds.

2.2. Canonical Correlation Analysis

Let us now assume that $n_S = n_T$ for the feature sets (called "views" here) in the source and target domains. The CCA is the procedure for obtaining the transformation matrices $\boldsymbol{\omega}_S$ and $\boldsymbol{\omega}_T$ which maximize the correlation coefficient between the two sets [50]:

$$\max_{\boldsymbol{\omega}_S, \boldsymbol{\omega}_T} \rho = \frac{\boldsymbol{\omega}_S^\dagger \Sigma_{ST} \boldsymbol{\omega}_T}{\sqrt{\boldsymbol{\omega}_S^\dagger \Sigma_{SS} \boldsymbol{\omega}_S} \sqrt{\boldsymbol{\omega}_T^\dagger \Sigma_{TT} \boldsymbol{\omega}_T}} \tag{1}$$

where $\Sigma_{ST} = \mathbf{X}_S\mathbf{X}_T^{\dagger}$, $\Sigma_{SS} = \mathbf{X}_S\mathbf{X}_S^{\dagger}$, $\Sigma_{TT} = \mathbf{X}_T\mathbf{X}_T^{\dagger}$, $\rho \in [0,1]$, and "$\dagger$" means the matrix transpose. In practice, $\boldsymbol{\omega}_S$ can be obtained by a generalized eigenvalue decomposition problem:

$$\Sigma_{ST}(\Sigma_{TT})^{-1}\Sigma_{ST}^{\dagger}\boldsymbol{\omega}_S = \eta(\Sigma_{SS})\boldsymbol{\omega}_S \tag{2}$$

where η is a constraint factor. Once $\boldsymbol{\omega}_S$ is obtained, $\boldsymbol{\omega}_T$ can be obtained by ${\Sigma_{TT}}^{-1}\Sigma_{ST}\boldsymbol{\omega}_S/\eta$. By adding the regularization terms $\lambda_S\mathbf{I}$ and into Σ_{SS} and Σ_{TT} to avoid overfitting and singularity problems, Equation (2) becomes:

$$\Sigma_{ST}(\Sigma_{TT} + \lambda_T\mathbf{I})^{-1}\Sigma_{ST}^{\dagger}\boldsymbol{\omega}_S = \eta(\Sigma_{SS} + \lambda_S\mathbf{I})\boldsymbol{\omega}_S \tag{3}$$

As a result, the source and target view data can be transformed into correlation subspaces by:

$$\mathbf{X}_S^C = \mathbf{X}_S\cdot\boldsymbol{\omega}_S, \boldsymbol{\omega}_S \in \Re^{d_S\times d} \tag{4}$$

$$\mathbf{X}_T^C = \mathbf{X}_T\cdot\boldsymbol{\omega}_T, \boldsymbol{\omega}_T \in \Re^{d_T\times d} \tag{5}$$

Note that one can derive more than one pair of transformation matrices $\{\omega_i^S\}_{i=1}^d$ and $\{\omega_i^T\}_{i=1}^d$, where $d = \min\{d_S, d_T\}$ is the dimension of the resulting CCA subspace. Once the correlation subspaces $\mathbf{X}_S^C$ and $\mathbf{X}_T^C$ spanned by $\boldsymbol{\omega}_S$ and $\boldsymbol{\omega}_T$ are derived, test data in the target view can be directly labeled by any model M_S^C that is trained using the source features $\mathbf{X}_S^C$.

2.3. Fusion Methods

If multiple "views" are available, then for each view, a label can be associated with each pixel used, for instance, CCA. If multiple labels are present, then they must be fused to obtain a single value using a so-called decision-based fusion procedure. Decision-based fusion aims to provide the final classification label for a pixel by combining the labels obtained, in this case, by multiple view analysis. This usually is obtained using two classes of procedures: weighted voting methods and meta-learning methods [51].

For weighted voting, the labels are combined using the weights assigned to each result. Many variants have been proposed in past decades. For the sake of comparison and because we must consider these options to evaluate the performance of the canonical correlation weighted voting (CCWV) scheme proposed in this paper, here, we consider only the following state-of-the-art techniques:

- Accuracy weighted voting (AWV), in which the weight of each member is set proportionally to its accuracy performance on a validation set [51]:

$$w_i = \frac{a_i}{\sum_{j=1}^{T} a_j} \tag{6}$$

  where a_i is a performance evaluation of the i-th classifier on a validation set.
- Best–worst weighted voting (BWWV), in which the best and the worst classifiers are given a weight of 1 or 0, respectively [51], and for the ones the weights are compute according to:

$$\alpha_i = 1 - \frac{e_i - \min\limits_i(e_i)}{\max\limits_i(e_i) - \min\limits_i(e_i)} \tag{7}$$

  where e_i is the error of the i-th classifier on a validation set.
- Quadratic best–worst weighted voting (QBWWV), that computes the intermediate weights between 0 and 1 via squaring the above-mentioned BWWV:

$$\alpha_i = \left(\frac{\max\limits_i(e_i) - e_i}{\max\limits_i(e_i) - \min\limits_i(e_i)}\right)^2 \tag{8}$$

3. The (Semi) Supervised Canonical Correlation Analysis Ensemble

3.1. Supervised Procedure

The idea of this procedure is to adopt MVL to decompose the target domain data into multiple disjoint or partial joint feature subsets (views), where each view is assumed to bring complementary information [52]. Next, these multiple views are used for DA, providing multiple matches between the source and the target domains. Eventually, the labeling task in the SD is transferred into the target domain through CCA, and the results of this "multi-view" CCA are combined to achieve a more efficient heterogeneous DA.

Specifically, without loss of generality, let us assume a heterogeneous DA from a low-dimensional $\mathbf{X}_S$ to a high-dimensional $\mathbf{X}_T$, with $d_S < d_T$, which requires that $\mathbf{X}_T$ is decomposed into N views, i.e., $\mathbf{X}_T = \{\mathbf{X}_T^i\}_{i=1}^N, \mathbf{X}_T^i \in \Re^{d_i \times n_T}, d_T = \sum_{i=i}^N d_i$. In this case, the implementation of MVCCA corresponds to searching for the following:

$$\underset{(\boldsymbol{\omega}_S^i,\boldsymbol{\omega}_T^i),\ldots,(\boldsymbol{\omega}_S^N,\ldots,\boldsymbol{\omega}_T^N)}{\operatorname{argmax}} (\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_N) = \sum_{i=1}^{N} \frac{\left(\boldsymbol{\omega}_S^i\right)^{\dagger} \Sigma_{ST}^i \boldsymbol{\omega}_T^i}{\sqrt{\left(\boldsymbol{\omega}_S^i\right)^{\dagger} \Sigma_{SS}^i \boldsymbol{\omega}_S^i} \sqrt{\left(\boldsymbol{\omega}_T^i\right)^{\dagger} \Sigma_{TT}^i \boldsymbol{\omega}_T^i}} \tag{9}$$

where $\Sigma_{ST}^i = \mathbf{X}_S \left(\mathbf{X}_T^i\right)^{\dagger}, \Sigma_{SS}^i = \mathbf{X}_S \mathbf{X}_S^{\dagger}$ and $\Sigma_{TT}^i = \mathbf{X}_T^i \left(\mathbf{X}_T^i\right)^{\dagger}$. Generalizing the standard CCA, Equation (9) can be rewritten as:

$$\begin{gathered}\underset{(\boldsymbol{\omega}_S^i,\boldsymbol{\omega}_T^i),\ldots,(\boldsymbol{\omega}_S^N,\ldots,\boldsymbol{\omega}_T^N)}{\operatorname{argmax}} (\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_N) = \sum_{i=1}^{N} \left(\boldsymbol{\omega}_S^i\right)^{\dagger} \Sigma_{ST}^i \boldsymbol{\omega}_T^i \\ s.t. \left(\boldsymbol{\omega}_S^1\right)^{\dagger} \Sigma_{ST}^1 \boldsymbol{\omega}_T^i = 1, \ldots, \left(\boldsymbol{\omega}_S^N\right)^{\dagger} \Sigma_{ST}^N \boldsymbol{\omega}_T^N = 1\end{gathered} \tag{10}$$

As a result, by using the solutions $\omega_S^i\big|_{i=1}^N$ and $\omega_T^i\big|_{i=1}^N$, we will have multiple transformed correlation subspaces, each one considering the SD and one of the target "views":

$$\mathbf{X}_S^{Ci} = \mathbf{X}_S \cdot \boldsymbol{\omega}_S^i, \boldsymbol{\omega}_S^i \in \Re^{d_S \times \hat{d}_i} \tag{11}$$

$$\mathbf{X}_T^{Ci} = \mathbf{X}_T^i \cdot \boldsymbol{\omega}_T^i, \boldsymbol{\omega}_T^i \in \Re^{d_T \times \hat{d}_i} \tag{12}$$

For any new instance of the target domain, i.e., $\mathrm{x} = \{\mathrm{x}_i\}\big|_{i=1}^N, \mathrm{x}_i \in \mathbf{X}_T^{Ci}$, the decision function of this SMVCCAE, trained with labeled training samples $\left\{ \left(\mathrm{x}_j^{SC}, y_j^S\right)\Big|_{j=1}^{n_S} \right\}, \mathrm{x}_j^{SC} \in \mathbf{X}_S^{Ci}, i = \forall N$, can be implemented via majority voting (MV):

$$\begin{aligned} H(\mathrm{x}) &= sign\left(\sum_{i=1}^{N} h_i(\mathrm{x}_i)\right) \\ &= \begin{cases} \varpi_l, \text{ if} \sum\limits_{i=1}^{N} h_i^l(\mathrm{x}_i) \succ \frac{1}{2} \sum\limits_{k=1}^{c} \sum\limits_{i=1}^{N} h_i^k(\mathrm{x}_i) \\ reject, \quad \text{otherwise} \end{cases} \end{aligned} \tag{13}$$

However, to further optimize the ensemble results, one can also recall that the canonical correlations $\boldsymbol{\rho} = \left\{ \{\rho_1, \ldots, \rho_j\}\big|_{j=1}^{\hat{d}_1}, \ldots, \{\rho_1, \ldots, \rho_j\}\big|_{j=1}^{\hat{d}_N} \right\}$ obtained together with the transformation matrices $\boldsymbol{\omega}_S^i$ and $\boldsymbol{\omega}_T^i$ provide information about correlation between the SD and each target view. Since larger values of $\forall \{\rho_j\}\big|_{j=1}^{\hat{d}_i} \in \{\boldsymbol{\rho}_i\}\big|_{i=1}^N$ show a greater correlation, this can also be considered a hint to obtain a better domain transfer ability for the corresponding view. We expect that poor correlation values (i.e., low values of $\sum_{j=1}^{\hat{d}_i} \rho_j$) will result in poor domain transfer abilities. Therefore, $\sum_{j=1}^{\hat{d}_i} \rho_j$ may be used to quantitatively evaluate the domain transfer ability of the transformation matrices $\boldsymbol{\omega}_S^i$ and

$\boldsymbol{\omega}_T^i$. Accordingly, we propose to include the following canonical correlation coefficient in the voting strategy of Equation (13):

$$H(\mathrm{x}) = sign\left(\sum_{i=1}^{N}\sum\nolimits_{j=1}^{\hat{d}_i}\rho_j h_i(\mathrm{x}_i)\right) \tag{14}$$

The algorithmic steps of the new algorithm (called Supervised MVCCA Ensemble, or SMVCCAE for short) are summarized in Algorithm 1.

Algorithm 1. Algorithmic details of SMVCCAE.

```
Inputs: SD $\mathbf{X}_S = \left[\mathrm{x}_1^S, \ldots, x_{n_S}^S\right] \in \Re^{d_S \times n_S}$; TD $\mathbf{X}_T = \left[\mathrm{x}_1^T, \ldots, x_{n_T}^T\right] \in \Re^{d_T \times n_T}$; id for labeled training samples
$\left\{ \left(x_j^S, y_j^S\right)\Big|_{j=1}^{n_S} \right\}, y_j^S \in \Omega = \{\varpi_l\}_{l=1}^{c}$ from $\mathbf{X}_S$, where the superscript $C$ represents the number of
class types;
Supervised classifier $\zeta$; N the number of views of the TD; and $\min(d_S, d_T) \leq \left\lfloor \frac{\max(d_S, d_T)}{N} \right\rfloor$.
Train: for i = 1 to N
    generate the target domain view $\mathbf{X}_T^i \in \Re^{d_i \times n_T}, d_T = \sum_{i=i}^{N} d_i$;
    return the transformation matrices $\boldsymbol{\omega}_S^i$ and $\boldsymbol{\omega}_T^i$ according to Equation (10);
    obtain the correlation subspaces $\mathbf{X}_S^{Ci}$ and $\mathbf{X}_T^{Ci}$ according to Equations (11) and (12);
    compute the transformed training samples $\left\{ \left(\mathrm{x}_j^{SC}, y_j^S\right)\Big|_{j=1}^{n_S} \right\}$ from $\mathbf{X}_S^{Ci}$ according to id;
    train the classifier $h_i = \zeta\left(x^{SC}, y^S\right)$;
end
Output: return the classifier pool $\{h_1, \ldots, h_N\}$;
Classification: For a given new instance $\mathrm{x} = \{\mathrm{x}_i\}\big|_{i=1}^{N}, \mathrm{x}_i \in \mathbf{X}_T^{Ci}$, predict the label according to
Equation (14).
```

3.2. Semi-Supervised Version

To implement a semi-supervised version of the proposed algorithm, the multiple speed-up SRKDA approach has been incorporated into the supervised procedure. SRDKA essentially improves the original idea of the spectral regression proposed in [53] for linear discriminant analysis (LDA), by transforming the eigenvector decomposition based discriminant analysis into a regression framework via spectral graph embedding [40]. For the sake of clarity, we briefly recall here the SRKDA notation before formalizing its implementation in the new procedure.

Given the labeled samples $\left\{ \left(\mathrm{x}_j^S, y_j^S\right)\Big|_{j=1}^{n_S} \right\}, y_j^S \in \Omega = \{\varpi_l\}_{l=1}^{c}$, the LDA objective function is:

$$\begin{aligned} \boldsymbol{a}_{LDA} &= \operatorname{argmax}\frac{\boldsymbol{a}^{\dagger}\boldsymbol{\psi}_b\boldsymbol{a}}{\boldsymbol{a}^{\dagger}\boldsymbol{\psi}_w\boldsymbol{a}} \\ \boldsymbol{\psi}_b &= \sum_{k=1}^{c} n_k\left(\boldsymbol{u}^{(k)} - \boldsymbol{u}\right)\left(\boldsymbol{u}^{(k)} - \boldsymbol{u}\right)^{\dagger} \\ \boldsymbol{\psi}_w &= \sum_{k=1}^{c}\left(\sum_{q=1}^{n_k}\left(\mathbf{x}_q^{(k)} - \boldsymbol{u}^{(k)}\right)\left(\mathbf{x}_q^{(k)} - \boldsymbol{u}^{(k)}\right)^{\dagger}\right) \end{aligned} \tag{15}$$

where $\boldsymbol{u}$ is the global centroid, n_k is the number of samples in the k-th class, $\boldsymbol{u}^{(k)}$ is the centroid of the k-th class, $\mathbf{x}_q^{(k)}$ is the q-th sample in the k-th class, and $\boldsymbol{\psi}_w$ and $\boldsymbol{\psi}_b$ represent the within-class scatter matrix and the between-class scatter matrix respectively, so that the total scatter matrix is computed as $\boldsymbol{\psi}_t = \boldsymbol{\psi}_b + \boldsymbol{\psi}_w$. The best solutions for Equation (15) are the eigenvectors that correspond to the nonzero eigenvalues of:

$$\boldsymbol{\psi}_b\boldsymbol{a}_{LDA} = \lambda\boldsymbol{\psi}_t\boldsymbol{a}_{LDA} \tag{16}$$

To address the nonlinearities, the kernel extension of this procedure maps the input data to a kernel Hilbert space through nonlinear positive semi-definite kernel functions, such as the Gaussian kernel $K(\mathbf{x}, y) = \exp\left(-\|\mathbf{x} - y\|^2/2\sigma^2\right)$, the polynomial kernel $K(\mathbf{x}, y) = \left(1 + \mathbf{x}^\dagger y\right)^d$ and the sigmoid kernel $K(\mathbf{x}, y) = \tanh\left(\mathbf{x}^\dagger y + a\right)$. Generalizing Equation (15), the projective function of KDA is therefore:

$$\begin{aligned} \boldsymbol{\upsilon}_{KDA} &= \operatorname{argmax} \frac{\boldsymbol{\upsilon}^\dagger \boldsymbol{\psi}_b^\phi \boldsymbol{\upsilon}}{\boldsymbol{\upsilon}^\dagger \boldsymbol{\psi}_t^\phi \boldsymbol{\upsilon}} \\ \boldsymbol{\psi}_b^\phi &= \sum_{k=1}^{c} n_k \left(\boldsymbol{u}_\phi^{(k)} - \boldsymbol{u}_\phi\right)\left(\boldsymbol{u}_\phi^{(k)} - \boldsymbol{u}_\phi\right)^\dagger \\ \boldsymbol{\psi}_w^\phi &= \sum_{k=1}^{c} \left(\sum_{q=1}^{n_k} (\phi(\mathbf{x}_q^{(k)}) - \boldsymbol{u}_\phi^{(k)})(\phi(\mathbf{x}_q^{(k)}) - \boldsymbol{u}_\phi^{(k)})^\dagger\right) \\ \boldsymbol{\psi}_t^\phi &= \boldsymbol{\psi}_b^\phi + \boldsymbol{\psi}_w^\phi \end{aligned} \tag{17}$$

where $\boldsymbol{\psi}_b^\phi$, $\boldsymbol{\psi}_w^\phi$, and $\boldsymbol{\psi}_t^\phi$ denote the between-class, within-class and total scatter matrices in the kernel space, respectively.

Because the eigenvectors of $\boldsymbol{\psi}_b^\phi \boldsymbol{\upsilon}_{KDA} = \lambda \boldsymbol{\psi}_t^\phi \boldsymbol{\upsilon}_{KDA}$ are linear combinations of $\phi(\mathbf{x}_q)$ [54], there is always a coefficient $\boldsymbol{\varepsilon}_q$ such as $\boldsymbol{\upsilon}_{KDA} = \sum_{q=1}^{n_k} \varepsilon_q \phi(\mathbf{x}_q)$. This constrain makes Equation (17) equivalent to:

$$\boldsymbol{\varepsilon}_{KDA} = \operatorname{argmax} \frac{\boldsymbol{\varepsilon}^\dagger \mathbf{KWK} \boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^\dagger \mathbf{KK} \boldsymbol{\varepsilon}} \tag{18}$$

where $\boldsymbol{\varepsilon}_{KDA} = \left[\varepsilon_1, \ldots, \varepsilon_{n_k}\right]^\dagger$. Then, the corresponding eigenproblem becomes:

$$\mathbf{KWK}\boldsymbol{\varepsilon}_{KDA} = \lambda \mathbf{KK}\boldsymbol{\varepsilon}_{KDA} \tag{19}$$

where $\mathbf{K}$ is the kernel matrix, and the affinity matrix $\mathbf{W}$ is defined using either HeatKernel [55] or the binary weight mode:

$$W_{i,j} = \begin{cases} 1/n_k, \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ both belong to the } k^{\text{th}} \text{ class;} \\ 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{20}$$

To efficiently solve the KDA eigenproblem in Equation (19), let us consider $\boldsymbol{\vartheta}$ to be the solution of $\mathbf{W}\boldsymbol{\vartheta} = \lambda\boldsymbol{\vartheta}$. Replacing $\mathbf{K}\boldsymbol{\varepsilon}_{KDA}$ on the left side of Equation (19) by $\boldsymbol{\vartheta}$, we have:

$$\mathbf{KWK}\boldsymbol{\varepsilon}_{KDA} = \mathbf{KW}\boldsymbol{\vartheta} = \mathbf{K}\lambda\boldsymbol{\vartheta} = \lambda \boldsymbol{K}\boldsymbol{\vartheta} = \lambda\mathbf{KK}\boldsymbol{\varepsilon}_{KDA} \tag{21}$$

To avoid singularities, a constant matrix δI is added to $\mathbf{K}$ to keep it positive definite:

$$\boldsymbol{\varepsilon}_{KDA} = (\mathbf{K} + \delta I)^{-1}\boldsymbol{\vartheta} \tag{22}$$

where $\boldsymbol{I}$ is the identity matrix, and $\delta \geq 0$ represents the regularization parameter. It can be easily verified that the optimal solution given by Equation (22) is the optimal solution of the following regularized regression problem [56]:

$$\min_{f \in F} \sum_{j=1}^{n_S} \left(f\left(\mathbf{x}_j\right) - y_j\right)^2 + \delta \|f\|_K^2 \tag{23}$$

where F is the kernel space associated with the kernel K, and $\|f\|_K$ is the corresponding norm.

According to Equations (19) and (21), the solution can be reached in two steps: (1) solve the eigenproblem $\mathbf{W}\boldsymbol{\vartheta} = \lambda\boldsymbol{\vartheta}$ to obtain $\boldsymbol{\vartheta}$; and (2) find a vector $\boldsymbol{\varepsilon}_{KDA}$ that satisfies $\mathbf{K}\boldsymbol{\varepsilon}_{KDA} = \boldsymbol{\vartheta}$. For Step 1, it is easy to check that the involved affinity matrix $\mathbf{W}$ has a block-diagonal structure:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} & 0 & \cdots & 0 \\ 0 & \mathbf{W}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{W}^{(c)} \end{bmatrix} \tag{24}$$

where $\left\{\mathbf{W}^{(k)}\right\}_{k=1}^{c}$ is an $n_k \times n_k$ matrix with all of the elements defined in Equation (19), and it is straightforward to show that $\mathbf{W}^{(k)}$ has the eigenvector $\mathbf{e}^{(k)}$ associated with $e^{(k)} = [1, 1, ..., 1]^{\dagger}$. In addition, there is only one nonzero eigenvalue of $\mathbf{W}^{(k)}$ because the rank of $\mathbf{W}^{(k)}$ is always 1. Thus, there are exactly c eigenvectors of $\mathbf{W}$ with the same eigenvalue 1:

$$\boldsymbol{\vartheta}_k = [\underbrace{0, ..., 0}_{\sum_{i=1}^{k-1} n_i}, \underbrace{1, ..., 1}_{n_k}, \underbrace{0, ..., 0}_{\sum_{i=k+1}^{c} n_i}]^{\dagger} \tag{25}$$

According to the theorem proven by Cai and He in [57], the kernel matrix is positive definite, and the $\boldsymbol{c}$-1 projective function of KDA gives exactly the same solutions as the $\boldsymbol{c}$-1 linear equations systems $\mathbf{K}\boldsymbol{\varepsilon}_{KDA}^{k} = \overline{\boldsymbol{\vartheta}}^{k}$. Then let $\boldsymbol{\Theta} = [\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_{c-1}]$ be the KDA transformation matrix which embeds the data into the KDA subspace:

$$\boldsymbol{\Theta}^{\dagger}\left[K(:, x_1), ..., K(:, x_{n_k})\right] = \overline{Y}^{\dagger} \tag{26}$$

where the columns of $\overline{Y}^{\dagger}$ are the embedding results. Accordingly, the data with the same label correspond to the same point in the KDA subspace when the kernel matrix is positive definite.

To perform SRKDA in a semi-supervised way, one straightforward solution is to use the label information to guide the construction of the affinity matrix $\mathbf{W}$, as in [57–59]. Let $G = (V, E)$ be a graph with set of vertices V, which is connected by a set of edges E. The vertices of the graph are the labeled and unlabeled instances $\left(x_j^S, y_j^S\right)\Big|_{j=1}^{n_S} \cup \left\{\left(x_j^T\right)\Big|_{j=1}^{n_T}\right\}$. An edge between two vertices (or labeled and unlabeled samples) i, j represents the similarity of two instances with an associated weight $\{W_{i,j}\}$. Then, the affinity matrix W is built using both labeled and unlabeled samples. To achieve this goal, $\boldsymbol{p}$-nearest neighbors, ε-neighbors, or fully connected graph techniques can be adopted, where 0–1 weighting, Gaussian kernel weighting, Polynomial kernel weighting and Dot-product weighting can be considered to establish the graph weights [57,58]. Usually, graph-based SSL methods compute the normalized graph Laplacian:

$$L = I - D^{-1/2} W D^{-1/2} \tag{27}$$

where D denotes a diagonal matrix defined by $D_{ii} = \sum_j W_{i,j}$ (see [59,60] (Chapter 5) for more details on different families of graph based SSL methods).

According to this procedure, and inserting the notation for DA using multiple view CCA, the new semi-supervised procedure follows the steps reported in Algorithm 2.

Algorithm 2. Algorithmic details of SSMVCCAE.

```
Inputs: SD $\mathbf{X}_S = \left[\mathrm{x}_1^S, \ldots, x_{n_S}^S\right] \in \Re^{d_S \times n_S}$; TD $\mathbf{X}_T = \left[\mathrm{x}_1^T, \ldots, x_{n_T}^T\right] \in \Re^{d_T \times n_T}$; $id_S^L$ for labeled training
samples $\left\{ \left(x_j^S, y_j^S\right)\Big|_{j=1}^{n_S} \right\}, y_j^S \in \Omega = \{\omega_l\}_{l=1}^{c}$ from $\mathbf{X}_S$, where superscript $C$ represents the number
of class
types; $id_T^U$ for unlabeled candidates $\left\{ \left(x_j^T\right)\Big|_{j=1}^{n_T} \right\}$ from $\mathbf{X}_T$ Semi-supervised classifier $\zeta_{SRKDA}$; $\boldsymbol{N}$ =
Number of views of the target domain; and $\min(d_S, d_T) \le \left\lfloor \frac{\max(d_S,d_T)}{N} \right\rfloor$.
Train: for i = 1 to N
        generate the target domain view $\mathbf{X}_T^i \in \Re^{d_i \times n_T}, d_T = \sum_{i=i}^{N} d_i$;
        return the transformation matrices $\boldsymbol{\omega}_S^i$ and $\boldsymbol{\omega}_T^i$ according to Equation (10);
        obtain the correlation subspaces $\mathbf{X}_S^{Ci}$ and $\mathbf{X}_T^{Ci}$ according to Equations (11) and (12);
        compute the transformed training samples $\left\{ \left(\mathrm{x}_j^{SC}, y_j^S\right)\Big|_{j=1}^{n_S} \right\}$ from $\mathbf{X}_S^{Ci}$ according to $id_S^L$ and the
        transformed unlabeled samples $\left\{ \left(x_j^{TC}\right)\Big|_{j=1}^{n_T} \right\}$ from $\mathbf{X}_T$ according to $id_T^U$;
        build the graph Laplacian $L_i$ according to Equation (27) using $\left(x_j^{SC}, y_j^{SC}\right)\Big|_{j=1}^{n_S} \cup \left\{ \left(x_j^{TC}\right)\Big|_{j=1}^{n_T} \right\}$;
        obtain the KDA transformation matrix $\boldsymbol{\Theta}_i$ according to the solutions of Equation (26) and
Equation (22);
        return the embedded results $\overline{Y}_i^{\dagger}$;
        end
Output: return the KDA transformation matrices $\{\boldsymbol{\Theta}_i\}_{i=1}^{N}$ and the full KDA subspace embedded results
$\left\{\overline{Y}_i^{\dagger}\right\}_{i=1}^{N}$;
Classification: For a given new instance $\mathrm{x} = \{\mathrm{x}_i\}\big|_{i=1}^{N}, \mathrm{x}_i \in \mathbf{X}_T^{Ci}$
for i = 1 to N
        first map $\mathbf{x}_i$ into RKHS with the specified kernel function $\phi(\mathbf{x}_i)$;
        obtain the embedded results $\overline{Y}_{iT}$ in KDA space according to Equation (26);
        return the decision function $h_i(x) = \operatorname{argmin} \sum_{j=1}^{c} \left(\|\overline{\mathbf{y}}_{iT} - \boldsymbol{u}_j\|^2\right), \overline{\mathbf{y}}_{iT} \in \overline{Y}_{iT}$, and $\boldsymbol{u}_j = \sum_{\mathbf{x}\in c_i} \mathbf{x} / \left|c_j\right|$,
which represents the class center of $c_i$ in the KDA embedded space.
end
obtain the final predicted label by a majority voting ensemble strategy using Equation (14).
```

Summing up algorithmic details of the SMVCCAE and SSMVCCAE as described in Sections 3.1 and 3.2, Figure 1 illustrate the general flowchart for the proposed heterogeneous DA algorithms for RS image classification.

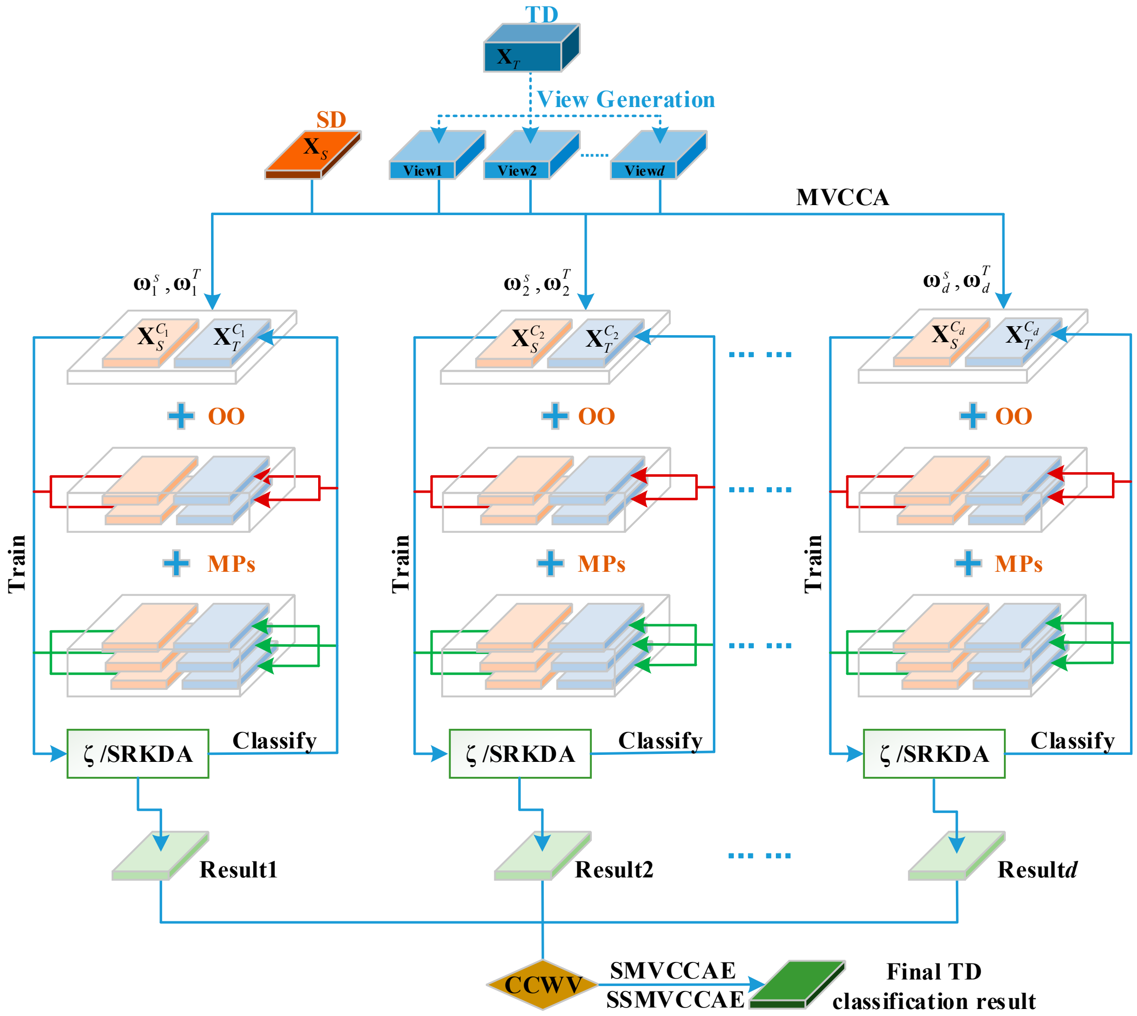


Figure 1. General flowchart for the proposed heterogeneous DA algorithms SMVCCAE and SSMVCCAE for RS image classification.

4. Data Sets and Setups

4.1. Datasets

For our analyses and evaluations, we consider two datasets, with different spatial and spectral resolutions. The first dataset is a 1.3 m spatial resolution image collected by the Reflective Optics Spectrographic Image System (ROSIS) sensor over the University of Pavia, with a size of 610 × 340 pixels (Figure 2). A total of 103 spectral reflectance bands that cover a region of the spectrum between 430 and 860 nm were retained for the analyses. The captured scene primarily represents a built-up setting with these thematic classes: asphalt, meadows, gravel, trees, metal sheets, bitumen, bare soil, bricks and shadows, as listed in Table 1. As described earlier, the main purpose of this article is to investigate the proposed methods in a heterogeneous DA problem. In this sense, the low-dimensional image is simulated by clustering the spectral space of the original ROSIS image. Specifically, the original bands of the original ROSIS image are clustered into seven groups using the K-Means algorithm, and the mean value of each cluster is considered as a new spectral band, providing a total of seven new bands. In the experiments, the new synthetic image is considered as the SD, whereas the original ROSIS image is considered as the TD.

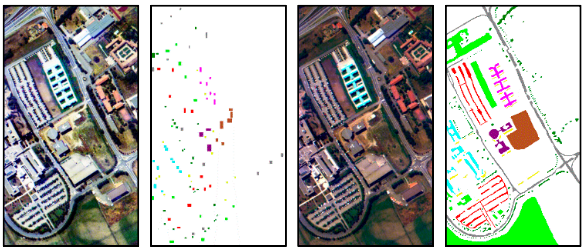

(**a**) (**b**) (**c**) (**d**)

Figure 2. (**a**–**d**) False color composite of the: synthetic low spectral resolution (**a**); and the original hyperspectral (**c**) images of the University campus in Pavia, together with: training (**b**); and validation (**d**) data sets (legend and sample details are reported in Table 1). False color composites are obtained and are displayed as R, G, and B bands 7, 5, and 4 for the synthetic, and bands 60, 30, and 2 for the original image, respectively.

Table 1. Class legend and sample details for the ROSIS University data set.

No.	**Class**	**Code**	**Source**	**Target**
			Train	**Test**
1	Asphalt	⬜	548	6631
2	Meadows	🟩	540	18649
3	Gravel	🟦	392	2099
4	Trees	🟩	524	3064
5	Metal sheets	🟪	265	1345
6	Bare soil	🟫	532	5029
7	Bitumen	🟪	375	1330
8	Bricks	🟥	514	3682
9	Shadows	🟨	231	947

The second dataset was gathered by the AVIRIS sensor over the Indian Pines test site in North-western Indiana in 1992, with 224 spectral reflectance bands in the wavelength range of 0.4 to 2.5 μm. It consists of 145 × 145 pixels with moderate spatial resolution of 20 m per pixel, and a 16-bit radiometric resolution. After an initial screening, the number of bands was reduced to 200 by removing bands 104–108, 150–163, and 220, due to noise and water absorption phenomena. This scene contains two-thirds agriculture, and one-third forest or other natural perennial vegetation. For the other Pavia data set, K-Means is used to simulate a low dimensional image with 10 bands. For illustrative purposes, Figure 3a,b shows false color composition of the simulated low dimensional and the original AVIRIS Indian Pines scene, whereas Figure 3b shows the ground truth map that is available for the scene, which is displayed in the form of a class assignment for each labeled pixel. In the experimenting stage, this ground truth map is subdivided into two parts for training and validation purposes, as detailed in Table 2.

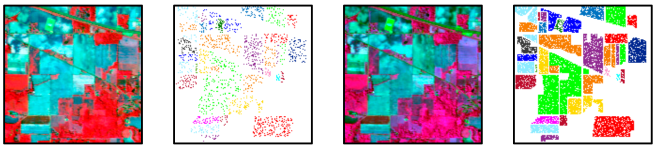

(**a**) (**b**) (**c**) (**d**)

Figure 3. (**a**–**d**) False color composites of the: simulated low spectral resolution (**a**); and original hyperspectral (**c**) images of Indian Pines data, together with: training (**b**); and validation (**d**) data sets (color legend and sample details are reported in Table 2). False color composites are obtained displaying as R, G, and B bands, 6, 4, and 5 for the synthetic, and bands 99, 51, and 21 for the original image, respectively.

Table 2. Class legend and sample details for the AVIRIS Indian Pines data set.

No.	Class	Code	Source	Target
			Train	Test
1	Alfalfa	🟪	23	23
2	Corn-notill	🟧	228	1200
3	Corn-mintill	🟦	130	700
4	Corn-notill	⬛	57	180
5	Grass-pasture	🟥	83	400
6	Grass-trees	🟨	130	600
7	Grass-pasture-mowed	🟦	14	14
8	Hay-windrowed	🟦	78	400
9	Oats	🟩	10	10
10	Soybean-notill	🟪	172	800
11	Soybean-mintill	🟩	255	2200
12	Soybean-clean	🟦	93	500
13	Wheat	🟪	55	150
14	Woods	🟥	265	1000
15	Buildings-grass-trees-drives	🟦	86	300
16	Stone-steel-towers	🟩	43	50

4.2. Experiment Setups

All of the experiments were performed using Matlab$^{\text{TM}}$ on a Windows 10 64-bit system with Intel$^{\circledR}$ Core$^{\text{TM}}$ i7-4970 CPU, @3.60 GHz, 32GB RAM. For the sake of evaluation and comparison, a Random Forest classifier (RaF) is considered as benchmark classifier for both the SMVCCAE and SVCCA approaches, because of its proven velocity, and its generalized and easy-to-implement properties [61,62]. The number of decision trees in RaF is set by default to 100, whereas the number of features is set by default to the floor of the square root of the original feature dimensionality.

For both the ROSIS and Indian Pines data sets, all of the initial and derived features have been standardized to a zero mean and unit variance. For incorporated object oriented (OO), five statistics are utilized, including the pixels' mean and standard deviation, area, orientation and major axis length of the segmented objects via K-Means clustering algorithm, whereas the spatial feature morphology profiles (MPs) are applied to the three transferred features that have the highest canonical correlation coefficients. Specifically, MPs are constructed by applying closing by reconstruction (CBR) with a circular element with a radius of 3–11 pixels, and opening by reconstruction (OBR) with an element with a radius of 3–6 pixels, refer to works carried out in [63,64]. Therefore, the feature dimensionality set in the experiments is 7 (10) vs. 103 (200) when using spectral features only for ROSIS (Indian Pines), 7 + 5 (10 + 5) vs. 103 + 5 (200 + 5) when using spectral features stacked with OO ones, 7 + 39 (10 + 39) vs. 103 + 39 (200 + 39) when using spectral features stacked with MPs features, and finally 7 + 5 + 39 (10 + 5 + 39) vs. 103 + 5 + 39 (200 + 5 + 39) when using all spectral, OO, and MPs features.

To assess the classification performances of the proposed semi-supervised approach, two state-of-the-art semi-supervised classifiers, Logistic label propagation (LLP) [65] and Laplacian support vector machine (LapSVM) [66] were considered. For the critical parameters of the semi-supervised technique (SRKDA), such as the regularization parameter δ and the number of neighbors NN used to construct the graph Laplacian L with HeatKernel [40], their values are obtained by a heuristic search in the (0.01–1) and (1–15) ranges, respectively. The parameter settings for LLP and LapSVM are instead reported in Table 3. Because LapSVM was originally proposed for binary classification problems, a one-against-all (OAA) scheme was adopted to handle the multiclass classification in our experiments.

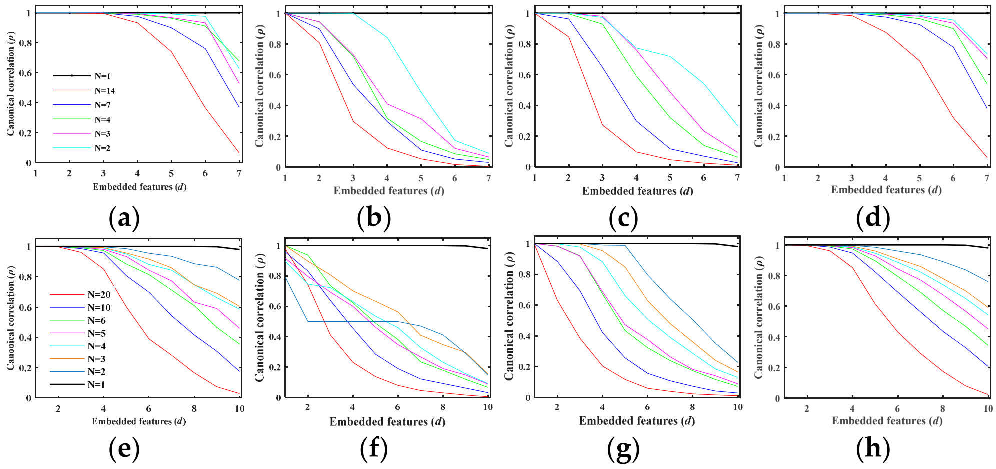
Table 3. Parameter details for LLP and LapSVM.

Classifier	Parameters	Meanings	Values
LLP	g	graph complete type	KNN
	τ	neighborhood type	Supervised
	N	neighbor size for constructing graph	5
	ω	weights for edge in graph	Heat Kernel
	σ	parameter for Heat Kernel	1
	C	regularization scale	0.001
	M	maxim iteration number	1000
	η	weight function for labeled samples	mean
LapSVM	γa	regularization parameter (ambient norm)	10^{-5}
	γi	regularization parameter (intrinsic norm)	1
	α	the initial weights	0
	κ	kernel type	RBF
	σ	RBF kernel parameter	0.01
	M	maximum iteration number	200
	c	LapSVM training type	primal
	η	Laplacian normalization	TRUE
	N	neighbor size for constructing graph	6

5. Experimental Results and Discussion

5.1. Domain Transfer Ability of MVCCA

As discussed in Section 3.1, each dimension in the derived CCA subspace is associated with a different canonical correlation coefficient which is a measure of its transfer ability. Moreover, in the MVCCA scenario, the transfer ability of each view and dimension is controlled not only by the number of views but also by the view generation technique. In this sense, Figure 4 presents the results of the average canonical correlation coefficient obtained using different view generation techniques, i.e., disjoint random sampling, uniform slice, clustering and partially joint random generation. Partially joint random view generation can apparently increase the chance of finding views with better domain transfer ability on the one hand, and to overcome the limitation ensemble techniques when the number of classifiers (equal to number of views in our case) is small on the other hand. Please note that for a more objective evaluation and comparison, each experiment was executed 10 times independently.



(**a**) (**b**) (**c**) (**d**)

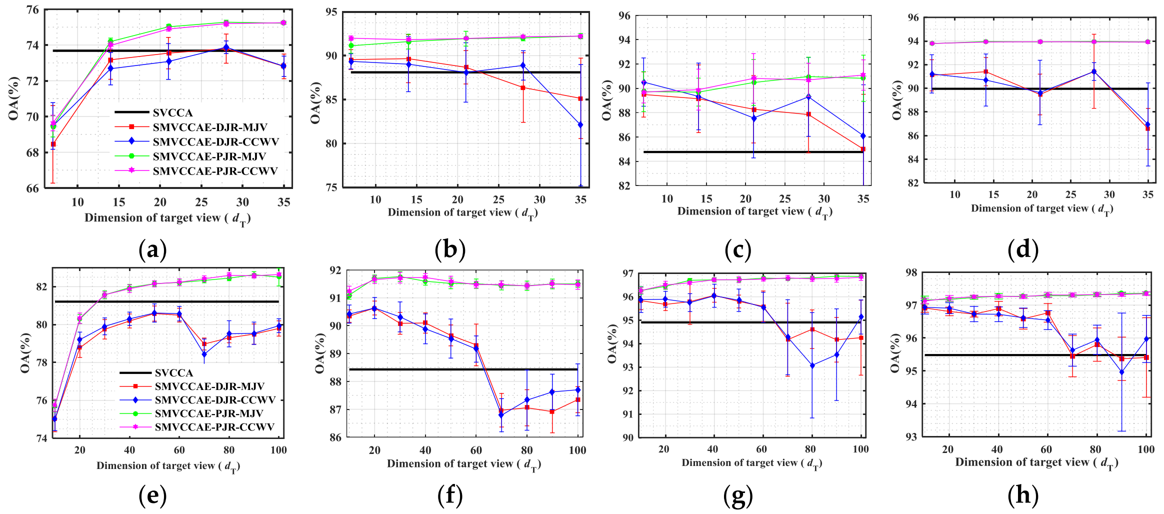
(**e**) (**f**) (**g**) (**h**)

Figure 4. Average canonical correlation coefficient versus embedded features for: ROSIS (**a**–**d**); and Indian Pines (**e**–**h**) data sets using different view generation techniques: disjoint random sampling (**a**,**e**); uniform slice (**b**,**f**); clustering (**c**,**g**); and partially joint random generation (**d**,**h**).

In Figure 4, we see that the embedded features with the highest canonical correlation coefficient are obtained by directly applying CCA without multi view generation (i.e., $n = 1$). However, single view CCA may still fail to balance potential mismatches across heterogeneous domains by overfitting, as demonstrated in the results reported in the following sections. Additionally, the decreasing trend of the canonical correlation coefficient with an increasing number of views is obvious because of the increasing mismatch between the source and target views. However, the decreasing rates of the canonical correlation coefficient for disjoint random and partially joint random generation techniques are lower than those from disjoint uniform slice and disjoint clustering view generations. Therefore, partially joint random and disjoint random view generation techniques have been selected for the following experiments.

5.2. Parameter Analysis for SMVCCAE

In Figure 5, we report the results of a sensitivity analysis of SMVCCAE that involves its critical parameters: the dimension of the target view $d_T^i = \frac{d_T}{N}$, the view generating strategies including disjoint random (DJR) and partially joint random (PJR) generation, as well as the ensemble approaches MJV and CCWV. Please note that the number of views for PJR based SMCCAE was set to 35, which is a number that will be discussed later in this paper.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

Figure 5. (**a**–**h**)Average OA values versus target view dimensionality for SMVCCAE with different fusion strategies using: spectral (**a**,**e**); spectral-OO (**b**,**f**); spectral-MPs (**c**,**g**); and spectral-OO-MPs (**d**,**h**) features on: ROSIS University (**a**–**d**); and Indian Pine datasets (**e**–**h**).

As illustrated in Figure 5 for the test data sets, the choice of PJR view generation with MJV and CCWV strategies allows the best overall accuracy values (OA curves in color green and pink). Concerning the dimensionality of the target views, they are different using different features. Specifically, for spectral features, the larger the dimensionality of the target views, the larger the OA values for PJR-based SMVCCAE because of the better domain transfer capacity with more ensemble classifiers. However, a dimensionality that is too large leads to too few view splits, i.e., a small number of ensemble elements, eventually resulting in a degraded performance. For example, when target view dimensionality is larger than four times the source view (7) dimensionality for ROSIS and larger than six times this value for Indian Pines, the OA value exhibits a decreasing trend (Figure 5a,e). Among the different types of features, (e.g., spectral and object-oriented features (labeled "spectral-OO"), spectral and morphological profile features (labeled "spectral-MPs"), and all of them together (labeled "spectral-OO-MPs"), the outcome is as expected, which is that the best results are obtained using spectral-OO-MPs. Interestingly, whereas the classification performances of the PJR-based approach are quite stable with respect to the dimensionality of the target views, the DJR-based results show a negative trend with an increasing number of target views. This finding is especially true when spatial (i.e., OO and morphological profiles) features are incorporated. This result can be explained by the trade-off between the diversity, OA and number of classifiers in an ensemble system. Specifically, the statistical diversity among spectral and spatial features tends to enhance the classification accuracy diversities more than using any view splitting strategy. As a result, the final classification performance could be limited or even degraded, especially when the number of classifiers is small.

Finally, in Figure 6, we focus on the computational complexity of the proposed approach by presenting OA, kappa statistics and CPU time values with respect to the number of views and the various fusion strategies.

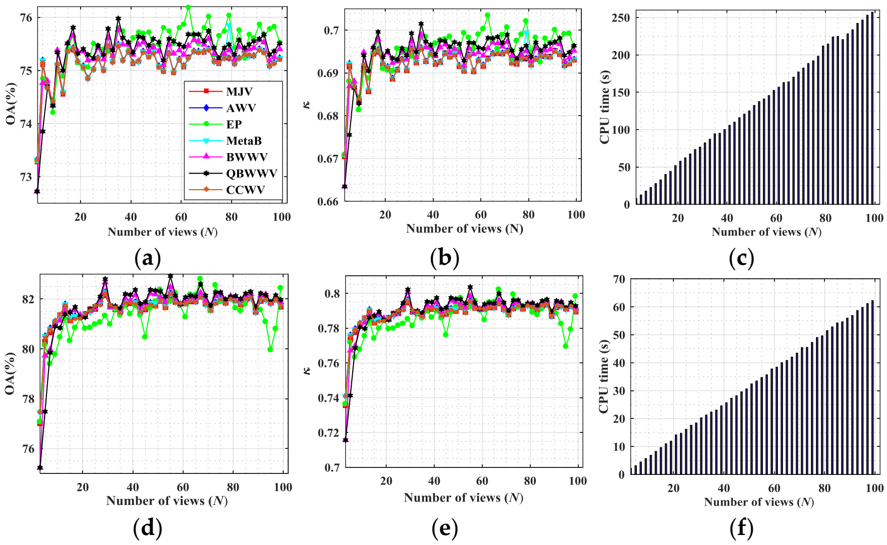


Figure 6. Average OA, Kappa (*κ*) and CPU time in seconds vs. the number of views for SMVCCA with PJR view generation and various fusion strategies applied to spectral features of ROSIS: University (**a**–**c**); and Indian Pines datasets (**d**–**f**).

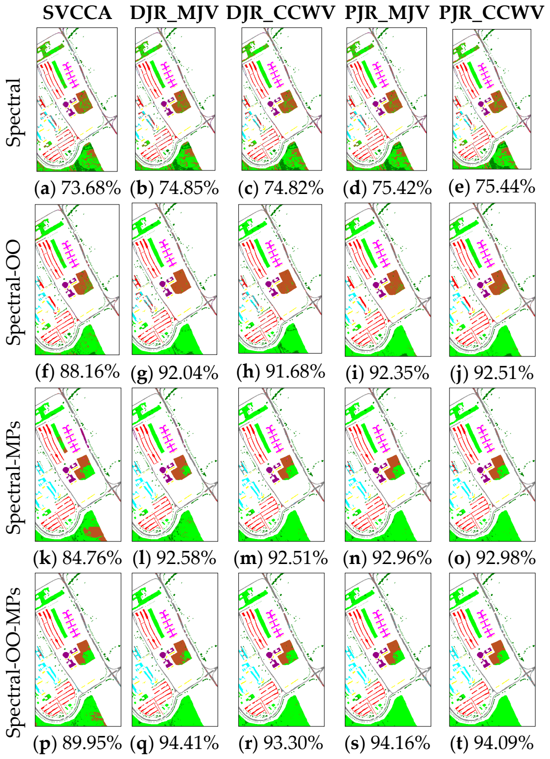
According to Figure 6, the proposed CCWV fusion technique is effective as the other fusion techniques. Apparently, with regard to the improvements in the OA values (see Figure 6a,b,d,e), and the computational burden from the number of views (see Figure 6c,f), views between 30 and 40 produce the best tradeoff between computational burden and classification accuracy.

In summary, in a scenario in which low-dimensional and high-dimensional data sets require DA, a well-designed SMVCCAE requires us to set the dimensionality of each target view to three or four times the dimensionality of the source view, and to use a PJR view generation technique.

5.3. Validation of SMVCCAE

Figure 7 provides the SMVCCAE heterogeneous cross-domain classification maps with OA values for the ROSIS University dataset using spectral, spectral-OO, spectral-MPs and spectral-OO-MPs features. Compared with the maps produced by a single-view canonical correlation analysis (SVCCA) approach, the thematic maps obtained by SMVCCAE using the associated features are better, specifically with adequate delineations of the bitumen, gravel and bare soil areas (see the numbers in Table 4). These results experimentally verify our earlier assumptions that single view CCA could fail to balance potential mismatches across heterogeneous domains by overfitting. Additionally, the most accurate result is obtained with spectral-OO-MPs by SMVCCAE using the PJR view generation strategy, as shown by the results in Figure 7 and the numbers in bold in Table 4.

For the Indian Pines dataset, Figure 8 shows the thematic maps with OA values, whereas Table 5 reports the classification accuracies (Average accuracy (AA) and OA), and kappa statistics (k) with respect to various features. Once again, the thematic maps with larger OA values produced by SMVCCAE are better than the results produced by SVCCA, especially when the OO and MPs are incorporated. The numbers in bold in Table 5 show that the largest accuracies for various class types are obtained by the SMVCCAE with the PJR technique using spectral-OO-MPs features.



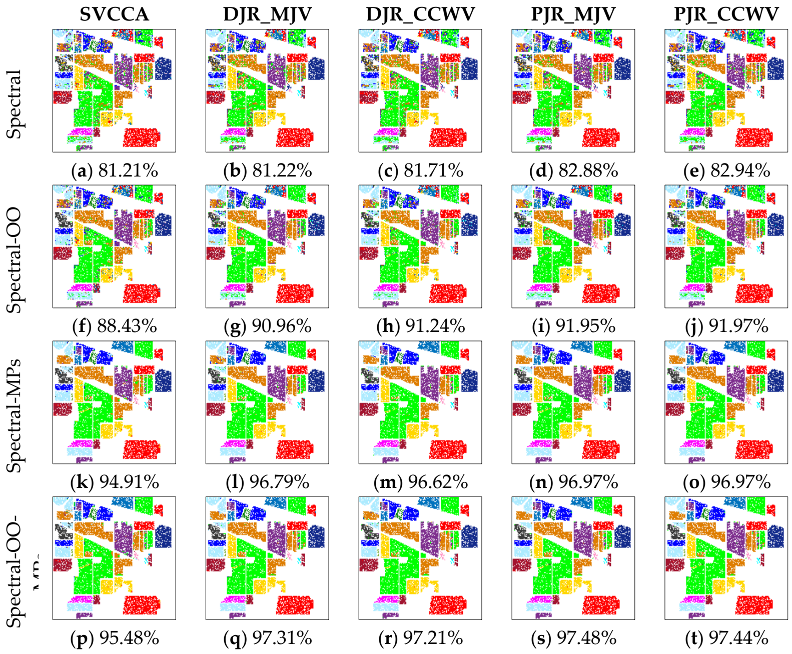
(**a**) 73.68% (**b**) 74.85% (**c**) 74.82% (**d**) 75.42% (**e**) 75.44%

(**f**) 88.16% (**g**) 92.04% (**h**) 91.68% (**i**) 92.35% (**j**) 92.51%

(**k**) 84.76% (**l**) 92.58% (**m**) 92.51% (**n**) 92.96% (**o**) 92.98%

(**p**) 89.95% (**q**) 94.41% (**r**) 93.30% (**s**) 94.16% (**t**) 94.09%

Figure 7. (**a**–**t**) Summary of the best classification maps with OA values for SMVCCAE with different fusion strategies using spectral, OO and MPs features of ROSIS University.



(**a**) 81.21% (**b**) 81.22% (**c**) 81.71% (**d**) 82.88% (**e**) 82.94%

(**f**) 88.43% (**g**) 90.96% (**h**) 91.24% (**i**) 91.95% (**j**) 91.97%

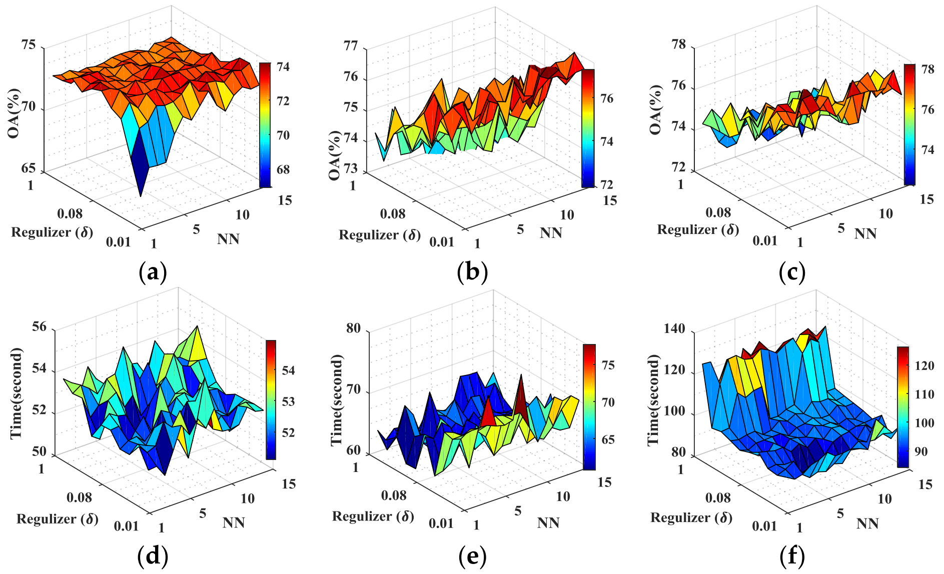
(**k**) 94.91% (**l**) 96.79% (**m**) 96.62% (**n**) 96.97% (**o**) 96.97%

(**p**) 95.48% (**q**) 97.31% (**r**) 97.21% (**s**) 97.48% (**t**) 97.44%

Figure 8. (**a**–**t**) Summary of the best classification maps with OA values for SMVCCAE with different fusion strategies using spectral, OO and MPs features of Indian Pines.

5.4. Parameter Analysis for the Semi-Supervised Version of the Algorithm

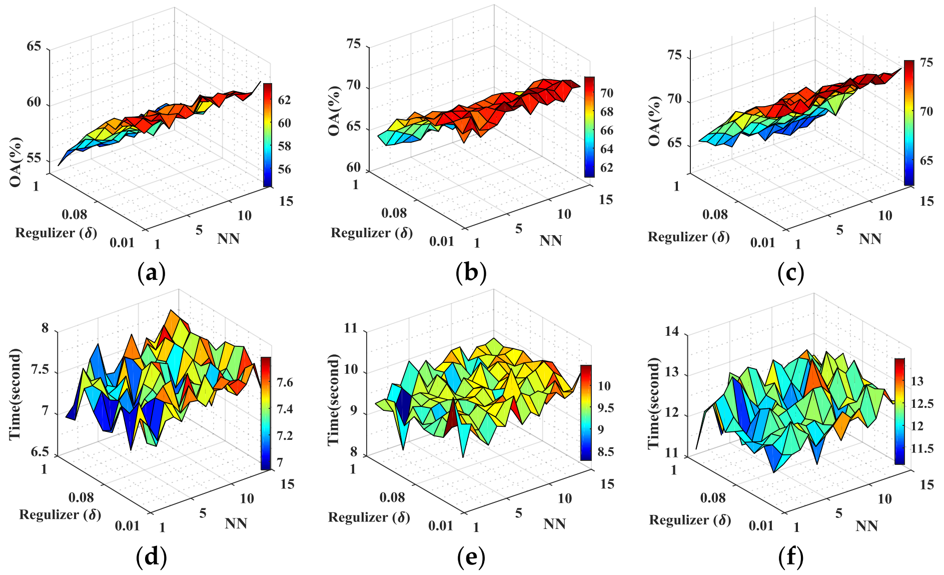
In Figures 9 and 10, we report the results of the sensitivity analysis for SSMVCCAE while considering the two critical parameters from the adopted SRKDA technique: (1) the regularization parameter δ; and (2) the number of neighbors *NN* used to construct the graph Laplacian *L*. The other parameters, such as the target view dimensionality, d_T^i and the number of total views *N* (i.e., the ensemble size), are set by default to $d_T^i = 4 \times d_s$ and *N* = 35, according to our previous experimental analysis for the supervised version of the same technique.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 9. (**a**–**f**) OA values and CPU time (in seconds) versus the regularization parameter (δ) and nearest neighborhood size (*NN*) set of SSMVCCAE with DJR view generation strategy for ROSIS University using different sizes of labeled samples: 10 pixels/class (**a**,**d**); 50 pixels/class (**b**,**e**); and 100 pixels/class (**c**,**f**).



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 10. (**a**–**f**) OA values and CPU time (in seconds) versus the regularization parameter (δ) and nearest neighborhood size (*NN*) set of SSMVCCAE with the DJR view generation strategy for Indian Pine using different size of labeled samples: 10 pixels/class (**a**,**d**); 30 pixels/class (**b**,**e**); and 55 pixels/class (**c**,**f**).

Table 4. Classification accuracy values for the SVCCA and SMVCCAE ($d_T = 4 \times d_S$) methods for ROSIS University. Considered metrics: Overall accuracy (OA), Average accuracy (AA), Kappa statistic (Kappa).

Methods	**Strategy**	**Features**	**Asphalt**	**Meadows**	**Gravel**	**Trees**	**Metal Sheets**	**Bare Soil**	**Bitumen**	**Bricks**	**Shadows**	**AA (%)**	**OA (%)**	**Kappa**
SVCCA	~	F1	94.96	91.61	66.17	52.56	98.68	45.70	70.63	69.50	99.26	76.56	73.68	0.67
		F2	98.02	95.88	93.04	71.59	98.68	71.13	90.88	76.59	**100.00**	88.42	88.76	0.85
		F3	99.47	88.06	94.39	80.89	99.55	53.10	72.31	90.67	98.90	86.37	84.76	0.80
		F4	99.86	89.32	91.35	82.43	98.67	71.98	96.44	95.76	99.15	91.66	89.96	0.87
SMVCCAE	DJR_MJV	F1	94.72	93.38	73.72	51.80	99.40	45.96	73.73	68.49	99.62	77.87	73.80	0.68
		F2	95.35	97.59	77.93	61.68	99.61	82.74	81.94	68.45	99.90	85.02	86.37	0.82
		F3	99.36	90.58	95.88	74.91	99.37	72.69	94.79	85.28	99.21	90.23	88.16	0.85
		F4	99.49	92.05	96.76	79.38	99.77	88.84	96.45	86.15	99.26	93.13	91.44	0.89
	DJR_CCWV	F1	94.81	93.25	73.44	51.82	99.65	45.99	74.43	68.29	99.61	77.92	73.84	0.68
		F2	96.56	**98.56**	78.45	62.47	99.47	92.16	88.74	68.79	99.81	87.22	88.89	0.86
		F3	99.31	90.60	96.05	74.26	99.52	75.25	96.40	85.82	99.32	90.72	88.33	0.85
		F4	99.24	90.64	96.72	77.65	99.28	93.32	97.36	86.71	99.19	93.34	91.42	0.89
	PJR_MJV	F1	95.30	94.31	76.84	52.48	99.83	48.15	76.36	69.20	99.87	79.15	75.28	0.69
		F2	98.64	98.00	95.62	77.19	99.52	89.26	97.57	72.69	99.79	92.03	92.14	0.90
		F3	99.72	91.09	98.96	82.15	99.93	80.10	**99.62**	87.11	99.36	93.11	90.97	0.88
		F4	**99.89**	91.46	94.97	**84.63**	99.93	98.81	99.34	95.66	99.40	96.01	**93.97**	**0.92**
	PJR_CCWV	F1	95.26	94.33	77.45	52.44	99.83	47.96	76.56	69.06	99.86	79.19	75.20	0.69
		F2	98.55	98.02	95.55	77.33	99.54	89.40	97.57	72.62	99.79	92.04	92.16	0.90
		F3	99.71	90.90	**99.08**	82.14	**99.95**	77.94	99.56	86.99	99.36	92.85	90.69	0.88
		F4	**99.89**	91.36	95.16	84.40	**99.95**	**98.88**	99.35	**95.78**	99.38	**96.02**	93.92	**0.92**

F1: Spectral; F2: Spectral-OO; F3: Spectral-MPs; F4: Spectral-OO-MPs.

Table 5. Classification accuracy values (average) for the SVCCA and SMVCCAE methods for Indian Pines. Considered metrics: Overall accuracy (OA), Average accuracy (AA), Kappa statistic (Kappa).

Methods	**SVCCA**				**SMVCCAE**															
Strategy	**~**				**DJR_MJV**				**DJR_CCWV**				**PJR_MJV**				**PJR_CCWV**			
Features	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**	**F1**	**F2**	**F3**	**F4**
Alfalfa	69.57	91.30	86.96	86.96	45.22	91.30	95.22	**95.65**	48.26	91.30	**95.65**	**95.65**	46.09	91.30	**95.65**	**95.65**	46.09	91.30	95.65	**95.65**
Corn-notill	81.42	87.42	90.67	93.92	78.04	87.86	90.45	91.92	77.91	87.05	90.41	91.70	80.83	89.12	91.06	92.78	80.82	89.38	91.14	**92.90**
Corn-mintill	70.71	88.71	96.86	96.57	64.70	84.54	97.00	98.16	65.34	83.29	97.31	97.91	65.87	86.54	98.03	99.20	65.43	86.93	97.97	**99.24**
Corn-notill	75.56	94.44	95.56	96.67	62.83	83.89	94.11	95.33	62.94	83.44	93.11	95.61	65.56	86.17	95.56	95.89	65.61	86.28	95.56	**95.95**
Grass-pasture	79.00	81.50	88.75	88.75	72.98	86.48	93.13	93.60	72.95	86.38	92.55	93.70	75.90	88.65	93.63	93.90	75.95	88.43	93.73	**94.00**
Grass-trees	91.00	92.17	99.00	99.00	96.03	96.80	99.70	99.69	95.68	96.73	99.67	99.64	96.58	97.33	**99.67**	**99.67**	96.50	97.49	**99.67**	**99.67**
Grass-pasture-mowed	85.71	85.71	85.71	85.71	91.43	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**
Hay-windrowed	97.75	97.50	99.75	99.75	98.45	96.30	99.68	99.70	98.68	96.93	99.68	99.63	99.45	97.70	99.73	99.75	99.23	97.83	99.65	**99.75**
Oats	90.00	90.00	100.00	100.00	80.00	89.00	91.00	99.00	85.00	92.00	97.00	98.00	91.00	99.00	**100.**	**100.**	**89.00**	**98.00**	**100.**	**100.**
Soybean-notill	77.75	87.38	87.25	89.00	79.37	89.83	91.34	92.59	79.54	89.02	90.93	92.22	81.89	91.85	92.27	93.14	81.70	91.89	92.15	**92.92**
Soybean-mintill	78.55	87.45	95.77	95.64	76.94	91.20	97.35	98.44	77.20	91.22	97.54	98.16	78.52	92.82	98.16	98.54	78.74	92.88	98.25	**98.52**
Soybean-clean	71.60	82.80	93.80	93.80	73.42	81.86	95.20	96.36	73.90	82.72	95.18	95.94	77.90	86.10	97.14	97.16	77.50	86.16	96.90	**97.18**
Wheat	**99.33**	**99.33**	98.67	98.67	98.20	99.00	98.60	98.73	98.13	98.53	98.27	98.80	98.54	**99.33**	98.67	99.00	98.47	**99.33**	98.67	98.80
Woods	96.60	96.90	99.60	99.50	98.23	98.94	99.87	99.82	98.17	98.90	99.83	99.85	98.37	99.08	**99.91**	**99.91**	98.41	99.06	99.90	99.90
Buildings-grass-trees-drives	51.33	61.33	99.00	99.00	56.00	69.60	99.77	99.97	55.27	70.27	99.90	99.97	57.17	69.87	**100.**	**100.**	56.60	70.33	**100.**	**100.**
Stone-steel-towers	**100.**	**100.**	**100.**	**100.**	99.80	**100.**	**100.**	**100.**	99.60	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**
AA (%)	82.24	89.00	94.83	95.18	79.48	89.97	95.95	96.99	80.09	90.04	96.24	96.85	81.66	91.73	97.02	**97.34**	81.43	91.76	97.01	97.33
OA (%)	81.21	88.43	94.91	95.48	80.19	90.10	96.07	96.89	80.30	89.89	96.05	96.72	81.95	91.60	96.73	**97.27**	81.89	91.71	96.73	**97.27**
Kappa	0.78	0.87	0.94	0.95	0.77	0.89	0.95	0.96	0.77	0.88	0.95	0.96	0.79	0.90	0.96	**0.97**	0.79	0.90	0.96	**0.97**

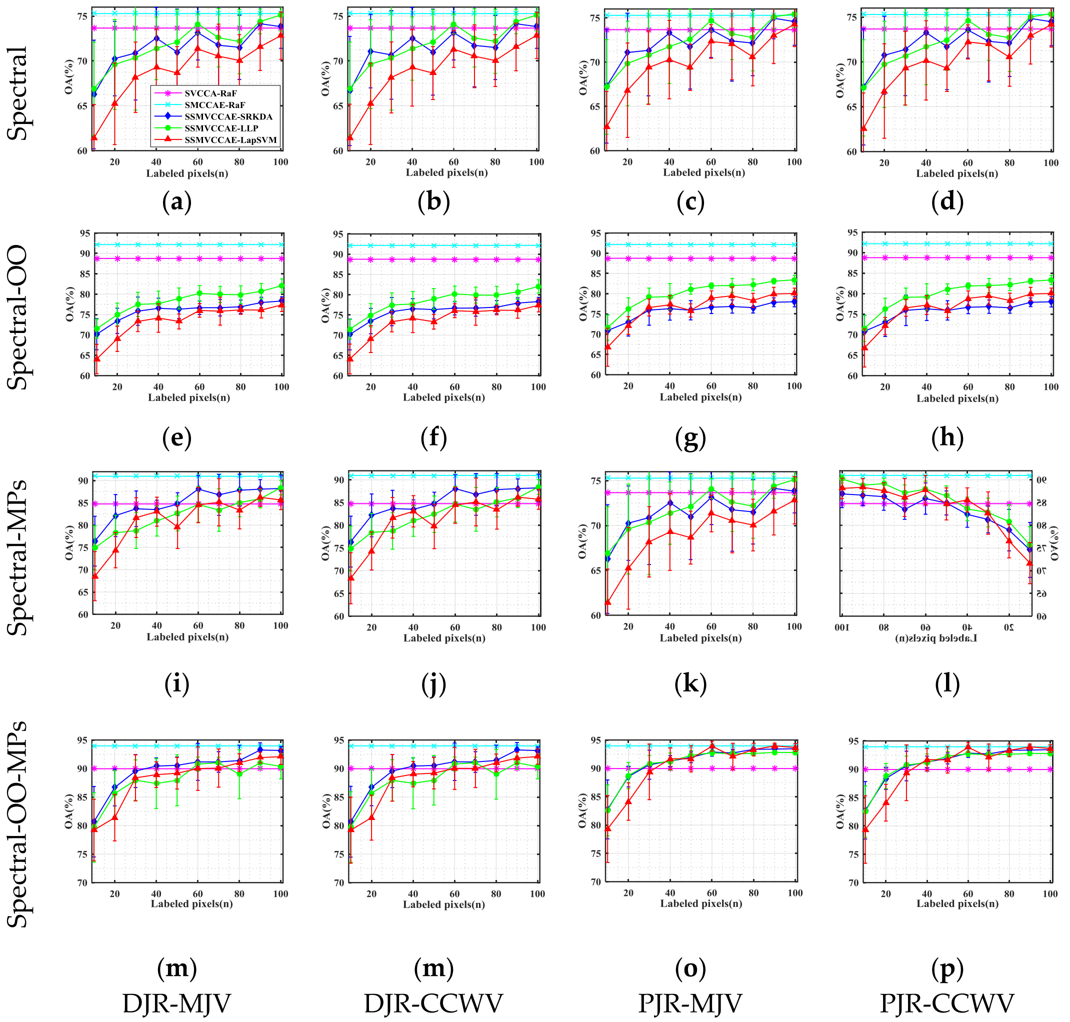
F1: Spectral; F2: Spectral-OO; F3: Spectral-MPs; F4: Spectral-OO-MPs.

According to the results, the smaller the regularization parameter *δ is* and the larger the number of neighbors *NN*, the larger the OA values. Thus, *δ* = 0.01 and *NN* = 12 were considered in all of the experiments. Computational complexity is primarily controlled by the labeled sample size (note the vertical axis in Figures 9d–f and 10d–f.

5.5. Validation of the Semi-Supervised MVCCAE

To validate the performances of the semi-supervised version of the proposed algorithm, comparisons with existing methods, specifically LLP and LapSVM, are presented for the ROSIS University data set, starting from a label set of increasing size.

Figure 11 shows the learning curves for SSMVCCAE, LLP, and LapSVM using different view generation and classifier ensemble strategies as a function of this size. Each point on the *x*-axis represents the size of the labeled samples (pixels) for each class type, while the *y*-axis represents the average overall classification accuracy. In Table 6, we report the average overall classification accuracies and kappa statistics (*κ*) over 10 independent runs, when a total of 100 labeled samples are considered for each class.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

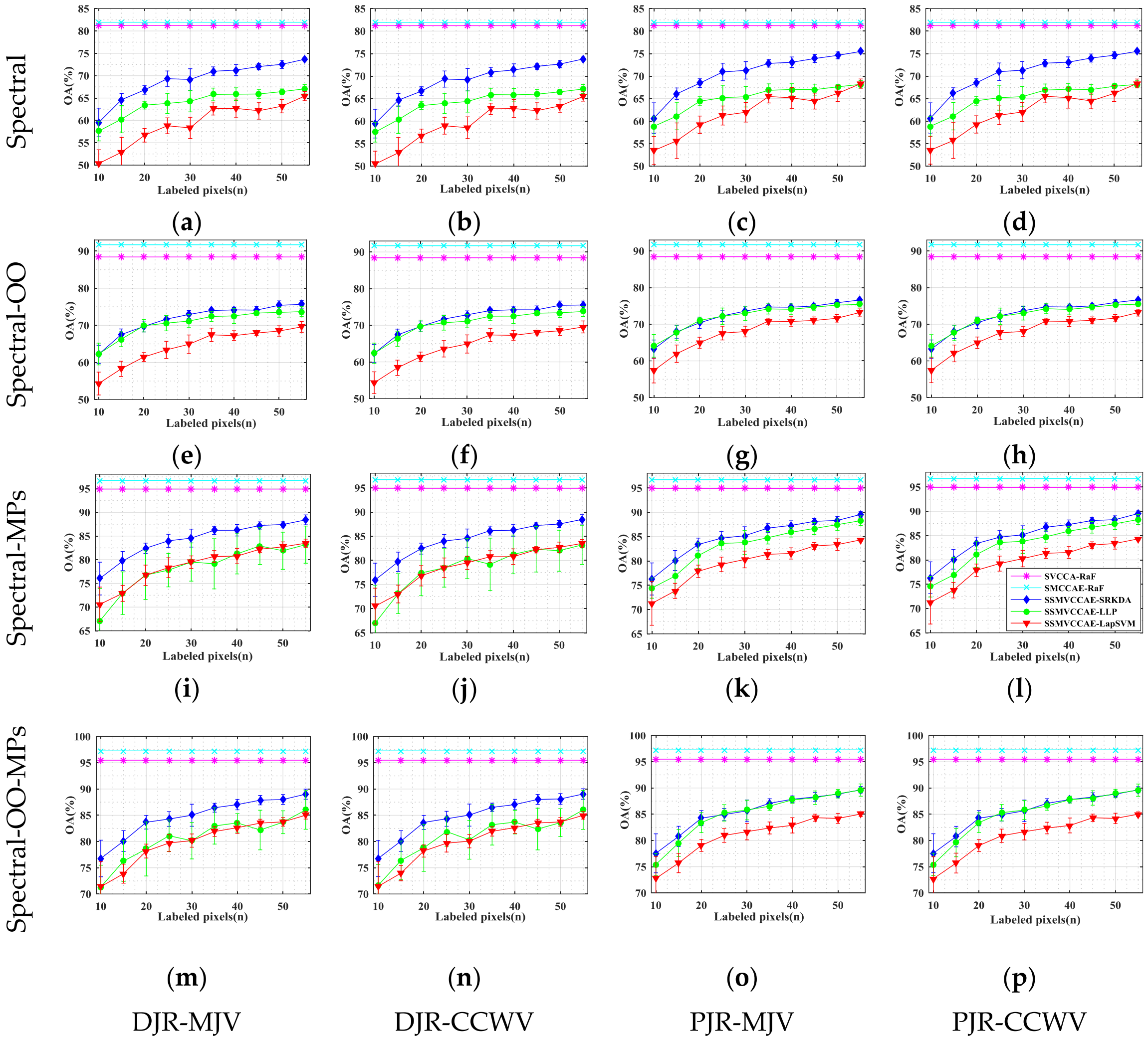
(**m**) (**m**) (**o**) (**p**)

DJR-MJV DJR-CCWV PJR-MJV PJR-CCWV

Figure 11. (**a**–**p**) Average OA values versus labeled pixels for SSMVCCAE with different view generation and fusion strategies for ROSIS University dataset.

According to the results in Figure 11 and Table 6, the proposed semi-supervised heterogeneous DA approach achieves comparable and sometimes better results in any case (see the learning curves in blue for SSMVCCAE-SRKDA vs. green for SSMVCCAE-LPP and red for SSMVCCAE-LapSVM in Figure 11). Moreover, larger OA values with faster convergence rates are shown by SSMVCCAE with PJR as opposed to DJR view generation, either by MJV fusion or by the CCWV fusion, especially using the spectral-OO-MPs features.

In Figure 12 and Table 7, the results of the same experiments are reported for the Indian Pines test set. Please note that because only a few samples are available for some classes in the Indian Pines case, class types that contain less than 70 pixels for training are not considered here. Even in this case, to obtain a more objective comparison and evaluation, each test is executed independently for 10 rounds.



(**a**) (**b**) (**c**) (**d**)

(**e**) (**f**) (**g**) (**h**)

(**i**) (**j**) (**k**) (**l**)

(**m**) (**n**) (**o**) (**p**)

Figure 12. (**a**–**p**) Average OA values versus labeled pixels for SSMVCCAE with different view generation and fusion strategies on Indian Pines data.

Figure 12 shows that better classification results are obtained by the SSMVCCAE with SRKDA, not only using the original spectral features but also using spectral features that incorporate OO and MPs features (see the learning curves in blue vs. those in green and red). Moreover, the best classification results are obtained by SSMVCCAE-SRKDA with the PJR view generation technique, and when considering the spectral-OO-MPs stacked features (see the numbers in bold in Table 7).

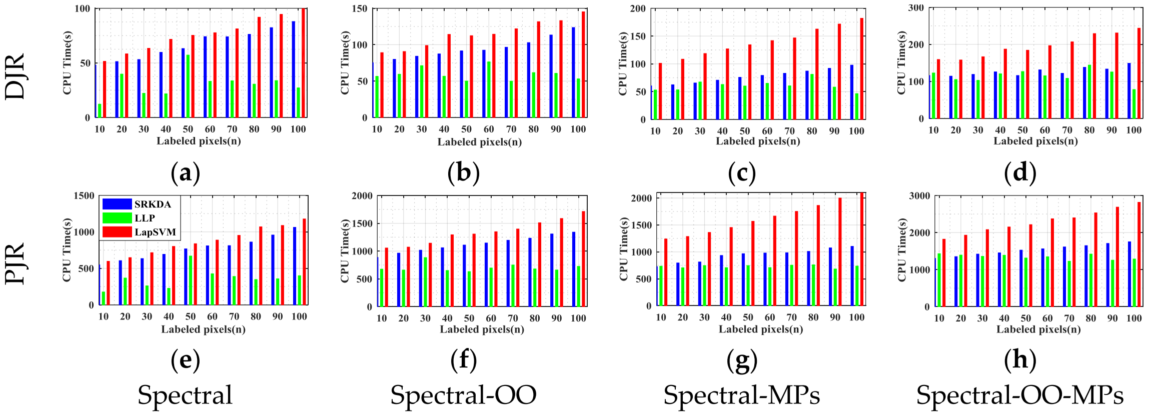
Table 6. Average overall classification accuracies and kappa statistics (κ) for SSMVCCAE with different semi-supervised classifiers for the ROSIS University data. Total 100 labeled samples are available for each class over 10 independent runs.

Classifier	**SSMVCCAE (SRKDA)**								**SSMVCCAE (LPP)**								**SSMVCCAE (LapSVM)**							
View Generation	**DJR**				**PJR**				**DJR**				**PJR**				**DJR**				**PJR**			
Voting	**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**	
Statistics	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ
Spectral	73.82	0.66	73.83	0.66	74.58	0.67	74.47	0.67	75.12	0.68	75.11	0.68	**75.45**	**0.68**	75.37	0.68	72.82	0.66	72.85	0.66	74.26	0.67	74.16	0.67
Spectral-OO	78.31	0.72	78.29	0.72	78.02	0.72	78.00	0.72	82.08	0.76	82.04	0.76	**83.44**	**0.78**	83.32	0.78	77.37	0.71	77.38	0.71	80.08	0.74	80.04	0.74
Spectral-MPs	88.23	0.85	88.24	0.85	86.73	0.82	86.89	0.83	88.44	0.85	88.45	0.85	90.16	0.87	**90.17**	**0.87**	85.65	0.82	85.72	0.82	88.14	0.85	88.12	0.85
Spectral-OO-MPs	93.17	0.91	93.14	0.91	93.47	0.91	93.46	0.91	90.36	0.87	90.31	0.87	92.78	0.90	92.77	0.90	92.08	0.90	92.10	0.90	**93.68**	**0.92**	93.67	0.92

Table 7. Average overall classification accuracies and kappa statistics (κ) for SSMVCCAE with different semi-supervised classifiers for the Indian Pines data. A total of 55 labeled samples are available for each class over 10 independent runs.

Classifier	**SSMVCCAE (SRKDA)**								**SSMVCCAE (LPP)**								**SSMVCCAE (LapSVM)**							
View Generation	**DJR**				**PJR**				**DJR**				**PJR**				**DJR**				**PJR**			
Voting	**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**		**MJV**		**CCWV**	
Statistics	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ	**OA**	κ
Spectral	73.67	0.70	73.81	0.70	75.49	0.72	**75.51**	**0.72**	67.05	0.62	67.13	0.62	68.18	0.64	68.17	0.64	65.43	0.61	65.48	0.61	68.34	0.64	68.38	0.64
Spectral-OO	75.64	0.72	75.53	0.72	76.61	0.73	**76.66**	**0.73**	73.69	0.70	73.75	0.70	75.42	0.72	75.47	0.72	69.59	0.65	69.54	0.65	73.20	0.69	73.18	0.69
Spectral-MPs	88.42	0.87	88.49	0.87	89.57	0.88	**89.54**	**0.88**	83.19	0.81	83.22	0.81	88.29	0.86	88.31	0.86	83.49	0.81	83.51	0.81	84.33	0.82	84.33	0.82
Spectral-OO-MPs	89.02	0.87	89.05	0.87	**89.66**	**0.88**	**89.66**	**0.88**	86.01	0.84	86.02	0.84	89.60	0.88	89.59	0.88	85.01	0.83	84.99	0.83	85.03	0.83	85.01	0.83

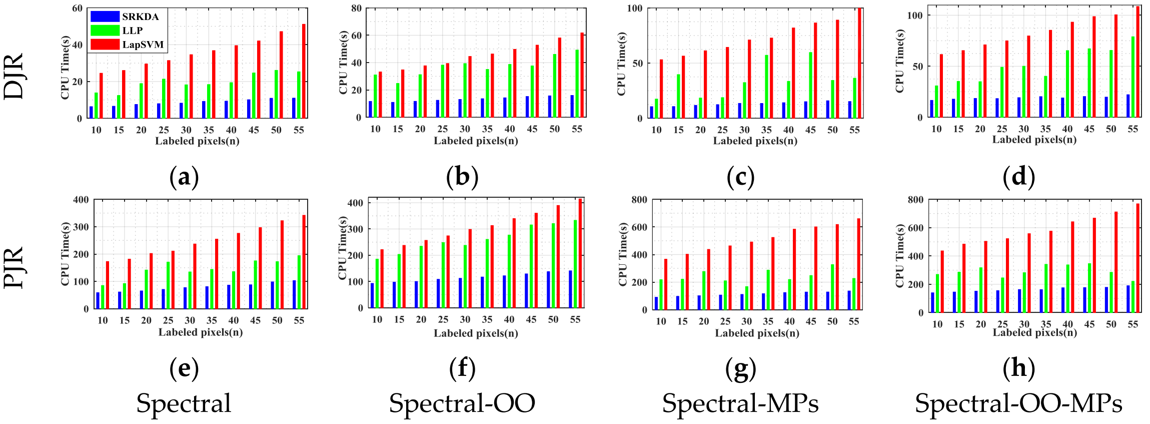
Finally, in Figures 13 and 14, the CPU time consumptions in seconds for the different implementations of the semi-supervised procedure are reported as a function of the labeled sample size for both Pavia and Indian Pines. According to the results, SSMVCCAE with SRKDA is only slightly more efficient than LapSVM for the ROSIS University data, but is much more efficient for the Indian Pines data. Moreover, the computational complexities of LapSVM and LLP increase linearly with the number of labeled samples, because they are more visible for the Indian Pines data, whereas the CPU time for SRKDA stays almost constant.



(**a**) (**b**) (**c**) (**d**)

(**e**) Spectral (**f**) Spectral-OO (**g**) Spectral-MPs (**h**) Spectral-OO-MPs

Figure 13. (**a**–**h**) CPU time consumption in seconds versus the size of the labeled samples for SSMVCCAE-SRKDA/-LLP/-LapSVM for the ROSIS University data.



(**a**) (**b**) (**c**) (**d**)

(**e**) Spectral (**f**) Spectral-OO (**g**) Spectral-MPs (**h**) Spectral-OO-MPs

Figure 14. (**a**–**h**) CPU time versus the size of the labeled samples for SSMVCCAE-SRKDA/ -LLP/-LapSVM for the Indian Pines data.

Summing the results presented in this section, it can be concluded that the novel proposed semi-supervised heterogeneous DA approach works properly and achieves satisfactory results better than the current state-of-the-art techniques when using a PJR view generation technique either with majority voting or with canonical correlation coefficient voting. A comparison of the results by SSMVCCAE with those by LLP and LapSVM shows that the performance of SRKDA is superior for both classification accuracy and computational efficiency. Finally, the computational burden caused by the sizes of the labeled samples and feature dimensionality is much smaller for SSMVCCAE with SRKDA, whereas it increases linearly with the sample size when using the other techniques.

6. Conclusions

In this paper, we have presented the implementation details, analyzed the parameter sensitivity, and proposed a comprehensive validation of two versions of an ensemble classifier that is suitable for heterogeneous DA and based on multiple view CCA. The main idea is to overcome the limitations of SVCCA by incorporating multi view CCA into EL. Superior results have been proven using two high dimensional (hyperspectral) images, the ROSIS Pavia University and the AVIRIS Indian Pine datasets, as high dimensional target domains, with synthetic low dimensional (multispectral) images as associated SDs. The best classification results were always obtained by jointly considering the original spectral features stacked with object-oriented features assigned to segmentation results, and the morphological profiles, which were subdivided into multiple views using the PJR view generation technique.

To further mitigate the marginal and/or conditional distribution gap between the source and the target domains, when few or even no labeled samples are available from the target domain, we propose a semi-supervised version of the same approach via training multiple speed-up SRKDA.

For new research directions, we are considering more complex problems, such as single SD vs. multiple TDs, as well as multiple SDs vs. multiple TDs supervised and semi-supervised adaptation techniques.

Acknowledgments: This work was partially supported by the Project funded by China Postdoctoral Science Foundation (2016M592872), the Xinjiang Uyghur Autonomous Region High Level Talents Introduction Project (Y648031) and the National Natural Science Foundation of China (No. 41601440, No. 41601354, and No. 41471098).

Author Contributions: Alim Samat developed the algorithms, executed all of the experiments, finished the original manuscript and the subsequent revisions, and provided part of the funding. Claudio Persello and Paolo Gamba offered valuable suggestions and comments, and carefully revised the original manuscript and its revisions. Jilili Abuduwaili provided part of the funding. Sicong Liu and Erzhu Li, contributed to revising of the manuscript.

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

References

1. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
2. Curlander, J.C. Location of spaceborne SAR imagery. *IEEE Trans. Geosci. Remote Sens.* **1982**, *3*, 359–364. [CrossRef]
3. Bruzzone, L.; Cossu, R. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 1984–1996. [CrossRef]
4. Torralba, A.; Efros, A.A. Unbiased Look at Dataset Bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1521–1528.
5. Khosla, A.; Zhou, T.; Malisiewicz, T.; Efros, A.A.; Torralba, A. Undoing the Damage of Dataset Bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 20–25 June 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 158–171.
6. Schott, J.R.; Salvaggio, C.; Volchok, W.J. Radiometric scene normalization using pseudoinvariant features. *Remote Sens. Environ.* **1988**, *26*, 1–16. [CrossRef]
7. Woodcock, C.E.; Macomber, S.A.; Pax-Lenney, M.; Cohen, W.B. Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. *Remote Sens. Environ.* **2001**, *78*, 194–203. [CrossRef]
8. Olthof, I.; Butson, C.; Fraser, R. Signature extension through space for northern landcover classification: A comparison of radiometric correction methods. *Remote Sens. Environ.* **2005**, *95*, 290–302. [CrossRef]
9. Rakwatin, P.; Takeuchi, W.; Yasuoka, Y. Stripe noise reduction in MODIS data by combining histogram matching with facet filter. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1844–1856. [CrossRef]
10. Inamdar, S.; Bovolo, F.; Bruzzone, L.; Chaudhuri, S. Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 1243–1252. [CrossRef]
11. Bruzzone, L.; Marconcini, M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 770–787. [CrossRef] [PubMed]
12. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]

13. Banerjee, B.; Bovolo, F.; Bhattacharya, A.; Bruzzone, L.; Chaudhuri, S.; Buddhiraju, K.M. A Novel Graph-Matching-Based Approach for Domain Adaptation in Classification of Remote Sensing Image Pair. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4045–4062. [CrossRef]
14. Matasci, G.; Volpi, M.; Kanevski, M.; Bruzzone, L.; Tuia, D. Semisupervised Transfer Component Analysis for Domain Adaptation in Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3550–3564. [CrossRef]
15. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
16. Samat, A.; Gamba, P.; Abuduwaili, J.; Liu, S.; Miao, Z. Geodesic Flow Kernel Support Vector Machine for Hyperspectral Image Classification by Unsupervised Subspace Feature Transfer. *Remote Sens.* **2016**, *8*, 234. [CrossRef]
17. Daumé, H., III; Kumar, A.; Saha, A. Frustratingly Easy Semi-Supervised Domain Adaptation. In Proceedings of the 2010 Workshop on Domain Adaptation Natural Language Processing, Uppsala, Sweden, 15 July 2010; pp. 53–59.
18. Li, W.; Duan, L.; Xu, D.; Tsang, I.W. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1134–1148. [CrossRef] [PubMed]
19. Patel, V.M.; Gopalan, R.; Li, R.; Chellappa, R. Visual domain adaptation: A survey of recent advances. *IEEE Sign. Process. Mag.* **2015**, *32*, 53–69. [CrossRef]
20. Gao, J.; Fan, W.; Jiang, J.; Han, J. Knowledge Transfer via Multiple Model Local Structure Mapping. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 283–291.
21. Jiang, J.; Zhai, C. Instance Weighting for Domain Adaptation in NLP. In Proceedings of the ACL, Prague, Czech Republic, 23–30 June 2007; Volume 7, pp. 264–271.
22. Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P.V.; Kawanabe, M. Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems*; Springer: Vancouver, BC, Canada, 2008; pp. 1433–1440.
23. Persello, C.; Bruzzone, L. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4468–4483. [CrossRef]
24. Bruzzone, L.; Persello, C. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3180–3191. [CrossRef]
25. Pan, S.J.; Kwok, J.T.; Yang, Q. Transfer Learning via Dimensionality Reduction. In Proceedings of the AAAI, 8, Stanford, CA, USA, 26–28 March 2008; pp. 677–682.
26. Long, M.; Wang, J.; Ding, G.; Pan, S.J.; Yu, P.S. Adaptation regularization: A general framework for transfer learning. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 1076–1089. [CrossRef]
27. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2011**, *22*, 199–210. [CrossRef] [PubMed]
28. Mihalkova, L.; Huynh, T.; Mooney, R.J. Mapping and Revising Markov Logic Networks for Transfer Learning. In Proceedings of the AAAI, 7, Vancouver, BC, Canada, 22–26 July 2007; pp. 608–614.
29. Yeh, Y.R.; Huang, C.H.; Wang, Y.C.F. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Trans. Image Proc.* **2014**, *23*, 2009–2018.
30. Gopalan, R.; Li, R.; Chellappa, R. Domain Adaptation for Object Recognition: An Unsupervised Approach. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 999–1006.
31. Gopalan, R.; Li, R.; Chellappa, R. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *6*, 2288–2302. [CrossRef] [PubMed]
32. Duan, L.; Xu, D.; Tsang, I.W. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 504–518. [CrossRef] [PubMed]
33. Wang, W.; Zhou, Z.H. A New Analysis of Co-Training. In Proceedings of the 27th International Conference on Machine Learning, (ICML-10) 2010, Haifa, Israel, 21–24 June 2010; pp. 1135–1142.
34. Di, W.; Crawford, M.M. View generation for multiview maximum disagreement based active learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 1942–1954. [CrossRef]

35. Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.* **2013**, *23*, 2031–2038. [CrossRef]
36. Kuncheva, L.I.; Rodríguez, J.J.; Plumpton, C.O.; Linden, D.E.; Johnston, S.J. Random subspace ensembles for fMRI classification. *IEEE Trans. Med. Imaging* **2010**, *29*, 531–542. [CrossRef] [PubMed]
37. Samat, A.; Du, P.; Liu, S.; Li, J. E2LMs: Ensemble extreme learning machines for hyperspectral image classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 1060–1069. [CrossRef]
38. Hady, M.F.A.; Schwenker, F. Semi-Supervised Learning. In *Handbook on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 215–239.
39. Kulis, B.; Saenko, K.; Darrell, T. What you Saw Is not What you Get: Domain Adaptation Using Asymmetric Kernel Transforms. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1785–1792.
40. Kumar, A.; Saha, A.; Daume, H. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*; Springer: Vancouver, BC, Canada, 2010; pp. 478–486.
41. Cai, D.; He, X.; Han, J. Speed up kernel discriminant analysis. *VLDB J. Int. J. Very Large Data Bases* **2011**, *20*, 21–33. [CrossRef]
42. Dai, W.; Chen, Y.; Xue, G.R.; Yang, Q.; Yu, Y. Translated learning: Transfer learning across different feature spaces. In *Advances in Neural Information Processing Systems*; Springer: Vancouver, BC, Canada, 2008; pp. 353–360.
43. Yang, Q.; Chen, Y.; Xue, G.R.; Dai, W.; Yu, Y. Heterogeneous Transfer Learning for Image Clustering via the Social Web. In Proceedings of the Joint Conference 47th Annual Meeting of the ACL and the 4th International Joint Conference Natural Language Process, AFNLP, Singapore, 2–7 August 2009; Volume 1, pp. 1–9.
44. Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S.J.; Xue, G.R.; Yu, Y.; Yang, Q. Heterogeneous Transfer Learning for Image Classification. In Proceedings of the AAAI, San Francisco, CA, USA, 7–11 August 2011.
45. Evangelopoulos, N.; Zhang, X.; Prybutok, V.R. Latent semantic analysis: Five methodological recommendations. *Eur. J. Inf. Syst.* **2012**, *21*, 70–86. [CrossRef]
46. Hong, L. A Tutorial on Probabilistic Latent Semantic Analysis. Available online: https://arxiv.org/pdf/1212.3900.pdf (accessed on 21 December 2012).
47. Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory* **2001**, *47*, 1902–1914. [CrossRef]
48. Gong, B.; Shi, Y.; Sha, F.; Grauman, K. Geodesic Flow Kernel for Unsupervised Domain Adaptation. In Proceedings of the 2012 IEEE Conference Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 2066–2073.
49. Shi, X.; Liu, Q.; Fan, W.; Yu, P.S.; Zhu, R. Transfer Learning on Heterogenous Feature Spaces via Spectral Transformation. In Proceedings of the IEEE 10th International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 1049–1054.
50. Zhou, J.T.; Tsang, I.W.; Pan, S.J.; Tan, M. Heterogeneous Domain Adaptation for Multiple Classes. In Proceedings of the AISTATS, Reykjavik, Iceland, 22–25 April 2014; pp. 1095–1103.
51. Hardoon, D.R.; Szedmak, S.; Shawe-Taylor, J. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **2004**, *16*, 2639–2664. [CrossRef] [PubMed]
52. Rokach, L. *Pattern Classification Using Ensemble Methods*; World Scientific Publishing Company: Singapore, 2010; Volume 75.
53. Xu, C.; Tao, D.; Xu, C. Multi-View Learning with Incomplete Views. *IEEE Trans. Image Proc.* **2015**, *24*, 5812–5825. [CrossRef] [PubMed]
54. Cai, D.; He, X.; Han, J. Spectral Regression: A Unified Subspace Learning Framework for Content-Based Image Retrieval. In Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, 24–29 September 2007; pp. 403–412.
55. Baudat, G.; Anouar, F. Generalized discriminant analysis using a kernel approach. *Neural Comput.* **2000**, *12*, 2385–2404. [CrossRef] [PubMed]
56. Vassilevich, D.V. Heat kernel expansion: User′s manual. *Phys. Rep.* **2003**, *388*, 279–360. [CrossRef]
57. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
58. Cai, D.; He, X.; Han, J. Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1624–1637. [CrossRef]

59. Subramanya, A.; Talukdar, P.P. Graph-based semi-supervised learning. *Synth. Lect. Arti. Intell. Mach. Learn.* **2014**, *8*, 1–125. [CrossRef]
60. Camps-Valls, G.; Marsheva, T.V.B.; Zhou, D. Semi-supervised graph-based hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 3044–3054. [CrossRef]
61. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
62. Du, P.; Samat, A.; Waske, B.; Liu, S.; Li, Z. Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 38–53. [CrossRef]
63. Benediktsson, J.A.; Palmason, J.A.; Sveinsson, J.R. Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 480–491. [CrossRef]
64. Fauvel, M.; Benediktsson, J.A.; Chanussot, J.; Sveinsson, J.R. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3804–3814. [CrossRef]
65. Kobayashi, T.; Watanabe, K.; Otsu, N. Logistic label propagation. *Pattern Recognit. Lett.* **2012**, *33*, 580–588. [CrossRef]
66. Melacci, S.; Belkin, M. Laplacian support vector machines trained in the primal. *J. Mach. Learn. Res.* **2011**, *12*, 1149–1184.