Information Fusion 78 (2022) 20–39

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Multimodal Earth observation data fusion: Graph-based approach in shared latent space

P.V. Arun [a,b], R. Sadeh [b], A. Avneri [b], Y. Tubul [b], C. Camino [c], K.M. Buddhiraju [a], A. Porwal [a], R. N. Lati [d], P.J. Zarco-Tejada [e,f], Z. Peleg [b], I. Herrmann [b,*]

[a] Indian Institute of Technology Bombay, India
[b] The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot, 7610001, Israel
[c] European Commission (EC), Joint Research Centre (JRC), Ispra (VA), Italy
[d] Department of Plant Pathology and Weed Research, Newe Ya'ar Research Center, Agricultural Research Organization, Ramat Yishay, 30095, Israel
[e] School of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences (FVAS) & Department of Infrastructure Engineering, Melbourne School of Engineering (MSE), University of Melbourne, Parkville, Australia
[f] Instituto de Agricultura Sostenible (IAS), Consejo Superior de Investigaciones Científicas (CSIC), Alameda del Obispo s/n, 14004, Cordoba, Spain

ARTICLE INFO

Key words:
Convolutional neural networks
Fusion
Ground measured spectra
Multispectral UAV
Hyperspectral

ABSTRACT

Multiple and heterogenous Earth observation (EO) platforms are broadly used for a wide array of applications, and the integration of these diverse modalities facilitates better extraction of information than using them individually. The detection capability of the multispectral unmanned aerial vehicle (UAV) and satellite imagery can be significantly improved by fusing with ground hyperspectral data. However, variability in spatial and spectral resolution can affect the efficiency of such dataset's fusion. In this study, to address the modality bias, the input data was projected to a shared latent space using cross-modal generative approaches or guided unsupervised transformation. The proposed adversarial networks and variational encoder-based strategies used bi-directional transformations to model the cross-domain correlation without using cross-domain correspondence. It may be noted that an interpolation-based convolution was adopted instead of the normal convolution for learning the features of the point spectral data (ground spectra). The proposed generative adversarial network-based approach employed dynamic time wrapping based layers along with a cyclic consistency constraint to use the minimal number of unlabeled samples, having cross-domain correlation, to compute a cross-modal generative latent space. The proposed variational encoder-based transformation also addressed the cross-modal resolution differences and limited availability of cross-domain samples by using a mixture of expert-based strategy, cross-domain constraints, and adversarial learning. In addition, the latent space was modelled to be composed of modality independent and modality dependent spaces, thereby further reducing the requirement of training samples and addressing the cross-modality biases. An unsupervised covariance guided transformation was also proposed to transform the labelled samples without using cross-domain correlation prior. The proposed latent space transformation approaches resolved the requirement of cross-domain samples which has been a critical issue with the fusion of multi-modal Earth observation data. This study also proposed a latent graph generation and graph convolutional approach to predict the labels resolving the domain discrepancy and cross-modality biases. Based on the experiments over different standard benchmark airborne datasets and real-world UAV datasets, the developed approaches outperformed the prominent hyperspectral panchromatic sharpening, image fusion, and domain adaptation approaches. By using specific constraints and regularizations, the network developed was less sensitive to network parameters, unlike in similar implementations. The proposed approach illustrated improved generalizability in comparison with the prominent existing approaches. In addition to the fusion-based classification of the multispectral and hyperspectral datasets, the proposed approach was extended to the classification of hyperspectral airborne datasets where the latent graph generation and convolution were employed to resolve the domain bias with a small number of training samples. Overall, the developed transformations and architectures will be useful for the semantic interpretation and analysis of multimodal data and are applicable to signal processing, manifold learning, video analysis, data mining, and time series analysis, to name a few.

* Corresponding author.

https://doi.org/10.1016/j.inffus.2021.09.004
Received 10 October 2020; Received in revised form 5 July 2021; Accepted 15 September 2021
Available online 20 September 2021

1. Introduction

Airborne and spaceborne images are essential tools for environmental monitoring and management in Earth-science disciplines such as lithology, pedology, agriculture, ecology, and forestry. The conventional high-altitude aerial data sources are supplemented by unmanned aerial vehicles (UAVs) as they are easily available, more cost-effective and allow better data flexibility (Maes and Steppe, [26]). However, satellite and manned or unmanned airborne platforms are still being widely used owing to the operational benefits and wide aerial coverage. Reflectance spectroscopy using ground-based instruments such as spectroradiometers are also essential for studying the fine spectral characteristics of the targets. The issue of handling multimodal data is critical for effectively utilizing the data available from various sources, viz. UAV, satellite, airborne, or ground measured (Hong et al., [16]a). Most of these multi-source datasets do not lie in the same feature space and do not follow independent identical distribution, thereby affecting the applicability of conventional machine learning algorithms (Chlaily et al., [6]). The data biases and domain discrepancies constitute a major obstacle in training predictive models across the domains ([45]; Rasti et al., [29]). The use of ground measured spectra, having a very high spectral resolution, as training data to improve the detection capability of multispectral imagery (e.g., UAV and satellite imagery), have not been well explored owing to the former being a point data and also due to the lack of proper correlation between both.

Deep machine learning (ML) approaches, which learn abstract representations to transform inputs to intrinsic manifolds in an unsupervised manner, have reported better results than the conventional ML approaches for various Earth observation (EO) data applications. Convolutional neural networks (CNNs) are supervised algorithms for deep ML and their numerous variants have been widely used for modeling the spatial and spectral features of remote sensing images (Ji et al., 2013; [2, 3]). Graph CNNs have provided a powerful means for graph-based semi-supervised tasks (Wang et al., [40, 42]; Bacciu et al., [4]; Luo et al., [25]). Capsule networks and sparse coding have facilitated the development of interpretable latent spaces with regard to the specific objectives of resolution enhancement, fusion and classification of EO datasets (Sabour et al., [32]; Arun et al., [1]). Although deep ML-based approaches have reported state-of-the-art results for EO data fusion, the requirement of a large amount of cross-domain training samples limits their practical applicability.

Inspired by the recent developments in deep ML, the current study attempted to resolve the issues prevalent in the fusion of multimodal EO datasets having significant differences in spatial and spectral resolutions. In this regard, deep ML approaches were explored to project the multi-source EO data to a shared latent space with minimum cross-domain correspondence. To the best of our knowledge based on the available literature, the proposed research problem of using generative and covariance guided transformations to fuse a point spectral data with multispectral image patch using minimal or no cross-domain samples have not been attempted so far. In addition, unlike the existing graph-based approaches, the proposed approach dynamically learns the graph considering the cross-modal similarity of the samples. The shared latent space projection and the use of latent graph generation and graph convolutions are the main characteristics of the proposed approach. The main contributions of the proposed approach can be summarized as:

- Labelled and unlabeled samples based generative cross-modal transformation and guided shared latent space projection without the requirement of cross-domain correspondence.
- Latent graph generation and graph convolution considering the cross-modality and data bias with a minimum number of training samples.
- Regularizations, loss functions and architectures for resolving the differences in the spatial and spectral resolution of the input training samples as well as the domain bias between the source and target domains.

2. Related works

2.1. Deep machine learning based fusion techniques

Although different deep ML-based approaches have been explored for the fusion of multi-modal EO datasets, the significant differences in spatial and spectral resolutions affected the spectral and spatial fidelity of the reconstructions (Loncan et al., [24]; Marcello et al., 2019; Vivone and Chanussot, [37]; Restaino et al., [31]; Deng et al., [9]). Recently, He et al. [13] proposed a spectrally predictive structure embedded in the network to address the spectral range gap for resolving the spectral distortion and spatial blurring prevalent in conventional approaches. Zheng et al. [50] employed deep ML hyperspectral prior along with several channel-spatial attention residual blocks to model informative features of spectral channels and spatial locations to boost the fusion accuracy. However, these approaches were highly sensitive to co-registration errors and required a large number of training samples. Rostami et al. (2019) used two deep encoders such that the empirical distribution discrepancy between the two domains was minimized in the shared output of the deep encoders. Based on the adaptive degrees, Xu et al. [44] employed a network to preserve the adaptive similarity between the fusion result and source images. Xie et al. [43] employed a 3D-generative adversarial network (GAN) for the fusion of hyperspectral and multispectral images. Ramirez et al. (2020) proposed an alternating direction method of multipliers (ADMM) for the fusion of multi-sensor features where sparsity and total variation (TV) regularization constraints were employed to improve the classification performance. Marcello et al. (2019) have evaluated prominent hyperspectral sharpening methods for simulated as well as real multi-source datasets with different spatial resolution ratios and registration errors. Dian et al. [11] proposed a nonlocal sparse tensor factorization approach for the semi-blind fusion of hyperspectral and multispectral images and the approach was blind with respect to the point spread function (PSF) and copes with spatially variant PSFs. Most of these prominent EO data fusion requires a sufficient number of cross-domain samples which may not be always available due to practical difficulties. In addition, the fusion of ground measured spectra, which is point data, with images have been least explored. Unlike the existing deep ML-based fusion techniques, the proposed approach projected the multi-source data to a shared latent space with minimal or even no cross-domain samples.

2.2. Deep machine learning based domain adaptation techniques

The domain adaptation (DA) approaches that address the domain imbalance has also been employed for addressing the resolution trade-off in the remote sensing domain. Recent DA algorithms learn domain invariant and agnostic features, by exploiting the intrinsic structure of the data, to improve the performance of cross-domain classifiers ([45]; Zini et al., [54]; He et al., [14]; Zhuang et al., [53]; Zhang and Zhang, 2016; [47]; Zhang et al., [46]). However, most of the prominent DA approaches depend on the reconstruction and transformation matrix resulting in a possible negative transfer effect (Zhang et al., [46]). Also, the one-stage formulation adopted by most of these approaches generally fails to get the optimal projection and does not work well if the domain disparity is large. In this regard, deep domain adaptation (DDA) approaches leverage deep networks to learn more transferable representations as compared to conventional approaches by embedding domain adaptation in the pipeline of DL (Wang and Deng, [38]). Among the various DDA approaches, some are based on building domain-invariant feature spaces through generative learning (such as autoencoders, adversarial training) while others are based on the analysis of higher-order statistics or self-ensembling methods based on implicit discrepancy (Perone et al., 2019). GANs have been widely used in

(Zhu et al., [52]; Hoffman et al., [15]; Sankaranarayanan et al., [33]) to remap the distribution from the source to the target dataset thereby learning aligned embedding for both domains. Graph convolution-based approaches have also been proposed to consider the source and target domain discrepancy in terms of edge adjacency matrices to formulate an effective strategy to address the source and target domain discrepancy [41]. Most of the DDA approaches only consider the domain adaptation at the data level and share a common projection matrix for both domains which affects the measure of the difference between the domain-specific subspaces [47]. In addition, sufficient training samples with cross-domain correlation is essential for most of the DDA approaches, and practical difficulty in obtaining the same limits their application for multi-source EO data fusion. The prominent domain adaptation techniques, such as the ones discussed, generally attempts to address the domain bias between training and testing samples and are inconsistent to multisource modalities having significant resolution differences. In this regard, the proposed approach employed a dynamic time wrapping (DTW) based graph generation and graph convolution approach which can be trained in an end-to-end manner. The use of DTW-based network layers has significantly addressed the cross-modal similarity measurement issues associated with the prominent DL-based DA approaches.

2.3. Deep machine learning based image translation techniques

The cross-modality synthesis or image translation approaches, which are generative ones across multiple domains, project the data into either coupled or common subspace to associate images between both domains (Sarfraz et al., [34]). Deep ML-based image translation techniques use CNN-derived high-level feature subspace to generate new images by seeking the matched feature representations close to the input one while providing a correlation map for emphasizing the domain information [12]. GAN-based approaches synthesize samples conditioned on either image attributions, textures, or class labels to achieve more realistic synthesis (Liu and Tuzel, [21], Reed et al., [30]; Wang and Gupta, [39]; Zhu et al., [51]). Recently, dual learning along with GANs are being explored to form a closed loop between dual domains to generate informative feedbacks to loosen the requirement of paired training samples (Zhu et al., [52]; Kim et al., [20]; Tang et al., 2019). Although the image translation approaches resolve the domain biases and address the cross-modality discrepancies, the specific characteristics of remote sensing images and the need to ensure spectral fidelity during translation affect their results for EO data, particularly when the resolution differences of the multiple sources are significant. The use of generative approaches and guided transformations, proposed in this study, along with spectra-specific losses and constraints, addressed this issue effectively even with limited cross-domain samples. The proposed latent graph generation and feature learning successfully modelled the semantic similarity across different modalities.

3. Description of datasets

3.1. Standard benchmark datasets

Details of the standard benchmark datasets used in this study can be referred from [7]. The AVIRIS sensor-acquired Indian Pines dataset has a spatial resolution of 20 m and 200 spectral bands, covering 16 vegetation and urban land cover classes. The Salinas dataset, acquired with AVIRIS sensor, covers 16 vegetation land cover classes and has a spatial resolution of 3.7 m and a spectral dimension of 200 bands. The KSC dataset contains AVIRIS sensor-acquired data having a spectral dimension of 224 bands and a spatial resolution of 18 m. For classification purposes, 13 classes representing the various land cover types were defined for the site.

The standard hyperspectral benchmark datasets such as Indian Pines, Salinas, and KSC were spectrally and spatially downscaled to simulate multispectral data and corresponding ground-measured spectral readings. The original high spectral resolution spectra simulated the ground measured spectra while the spectrally downscaled version simulated the UAV spectra. Different downscaling strategies, such as bilinear, bi-cubic, and nearest-neighbor interpolation were employed to generate training and testing patches. Multiple downscaling strategies were adopted to avoid the bias of the trained network towards a particular approach. Spectra-specific augmentation techniques ([48]; Haut et al., Nalepa et al., [28]) were also employed to increase the number of training and testing samples. The simulated multispectral and hyperspectral spectra having dimensions of 1×15 and 1×120, respectively, were used to train the proposed models. Among the simulated and augmented spectra, generated from the standard datasets, 7500 samples were employed for training and testing the different frameworks discussed in this study.

3.2. Multispectral UAV and ground measured hyperspectral datasets

The multispectral data were collected over chickpea (*Cicer arietinum*) experimental plots in two sites located in the northwest Negev, Israel: the Gilat Research Center (31°20′N, 34°40′E) and Kibbutz Or-Haner (31°33′N; 34°35′E), hereafter defined as Plot-1 and Plot-2, respectively. The experiment was conducted during the 2019 growing season (January-May) and focused on assessing various plant traits and the outcome seed yield in response to five irrigation regimes. In Plot-1, each replicate was 3×25 m while in Plot-2, each replicate ranged between 15×15 m to 30×30 m. The multispectral data of Plot-1 were acquired by a MicaSense RE (MicaSense, Inc., Seattle, WA, USA) five-band camera (blue, green, red, red-edge, and near-infrared) mounted on a Tarot T-960 (FoxTech, Tianjin, China) UAV. The RGB images of Plot-2 were obtained by a Sony A5100 (Sony Corporation, Tokyo, Japan) camera mounted on a Phantom DJI 4 (SZ DJI Technology Co. Ltd., Shenzhen, China). Both UAVs acquired images at an altitude of 100 m above the ground, and the mosaics resulted in a pixel size of 7 cm and 3 cm for the multispectral and RGB images, respectively.

The hyperspectral reflectance of the chickpea canopy was obtained at ground level by an ASD FieldSpec4HR spectroradiometer (ASD Inc., Longmont, CO, USA). Spectral data were acquired in the range of 350–2500 nm, full width at half maximum (FWHM) of 3 nm in the range of 350–1000 nm, and 8 nm in the range of 1000–2500 nm. The bare fiber field of view was 25° and it was located ~1.5 m above ground level. The ground measured spectra obtained were resampled using bicubic interpolation to 128 spectral bands.

The multispectral UAV patches and corresponding ground measured hyperspectral spectra were employed to analyze the effectiveness of the proposed approach in generating a more discriminative high-dimensional feature space from the multispectral patches. Differential Global Positioning System (DGPS)-based coordinates of the ground measured spectra and multispectral images were adopted for the geo-registration of both the datasets. Ground control points (GCPs) marked by iron stakes were placed on the borders and inside the experimental field for the entire growing season, and geolocated using a Topcon GRS1 (Topcon Positioning Systems Inc., Livermore, CA, USA) real-time kinematic (RTK) global navigation satellite system (GNSS). Plates with crosses were placed on the ground with each of the iron stakes in their center for each of the imaging dates. Each multispectral band of the UAV dataset was separately geo-corrected by the GCPs in ERDAS Imagine 2018 (Hexagon Geospatial, Norcross, GA, USA) with an accuracy of root mean square error (RMSE) less than 0.15. The RGB images, of the same area and date, were processed to orthophoto mosaics in a Pix4D mapper software (Pix4D SA, Lausanne, Switzerland) environment. A shapefile, with one polygon for each plot delineated from an RGB orthophoto was also used to geo-rectify the images. To avoid the co-registration errors, multispectral UAV patches of dimension $5 \times 5 \times 5$ were mapped to the corresponding (resampled) ground measured spectra of dimension 1×128. The UAV and ground measured spectra samples were collected away from the boundaries. In addition to the collected multispectral and

ground measured spectra, the samples generated through augmentation techniques ([48]; Haut et al., Nalepa et al., [28]) were also employed to train and test the networks. A total of 6300 samples were employed for analyzing the different frameworks discussed in this study.

3.3. Airborne hyperspectral dataset

The hyperspectral data collection was conducted in July and August 2015, at three experimental plots: fully irrigated control, highly regulated deficit-irrigation, regime and rainfed. The data were collected using a micro-hyperspectral imager (Micro-Hyperspec VNIR model, Headwall Photonics, Fitchburg, MA, USA) set in tandem on board a Cessna aircraft operated at 200 m altitudes. The Micro-Hyperspec VNIR was set up with a configuration of 260 spectral bands acquired at 1.85 nm/pixel and 12-bit radiometric resolution in the 400–885 nm spectral region, yielding a 6.4 nm FWHM with a 25 μm slit. The hyperspectral imagery was atmospherically corrected using the irradiance measured during the flight by an ASD Field Spectrometer (FieldSpec Handheld Pro, ASD Inc., Longmont, CO, USA) with 3 nm bandwidth and a cosine corrector-diffuser probe.

The 260-band airborne hyperspectral imagery was used to analyze the effectiveness of the proposed approaches, especially the graph embedding and convolutional graph-based classification, in classifying the real-world data with a minimum number of training samples. The co-registration approaches discussed in Section 3.2 were employed to register the ground-measured spectra with the airborne hyperspectral imagery. The collected hyperspectral samples along with the ones generated through augmentation, numbering 4560, were employed to train and test the different frameworks with regard to hyperspectral classification.

4. Proposed method

Let $x^M \in R^{m \times n \times s}$ be a multispectral image having $p \times q$ pixel vectors, each denoted as $x_{ij} \in R^s$. Let there be p ground-measured spectral readings ($p \ll m*n$) of the same area, each having a dimension of h, denoted as $x^G{}_i \in R^h$. The proposed frameworks classify each $x^M{}_{ij}$ based on the prior information derived from the mapping between several available ground-measured spectra and corresponding pixel spectra. It may be noted that x^G was used only for training. We investigated to reduce the number of cross-domain samples for transformation to a cross-domain latent space (z) which is more separable or discriminative than R^s. In addition, to further resolve the issues of domain shift, a latent graph generator-based classifier was proposed to use both labelled and unlabeled samples for prediction. A summary of the workflow of the proposed approach is depicted in Fig. 1. For further clarity, a block diagram depicting the sub-module-based pseudocodes is presented in Fig. 2.A detailed description of each block of Figs. 1 and 2 is presented in the following sub-sections.

4.1. Cross-modal generation

The multi-source datasets need to be transformed to a shared latent space such that the projected latent representation should have better class separability as compared to the source feature spaces. The cross-modal generative architecture is presented in Fig. 3(a). The details of the generators for ground hyperspectral spectra and UAV spectra are presented in Fig. 3(b) and (c), respectively. It may be noted that the generators were cross-connected encoder-decoder networks coupled with a feature mapping layer in between the encoder and decoder streams. The discriminator distinguished the domains of the ground measured spectra and the learned latent representation to make the manifold more discriminative. It may be noted that the framework significantly reduced the requirement of cross-domain training samples as the multi-spectral domain was mapped adversarially to a more discriminative manifold.

The proposed cross-modal generation framework adopted a cyclic adversarial encoding approach to learn a shared latent space. The

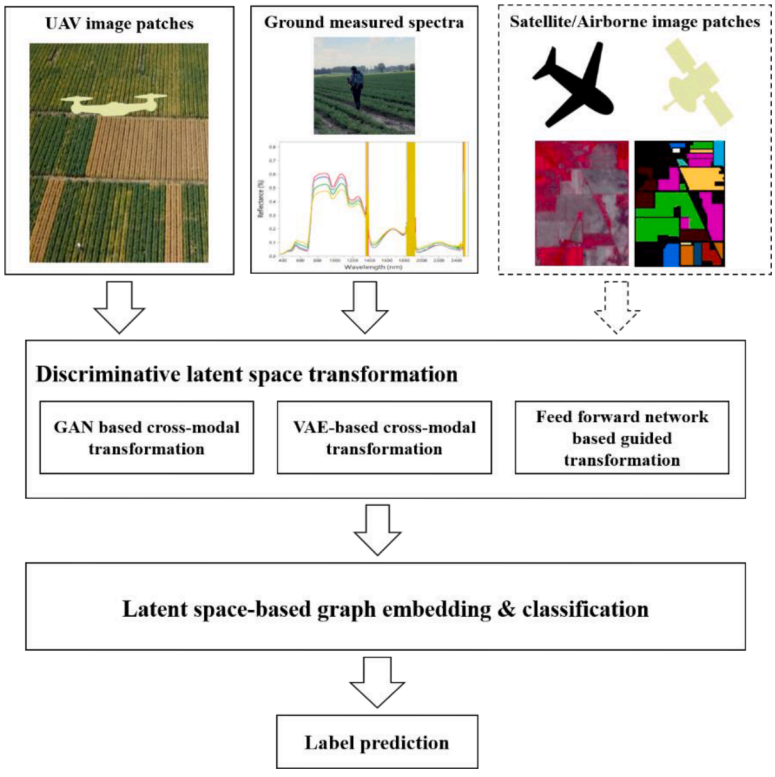

Fig. 1. Schematic diagram of the proposed approach (dotted box represents optional source). (VAE stands for variational autoencoder; and GAN stands for generative adversarial network.).

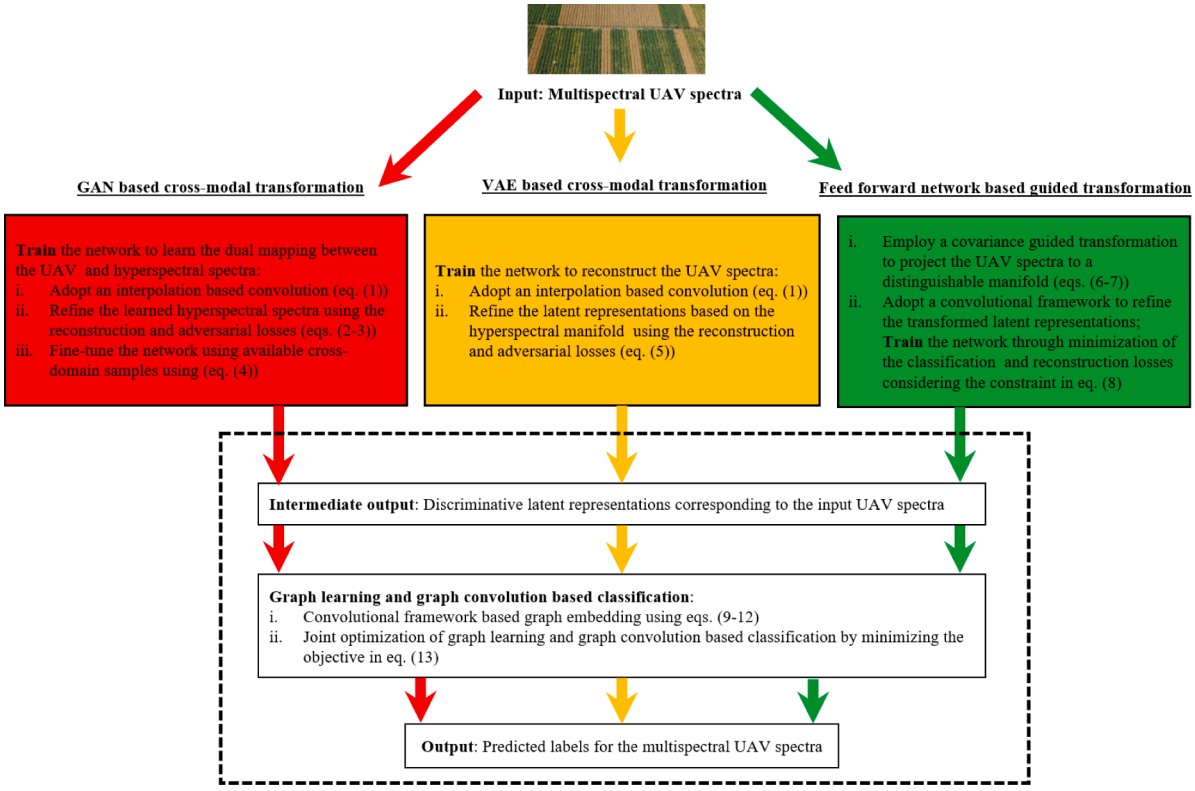

Fig. 2. Detailed pseudocode-based block diagram of the proposed framework. Different colored arrows indicate mutually exclusive computing paths either one of which can be adopted; white blocks in the dashed frame are used by each of the computing paths (VAE stands for variational autoencoder; UAV stands for unmanned aerial vehicle; and GAN stands for generative adversarial network.).

generative network facilitated the projection of the multispectral spectra to a higher dimensional space. An interpolation-based convolution was adopted to effectively capture the features of the ground spectra and to incorporate the shift or scale effects. The interpolated convolution centered at the location $\tilde{p}$ of the ground measured spectra (x^G) was implemented as:

$$x^G * \kappa(\tilde{p}) = \sum_{p'} \frac{1}{N_{p'}} \sum_{p_\delta} \varphi(p_\delta, p')x^G(\tilde{p}+p_\delta).\omega(p') \quad (1)$$

where κ is the kernel composed of kernel weights ω. Each kernel weight vector $\omega(p')$ has a coordinate location p' relative to the kernel center, and its weight is initialized and updated during training. The vector coordinate p' can either be fixed or updated during training. The interpolation function $\varphi(p_\delta, p') = e^{\frac{\|p_\delta - p'\|^2}{\sigma^2}}$ takes the coordinate p′ of a kernel weight vector ω(p') and the coordinate p_δ of a neighboring input point and computes a weight by the Gaussian interpolation algorithm [2]. The hyperparameter σ of the interpolation function controls the decay rate. To make convolutions sparsity invariant, a density normalization term Np′, which sums the interpolation weights or number of input points in the neighborhood of p′, is employed for each kernel weight vector $\omega(p')$.

The adversarial losses of both the mapping functions, i.e., projection from the ground measured spectra (x^G) to multispectral ones (x^M) and vice versa (x^M to x^G), were jointly expressed for obtaining the shared latent space. The jointly adapted bi-directional loss is defined as:

$$L_d(D_F, G, F) = E_{x^G \sim pr(x^G)}\left[\log D_F\left(x^G\right)\right] + E_{x^M \sim pr(x^M)}\left[\log\left(1 - D_F\left(F\left(x^M\right)\right)\right)\right] + E_{x^M \sim pr(x^M)}\|x^M - F\left(G\left(x^M\right)\right)\|_2 \quad (2)$$

where x^G and x^F respectively denote ground hyperspectral spectra and the UAV spectra respectively. The functions G: $x^G \rightarrow x^M$ and F: $x^M \rightarrow x^G$ are respectively the dual mappings and D_F denotes the discriminator for distinguishing real and reconstructed UAV spectra. To enforce the cross-domain generations to have better class separability, the generations were fed to the classifiers C_G and C_F respectively. Hence, the discriminator was conditioned on classification and the responsibility of the discriminator was to guide the synthesis such that the classes were separable. Similar to the approach of Huang et al. [19], the pseudo inputs inferred in the unsupervised dual learning problem was used to enforce a cycle-consistency constraint. In addition, the unbiased estimation of multi-kernel maximum mean discrepancy (MK-MMD) was employed to reduce the domain bias. Hence the loss function for the network was formulated as:

$$L_c\left(x^G, G, x^M, F\right) = L_d(D_F, G, F) + \beta\left(E_{Pdata\left(F\left(x^M\right)\right)}\delta\left(A^G\right) - E_{Pdata\left(G\left(x^G\right)\right)}\delta\left(A^M\right)\right) \quad (3)$$

where x^G and x^F respectively denote ground hyperspectral spectra and the UAV spectra respectively, A is the set of positive definite kernels, and δ is the non-linear mapping. The functions G: $x^G \rightarrow x^M$ and F: $x^M \rightarrow x^G$ are respectively the dual mappings and D_F denotes the discriminator for distinguishing the real and reconstructed UAV spectra. It may be noted that the last term in Eq. (3) matches all orders of statistics between x^M and x^G. The loss function in Eq. (3) was used to train the network without the need for cross-domain sample correspondence. For further

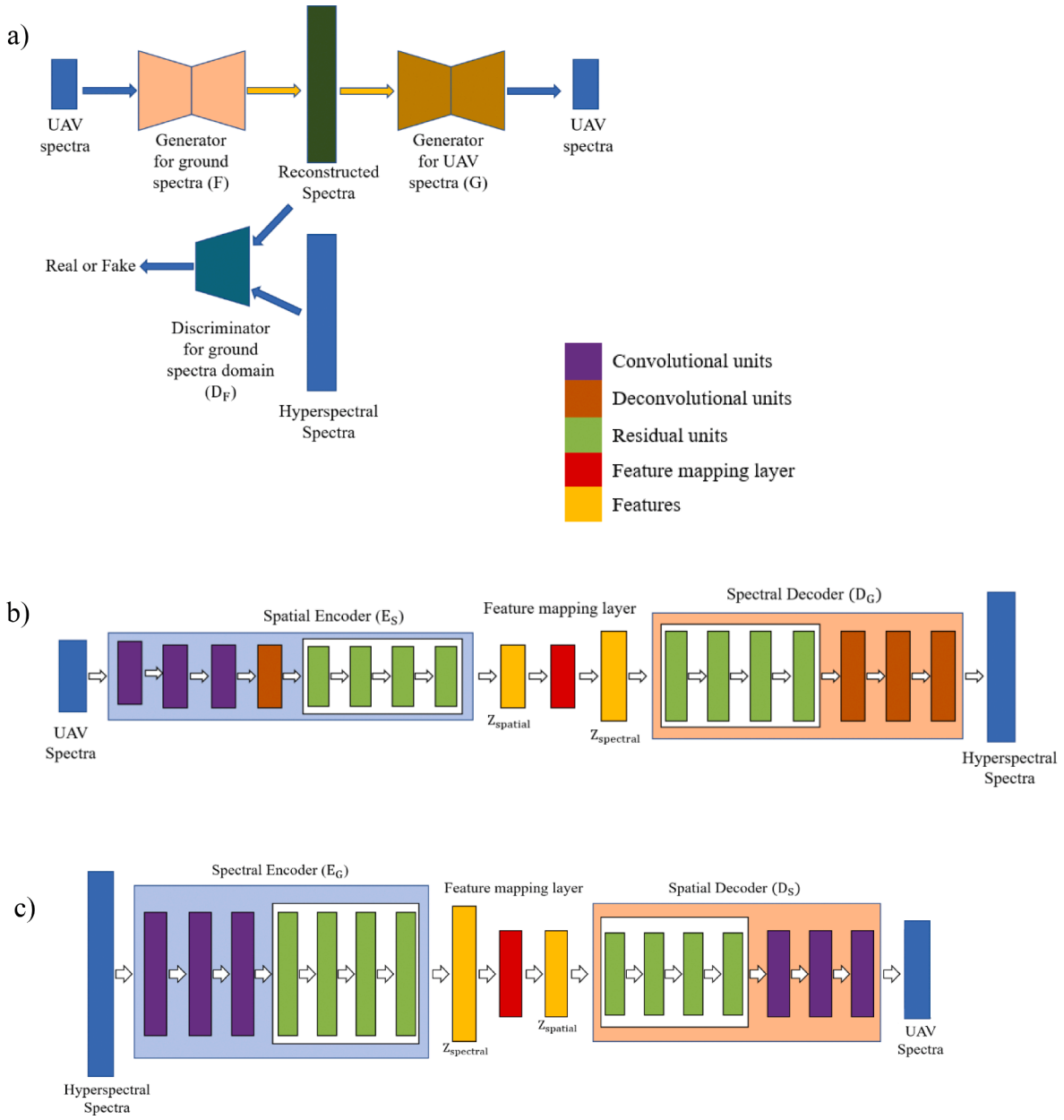

Fig. 3. Proposed GAN-based shared latent space projection (a) GAN-based framework; (b) Generator for Hyperspectral ground spectra; and (c) Generator for multispectral UAV spectra.

training the network with available paired cross-domain samples, the reconstruction of the ground spectra was constrained to be similar to the corresponding input hyperspectral spectra as:

$$E_{x^G \sim pr(x^G)} \| x^G - F\left(G\left(x^G\right)\right) \|_1 \tag{4}$$

where x^G and x^F respectively denote ground hyperspectral spectra and the UAV spectra, and the function $||.||_1$ is the L1 distance to quantitatively compare the input data and the reconstructed pseudo. The functions G: $x^G \rightarrow x^M$ and F: $x^M \rightarrow x^G$ denote the dual mappings. It may be noted that the discriminator network (Fig. 1(a)) was pre-trained using unlabeled hyperspectral samples and did not require any cross-domain correspondence.

4.2. Cross-modal variational encoding

Variational autoencoder (VAE) can also be employed to transform the data, having spatial spectral resolution difference, to shared latent space. An adversarial encoding architecture as presented in Fig. 4 was adopted to transform the multispectral inputs to a high-dimensional latent space based on the available training samples. It is worth pointing that the training samples did not need to have cross-domain correspondence as the approach learned to map the multi-spectral domain adversarially to a more discriminative manifold.

In the proposed approach, the latent representations (z) were considered as a combination of modality specific (s) and shared modality-independent space (c). To tractably maximize the marginal likelihood of the data, the true unknown posterior was approximated by a variational posterior which allowed optimizing an evidence lower bound (Shi et al., 2019) through stochastic gradient descent. The joint variational posterior was computed as a combination of unimodal posteriors using a mixture of experts as:

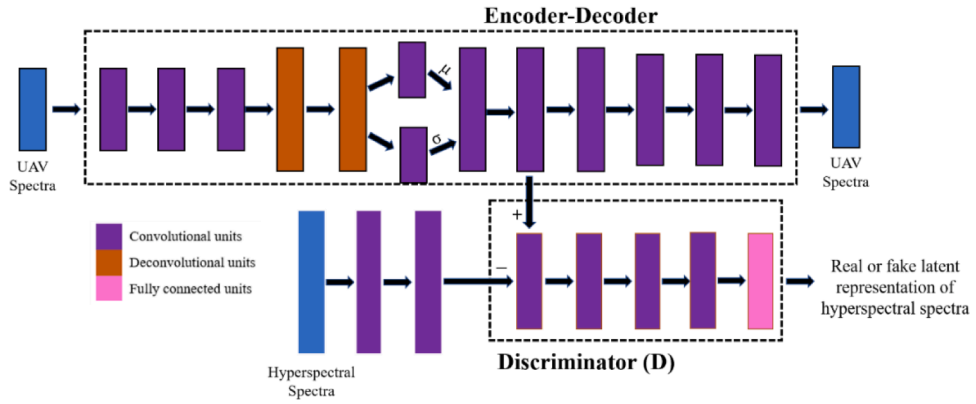

Fig. 4. Proposed variational encoding framework for cross-modal projection.

$$L_{VAE}(\Theta,\ \varphi;x)=\sum_{j=1}^{M}E_{q_{\varphi_c}(c|x)}\left[E_{q_{\varphi_c}(s_j|x_j)}\left[\log p_\Theta(x_j|s_j,c)\right]\right]-\sum_{j=1}^{M}D_{KL}\left(q_{\varphi_{s_j}}(s_j|x_j)\|p_\Theta(s_j)\right)-\Gamma_\pi^{M+1}\left(\left\{q_{\varphi_{c_j}}(c|x_j)\right\}_{j=1}^{M},p_\Theta(c)\right) \tag{5}$$

where Γ is the JS (Jensen–Shannon)-divergence as discussed in (Sutter et al., 2019), x denotes the multimodal input data ($x \subseteqq x^M \cup x^G$), π denote the distribution weights ($\Sigma\pi_i = 1$), D_{KL} denotes the Kullback–Leibler (KL) divergence, E(.) denotes the expectation, and M denotes the modality. It may be noted that the JS-divergence was used only for the multimodal latent factors c while modality-independent terms s_j were part of KL-divergence measures. Also, the variational approximation functions $q_{\varphi_{c_j}}(c_j|x_j)$ and $q_{\varphi_{s_j}}(s_j|x_j)$, and the generative model $p_\Theta(x_j|s_j,\ c)$ were implemented as neural network encoders. Similar to cross-modal generation using GAN, the VAE based approach also employed interpolated convolution for encoding the ground spectra.

4.3. Feed forward network-based guided transformation

The shared latent space projection, using GAN and VAE, required enough labelled samples without requiring them to the cross-correlated. In this regard, an alternate strategy was proposed to reduce the requirement of labelled samples. The approach adopted the DTW measure (Baumann et al., 2017) to estimate the cross-modal similarity of the samples without using very deep networks. The proposed latent space transformation was guided using the prior information regarding the labels of the multimodal samples. In other words, transformed space facilitated the intra cluster similarity and inter-cluster differences of the labelled samples. In this regard, a covariance guided projection was employed for the source and target projection as:

$$P_M=\max_{P_M}\left(Tr\left(P_M S_M P_M^T\right)\right) \tag{6}$$

$$P_G=\max_{P_G}\left(Tr\left(P_G S_G P_G^T\right)\right) \tag{7}$$

where S_M and S_G are respectively the interclass variance matrices of the multispectral and hyperspectral domains, Tr(.) denotes the matrix trace, and P_M and P_G are the projection matrices of the multispectral and hyperspectral domains respectively. The latent representation (z) obtained using P_M and P_G are transformed using a feed-forward network (having network weight β) with an additional constraint on the network weights as:

$$L_c=Tr\left(\beta^T K\beta\right)+Tr\left(\beta^T KMK\beta\right) \tag{8}$$

where K is the kernel matrix and M is the indicator matrix (discussed in Pan et al. (2011) and Zhao et al. [49]). The constraint L_c used classification prior to further reduce the modality and domain bias in the transformed space. It is noteworthy that the guided transformation required only a limited number of training samples as the approach did not employ any complex adversarial networks.

4.4. Graph embedding and label prediction

The latent representations obtained from multimodal inputs (using either of the approaches discussed in Sections 4.1, 4.2 or 4.3) were used to dynamically construct a graph. Given the latent representations $Z = (z_1, z_2 \bullet\bullet\bullet z_n) \in R^{n\times p}$, a feedforward network was employed to learn a function $S_{ij} = g(z_i, z_j)$ that represented the pairwise relationship between data z_i and z_j. The neural network layer is parameterized by the weight vector $a = (a_1, a_2, \bullet\bullet\bullet a_p)^T \in R^{p\times 1}$. The proposed framework for joint optimization of graph embedding and label prediction is shown in Fig. 5.

To make the prediction of the discrete graph structure differentiable and to train the graph learning end-to-end, a weighted adjacency matrix with probabilistic similarity measure was employed. For considering the multimodal nature of the training samples, the edge probability between two nodes p_i and p_j was computed in terms of the soft DTW similarity measure (Xingyu et al., 2019). In this regard, the graph embedding layer was employed to learn the probabilistic graph matrix S as:

$$S_{ij}=g\left(p_i,p_j\right)=\frac{e^{ReLU\left(a^T DTW\left(p_i,p_j\right)\right)}}{\sum_{j=1}^{n}e^{ReLU\left(a^T DTW\left(p_i,p_j\right)\right)}} \tag{9}$$

where p_i and p_j are respectively the projected latent representations, n is the total number of data vectors, DTW(.,.) is the dynamic time wrapping function, and ReLU(•) = max(0,•) is an activation function which guaranteed that S_{ij} was positive. In addition, to projecting the latent representations to a lower-dimensional manifold, the feed-forward layer preceding the graph embedding layer was pre-trained to minimize the maximum mean discrepancy (Ω(., .)) between the learned source and target feature representations (p^s and p^t) respectively to improve the generalizability and resolve the issue of domain discrepancy. In the current study, Ω(p^s, p^t) between the learned source and target representations was formulated as:

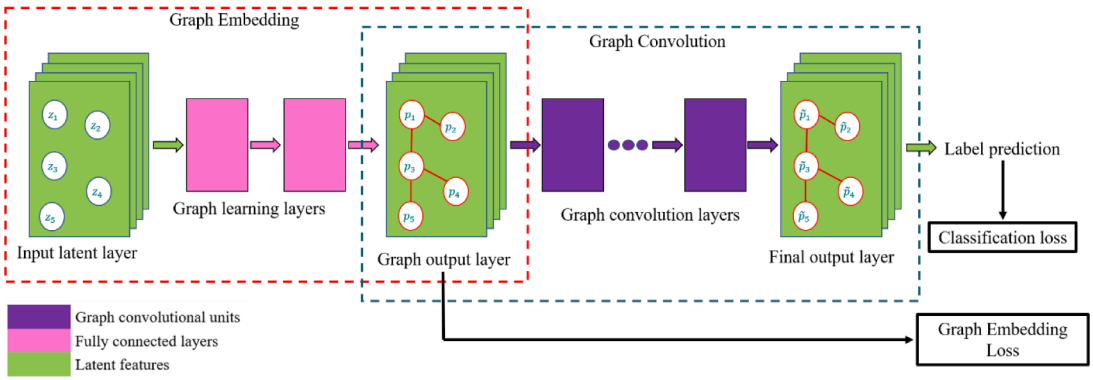

Fig. 5. Proposed graph embedding and graph convolutional framework.

$$\Omega(p_s,\ p_t) = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}K\left(p_i^s,p_j^s\right) + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}K\left(p_i^t,p_j^t\right) - \frac{2}{n_s n_t}\sum_{i=1}^{n_t}\sum_{j=1}^{n_s}K\left(p_i^s,p_j^t\right) \tag{10}$$

where n_s and n_t are respectively the number of sources and target samples, and K(.,.) is a kernel similarity measure

The learned graph S was fed to the sequence of graph convolutional layers where the k^{th} GC layer was defined as:

$$P^{k+1} = \sigma\left(D^{-\frac{1}{2}}SD^{-\frac{1}{2}}P^kW^k\right) \tag{11}$$

where D is a normalization matrix with $d_i = \sum_{j=1}^{n} S_{ij}$, P^k denotes the k^{th} layer features, σ(•) is the activation function, and W^k represents the model filters to be learnt. The GC layers were followed by a fully connected SoftMax layer where the output was formulated as:

$$O = SoftMax\left(D^{-\frac{1}{2}}SD^{-\frac{1}{2}}P^fW^f\right) \tag{12}$$

where D is a normalization matrix with $d_i = \sum_{j=1}^{n} S_{ij}$, P^k denotes the final layer features, W^f denotes the weight matrix, and final output $O \in R^{n\times c}$ denotes the label prediction for all the n data vectors. It may be noted that the entire model (latent-graph learning and graph convolution classifier) was trained in an end-to-end manner backpropagating directly through the graph adjacency. The loss function for jointly optimizing the graph learning and graph convolution networks was modelled as:

$$L_{graph} = \sum_{i\in L}\sum_{j=1}^{c}Y_{ij}\ln O_{ij} + \sum_{i,j=1}^{L}\|p_i - p_j\|_2^2 S_{ij} + \Upsilon\ \|\ S\|_F \tag{13}$$

where L indicates the set of labelled nodes, c is the number of classes, Y_{ij} denotes the label of the i^{th} sample for the j^{th} class, and O is the output of the final perceptron layer, S denotes the graph matrix, ϒ denotes the scaling factor, $\|\ .\|_2$ denotes the L_2 norm, $\|\ .\|_F$ is the Frobenius norm, and p_i and p_j respectively denote the projected latent vector representations of the i^{th} and j^{th} data points.

5. Experimental setup

The proposed approach experimented on different simulated and real datasets. The architectures, as well as hyper parameters, were fine-tuned in accordance with the datasets. The labelled and unlabeled samples were used for training the GAN- and VAE-based discriminative latent space transformations as well as the guided transformation-based approach. The transformed labelled cross-modal samples were used for training the graph generation and convolution.

5.1. Fusion of simulated multispectral and hyperspectral dataset

The implementation of the proposed frameworks for simulated datasets is summarized in Table 1. The optimal configuration of each of the proposed latent space transformations (GAN- based, VAE-based and feed forward network-based guided transformation) and graph-embedding were estimated through hyper-parameter optimization proposed in Bochinski et al. [5]. An early stopping framework using k-fold validation formed the basis of the parameter selection. The general convolutional layers were followed by spatial batch normalizations and parametric ReLU (PReLU) nonlinearity while the output layer applied the sigmoid activation. Besides, the convolutional layers adopted a stride of one in all cases except for downscaling where the stride was set to two.

The implementation of the proposed GAN-based cross-modal transformation (Section 4.1.) for simulated datasets constituted of bi-directional convolutional generators and a CNN-based discriminator as shown in Table 1. The generators were implemented as consisting of a 14-layer encoder-decoder network separated by a feature mapping layer. The transformation of multispectral UAV samples to hyperspectral spectra was implemented using a sequence of convolutional, deconvolutional and residual units. The encoder-decoder network, employed for the transformation of hyperspectral spectra to multispectral spectra, constituted of a series of convolutional and residual units. The residual blocks had 2 convolutional layers each followed by batch norm and PReLU activation layers. For the discriminative network, a stack of convolutional layers followed by a fully connected network were employed. The network was trained for 300 epochs with an initial learning rate of 0.01 and a decay rate 0.8 every 100 epochs with a batch size 100.

The variational approach, discussed in Section 4.2, adopted a joint encoding decoding network, as summarized in Table 1, to transform the multispectral image patches to a high dimensional latent space through adversarially matching the generated manifolds with that of the ground-measured spectra. The proposed implementation for simulated datasets constituted of 2D convolutional units followed by a sequence of 1D

Table 1
Implementation details of the proposed latent space transformations for the fusion of simulated datasets.

Network Steam	Network Configuration					
GAN-based latent space transformation	Stream	Encoder				
	Generator for ground spectra samples	Layer	Type	# Features In	# Features Out	Kernel Size
		1	Conv2D	5 × 5 × 15	1 × 120	5 × 5
		2	Conv1D	1 × 120	256 × 60	1 × 5
		3	Conv1D	256 × 60	512 × 60	1 × 5
		4	DeConv1D	512 × 60	256 × 120	1 × 5
		5	Residual units	256 × 120	256 × 120	1 × 5
		6	Residual units	256 × 120	256 × 120	1 × 5
		7	Residual units	256 × 120	256 × 120	1 × 5
		8	Residual units	256 × 120	256 × 60	1 × 5
		Feature mapping Layer				
		1	DeConv1D	256 × 60	128 × 120	1 × 5
		2	Conv1D	128 × 120	128 × 120	1 × 5
		Decoder				
		Layer	Type	# Features In	# Features Out	Kernel size
		1	Residual units	128 × 120	128 × 120	1 × 5
		2	Residual units	128 × 120	128 × 120	1 × 5
		3	Residual units	128 × 120	128 × 120	1 × 5
		4	Residual units	128 × 120	128 × 60	1 × 5
		5	DeConv1D	128 × 60	64 × 120	1 × 5
		6	DeConv1D	64 × 120	16 × 240	1 × 5
		7	Conv1D	16 × 240	1 × 120	1 × 5
	Generator for UAV spectra samples	Encoder				
		Layer	Type	# Features In	# Features Out	Kernel Size
		1	Conv1D	1 × 120	64 × 120	1 × 5
		2	Conv1D	64 × 120	128 × 120	1 × 5
		3	Conv1D	128 × 120	256 × 120	1 × 5
		4	Residual units	256 × 120	256 × 120	1 × 5
		5	Residual units	256 × 120	256 × 120	1 × 5
		6	Residual units	256 × 120	256 × 120	1 × 5
		7	Residual units	256 × 120	256 × 120	1 × 5
		Feature mapping Layer				
		1	Conv1D	256 × 120	256 × 60	1 × 5
		2	Conv1D	256 × 60	128 × 60	1 × 5
		Decoder				
		Layer	Type	# Features In	# Features Out	Kernel size
		1	Residual units	128 × 60	128 × 60	1 × 5
		2	Residual units	128 × 60	128 × 60	1 × 5
		3	Residual units	128 × 60	128 × 60	1 × 5
		4	Residual units	128 × 60	64 × 60	1 × 5
		5	Conv1D	64 × 60	32 × 30	1 × 5
		6	Conv1D	32 × 30	16 × 15	1 × 5
		7	Conv1D	16 × 15	1 × 15	1 × 5
	Discriminator	1	Conv1D	1 × 120	128 × 120	1 × 5
		2	Conv1D	128 × 120	256 × 60	1 × 5
		3	Conv1D	256 × 60	64 × 30	1 × 5
		4	Conv1D	64 × 30	32 × 30	1 × 5
		5	Fully connected	–	–	–
VAE-based latent space transformation		Encoder-Decoder				
		Layer	Type	# Features In	# Features Out	Kernel Size
		1	Conv2D	5 × 5 × 15	1 × 120	5 × 5
		2	Conv1D	1 × 120	128 × 120	1 × 5
		3	Conv1D	128 × 120	256 × 120	1 × 5
		4	DeConv1D	256 × 120	128 × 240	1 × 5
		5	DeConv1D	128 × 240	128 × 240	1 × 5
		6	Conv1D	128 × 120	256 × 120	1 × 5
		7	Conv1D	256 × 120	256 × 60	1 × 5
		8	Conv1D	256 × 60	128 × 30	1 × 5
		9	Conv1D	128 × 60	64 × 30	1 × 5
		10	Conv1D	64 × 30	16 × 15	1 × 5
		11	Conv1D	16 × 15	1 × 15	1 × 5
		Discriminator				
		Layer	Type	# Features In	# Features Out	Kernel size
		1	Conv1D	256 × 60	128 × 60	1 × 5
		2	Conv1D	128 × 60	64 × 30	1 × 5

(*continued on next page*)

Table 1 (*continued*)

Network Steam	Network Configuration				
	3	Conv1D	64 × 30	32 × 15	1 × 5
	4	Conv1D	32 × 15	16 × 15	1 × 5
	5	Fully connected	16 × 15	–	–
Feed forward network-based guided transformation	1	Conv1D	1 × 30	256 × 30	1 × 5
	2	DeConv1D	256 × 30	256 × 30	1 × 5
	3	Conv1D	256 × 30	256 × 60	1 × 5
	4	Conv1D	256 × 60	128 × 60	1 × 5
	5	Conv1D	128 × 60	64 × 60	1 × 5
	6	Conv1D	64 × 60	32 × 60	1 × 5
	7	Fully connected	–	–	–

Table 2
Implementation details of the proposed latent graph generation and convolution module for the fusion of simulated datasets.

Network Steam	Network Configuration				
Graph embedding stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Fully connected network layer	–	–	–
	2	Fully connected network layer	–	–	–
	3	Fully connected network layer	–	–	–
Graph convolution stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Graph Convolution	32 × 32 × 64	32 × 32 × 16	5 × 5
	2	Graph Convolution	32 × 32 × 16	16 × 16 × 16	5 × 5
	3	Graph Convolution	16 × 16 × 16	16 × 16 × 8	5 × 5
	4	Graph Convolution	16 × 16 × 8	8 × 8 × 4	5 × 5
	5	Fully connected network	8 × 8 × 4	–	–

convolutions and deconvolutions. The discriminator for adversarially refining the latent space constituted of a sequence of 1D convolutions followed by a fully connected feed-forward network. The optimal number of epochs, learning rate, decay rate and batch size were respectively set to 200, 0.01, 0.8 and 100.

The guided transformation (Section 4.3) used a 6-layered convolutional network followed by a fully connected stream to map the UAV-based latent space to the ground-measured spectra-based latent manifold. The PReLU was used as the activation function except for the output layers which used a sigmoid activation. The network was trained for 250 epochs with an initial learning rate of 0.01 and a decay rate 0.5 every 100 epochs with a batch size 50.

The graph generation and graph convolutional layers, employed to process the latent representations (as discussed in Section 4.4), were respectively set to have 3 and 5 layers. The projection, graph generation and graph convolution layers were trained in an end-to-end manner. The configuration of the adopted model for the simulated datasets is summarized in Table 2. For the graph convolution layers, a 5-neighbor approach was adopted. The model was trained for 300 epochs optimizing the loss using Adam Optimizer with a learning rate of 0.1 reduced to 0.001 at the intervals of 100 epochs in a piecewise constant fashion.

5.2. Fusion of multispectral UAV data with ground measured hyperspectral spectra

The implementation of the proposed GAN- and VAE-based frameworks for the fusion of multispectral UAV data with ground measured hyperspectral is summarized in Table 2. The implementation details and architectures are almost similar to the ones adopted for the fusion of the simulated data (Tables 3 and 5).

The proposed feed-forward network-based guided transformation (Section 4.3) used an 8-layered feed-forward network to map the UAV-based latent space to the ground-measured spectra-based latent manifold. The PReLU was used as the activation function expect for the output layers which used a sigmoid activation. The network was trained for 300 epochs with an initial learning rate of 0.01 and a decay rate 0.5 every 100 epochs with a batch size 100.

The depth of the graph generation and graph convolutional layers, employed to process the latent representations (as discussed in Section 4.4), were respectively set to 3 and 6. The details of the adopted network configuration is presented in Table 4. A 7-neighbor approach was adopted for implementing the graph convolution stream. The model was trained in an end-to-end manner for 200 epochs using Adam Optimizer with a learning rate of 0.1 reduced to 0.001 at the intervals of 100 epochs in a piecewise constant fashion.

5.3. Classification of airborne hyperspectral data

The airborne datasets were mainly used to evaluate the effectiveness of the proposed approach in addressing domain bias. The latent graph generation (discussed in Section 4.4) was proposed to address the domain bias when the training samples were scarce and were of different distribution. In this regard, a few labelled samples ($\leq$ 5%) were employed for training. The graph generation and graph convolutional layers, employed to process the latent representations, were set to have 4 and 6 layers respectively. The graph embedding and graph convolutional layers were trained in an end-to-end manner. A 5-neighbor approach was adopted with a learning rate of 0.1 reduced to 0.001 at the intervals of 100 epochs in a piecewise constant fashion. The model was trained for 200 epochs optimizing the loss using Adam Optimizer.

6. Results and discussion

To verify the effectiveness of the proposed methods, extensive experiments were conducted on the standard benchmark as well as real-world datasets. The Sections 6.1.1, 6.2.1 and 6.3.1 present detailed parameter analyses of the proposed approach for different datasets, Sections 6.1.2, 6.2.2 and 6.3.2 present a detailed analysis of the proposed constraints and regularizations over different datasets, and comparisons with some baseline methods for each dataset are discussed in Sections 6.1.3, 6.2.3 and 6.3.3.

The proposed approach was analyzed based on their effectiveness in classifying the spectrally coarse multispectral images using the prior

Table 3
Implementation details of the proposed latent space transformations for the fusion of the ground-measured hyperspectral spectra and UAV multispectral spectra.

Latent space transformation	Network Configuration				
GAN-based latent space transformation	Stream	Encoder			
	Generator for ground spectra samples				
	Layer	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Conv2D	5 × 5 × 5	1 × 256	5 × 5
	2	Conv1D	1 × 256	256 × 128	1 × 5
	3	Conv1D	256 × 128	512 × 64	1 × 5
	4	DeConv1D	512 × 64	256 × 128	1 × 5
	5	DeConv1D	256 × 128	128 × 256	1 × 5
	6	Residual units	128 × 256	128 × 256	1 × 5
	7	Residual units	128 × 256	128 × 256	1 × 5
	8	Residual units	128 × 128	128 × 64	1 × 5
	9	Residual units	128 × 64	256 × 64	5 × 5
	Feature mapping				
	Layer				
	1	DeConv1D	256 × 64	128 × 128	1 × 5
	2	DeConv1D	128 × 128	64 × 256	1 × 5
	3	Conv1D	64 × 256	64 × 128	1 × 5
	Decoder				
	Layer	**Type**	**# Features In**	**# Features Out**	**Kernel size**
	1	Residual units	64 × 128	64 × 128	1 × 5
	2	Residual units	64 × 128	64 × 128	1 × 5
	3	Residual units	64 × 128	64 × 128	1 × 5
	4	Residual units	64 × 128	64 × 64	1 × 5
	5	DeConv1D	64 × 64	64 × 128	1 × 5
	6	DeConv1D	64 × 128	64 × 256	1 × 5
	7	Conv1D	64 × 256	32 × 128	1 × 5
	8	DeConv1D	32 × 128	1 × 128	1 × 5
	Generator for multispectral UAV samples	**Encoder**			
	Layer	**Layer**	**Layer**	**Layer**	**Layer**
	1	Conv1D	1 × 128	64 × 128	1 × 5
	2	Conv1D	64 × 128	128 × 128	1 × 5
	3	Conv1D	128 × 128	256 × 64	1 × 5
	4	Residual units	256 × 64	256 × 64	1 × 5
	5	Residual units	256 × 64	256 × 64	1 × 5
	6	Residual units	256 × 64	256 × 64	1 × 5
	7	Residual units	256 × 64	256 × 64	1 × 5
	Feature mapping				
	Layer				
	1	Conv1D	256 × 64	128 × 64	1 × 5
	2	Conv1D	128 × 64	64 × 64	1 × 5
	Decoder				
	Layer	**Layer**	**Layer**	**Layer**	**Layer**
	1	Residual units	64 × 64	64 × 32	1 × 5
	2	Residual units	64 × 32	64 × 32	1 × 5
	3	Residual units	64 × 32	64 × 32	1 × 5
	4	Residual units	64 × 32	64 × 32	1 × 5
	5	Conv1D	64 × 32	128 × 16	1 × 5
	6	Conv1D	128 × 16	64 × 16	1 × 5
	7	Conv1D	64 × 16	1 × 8	1 × 5
	8	Fully connected network	1 × 8	1 × 5	–
Discriminator	1	Conv1D	1 × 128	128 × 128	1 × 5
	2	Conv1D	128 × 128	64 × 128	1 × 5
	3	Conv1D	64 × 128	32 × 64	1 × 5
	4	Conv1D	32 × 64	16 × 32	1 × 5
	5	Fully connected	–	–	–
VAE-based latent space transformation		**Encoder-Decoder**			
		Layer	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
		1	Conv2D	5 × 5 × 5	1 × 256	5 × 5
		2	Conv1D	1 × 256	128 × 256	1 × 5
		3	Conv1D	128 × 256	256 × 128	1 × 5
		4	DeConv1D	256 × 128	128 × 256	1 × 5
		5	DeConv1D	128 × 256	128 × 256	1 × 5
		6	Conv1D	128 × 256	64 × 256	1 × 5
		7	Conv1D	64 × 256	32 × 128	1 × 5
		8	Conv1D	32 × 128	32 × 64	1 × 5
		9	Conv1D	32 × 64	32 × 32	1 × 5
		10	Conv1D	32 × 32	16 × 32	1 × 5

(*continued on next page*)

Table 3 (*continued*)

Latent space transformation	Network Configuration				
	11	Conv1D	16 × 32	16 × 16	1 × 5
	12	Conv1D	16 × 16	1 × 16	1 × 5
	13	Fully connected network	1 × 16	–	–
	Discriminator				
	Layer	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Conv1D	32 × 128	256 × 64	1 × 5
	2	Conv1D	256 × 64	128 × 32	1 × 5
	3	Conv1D	128 × 32	64 × 16	1 × 5
	4	Conv1D	64 × 16	32 × 16	1 × 5
	5	Fully connected	32 × 16	–	–
Feed forward network-based guided transformation	1	Conv1D	1 × 5	256 × 15	5 × 5
	2	DeConv1D	256 × 15	128 × 30	1 × 5
	3	DeConv1D	128 × 30	128 × 60	1 × 5
	4	Conv1D	128 × 60	64 × 60	1 × 5
	5	Conv1D	64 × 60	32 × 60	1 × 5
	6	Conv1D	32 × 60	8 × 60	1 × 5
	7	Fully connected	–	–	–

Table 4
Implementation details of the proposed latent graph generation and convolution module for the multispectral UAV datasets with hyperspectral data.

Stream	Network Configuration				
Graph embedding stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Fully connected network layer	–	–	–
	2	Fully connected network layer	–	–	–
	3	Fully connected network layer	–	–	–
Graph convolution stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Graph Convolution	32 × 32 × 64	32 × 32 × 32	5 × 5
	2	Graph Convolution	32 × 32 × 32	16 × 16 × 32	5 × 5
	3	Graph Convolution	16 × 16 × 32	16 × 16 × 8	5 × 5
	4	Graph Convolution	16 × 16 × 8	8 × 8 × 8	5 × 5
	5	Graph Convolution	8 × 8 × 8	8 × 8 × 4	5 × 5
	6	Fully connected network	8 × 8 × 4	–	–

Table 5
Implementation details of the proposed graph embedding and graph convolutional for the classification of airborne hyperspectral spectra.

Stream	Network Configuration				
Graph embedding stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Fully connected network layer	–	–	–
	2	Fully connected network layer	–	–	–
	3	Fully connected network layer	–	–	–
	4	Fully connected network layer	–	–	–
Graph convolution stream	**Layer**	**Type**	**# Features In**	**# Features Out**	**Kernel Size**
	1	Graph Convolution	32 × 32 × 64	32 × 32 × 128	5 × 5
	2	Graph Convolution	32 × 32 × 128	32 × 32 × 64	5 × 5
	3	Graph Convolution	32 × 32 × 64	16 × 16 × 64	5 × 5
	4	Graph Convolution	16 × 16 × 64	16 × 16 × 16	5 × 5
	5	Graph Convolution	16 × 16 × 16	8 × 8 × 4	5 × 5
	6	Fully connected network	8 × 8 × 4	–	–

derived from the spectrally fine hyperspectral spectra. Hence, confusion matrix-based Kappa statistics and overall accuracy were used for evaluating the accuracy of different approaches. Kappa is similar to overall accuracy except that it is normalized at the baseline of the chance agreement for the given dataset. Kappa ranges from 0 to 1. A higher value of the kappa and overall accuracy indicate a better classification (McHugh, [27]). As the proposed approach focuses on the use of multi-sensor data for improving the classification performance, the experiments adopted in this study used classification-based accuracy measures. For comparing the approaches, which do not give classification results, with the proposed approaches, a support vector machine (SVM)-based classifier was used to classify the fused data. The SVM implementations, reported in this study, adopted RBF kernels and one-vs-one multi-class classification strategy. A Bayesian hyper-parameter optimization (Czarnecki et al., [8]) was used to finetune the values of parameters such as the regularization strength (C) (in the range $[10^{-2}$ - $10^{4}]$) and the width of the RBF kernel (γ) (in the range $[2^{-5} – 2^{4}]$). The number of iterations for hyperparameter optimization was set to 200.

6.1. Results on simulated datasets

The spectrally and spatially downscaled standard datasets along with their corresponding original high spectral resolution samples (as discussed in Section 3.1) were used for analyzing the proposed approaches.

6.1.1. Network parameter analysis

Analysis of the sensitivity of the latent space transformation

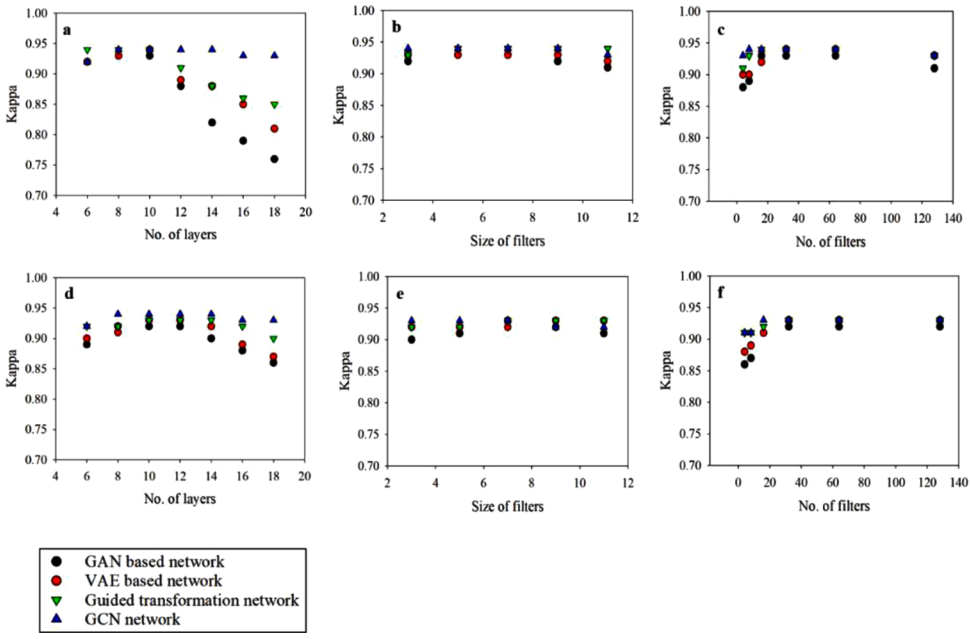

Fig. 6. Analysis of the sensitivity of the approach towards the depth of the network layers, Size of filters, and the number of filters for the Indian Pines dataset (a-c) and Pavia dataset (d -f) (in some cases the symbols are overlapping) *.*GAN, VAE and GCN denote the generative adversarial networks, variational autoencoder and graph convolutional layers respectively.

networks (GAN, VAE, and guided latent space projection) showed that increase in network depth, the number of filters and filter-size improved the accuracy to a limit beyond which it deteriorated or saturated. The use of multi-sized kernels was found to be a viable alternative as it significantly improved the results without significantly affecting the execution time. An illustration of the sensitivity analysis of the network layers toward network parameters for the Indian Pines and Pavia dataset is presented in Fig. 6. Experiments indicated that among the proposed cross-modal generative approaches, VAE based approach was less sensitive to the network depth as compared to GAN. The graph generation and graph convolutional network were less sensitive to the network parameters. The increase in kernel size as well as number of kernels improved the accuracy to a limit, but the trend saturated gradually.

Table 6
Analysis of loss functions and constraints.

Dataset	Losses/Constraints	Kappa statistics (K)	Overall Accuracy
Indian Pines*	GAN without cycle consistency constraint	0.88	91.23
	GAN without MK-MMD	0.89	93.40
	VAE without JS divergence	0.90	94.57
	GCN without MMD**	0.92	95.08
	Proposed Implementation* **	**0.94**	**97.92**
Salinas*	GAN without cycle consistency constraint	0.89	92.34
	GAN without MMD	0.91	93.46
	VAE without JS divergence	0.93	95.08
	GCN without MMD**	0.95	96.90
	Proposed Implementation* **	**0.98**	**99.28**
Pavia centre*	GAN without Cycle consistency constraint	0.86	88.56
	GAN without MMD	0.88	90.19
	VAE without JS divergence	0.91	93.40
	GCN without MMD**	0.92	94.56
	Proposed Implementation* **	**0.93**	**97.84**
KSC*	GAN without Cycle consistency constraint	0.88	90.34
	GAN without MMD	0.91	94.46
	VAE without JS divergence	0.93	95.70
	GCN without MMD**	0.93	95.19
	Proposed Implementation* **	**0.95**	**98.87**

*Data spectrally downscaled to 45 bands.
**GCN denotes the graph convolutional and generation layers.
***Generative adversarial network (GAN), variational autoencoder (VAE) and guided projection resulted in similar results for more than 60% cross-domain training samples.

6.1.2. *Analysis of the effect of loss functions and constraints*

In this study, different alternate regularizations and losses experimented with respect to the proposed approach and the results are summarized in Table 6. The use of cycle consistency constraint and MK-MMD, for GAN-based latent space transformation, had significantly improved the accuracy. The approach improved the generalizability as the discrepancy in source and target samples along with domain bias in input modality was resolved to an extent. The results also indicated that the use of JS divergence, in computing the VAE based shared latent space projection, improved the accuracy when compared to the use of KL divergence. The refinement of graph convolutional networks using MMD had significantly reduced the misclassifications. In addition, the use of DTW instead of Euclidean distance measure, in graph embedding, considered the multimodal nature of the samples and resulted in improved classification accuracy.

6.1.3. *Comparison with the state of the art*

The prominent approaches, applicable to multimodal data classification, were compared with the proposed approaches. The implementation details of all the benchmark approaches were adopted from the corresponding literatures. It may be noted that some of the benchmark approaches were modified to consider multispectral images and point hyperspectral data (ground measured spectra). This was accomplished using additional convolutional streams for transforming the multispectral patches to 1D spectra. Hyper-parameter optimization, proposed in Bochinski et al. [5], was employed to find the optimal parameter settings of the different approaches considered in this study. For the fusion techniques which did not yield classification results, an SVM classifier was adopted to classify the fused images.

The prominent hyperspectral PAN sharpening approaches such as Vivone, and Chanussot [37], Hong et al. [[17]b], He et al. [13], Restaino et al. [31], and Zheng et al. [50] were taken as the benchmarks. The

Table 7
Comparative evaluation of the proposed approach with prominent baseline approaches on simulated datasets for 60% of the training samples*#$.

Dataset	Method	K	OA	T	Method	K	OA	T
[45]	Zhang et al. [46]	].79	83.42	128	Ding and Fu [10]	0.90	92.14	298
	Zhang et al. (2019a)	0.81	84.36	296	Liu and Qin [22]	0.90	93.20	352
	Liu et al. [23]	0.78	81.23	386	Shi et al. [35]	0.89	92.38	456
	Deng et al. [9]	0.86	89.09	270	He et al. [14]	0.91	93.67	569
	Sutter et al. [36]	0.88	90.15	387	Hong et al. (2020b)	[17]	92.49	407
	Hong et al. [18]	0.85	88.19	254	He et al. [13]	0.88	90.06	380
	Huang et al. [14]	0.87	90.86	393	Restaino et al. (2020)	0.89	92.69	423
	Vivone, and Chanussot [37]	0.90	92.14	465	Zheng et al. [50]	0.91	94.73	305
	Zhao et al. (2020)	0.91	93.58	580	**Proposed approach*** **	**0.94**	**97.92**	189
[45]	Zhang et al. [46]	].85	90.45	94	Ding and Fu [10]	0.94	95.60	163
	Zhang et al. (2019a)	0.86	88.91	186	Liu and Qin (2020)	0.93	95.17	327
	Liu et al. [23]	0.85	89.57	251	Hong et al. [18]	0.89	92.36	209
	Deng et al. (2020)	0.86	88.09	153	He et al. [13]	0.93	94.67	265
	Sutter et al. (2020)	0.90	92.34	247	Hong et al. (2020b)	[17]	92.34	347
	Hong et al. [18]	0.92	94.56	285	He et al. (2020)	0.92	95.46	189
	Huang et al. (2020)	0.92	95.09	336	Restaino et al. (2020)	0.91	93.90	234
	Vivone, and Chanussot (2020)	0.90	92.19	309	Zheng et al. (2020)	0.94	95.63	292
	Zhao et al. (2020)	0.93	95.67	412	**Proposed approach*** **	**0.98**	**99.28**	128
[45]	Zhang et al. [46]	].84	88.34	164	Ding and Fu [10]	0.90	92.36	296
	Zhang et al. (2019a)	0.85	89.56	235	Liu and Qin (2020)	0.91	93.67	417
	Liu et al. [23]	0.87	90.43	391	Hong et al. [18]	0.89	92.44	329
	Deng et al. (2020)	0.88	92.90	261	He et al. (2020)	0.87	91.08	308
	Sutter et al. (2020)	0.89	91.08	386	Hong et al. (2020b)	[17]	90.64	461
	Hong et al. [18]	0.90	92.34	345	He et al. [13]	0.91	93.45	260
	Huang et al. (2020)	0.91	93.56	489	Restaino et al. (2020)	0.90	92.18	362
	Vivone, and Chanussot (2020)	0.89	91.08	396	Zheng et al. (2020)	0.91	93.60	391
	Zhao et al. (2020)	0.88	90.47	567	**Proposed approach*** **	**0.93**	**97.84**	215
[45]	Zhang et al. (2020)	].89	92.13	207	Ding and Fu [10]	0.91	93.16	342
	Zhang et al. (2019a)	0.89	94.36	359	Liu and Qin [22]	0.93	94.50	589
	Liu et al. [23]	0.90	93.53	436	Hong et al. [18]	0.92	94.82	416
	Deng et al. [9]	0.87	92.41	340	He et al. (2020)	0.89	91.56	489
	Sutter et al. [36]	0.89	94.16	457	Hong et al. (2020b)	[17]	94.48	587
	Hong et al. [18]	0.90	93.72	389	He et al. [13]	0.92	95.06	345
	Huang et al. [19]	0.89	94.18	493	Restaino et al. [31]	0.93	96.42	468
	Vivone, and Chanussot (2020)	0.92	93.45	581	Zheng et al. [50]	0.93	95.09	471
	Zhao et al. [49]	0.91	93.60	673	**Proposed approach*** **	**0.95**	**98.95**	367

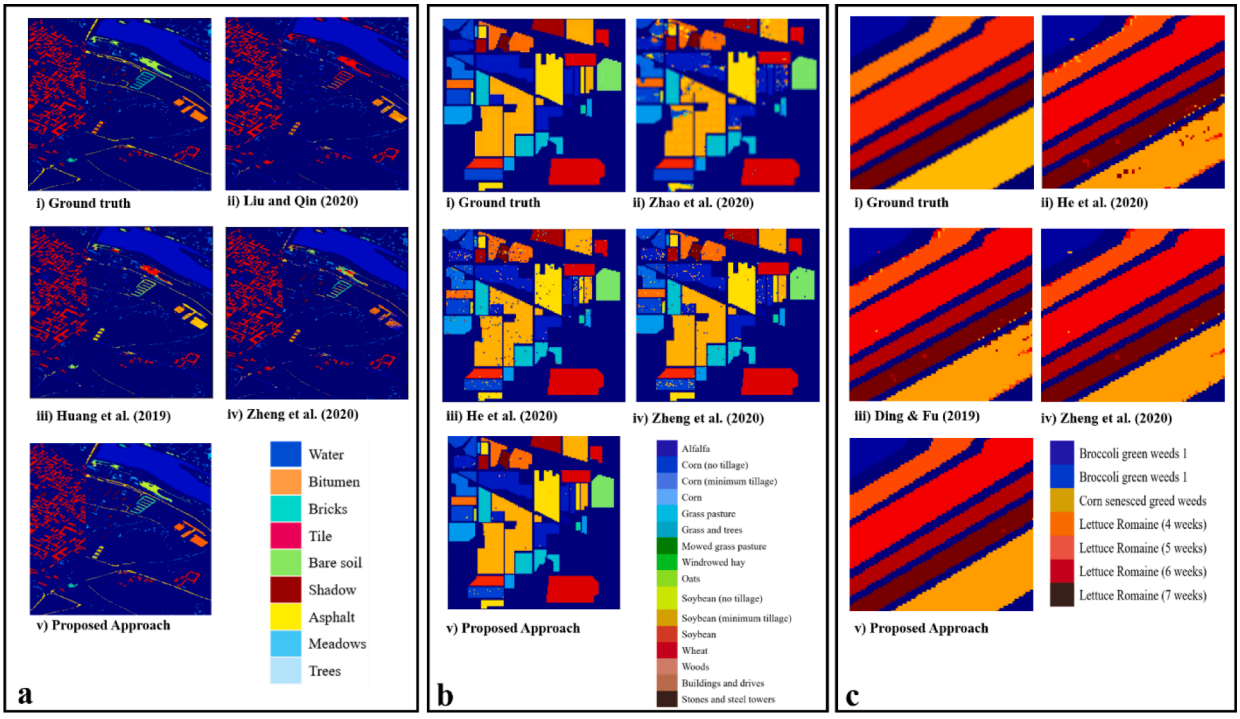

Fig. 7. Comparative analysis of the proposed approach on (a) Pavia dataset, (b) Indian Pines dataset and (c) Salinas dataset.*Methods implemented based on the available GitHub implementations and were fine-tuned with respect to the related publications. **Data spectrally downscaled to 15 bands. ***Latent space transformation implemented using cross-modal generation. #Support vector machine (SVM) classifier is used for transforming the output of general fusion approaches to classified maps. $ K denotes kappa statistics, OA denotes overall accuracy, and T denotes the running time.

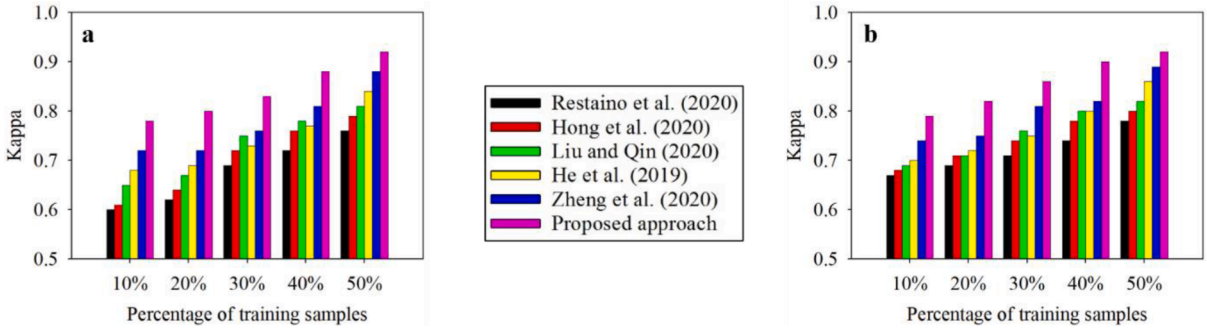

Fig. 8. Comparative analysis of the proposed approach on (a) Indian Pines and (b) KSC datasets.

selected approaches have reported state-of-the-art results and most of them employed CNN-based architectures. In addition, an SVM-based classifier was used for transforming the fused data to classified maps. The results of the comparative analyses are summarized in Table 7 and Figs. 7 and 8. As is evident, the proposed approach gave better results than the baseline approaches over all the datasets considered. These results can be attributed to their ability to address the differences in spatial and spectral resolutions of the source and target domains. The proposed approach, specifically the use of guided transformation discussed in Section 4.3, successfully fused point spectral data with multispectral or hyperspectral image patches even when cross-domain correlation was limited. This can be attributed to the use of guided transformation avoiding complex generative strategies. The graph-based approach, proposed in this study, generated the graphs dynamically and predicted the labels by considering both the labelled and unlabeled data across different modalities. The source and target domain discrepancy, as well as the input modality differences, were effectively resolved with a minimal number of training samples.

In addition to prominent hyperspectral PAN sharpening approaches, some of the recent multimodal classifiers for remote sensing data such as Ding and Fu [10], Liu and Qin [22], Zhang et al. [[47]b], Hong et al. [18] and He et al. [14] were also compared with the proposed approaches. Most of these approaches tried to resolve the cross-domain biases and required extensive training samples. As is evident from the results in Table 7 and Figs. 7 and 8, the proposed strategies considered multimodal nature as well as domain discrepancy, and achieved better results as compared to the existing ones. These results can be attributed to the shared latent space transformations and to the dynamic graph generation and convolutions. The constraints and transformations used also significantly reduced the requirement of cross-domain training samples. Additionally, the incorporation of DTW as a similarity measure resolved the issues of spectral resolution differences and facilitated effective comparisons.

The prominent multimodal fusion and domain adaptation techniques, in other domains, were also modelled for the remote sensing data for an effective comparison. In this regard, approaches such as Zhang et al. [46], Deng et al. [9], Ding and Fu (2020), Huang et al. [19], and Sutter et al. [36] were also compared with the proposed approaches. However, as is evident from Table 7 and Figs. 7 and 8, a simple adaptation of these approaches to the remote sensing domain was not sufficient to consider the specific characteristics of the EO data, especially to resolve the resolution differences as well as the point and patch nature of the different modalities. The use of proposed architectures, end-to-end training strategy, latent space transformations, latent graph generation, and graph-based labeling had given better results, even addressing the source and target domain discrepancies. The proposed guided transformation (Section 4.3) significantly addressed the issue of the scarcity of training samples as it does not involve complex generative architectures.

A comparative analysis of the proposed latent space projection approaches is presented in Fig. 9. The guided latent space projection was found to be effective when cross-domain training samples were scarce. However, VAE or GAN may be preferred when a sufficient number of training samples are available. Also, the guided transformation required the samples to be labelled while VAE and GAN used correlated but unlabeled samples.

6.2. *Fusion of multispectral UAV datasets with ground measured spectra*

To study the effectiveness of the proposed approach for real-world scenarios, multispectral UAV data, ground-measured spectral data and satellite data collected over agriculture plots were used. The UAV datasets collected over different plots along with a few ground measured spectra were used to classify different irrigation treatments. These image

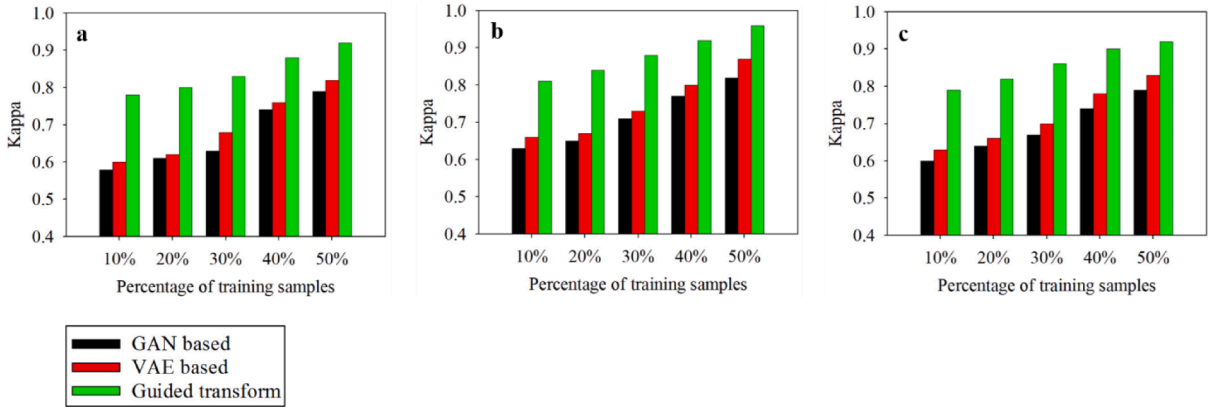

Fig. 9. Comparative analysis of the proposed latent space projection approaches for (a) Indian Pines dataset, (b) Salinas dataset and (c) KSC dataset*. *GAN and VAE denote the generative adversarial networks and variational autoencoder respectively.

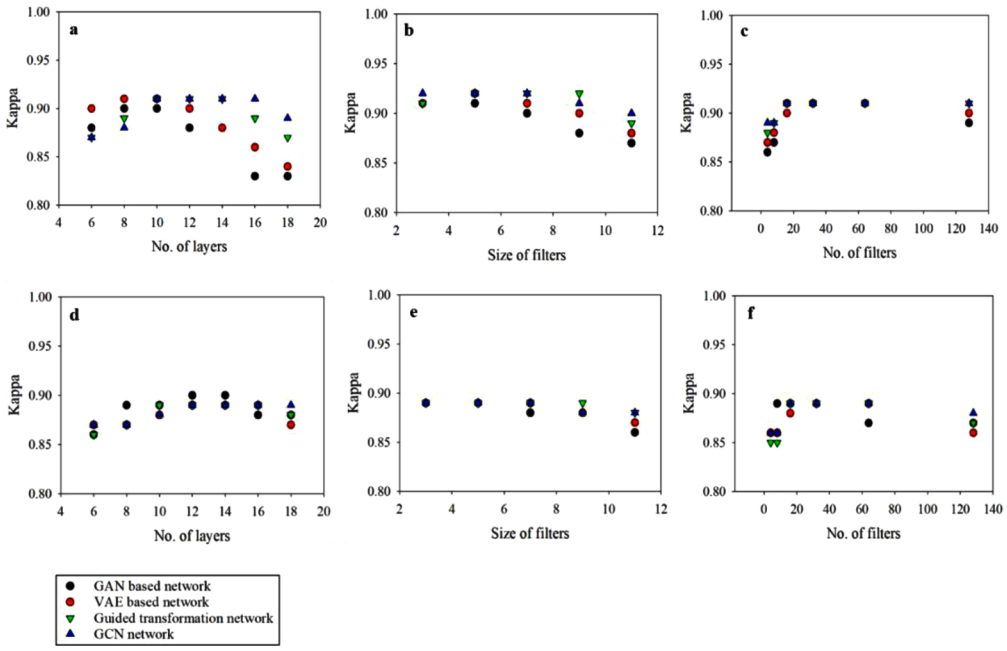

Fig. 10. Analysis of the sensitivity of the approach towards depth of the network layers, size of filters, and number of filters for Plot-1 (a-c) and Plot-2 (d-f) datasets (in some cases the symbols are overlapping)*. *GAN, VAE and GCN denote the generative adversarial networks, variational autoencoder and graph convolutional layers.

datasets covering different irrigation traits in four spectral bands helped to evaluate the proposed frameworks in their capability in distinguishing the irrigation traits (which is otherwise possible only using an exceptionally fine spectral resolution).

6.2.1. *Network parameter analysis*

Analysis of the sensitivity of the network layers with respect to the network parameters for Plot-1 and Plot-2 are summarized in Fig. 10. As is evident from the results, an increase in network depth improved the accuracy to a limit beyond which it deteriorated. Similarly, an increase in kernel size, as well as number of kernels, improved the accuracy, but the trend saturated gradually. The guided latent space transformation network was less sensitive to parameter variations as compared to the VAE and GAN-based approaches. The graph generation and graph convolutional networks were also less sensitive to the network parameters.

6.2.2. *Analysis of the effect of loss functions and constraints*

As is evident from Table 8, the GAN-based implementation of latent space projection had resulted in improved Kappa statistics when the cycle consistency constraint was employed. In addition, the use of MK-MMD had also significantly improved the accuracy as it addresses the source and target discrepancy (training and testing domain) along with the domain bias (difference in the modality of input data). The use of JS divergence has resulted in better results when compared to the KL divergence-based discrepancy measure. As the inputs consisted of multiple modalities, the use of MMD along with DTW-based similarity measure has significantly improved the generalization as well as the domain and source-target discrepancy. The approach had significantly reduced the requirement of training samples and gave a notable improvement in terms of Kappa measures in comparison with the state-of-the-art approaches, even when training samples were scarce.

Table 8
Analysis of loss functions and constraints.

Dataset	Losses/Constraints	Kappa statistics	Overall accuracy
Plot-1	GAN without cycle consistency constraint	0.90	92.08
	GAN without MK-MMD	0.90	93.34
	VAE without JS divergence	0.92	95.16
	GCN without MMD*	0.94	96.59
	Proposed Implementation**	**0.96**	**98.20**
Plot-2	GAN without cycle consistency constraint	0.88	90.74
	GAN without MK-MMD	0.92	94.61
	VAE without JS divergence	0.92	95.03
	GCN without MMD*	0.93	95.80
	Proposed Implementation**	**0.96**	**97.26**

*GCN denotes the graph convolutional and generation layers.
**GAN, VAE and guided projection resulted in similar results for more than 60% training samples.

6.2.3. *Comparison with the state of the art*

A comparison of the proposed approach with the prominent hyperspectral PAN sharpening approaches, multimodal classification algorithms and fusion techniques are presented in Table 9 and Fig. 11. The implementation details of the benchmark approaches are similar to the discussion in Section 6.1.3. For benchmark fusion and sharpening techniques, an SVM classifier was employed to transform the results into classification maps. As is evident from the results and similar to the discussions in Section 6.1.3, the proposed approach gave better results than the baseline approaches. The superior performance of the proposed approach was evident especially when the number of cross-domain training samples were scarce. Even when the cross-domain samples were absent, the proposed guided latent transformation coupled with graph generation and convolution yielded accurate results. Furthermore, this characteristic made the proposed strategy resilient to the co-registration errors prevalent in the multi-modal datasets.

A comparative analysis of the proposed latent space projection approaches is presented in Fig. 12. The guided transformation employed a smaller number of labelled uncorrelated samples of both domains for learning the transformation. The VAE- or GAN-based transformation may be preferred when enough training samples are available and labelled samples are scarce.

Table 9
Comparative evaluation of the proposed approach with prominent baseline approaches for 60% training samples*.

Dataset	Method	Kappa statistics	Overall accuracy	Time (s)	Method	Kappa statistics	Overall accuracy	Time (s)
[45]	Zhang et al. [46]	].78	80.15	283	Ding and Fu [10]	0.87	90.44	1509
	Zhang et al. (2019a)	0.79	82.54	612	Liu and Qin [22]	0.85	88.72	689
	Liu et al. [23]	0.83	86.18	856	Shi et al. [35]	0.85	88.49	915
	Deng et al. [9]	0.86	89.05	780	He et al. [14]	0.87	90.09	593
	Sutter et al. [36]	0.84	87.56	932	Hong et al. (2020b)	[17]	95.27	328
	Hong et al. [18]	0.83	85.92	619	He et al. [13]	0.92	94.15	216
	Huang et al. [19]	0.87	89.04	823	Restaino et al. [31]	0.93	95.06	409
	Vivone, and Chanussot (2020)	0.85	88.15	905	Zheng et al. (2020)	0.94	95.71	364
	Zhao et al. (2020)	0.87	90.36	1285	**Proposed approach****	**0.96**	**98.09**	390
[45]	Zhang et al. (2020)	].86	89.83	189	Ding and Fu [10]	0.87	89.29	1150
	Zhang et al. (2019a)	0.84	87.34	462	Liu and Qin (2020)	0.86	88.15	532
	Liu et al. [23]	0.85	87.90	690	Hong et al. [18]	0.87	90.40	764
	Deng et al. (2020)	0.87	89.46	580	He et al. (2020)	0.88	91.08	312
	Sutter et al. [36]	0.84	87.32	779	Hong et al. (2020b)	[17]	93.36	408
	Hong et al. [18]	0.85	88.67	441	He et al. [13]	0.89	91.10	174
	Huang et al. (2020)	0.89	91.56	657	Restaino et al. [31]	0.90	92.45	319
	Vivone, and Chanussot (2020)	0.84	87.43	708	Zheng et al. (2020)	0.92	94.68	290
	Zhao et al. (2020)	0.86	89.05	345	**Proposed approach****	**0.96**	**97.26**	201

*Methods implemented based on the available GitHub implementations and were fine-tuned with respect to the related publications.
**Latent space transformation implemented using cross-modal generation.
$ K denotes kappa statistics, OA denotes overall accuracy, and T denotes the running time.

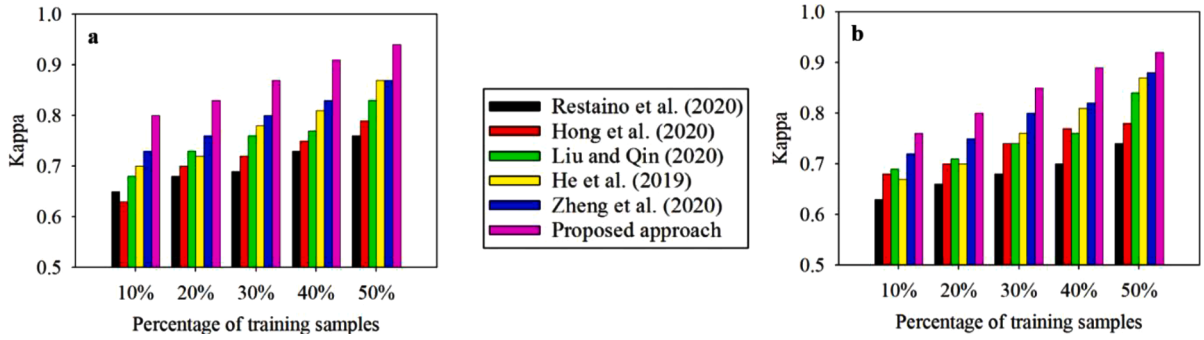

Fig. 11. Comparative analysis of the proposed approach over the (a) Plot-1 and (b) Plot-2 data.

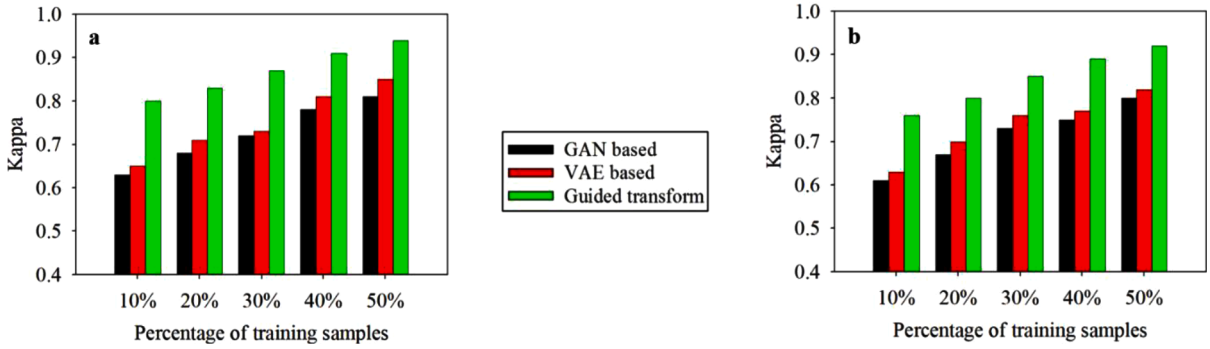

Fig. 12. Comparative analysis of the proposed latent space projection approaches for (a) Plot-1 dataset (b) Plot-2 dataset.

6.3. *Fusion of airborne dataset with ground measured spectra*

The hyperspectral airborne data collected over an almond orchard, located in Cordoba, southern Spain, at the Alameda del Obispo Research Station (37° 52′N, 4° 49′W) (Plot-3), was also employed to analyze the effectiveness of the proposed frameworks. This subsection analyzed the proposed approach for hyperspectral image classification when the training samples were limited with significant resolution differences.

6.3.1. *Network parameter analysis*

Analysis of the sensitivity of the network layers with respect to the network parameters for plot-3 is summarized in Fig. 13. As is evident from the results, an increase in network depth improved the accuracy to a limit beyond which it deteriorated. The increase in kernel size, as well as number of kernels, improved the accuracy to a limit, but the trend saturated gradually. As the graph generation and graph convolutional networks were only employed preceded by a two-stream shallow layer, the network was less sensitive to the parameters.

6.3.2. *Analysis of the effect of loss functions and constraints*

An illustration of the analysis of loss functions and constraints is presented in Table 10. The use of DTW improved the results significantly

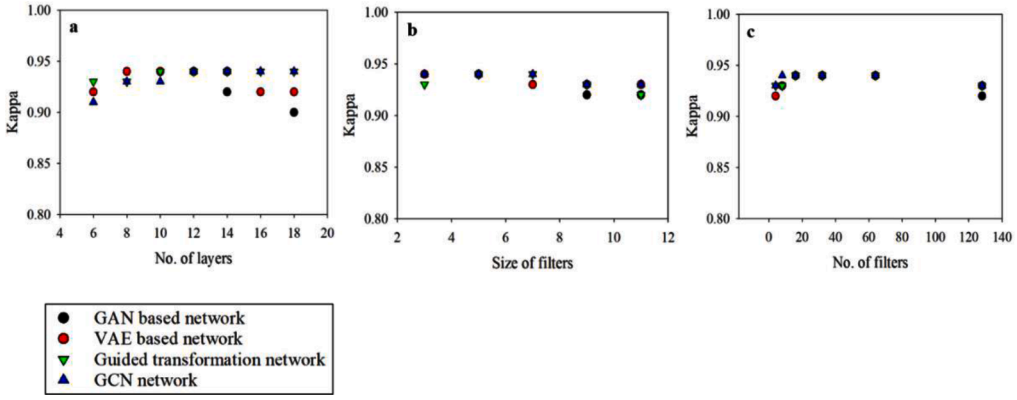

Fig. 13. Analysis of the sensitivity of the approach towards (a) depth of the network layers, (b) size of filters, and (c) number of filters for Plot-3 dataset (in some cases the symbols are overlapping)*.*GAN, VAE and GCN denote the generative adversarial networks, variational autoencoder and graph convolutional layers.

Table 10
Analysis of loss functions and constraints.

Losses/Constraints	Kappa statistics	Overall Accuracy
GCN with Euclidean distance instead of DTW	0.87	90.05
GCN without MMD*	0.88	92.47
Proposed Implementation	**0.93**	**96.84**

*GCN denotes the graph generation and convolutional layers.

as it effectively resolves the resolution biases. In addition, the consideration of MMD to resolve the source-target discrepancy, while embedding the graph in the latent space, also improved the classification accuracy.

6.3.3. Comparison with the state of the art

A comparison of the prominent benchmark approaches with the proposed approach with regard to the classification of airborne hyperspectral imagery is presented in Fig. 14 and Table 11. The baseline approaches were implemented in accordance with the corresponding literature and a brief discussion of the same is presented in Section 6.1.3. It may be noted that the results of some baseline fusion approaches were transformed using an SVM classifier to compare them with the proposed approaches. The proposed approach gave better results even with a limited number of training samples and can be attributed to the proposed non-generative architecture (Section 5.3). It may be noted that the training samples (source domain) had different spectral resolution as compared to the data to be classified. The improvement in classification results can be attributed to the proposed dynamic graph generation and

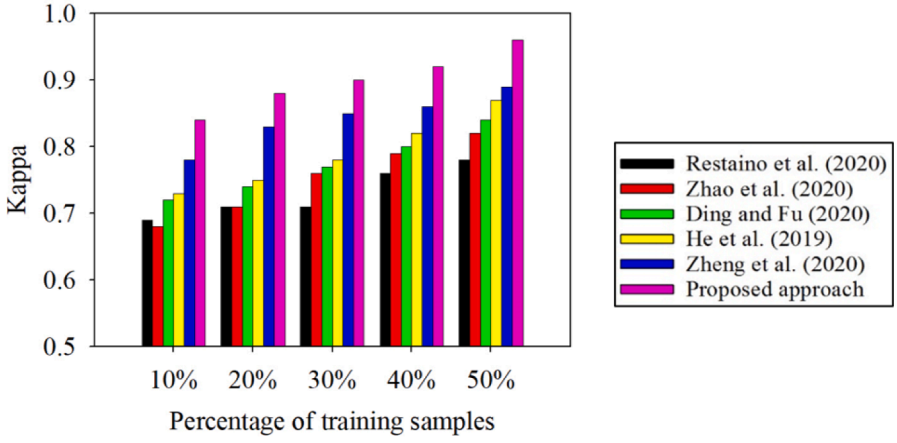

Fig. 14. Comparative analysis of the proposed approach for hyperspectral airborne (plot-3) data.

Table 11
Comparative evaluation of the proposed approach with prominent baseline approaches for 60% of the cross-domain samples*.

Method	Kappa statistics	Overall accuracy	Time (s)	Method	Kappa statistics	Overall accuracy	Time (s)
Zhang et al. [46]	0.83	86.82	**118**	Ding and Fu [10]	0.93	95.18	380
Zhang et al. (2019a)	0.84	87.08	154	Liu and Qin (2020)	0.86	89.53	245
[[45]t al. ([23]]	0.82	85.67	206	Shi et al. [35]	0.87	90.28	312
Deng et al. (2020)	0.83	86.74	180	He et al. (2020)	0.94	95.91	290
Sutter et al. (2020)	0.85	88.90	132	Hong et al. (2020b)	[17]	91.23	178
Hong et al. [18]	0.84	86.23	179	He et al. [13]	0.94	96.69	216
]uang et al. (2020)	0.85	88.90	203	Restaino et al. (2020)	0.93	95.14	169
Vivone, and Chanussot (2020)	0.86	89.26	253	Zheng et al. (2020)	0.94	96.25	207
Zhao et al. (2020)	0.93	95.47	465	**Proposed approach****	**0.98**	**99.44**	190

*Methods implemented based on the available GitHub implementations and were fine-tuned with respect to the related publications.
**Latent space transformation implemented using guided projection.

DTW-based graph convolution along with other constraints which could effectively model the relations between the unlabeled and labelled source and target samples. Besides, the proposed approach did not require the multispectral and hyperspectral samples to have a perfect cross-domain correspondence and were resilient to the co-registration errors.

7. Conclusions

This study proposed shared latent space projection approaches for multimodal datasets resolving the issues of source-target and multi-modality biases. A latent graph generation and graph convolutional-based approach was also proposed to accurately predict the class labels by considering labelled and unlabeled samples. The proposed approach was performing with high quality even when the training samples were scarce. The cross-modal generative approaches using GAN and VAE, proposed in this study, performed well even with a smaller number of cross-domain samples as compared to the existing shared latent space projection approaches. This can be attributed to the fact that the generative frameworks can be trained with the unlabeled samples. The cycle consistency loss and the use of MMD measures improved the generalizability of the GAN-based transformations. The use of JS measure and the concept of shared and discriminate latent spaces in VAE based approach improved the results. The generative transformation models employed unlabeled and a few labelled samples while the proposed covariance guided transformation required labelled samples. The convolutional layers, used in the proposed approaches, adopted an interpolation-based convolution to process the ground spectra effectively. It may be noted that the proposed covariance guided transformation eliminated the need for cross-domain training samples (samples having cross-domain correspondence) without much affecting the accuracy. The use of DTW-based similarity measure in graph generation and graph convolutional layers suggested in this study, effectively addressed the source-target and multimodal domain mismatches. Experiments on simulated and real datasets illustrated that the proposed architectures and regularizations resolved the issue of cross-domain sample requirement which was a critical issue in the fusion of ground spectra with UAV or satellite images. In addition, the proposed approach outperformed the baseline PAN sharpening, fusion and domain adaptation methods considered in this study owing to its capability in handling multimodal data and domain biases. The proposed approach can be extended to various analyses and applications requiring the fusion of multi-source data sets particularly when it is difficult to have training samples with cross-domain correlation. The interpolated convolution and DTW-based latent graph generation, adopted in this study, can be used for various time series or signal analyses.

CRediT authorship contribution statement

P.V. Arun: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **R. Sadeh:** Data curation. **A. Avneri:** Data curation. **Y. Tubul:** Data curation, Project administration. **C. Camino:** Data curation. **K.M. Buddhiraju:** Funding acquisition. **A. Porwal:** Funding acquisition. **R.N. Lati:** Data curation, Funding acquisition. **P.J. Zarco-Tejada:** Data curation, Funding acquisition. **Z. Peleg:** Project administration, Writing – review & editing. **I. Herrmann:** Conceptualization, Supervision, Resources, Project administration, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the: Study in Israel Sandwich Ph.D. Scholarship Program for partially supporting the study; Dr. David J. Bonfil for the Gilat field experiment. This research was partly supported by the Hebrew University of Jerusalem Intramural Research Found Career Development, Association of Field Crop Farmers in Israel and the Chief Scientist of the Israeli Ministry of Agriculture and Rural Development (projects 20-02-0087 and 12-01-0041).

References

[1] P.V. Arun, K.M. Buddhiraju, A. Porwal, Capsulenet-based spatial–spectral classifier for hyperspectral images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (6) (Jun. 2019) 1849–1865.
[2] P.V. Arun, K.M. Buddhiraju, A. Porwal, J. Chanussot, CNN-based super-resolution of hyperspectral images, IEEE Trans. Geosci. Remote Sens. 58 (9) (2020) 6106–6121. Sep.
[3] P.V. Arun, K.M. Buddhiraju, A. Porwal, J. Chanussot, CNN based spectral super-resolution of remote sensing images, Signal Processing 169 (4) (2020), 107394. Apr.
[4] D. Bacciu, F. Errica, A. Micheli, M. Podda, A gentle introduction to deep learning for graphs, Neural Netw. 129 (9) (Sep. 2020) 203–221.
[5] E. Bochinski, T. Senst, T. Sikora, Hyper-parameter optimization for convolutional neural network committees based on evolutionary algorithms, in: proceedings of the International Conference on Image Processing (ICIP), 2018, pp. 3924–3928.
[6] S. Chlaily, M.D. Mura, J. Chanussot, C. Jutten, P. Gamba, A. Marinoni, Capacity and limits of multimodal remote sensing: theoretical aspects and automatic information theory-based image selection, IEEE Trans. Geosci. Remote Sens. (Aug. 2020) 1–21.
[7] Computational Intelligence Group, Hyperspectral Remote Sensing Scenes, Basque University, 2020. Retrieved from the website: "http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes.
[8] W.M. Czarnecki, S. Podlewska, A.J. Bojarski, Robust optimization of SVM hyperparameters in the classification of bioactive compounds, J Cheminform 7 (38) (Aug. 2015).
[9] C. Deng, X. Liu, J. Chanussot, Y. Xu, B. Zhao, Towards perceptual image fusion: a novel two-layer framework, Inform. Fusion 57 (May 2020) 102–114.
[10] Z. Ding, Y. Fu, Deep transfer low-rank coding for cross-domain learning, IEEE Trans Neural Netw Learn Syst 30 (6) (Jun. 2019) 1768–1779.
[11] R. Dian, S. Li, L. Fang, T. Lu, J.M. Bioucas-Dias, Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion, IEEE Trans Cybern (November 2019) 1–12.
[12] L.A. Gatys, A.S. Ecker, M. Bethge, Image style transfer using convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414–2423.
[13] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, B. Li, HyperPNN: hyperspectral pansharpening via spectrally predictive convolutional neural networks, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12 (8) (Aug. 2019) 3092–3100.
[14] X. He, Y. Chen, P. Ghamisi, Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network, IEEE Trans. Geosci. Remote Sens. 58 (5) (May 2020) 3246–3263.
[15] J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, CyCADA: cycle-consistent adversarial domain adaptation, in: In Proceedings of the International Conference on Machine Learning Research 81, Jul. 2018, pp. 1989–1998.
[16] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, B. Zhang, More diverse means better: multimodal deep learning meets remote-sensing imagery classification, IEEE Trans. Geosci. Remote Sens. (Aug. 2020) 1–15, a.
[17] D. Hong, N. Yokoya, G.S. Xia, J. Chanussot, X.X. Zhu, X-ModalNet: a semi-supervised deep cross-modal network for classification of remote sensing data, ISPRS J. Photogramm. Remote Sens. 167 (2020) 12–23. Sep.
[18] D. Hong, N. Yokoya, J. Chanussot, X.X. Zhu, CoSpace: common subspace learning from hyperspectral-multispectral correspondences, IEEE Trans. Geosci. Remote Sens. 57 (7) (Jul. 2019) 4349–4359.
[19] Y. Huang, F. Zheng, R. Cong, W. Huang, M.R. Scott, L. Shao, MCMT-GAN: multi-task coherent modality transferable GAN for 3D brain image synthesis, IEEE Trans. Image Process. 29 (Jul. 2020) 8187–8198.
[20] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: In Proceedings of the International Conference on Machine Learning 70, Apr. 2017, pp. 857–1865.
[21] M.Y. Liu, O. Tuzel, Coupled generative adversarial networks, in: Proceedings of the Advanced Neural Information Processing System, Dec. 2016, pp. 469–477.
[22] W. Liu, R. Qin, A multikernel domain adaptation method for unsupervised transfer learning on cross-source and cross-region remote sensing data classification, IEEE Trans. Geosci. Remote Sens. 58 (6) (Jun. 2020) 4279–4289.
[23] X. Liu, C. Deng, J. Chanussot, D. Hong, B. Zhao, StfNet: a two-stream convolutional neural network for spatiotemporal image fusion, IEEE Trans. Geosci. Remote Sens. 57 (9) (Sep. 2019) 6552–6564.
[24] L. Loncan, L.B. de Almeida, J.M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, Hyperspectral pansharpening: a review, IEEE Geosci. Remote Sens. Mag. 3 (3) (Sept. 2015) 27–46.

[25] Y. Luo, R. Ji, T. Guan, J. Yu, P. Liu, Y. Yang, Every node counts: self-ensembling graph convolutional networks for semi-supervised learning, Pattern Recognit. 106 (10) (Oct. 2020), 107451.
[26] W.H. Maes, K. Steppe, Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture, Trend. Plant Sci. 24 (2) (Feb. 2019) 152–164.
[27] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. (Zagreb) 22 (3) (Mar. 2012) 276–282.
[28] J. Nalepa, M. Myller, M. Kawulok, Training- And Test-Time Data Augmentation for Hyperspectral Image Segmentation, IEEE Geosci. Remote Sens. Lett. 17 (2) (Feb. 2020) 292–296.
[29] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, J.A. Benediktsson, Feature extraction for hyperspectral imagery: the evolution from shallow to deep (overview and toolbox), IEEE Geosci. Remote Sens. Mag. (April 2020).
[30] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, Proceed. Int. Confer. Mach. Learn. Res. 48 (Jun. 2016) 1060–1069.
[31] R. Restaino, G. Vivone, P. Addesso, J. Chanussot, Hyperspectral sharpening approaches using satellite multiplatform data, IEEE Trans. Geosci. Remote Sens. (Jun. 2020) 1–19.
[32] S. Sabour, N. Frosst, G.E. Hinton, Dynamic routing between capsules, in: In P*roceedings of the Advances in Neural Information Processing Systems*, Oct. 2017, pp. 3859–3869.
[33] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: aligning domains using generative adversarial networks, in: In Proc*eedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 8503–8512.
[34] M.S. Sarfraz, R. Stiefelhagen, Deep perceptual mapping for cross-modal face recognition, Int. J. Comput. Vis. 122 (1) (2017) 426–438. Jan.
[35] Y. Shi, N. Siddharth, B. Paige, P. Torr, Variational mixture-of-experts autoencoders for multi-modal deep generative models, in: In P*roceedings of the Advances in Neural Information Processing Systems*, 2019, pp. 15692–15703. Jun.
[36] T.M. Sutter, I. Daunhawer, J.E. Vogt, Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence, **I**n, Proceedings of the Advances in Neural Information Processing Systems 2020 (2020) 1–11. Dec.
[37] G. Vivone, J. Chanussot, Fusion of short-wave infrared and visible near-infrared WorldView-3 data, Inform. Fusion 61 (2020) 71–83. Sept.
[38] M. Wang, W. Deng, Deep visual domain adaptation: a survey, Neurocomputing 312 (10) (2018) 135–153. Oct.
[39] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 318–335. Oct.
[40] Z. Wang, J. Chen, S.C.H. Hoi, Deep learning for image super-resolution: a survey, IEEE Trans. Pattern Anal. Mach. Intell. (2020). Mar. 1-1.
[41] H. Wu, Y. Yan, Y. Ye, M.K. Ng, Q. Wu, Geometric Knowledge Embedding for unsupervised domain adaptation, Knowl. Based Syst. 191 (3) (2020), 105155. Mar.
[42] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–21. Mar.
[43] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, J. Li, HPGAN: hyperspectral pansharpening using 3-D generative adversarial networks, IEEE Trans. Geosci. Remote Sens. (May 2020) 1–15.
[44] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: a unified unsupervised image fusion network, IEEE Trans. Pattern Analy. Mach. Intell. (July 2020). Early Access, pp. 1-1.
[45] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, C. Shen, Hyperspectral classification based on lightweight 3-D-CNN with transfer learning, IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5813–5828. Aug.
[46] L. Zhang, J. Fu, S. Wang, D. Zhang, Z. Dong, C.L.P. Chen, Guide subspace learning for unsupervised domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (Sep. 2020) 3374–3388.
[47] L. Zhang, S. Wang, G. Huang, W. Zuo, J. Yang, D. Zhang, Manifold criterion guided transfer learning via intermediate domain generation, IEEE Trans. Neural Netw. Learn. Syst. 30 (12) (2019) 3759–3773. Dec.
[48] Z. Zhang, F. Duan, J. Sole-Casals, J. Dinares-Ferran, A. Cichocki, Z. Yang, Z. Sun, A novel deep learning approach with data augmentation to classify motor imagery signals, IEEE Access 7 (2019) 15945–15954, https://doi.org/10.1109/ACCESS.2019.2895133.
[49] J. Zhao, L. Li, F. Deng, H. He, J. Chen, Discriminant geometrical and statistical alignment with density peaks for domain adaptation, IEEE Trans. Cybern. (Jun. 2020).
[50] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, J. Chanussot, Hyperspectral pansharpening using deep prior and dual attention residual network, IEEE Trans. Geosci. Remote Sens. (Apr. 2020) 1–18.
[51] J.Y. Zhu, P. Krähenbühl, E. Shechtman, A.A. Efros, Generative visual manipulation on the natural image manifold, in: Proceedings of the European Conference on Computer Vision, Oct. 2016, pp. 597–613.
[52] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 2242–2251.
[53] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, in: Proceedings of the IEEE, Jul. 2020, pp. 1–34.
[54] J.E. Zini, Y. Rizk, M. Awad, A deep transfer learning framework for seismic data analysis: a case study on bright spot detection, IEEE Trans. Geosci. Remote Sens. 58 (5) (May 2020) 3202–3212.